Essays on nonparametric and high-dimensional econometrics

Masaya Takano

Doctor of Philosophy

Department of Economics
McGill University
Montréal, Québec, Canada

November, 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

ABSTRACT

This thesis consists of four essays which address key problems in nonparametric and high-dimensional econometrics. The first essay proposes a bound approach to nonparametric regression. We employ a possibly misspecified model to approximate the true model, and then bound the approximation error using concentration inequalities in order to build confidence sets for a conditional expectation of unknown form. In the second essay, we study hypothesis testing of linear and nonlinear restrictions on a finite-dimensional parameter, using generalized $C(\alpha)$-type statistics based on estimating functions (or moment equations), when the estimating functions or the nuisance parameter estimates converge at non-standard rates (which may be slower than the square root of the sample size n). We establish conditions under which the C-alpha-type statistics follow the usual chi-square distribution. The third essay proposes a representation of a stochastic process of probability distributions in a L2 space which incorporates mass points and a varying support. These features are important to study the dynamics of earnings, income and wealth distributions. In the fourth essay, we propose a semiparametric approach for testing independence between two infinite-order cointegrated vector autoregressive series based on the residuals of long autoregressions.

Résumé

Cette thèse comprend quatre essais sur des problèmes d'économétrie non-paramétrique et en grande dimension. Le premier essai propose une approche fondée sur des bornes pour la régression non-paramétrique $\mathbb{E}[Y \mid X]$. Afin d'approximer le vrai modèle, nous utilisons un modèle qui peut comporter une erreur de spécification, et nous montrons comment l'erreur d'approximation peut être bornée par des inégalités de concentration, de façon à construire des régions de confiance pour une espérance conditionnelle de forme inconnue. Dans le deuxième essai, nous étudions des tests de type $C(\alpha)$ pour des restrictions linéaires et non-linéaires sur un paramètre de dimension finie, en utilisant des fonctions d'estimation (ou des conditions de moments) générales, lorsque les fonctions d'estimation ou l'estimateur des paramètres du modèle convergent à un rythme non-standard (qui peut être plus lent que la racine carrée de la taille d'échnatillon n). Nous établissons des conditions sous lesquelles les statistiques de type C-alpha continuent à suivre asymptotiquement la loi chi-carré habituelle. Le troisième essai propose une représentation d'un processus stochastique de distributions de probabilités dans l'espace $L^2[0,1]$, lequel comporte des masses des probabilités ou un support variable dans le temps. Ces caratéristiques sont importantes pour l'étude de la dynamique des distributions de salaires, de revenus et de richesse. Dans le quatrième essai, nous proposons une approche semi-paramétrique pour tester l'indépendance entre deux séries cointégrées de type VAR basée sur les résidus de autorégressions longues.

Contributions

The first and second essays are based on both joint works with Jean-Marie Dufour. Professor Dufour conceived the presented idea of the first essay. The primitive results of the second project were developed as a byproduct of my independent work, which generalizes his and his coauthors' earlier work, Dufour, Trognon and Tuvaandorj (2016). For both projects, the literature review, main theoretical results, simulations, and empirical applications, and the writing are primarily my own work. Professor Dufour directed the projects and contributed to the writing and editing, in particular, of the introduction sections.

The third essay is based on a joint work with Professor Victoria Zinde-Walsh. Professor Zinde-Walsh conceived the idea and was in charge of overall direction of the research project. The literature review, the inference framework and results, and the design and implementation of simulations and empirical study are primarily my own work. The theoretical framework and results are Professor Zinde-Walsh's work. She also contributed to I took the lead in writing the manuscript, except for Section 4.2.3-4.2.5. Professor Zinde-Walsh edited my writing; she added detailed accounts of the assumptions and interpretation of the results in the inference section, and improved the exposition of the proofs.

The fourth essay is based on an article forthcoming in *Advances in Econometrics: Essays in Honor of Joon Y. Park*, published in 2022, joint with Chafik Bouhaddioui and Jean-Marie Dufour. My contributions to this essay are local power analysis provided in Section 5.5 as well as the exposition and editing of the paper. The framework, other theoretical results in Section 5.2- 5.4 and simulation study are work of Professor Bouhaddioui and Professor Dufour.

Acknowledgements

I am deeply indebted to my supervisors, Jean-Marie Dufour and Victoria Zinde-Walsh, for their mentorships throughout my doctoral studies at McGill University.

Professor Dufour's insight based on his fundamental understanding of econometrics and statistics and wealth of experience has been pivotal part of my growth as a researcher. Professor Zinde-Walsh's wisdom has been my inspiration and she has always gone and beyond to support me in so many ways. I cherish every advice I feel fortunate to have received from the two leading econometricians throughout my career going forward.

Besides my supervisors, I am grateful of Markus Poschke for my research assistantship under him. I have benefited from his expertise in macroeconomics and the opportunities to work on various problems in labor market, which helped me nourish my apprehension of applied and empirical economics. Francisco Alvarez-Cuadrado and Francesco Amodio, as placement directors, guided me through the challenging process of the job market. I would also like to thank for Angela Fotopoulos, Lisa Stevenson, and Andrew Stoten for their administrative assistance.

I have a lot of helpful feedback on the thesis during seminars and on other occasions and, in particular, would like to thank for Taoufik Bouezmarni, John Galbraith, and Sílvia Gonçalves,

Masao Ogaki and Taisuke Otsu have been extremely supportive of my academic aspiration since I was an undergraduate student at Keio University. During my time at Brown University, Eric Renault and Frank Kleibergen generously provided me with their guidance.

I appreciate my friendship with Victor Aguiar, with whom I enjoy sharing any research ideas and receiving feedback since we met at Brown.

Last but not the least, this endeavor would not have been possible without my family's support. My parents encouraged me to pursue college education they could not receive and have been always my biggest supporters. I owe a debt of gratitude to my partner, Valérie, for her constant encouragement and patience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical inference relies on a set of assumptions imposed on the underlying data generating process and its validity is in question when any of the assumptions is not satisfied. In this thesis, we study econometric and statistical problems in which such an issue posed for existing methods has significant practical relevance, but has not been yet properly addressed in the literature. The true mechanism which generates observed data is not known to the practitioner and is likely to be highly complex especially in non-experimental research. Then, any postulated model should be seen as an approximation of the true model and the approximation error needs to be incorporated in inference either by bounding the error (Chapter 2) or by increasing the complexity of the approximate model as the sample size increases so that the error is asymptotically negligible (Chapter 5). For functional data, model misspecification may be tackled by considering a transformation which incorporates key features without imposing arbitrary smoothness conditions on a function of interest (Chapter 4). Even when the model structure imposed is assumed to be true, valid inference based on the standard $\sqrt{n}$-asymptotics may not be feasible due to some feature of the data generating process. Hypothesis testing can be still conducted in this case by utilizing a transformed estimating function, the asymptotic distribution of which is insensitive to estimation error (Chapter 3). The thesis contributes to nonparametric and high-dimensional econometrics by proposing statistical methods which allow for flexible model structures that existing methods may not accommodate.

In Chapter 2, we propose a bound approach to nonparametric regression. The object of interest, the conditional expectation $\mathbb{E}[Y \mid X]$ is in general unknown and difficult to iden-

tify. In the spirit of the parsimony principle, we employ a simple parametric model to approximate the true model and then bound the approximation error using concentration inequalities to build confidence sets for $\mathbb{E}[Y \mid X]$. Our approach is valid under less stringent regularity assumptions than conventional nonparametric methods, such as kernel regression and the sieve method. In particular, our framework allows for incomplete identification of the regression function and inference takes the form of sets in a partial identification framework. We show that approximation bounds can be built using only moments of observables and discuss how shape restrictions (e.g. smoothness) can be exploited to improve such bounds. We study optimality and uniformity of the proposed bounds by using the concepts of sharpness and honesty criteria. Inference only requires estimation of a simple parametric model and moments of observables along with results from the theory of extremum estimation. Thus, it is easy to implement and enjoys favorable finite-sample properties. Our Monte Carlo simulation studies compare our method with alternative methods (local polynomial regression, the sieve method, random forest, LASSO, and neural network) in terms of the average widths and coverage probabilities of associated confidence sets and the mean squared error of point estimates. Results show that the proposed method delivers valid confidence sets in cases where the other methods fail or cannot provide confidence sets at all. As an empirical application, we apply our method to inference for auto miles-per-gallon based on car attributes, the dataset of which is available from the UCI machine learning repository. Our method yields confidence sets with the shortest width while maintaining the size and generates best out-of-sample predictions based on point estimates. These findings support our theoretical results on finite-sample properties. In another application, we demonstrate how our bound approach provides economically significant information regarding the shape of regression curves, using household consumption data.

In Chapter 3, we study hypothesis testing of linear and nonlinear restrictions on a finite-dimensional parameter, using general estimating functions (or moment equations), when nuisance parameters are estimated at a rate which may be slower than $n^{1/2}$ under the null hypothesis. We focus on generalized $C(\alpha)$ tests, which allow one to use a wide class of nuisance parameter estimators, under weak assumptions on the asymptotic distribution of the estimators. However, root-n consistency remains notably restrictive, because it precludes estimators which converge at a slow rate, e.g. many estimators based on nonparametric regression. In this paper, we first establish conditions under which generalized $C(\alpha)$-type

statistics follow the usual chi-square distribution under nonstandard convergence rates: we allow for a convergence rate slower than the usual $n^{1/2}$ rate for the restricted estimator of the parameter of interest, as well as nonstandard convergence rates for the estimating functions (or moment equations) and their derivatives. In particular, when the estimating function converges to its limit at rate $n^{1/2}$, we only require that the convergence rate of the restricted estimator be faster than $n^{1/4}$. Second, we consider the case where the primary estimating functions include (possibly unrestricted) nuisance parameters which are replaced by estimators based on an auxiliary estimating function which may converge at a different rate from the primary estimating functions. In these cases, we propose extended generalized $C(\alpha)$-type statistics [$EC(\alpha)$], and derive their asymptotic null distribution. For such statistics, the estimation error on the nuisance parameters is asymptotically negligible, and the asymptotic chi-square distribution holds regardless of the choice of the nuisance parameter estimate, as long as the convergence rate is faster than $n^{1/4}$. The generalized $C(\alpha)$ statistics suggested nest existing $C(\alpha)$-type statistics as special cases, and thus broadens the applicability of these statistics to problems involving nonstandard rates. Four applications are discussed: (1) testing the derivatives of a conditional expectation; (2) average treatment effects in regression discontinuity designs; (3) semiparametric stochastic discounting factors; (4) testing the homogeneity of regression functions when the two samples go to infinity at different rates.

In Chapter 4, we focus on dynamics of earnings, income and wealth distributions without removing such features as mass points and varying support that are important characteristics of these distributions. The contribution of this paper is three-fold. Firstly, we evaluate the importance of mass points and accounting for the support via several stylized examples. We demonstrate that trimming the data could result in misinterpreting the stochastic properties of the process of distributions. Our second contribution is a new transformation into the $L^2[0,1]$ space that accounts for mass points and the varying support of the distribution. We link our representation to that of the demeaned density process in Chang et al (2016) (for the case of an absolutely continuous distribution) and demonstrate that the dimension of that non-stationary subspace for the stochastic process is preserved by our transform in the absolutely continuous case. Third, we provide a direct comparison with the empirical results of that paper by using the same data set (extended in time). Our test results similarly give the dimension of the unit root subspace to be 2, although we get much stronger

statistical evidence for persistence and demonstrate that the dynamics of the support or top quantiles is the main driver for persistence.

In Chapter 5, we propose a semiparametric approach for testing independence between two infinite-order cointegrated vector autoregressive series [IVAR (∞)]. The procedures considered can be viewed as extensions of classical methods proposed by Haugh (1976, JASA) and Hong (1996, Biometrika) for testing independence between stationary univariate time series. The tests are based on the residuals of long autoregressions, hence allowing for computational simplicity, weak assumptions on the form of the underlying process, and a direct interpretation of the results in terms of innovations (or shocks). The test statistics are standardized versions of the sum of weighted squares of residual cross-correlation matrices. The weights depend on a kernel function and a truncation parameter. Multivariate portmanteau statistics can be viewed as a special case of our procedure based on the truncated uniform kernel. The asymptotic distributions of the test statistics under the null hypothesis are derived, and consistency is established against fixed alternatives of serial cross-correlation of unknown form. A simulation study is presented which indicates that the proposed tests have good size and power properties in finite samples.

Chapter 2

Approximation bounds for conditional expectations and nonparametric regressions: theory and inference

2.1. Introduction

The objective of empirical econometrics is to provide a better understanding of key aspects of the complex economic system in the form of statistical modeling and inference. The majority of research questions in this endeavor involves relationships between macro/micro economic variables, in particular, in causal contexts. Regression modeling has been the central tool as an apparatus to investigate such relationships through conditional expectations. Regression analysis models the conditional expectation $\mathbb{E}[Y|X]$ of some outcome variable Y given exogenous variables X. The conditional expectation has intuitive interpretation as the mean outcome of Y conditional on X and enjoys an optimality property as the best predictor of Y given X in terms of the mean squared error. While other characteristics of the conditional distribution of Y given X can be exploited, e.g. through conditional quantiles (Koenker and Hallock (2001)), the regression framework has been the most celebrated statistical tool in applied econometrics due to its attractive features.

The conditional expectation function $m(\cdot) = \mathbb{E}[Y|X = \cdot]$ (CEF, hereafter) is in general unknown and difficult to identify. Whether parametric or nonparametric, identification can be achieved only under restrictive assumptions on the underlying data generating process. Parametric identification is valid only on the premise that the true model belongs to a class of models indexed by some finite-dimensional parameter, specified by the practitioner. In practice, it is hard to provide a compelling argument to justify such parametric specification. When this presumption fails, any attempt to identify the model parameter does not result in identification of the true model and the model parameter identified under regularity conditions, e.g. rank conditions in the case of linear models, can be only interpreted as the pseudo true value of a misspecified model, which may fail to represent the key aspects of the true model. It might appear that nonparametric identification, where the functional form of the the true model is treated as an infinite-dimensional parameter does not pose such concerns regarding model misspecification. However, this type of identification approach suffers from analogous issues. Identification of a nonparametrically specified object requires that there exists an injective mapping from the space of the distributions of observed variables to the space of the parameter of interest. This requirement is typically achieved by imposing restrictions on the support and domain of the mapping. Thus, violation of such restrictions puts identifiability into question. Matzkin (2007) presents identification conditions for a class of additive models, which includes regression models as a special case. In particular, she assumes the function g of interest to be continuous while the dependent variable X is also assumed to be continuous, when X is continuously distributed, continuity of the function g is a necessary condition. Thus, it is not possible to uniquely determine the value of the function from the data distribution when some type of discontinuity may not be ruled out at a point in question. This problem cannot be avoided unless all points of discontinuity are known a priori, which is often difficult to justify in practice. These observations pose limitations of existing methods that rely on point identification.

Even when taking the identification problem out of the equation, inference for nonparametrically specified functions is a challenging problem as inference requires more stringent regularity conditions than what does identification. It also suffers from slower convergence rates and is also prone to the curse of dimensionality.

This paper proposes a simple bound approach which circumvents these issues of the existing methods. Instead of attempting to estimate the conditional expectation $m(X)$ pre-

cisely, we employ an approximate model $h(\cdot)$ which we acknowledge captures only certain features of $m(\cdot)$ and can be arbitrarily different from it and then bound the approximation error $|m(X) - h(X)|$ using concentration inequalities given a prespecified confidence level. We show that such approximation bounds can be constructed from easily accessible characteristics of the distribution of observables, such as moments. Then, the bounds combined with the approximate model $h(\cdot)$ are used to construct a confidence set for $m(X)$. We show that our bounds are sharp and thus cannot be improved without imposing additional restrictions on the data generating process. They are also uniform in the sense that they are valid for any regression function consistent with the marginal distribution Y and that of X. Such property is known as honesty (reference).

While our approach accommodates the use of any approximating models, we suggest the use of simple parametric models. While parsimonious models are subject to more stringent specification, they often have superior predictive powers than highly complex models in practice due to smaller estimation errors. Since specification errors are incorporated in our bounds, using simple models as approximation does not raise any issues and instead leads to favorable finite-sample performances in terms of both coverage probability of associated confidence set and predictive power of point estimates.

Our approach is easy to implement and inference enjoys favorable finite-sample properties. We study inference for misspecified paramaetric models in the framework of extremum estimation and observe that the estimators are typically $n^{1/2}$-consistent and asymptotically normal. We show that the proposed approximation bounds can be consistently estimated by their sample analogue under general conditions. An estimated approximate model and bounds are used to construct an estimator of a confidence set for $m(X)$, which achieves a desired level asymptotically.

We consider two empirical applications: inference for car models' miles-per-gallon (auto MPG) and the Engle curve for alcohol share.The dataset used in the exercise is frequently used in the machine learning literature and is available from the UCI machine learning repository (Dua and Graff (2017)). We compare the average width and coverage probability of confidence sets associated with our method awith two kernel estimators and random forest. Our method yields confidence sets with the shortest width with correct empirical size. We choose an approximate model using the stepwise regression and this ad-hoc choice outperforms the alternative methods. These findings echo our Monte Carlo

experiments. The second empirical application uses household expenditure data (Family Expenditure Survey (FES) from 2000-2001 collected by Office for National Statistics (2002)). We show how our method can be applied to investigate the shape of the Engle curve for alcohol share without imposing arbitrary shape restrictions. Our bound based approach appears to provide meaningful information regarding the shape of the curve without imposing a stringent structure on it.

We proceed as follows. We review the literature in Section 2.2. Section 2.3 presents impossbility results on identification and testability of noparametric regression. Section 2.4 introduces our framework in an appropriate Hilbert space. In Section 2.5, we propose approximation bounds based on observed error and study their optimality and uniformity. Section 2.6 considers conditional bounds and their links to inference for $m(x)$ for a fixed point x. Section 2.7 discusses how smoothness of the function $m(\cdot)$ can be utilized to form an alternative approximation bound. Section 2.8 provides our inference framework for confidence sets for $m(X)$. We conclude with Monte Carlo simulation and empirical applications in Section 2.9 and 2.10, respectively. Section 2.11. concludes.

2.2. Related literature

Our work builds upon the literature on various topics in nonparametric and semiparametric inference in econometrics and statistics. In this section, we review earlier relevant works and point out the connections to our method to highlight our contribution.

2.2.1. Nonparametric estimation of the conditional expectation

Kernel regression and the method of sieve are widely-known branches of nonparametric regression estimation frameworks and have been extensively studied in the literature of statistics and econometrics. Estimators based on kernel regression include the Nadaraya-Watson and Gasser-Műller estimators (Nadaraya (1964), Watson (1964), Gasser and Müller (1979)) and local polynomial estimators (Cleveland and Devlin (1988)). Kernel regression approximates a regression function locally by the weighted average of neighboring points and controls the degree of smoothing by the bandwidth parameter, which is assumed to shrink to zero at some rate as the sample size increases. The method of sieve (Grenander

(1981)) employs a sequence of approximating spaces (or sieve spaces) which is asymptotically dense in the space of regression functions and the sieve estimators are obtained by minimizing of appropriate objective functions associated with approximating spaces. For example, series estimators and are partitioning-based least squares nonparametric regression estimators (Cattaneo, Farrell and Feng (2020)) are sieve estimators. Asymptotic properties of kernel regression and the method of sieve are well-understood (Newey (1997), Pagan and Ullah (1999), Chen (2007), Li and Racine (2007), Horowitz (2009) among others) and they can be implemented, despite being computationally-intensive, relatively easily using common statistical packages. Any estimation method which attempts point-estimation of a parameter of interest is not robust to identification failure since the estimator converges to a single point while there are multiple values of the parameter that are consistent with the data generating process.

Furthermore, in the inference problem of nonparametric regression, additional regularity conditions are imposed to ensure that an estimator approximates the infinite-dimensional parameter arbitrarily well asymptotically. Kernel estimation of nonparametric regression typically imposes differentiability of the regression function up to some order and the existence of the density functions of the covariates. In the method of sieve, it is typically assumed that the regression function belongs to some Hőlder class of functions when commonly used linear sieves, such as power series, splines or wavelets, form approximating spaces. Due to the slow rates of convergence these estimators achieve, finite-sample performance is often unreliable and the situation worsens as the dimension of the covariates increases, the phenomenon known as the curse of dimensionality. Confidence sets based on normal approximation may fail to attain a desired level due to the presence of bias and remedies based on undersmoothing, bootstrap, and bias correction have been proposed in the literature (Hall (1992), Hall and Horowitz (2013), Calonico, Cattaneo and Farrell (2018), Cattaneo et al. (2020)). However, they also face practical challenges since they either require an unconventional choice of the tuning parameter with little guidance or require estimation of higher order derivatives, thus finite-sample performance in multi-dimensional covariates can be questionable.

Some machine learning algorithms, such as regularized least squares, tree based models, and neural network, can be applied for learning problems on nonparametric regression. The purpose of these methods is predictive modeling, which minimizes the prediction error

under a given sample size and thus they do not necessarily select the true model asymptotically. Furthermore, their asymptotic theory is still under-developed except for a few methods (Wager and Athey (2018), Farrell, Liang and Misra (2021)). Our approach requires substantially weaker assumptions than these methods and is valid for almost any class of regression functions. Our confidence sets can be estimated at a parametric rate and possess favorable finite-sample properties. As our simulation results and empirical applications indicate, confidence sets based on nonparametric methods or machine learning algorithms are often undersized even in a fairly large sample size in cases where the regression function is not smooth or irrelevant variables are included in the covariates. Our confidence sets deliver the correct coverage probability in these cases even under a relatively small sample size. Thus, our bound approach offers features that are of practical importance and provides practitioners with a robust and reliable inference method for regression models.

2.2.2. Set identification and inference based on bounds and confidence sets

The partial identification approach employs the notion of set identification and conducts inference for the parameter of interest without assuming identification in the standard sense (point identification); Manski (1990), Manski (2003), Imbens and Manski (2004), Chernozhukov, Hong and Tamer (2007*a*), Romano and Shaikh (2010), Santos (2012).

Our framework is in line with the spirit of the partial identification framework in that we avoid restrictive assumptions needed to achieve point identification and instead deliver valid inference using bounds and confidence sets. Our confidence set for $\mathbb{E}[Y|X]$ is valid under identification failure of the conditional expectation function. We postulate and estimate a statistical model only for the purpose of approximating a true regression function and forming a confidence set for $\mathbb{E}[Y|X]$. This is in contrast with conventional nonparametric approaches which attempt to estimate each point of a regression function.

2.2.3. Use of misspecified parametric models for inference

Our work contributes to the literature on model misspecification by proposing a valid inference framework for an unknown true model by complementing general misspecified models with approximation bounds. In the series of his work in 1980's (White (1980*a*),

White (1980*b*), White (1982)), Halbert White and his coauthors built modern foundation of inference for misspecified models. White (1982) provides a robust view of maximum likelihood estimation to misspecification and shows that the maximizer of the quasi likelihood still exists as a minimizer to the Kullbuck-Leibler distance to the true density. The heteroskedasticity-robust covariance estimator (White (1980*a*)) consistently estimates the asymptotic covariance matrix of the slope parameter regardless of whether the true regression function is indeed linear with respect to the covariates[1]. Buja, Berk, Brown, George, Pitkin, Traskin, Zhao and Zhang (2015) revisit and elaborate further his work. They note that the coefficients of a misspecified linear model depend on the regressor distribution and make an argument against conditioning on regressors when the linear model is only an approximation. In spite of the extensive literature, there are few papers which take into account the specification error of an arbitrary misspecified model and deliver a valid inference for the underlying true model. As a notable exception, Glad (1998) proposes to employ a simple parametric model to approximate the regression function to reduce the bias and then nonparametrically estimate the correction factor, a function which corrects model misspecification. Her approach known as parametrically guided nonparametric regression shares certain similarities with our method in the use of misspecified models to approximate a regression function. However, her method involves nonparametric estimation of an infinite-dimensional parameter and hence inherits features of the nonparametric approach which we discussed in Section 2.2.1 including a slow convergence rate. We only bound the approximation error instead of estimating it and such a bound is typically constructed from the moments of the data distribution and thus benefits from the favorable finite sample properties of parametric estimation.

2.2.4. Principle of parsimony in statistical modeling

We suggest the use of simple parametric models to approximate the regression function in our bound approach. The principle of parsimony, also known as Occam's razor, is a general philosophical rule which favors simpler explanations, coined by an English logician William of Ockham in the 14th century. Its justification has been provided theoretically and empirically in various disciplines, including statistics and econometrics (Rissanen

[1] When linearity of the regression function does not hold, the misspecified linear model can be still interpreted as a linear projection model.

(1978), Rissanen (1982), Rissanen (1987), Rissanen (1987), Ploberger and Phillips (2001), Ploberger and Phillips (2003)). In empirical economics, parsimonious modeling, especially reliance on a linear structure, has been common. In nonlinear dynamic stochastic general equilibrium (DSGE) models, it is common to obtain the (approximate) solution by considering the Taylor expansion of policy functions (Kydland and Prescott (1982), King, Plosser and Rebelo (1988), Schmitt-Grohé and Uribe (2004)). In survival analysis, the two dominating empirical models, the Cox proportional hazards model (Cox (1972)) and the accelerated failure time model (Wei (1992)) both impose a parametric form to capture the dependence between the dependent variable and covariates. Inference for random choice models, notably demand in differentiated-product markets (Berry (1994), Berry, Levinsohn and Pakes (1995)) typically assumes that the random utility depends on covariates linearly. Such specification is often ad-hoc and should be seen as only an approximation of the true data generating process. The prevalence of parsimonious modeling can be seen from the popularity of the Akaike information criterion (AIC, Akaike (1974)) While AIC achieves optimality in the MSE sense, it is known that it is not necessarily consistent, that is, it does not choose the correctly specified model (Vrieze (2012)). While the inference framework for misspecified models has been established in the literature we review in the previous section, little attention has been paid to the question of how misspecified models can be used for inference for the true model. This paper is the first work to provide justification to the use of misspecified empirical models by explicitly taking into account model specification bias and deliver asymptotically valid inference for nonparametric regression. In order to implement our method, practitionerrs would be required in addition to using the standard estimation procedures (1) to compute heteroskedasticity-consistent standard errors of estimators of the approximate model parameters and (2) to estimate the approximation bounds as we discuss in detail in Section 2.8.

2.3. Some negative results on identification and testability of nonparametric regression

This section presents some impossibility results concerning identification and testability of nonparametric regression. We revisit Matzkin (2007)'s results on the nonparametric

identification of the regression function m (or $m(\cdot)$ evaluated at a given point x_0: $m(x_0)$) and observe that continuity of the function m plays an essential role in her results. We show that when m is arbitrarily specified, $m(x_0)$ is not identified for every x_0 in the support of $\mathscr{X}$. In such a case, any point estimator of $m(x_0)$ is inconsistent. Furthermore, if the confidence set of $m(x_0)$ based on such a point estimate only incorporates estimation uncertainty, its size is zero regardless of the nominal level. We further observe that even when $m(x_0)$ is identified, if there is no restriction on the underlying data generating process, meaningful inference on $m(x_0)$ may not be feasible in the sense that any testing procedure for the hypothesis of the form: $H_0(\mu_0) : m(x_0) = \mu_0$ for a fixed value $\mu_0 \in \mathbb{R}$ has trivial power. A more formal account of the nonparametric identification is provided in Appendix: 2.A. For a pair of random variables (Y,X) such that $\mathbb{E}|Y| < \infty$, we consider the nonparametric regression setup:

$$Y \quad = \quad m(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0. \tag{2.3.1}$$

Consider the class of functions $\mathscr{M}$ defined as

$$\mathscr{M} = \{m^* : \mathscr{X} \to \mathscr{Y} \ : \ m^*(X) \text{ is integrable.}\} \tag{2.3.2}$$

We assume that X is continuously distributed and make the following assumption.

Assumption 2.3.1 DENSITY f_X OF X. *For the marginal distribution F_X of X, the density f_X exists, and is continuous. Furthermore, $f_X(x) > 0$ for any $x \in \mathscr{X}$.*

The same assumption is made in Matzkin (2007) for identification of $m(\cdot)$. We shall cover the case where X is discrete in Lemma 2.3.2. The following assumption states that there is essentially no restriction made on the functional form of the function m.

Assumption 2.3.2 FUNCTION CLASS OF m. *The regression function m is in $\mathscr{M}$: $m \in \mathscr{M}$.*

This assumption differs from Matzkin (2007), who only considers the class of continuous functions. When such a general class of functions is allowed, $m(\cdot)$ is not identified at any point of the support.

Proposition 2.3.1 *Assumption 2.3.1-2.3.2 hold. Then, for any $x_0 \in \mathscr{X}$, $m(x_0)$ is not identified.*

As a corollary, we show that when x_0 is a point of probability mass, $m(x_0)$ is identified.

Corollary 2.3.2 *Suppose Assumption 2.3.2 holds and $Pr(X = x_0) > 0$. Then, $m(x_0)$ is identified.*

The following result is a generalization of Bahadur and Savage (1956) on testability of the mean of a random variable.

Proposition 2.3.3 *Suppose $\{(y_i, x_i)\}_{i=1}^n$ is i.i.d. and $\mathbb{E}|y_1| < \infty$. For a given point $x_0 \in \mathscr{X}$ and $\mu_0 \in \mathscr{Y}$, consider the null hypothesis:*

$$H_0(\mu_0; x_0) : m(x_0) = \mu_0. \tag{2.3.3}$$

Let

$$\mathscr{H}(\mu_0; x_0) = \{\textit{Distribution functions } F_n \in \mathscr{F}_n \textit{ such that m}(x_0) \textit{ is identified and } H_0(x_0, \mu_0) \textit{ holds}\} \tag{2.3.4}$$

where $\mathscr{F}_n$ is the family of all probability distributions of $\{(y_i, x_i)\}_{i=1}^n$. Then, $H_0(\mu_0; x_0)$ is not testable, i.e. if a test has level α ($\alpha \in (0,1)$) for $H_0(\mu_0; x_0)$, that is

$$P_{F_n}(\textit{Rejecting } H_0(\mu_0; x_0)) \leq \alpha \qquad \textit{for all } F_n \in \mathscr{H}(\mu_0; x_0), \tag{2.3.5}$$

then for any $\mu_1 \neq \mu_0$, for all $F_n \in \mathscr{H}(\mu_1; x_0)$,

$$P_{F_n}(\textit{Rejecting } H_0(\mu_0; x_0)) \leq \alpha. \tag{2.3.6}$$

Further, if there is at least one value $\mu_1 \neq \mu_0$ such that for at least one $F_n \in \mathscr{H}(\mu_1; x_0)$,

$$P_{F_n}(\textit{Rejecting } H_0(\mu_0; x_0)) \geq \alpha \tag{2.3.7}$$

then for all $\mu_1 \neq \mu_0$,

$$P_{F_n}(\textit{Rejecting } H_0(\mu_0; x_0)) = \alpha \qquad \textit{for all } F_n \in \mathscr{H}(\mu_1; x_0). \tag{2.3.8}$$

Proposition 2.3.3 states that when $m(x_0)$ is identified, the power of any test on the value of $m(x_0)$ does not exceed its nominal level. Along with Proposition 2.3.1, it entails that under only minimal assumptions imposed on the data generating process, inference on $m(x_0)$, in particular based on a point estimate, is challenging. On the other hand, a bound approach for $m(X)$ considered here is valid under such circumstances. In particular, we allow for $m(\cdot)$ to be weakly identified.

2.4. Framework

2.4.1. Brief description of the framework

We consider the standard nonparametric regression setup:

$$Y = m(X) + \varepsilon \tag{2.4.9}$$

where $m(x)$ is the conditional expectation of Y given $X = x$ and the error $\varepsilon \equiv Y - m(X)$ is additive and satisfies the mean-independence of X:

$$\mathbb{E}[\varepsilon|X] = 0. \tag{2.4.10}$$

Note that other characteristics of the distribution of ε given X remain unspecified in this general setup. In particular, ε^2 is not mean-independent of X in general, i.e. $\mathrm{Var}(\varepsilon|X)$ can depend on X. The conditional moment restriction (2.4.10) implies that

$$\mathrm{Var}(Y - m(X)) \leq \mathrm{Var}(Y - h(X)) \tag{2.4.11}$$

for any square-integrable transformation $h(X)$ of X and the equality holds if and only if $m(X) \neq h(X)$ with probability zero. Then, it is tempting to evaluate a given approximate model h for m in terms of the mean squared deviation; h is deemed to be better approximation for m when

$$\mathbb{E}|m(X) - h(X)|^2 \tag{2.4.12}$$

is smaller. When one chooses the best approximate model based on this criterion among a family $\mathscr{H}$ of candidate models,the best model $h^* \in \mathscr{H}$ can be interpreted as a projection

onto $\mathscr{H}$ in an appropriate Hilbert space and its existence and uniqueness is guaranteed, by the Hilbert projection theorem, provided $\mathscr{H}$ is a nonempty convex closed set. Furthermore, $h^* \in \mathscr{H}$ minimizes (2.4.12) if and only if it minimizes $\mathbb{E}|Y - h(X)|^2$, the moment only involving observables, in particular, it is then free of the unknown object of interest, $m(X)$. In the next part of this section, we formulate the framework in the Hilbert space $\mathscr{L}^2$ and formalize the discussion here.

2.4.2. Detailed framework

We first define an appropriate Hilbert space.

Definition 2.4.1 L^2 SPACE. *Consider a probability space $(\Omega, \mathscr{F}, P)$, where Ω is a sample space, $\mathscr{F}$ is a sigma-algebra of subsets of Ω, and P is a probability measure on the measurable space $(\Omega, \mathscr{F})$. $L^2 = L^2(\Omega, \mathscr{F}, P)$ is a real Hilbert space of univariate random variables with mean zero equipped with inner product $\langle \cdot, \cdot \rangle : L^2 \times L^2 \to \mathbb{R}$ defined as*

$$\langle Z, W \rangle = \mathbb{E}[ZW] \tag{2.4.13}$$

for $Z, W \in L^2$ where expectation is taken with respect to P. The induced norm $\|\cdot\| : L^2 \to \mathbb{R}$ on this space is defined by

$$\|Z\| = \sqrt{\langle Z, Z \rangle} = \sqrt{Var(Z)} \tag{2.4.14}$$

for $Z \in L^2$. Any element Z in L^2 is a measurable mapping from $(\Omega, F) \to (\mathscr{Z}, \mathscr{B}(\mathscr{Z}))$ where $\mathscr{Z} \subset \mathbb{R}$ is the support of Z and $\mathscr{B}(\mathscr{Z})$ is the Borel sigma-algebra for $\mathscr{Z}$ and Z is square-integrable, i.e.

$$\mathbb{E}Z^2 \equiv \int (Z(\omega))^2 \, dP(\omega) < \infty. \tag{2.4.15}$$

We say that two elements of L^2, Z and W, are orthogonal, denoted by $Z \perp W$, if

$$\langle Z, W \rangle = 0. \tag{2.4.16}$$

Similarly, Z is said to be orthogonal to a set $\mathscr{W}$, denoted by $Z \perp \mathscr{W}$, if Z is orthogonal to

any element of $\mathscr{W}$:

$$\langle Z, W \rangle = 0, \quad \forall W \in \mathscr{W} \tag{2.4.17}$$

Definition 2.4.2 PROJECTION. *For $\boldsymbol{C} \subset L^2$, the projection $P_{\mathbf{C}}; L^2 \to \boldsymbol{C}$ of Y onto* **C** *is an element q in* **C** *such that*

$$\|Y - q\| \leq \|Y - c\| \qquad , \forall c \in \boldsymbol{C}. \tag{2.4.18}$$

Note that the definition of projection in Definition 2.4.2 is more general than projection onto a subspace, often called linear projection. While linear projection possesses intuitive characterization through the orthogonality condition discussed in Lemma 2.4.3, it fails to subsume many interesting econometric models in which the index parameter is not linear in regressors. As stated in Lemma 2.4.1, closedness and convexity of a set is necessary and sufficient for the existence and uniqueness of projection for any element in L_2, which implies that projection must be linear. However, we are often interested in projection of a specific element of L_2, e.g. a conditioned random variable Y and in such a case, orthogonality condition stated in Lemma 2.4.2 guarantees the existence and uniqueness without imposing convexity on the set.

Lemma 2.4.1 CHEBYSEV SET AS A NONEMPTY CLOSED SET. *A set $\boldsymbol{C} \subset L_2$ is nonempty, closed, and convex if and only if* **C** *is a Chebyshev set, i.e. for any $Y \in L^2$, $P_{\mathbf{C}}Y$ exists and is unique. Furthermore,*

$$\langle Y - q, q \rangle \geq \langle Y - q, c \rangle, \quad \forall c \in \boldsymbol{C} \tag{2.4.19}$$

if and only if

$$q = P_{\mathbf{C}}Y. \tag{2.4.20}$$

Lemma 2.4.1 implies that in order to guarantee the existence and uniqueness of the projection of "every" square integrable random variable, projection has to be linear. The next lemma allows one to consider a more general class of projection by providing conditions under which projection of a specific random variable exists and is unique.

Lemma 2.4.2 EXISTENCE AND UNIQUENESS OF PROJECTION THROUGH ORTHOGONALITY. *For a set $\boldsymbol{C} \subset L_2$, suppose for $Y \in L_2$, there exists some $q \in \boldsymbol{C}$ such that*

$$\langle Y - q, c \rangle = 0 \quad \forall c \in \boldsymbol{C}. \tag{2.4.21}$$

Then, $P_{\mathbf{C}}Y$ exists and is unique.

The following lemma states that projection of any element in L^2 onto a subspace is also characterized by the same orthogonal conditions as in Lemma 2.4.2.

Lemma 2.4.3 ORTHOGONAL PROJECTION ONTO A SUBSPACE. *Let $\mathbf{C}$ be a subspace in L_2. Then, for each $Y \in L_2$, $P_{\mathbf{C}}Y$ exists as a unique element in $\mathbf{C}$. Further,*

$$\langle Y - q, c \rangle = 0, \quad \forall c \in \boldsymbol{C} \tag{2.4.22}$$

if and only if

$$q = P_{\mathbf{C}}Y. \tag{2.4.23}$$

In the L^2-space, the conditional expectation $\mathbb{E}[Y|X]$ can be identified as the projection onto the subspace $\mathscr{M}(X)$ of L_2 consisting of all square-integrable functions of X, i.e.

$$E[Y|X] = \arg \inf_{q \in \mathscr{M}(X)} \|Y - q\|. \tag{2.4.24}$$

By Lemma 2.4.3, such an element is unique almost everywhere. Further, if we let $\varepsilon \equiv Y - \mathbb{E}[Y|X]$, ε is orthogonal to $\mathscr{M}(X)$:

$$\varepsilon \perp \mathscr{M}(X) \tag{2.4.25}$$

or equivalently

$$\mathbb{E}[\varepsilon|X] = 0. \tag{2.4.26}$$

Note that the conditional variance of ε

$$\sigma_\varepsilon^2(X) \equiv \mathbb{E}\left[\varepsilon^2|X\right] \tag{2.4.27}$$

could depend on X. Representation of $m(X)$ as the projection of Y onto $\mathscr{M}(X)$ in (2.4.24)

indicates a complication in identifying and estimating the functional form of m since the cardinality of $\mathscr{M}(X)$ is $\left|\mathscr{Y}^{\mathscr{X}}\right|$, where $\mathscr{X}$ and $\mathscr{Y}$ are the supports of X and Y, respectively: the minimization problem in the RHS in general involves optimization over a uncountably infinite-dimensional parameter q. On the other hand, the projection onto a smaller subset of $\mathscr{M}(X)$ is often more tractable, especially when the set is indexed by some finite-dimensional parameter. e.g. $C(X;\Theta) = \{h(X;\theta) \in L_2, \theta \in \Theta\}$ where h is known up to the finite-dimensional parameter $\theta \in \Theta$. The set $C(X) \subset \mathscr{M}(X)$ can be interpreted as a collection of approximate models for $m(X)$ and the projection $P_{\mathbf{C}}m(X)$ of $m(X)$ onto $\boldsymbol{C}(X)$ is the best approximate model of $m(X)$ in the MSE (mean squared error) sense among $\boldsymbol{C}(X)$. The next lemma shows that in order to find $P_{\mathbf{C}}m(X)$, we only need to find the projection $P_{\mathbf{C}}Y$ of Y onto the same space $\mathbf{C}$, a problem involving only observables.

Lemma 2.4.4 IDENTITY OF PROJECTIONS OF Y AND THE REGRESSION FUNCTION $m(X)$. *For the set $\boldsymbol{C} := \boldsymbol{C}(X) \subset \mathscr{M}(X)$, define the projection operator $P_{\mathbf{C}} : L^2 \to \boldsymbol{C}$ onto the set $\mathbf{C}$ as in 2.4.2. Assume $P_{\mathbf{C}}Y$ is unique. Then, the following identify holds:*

$$P_{\mathbf{C}}Y = P_{\mathbf{C}}m(X). \tag{2.4.28}$$

If $\boldsymbol{C}(X)$ is further assumed to be indexed by some parameter space Θ of finite dimension, then $\boldsymbol{C} = \boldsymbol{C}(X;\Theta)$ can be interpreted as a collection of parametric models. While our general framework does not restrict approximation models to be parametric and any square-integrable function of X may be be employed, this class of models is of particular interest due to its tractability in estimation and inference. Inference for a parametric model is based on asymptotic properties of a finite-dimensional parameter θ_0, the unique element of Θ such that it is associated with the projection $P_{\mathbf{C}}m(X)$ for which a $\sqrt{n}$ estimator of θ_0 is available under general assumptions as we discuss in Section 2.8. Note that uniqueness of the projection $P_{\mathbf{C}}m(X)$ does not imply identification of θ_0. Assumption 2.4.1 states that there is no redundancy in Θ in the sense that two distinct elements of Θ correspond to two distinct approximate models.

Assumption 2.4.1 IDENTIFICATION. *Consider a set $\boldsymbol{C}(X;\Theta) \subset \mathscr{M}(X)$ which is generated by a square integrable function $h(X;\theta)$ indexed by a finite-dimensional parameter*

$\theta \in \Theta \subset \mathbb{R}^k$:

$$\boldsymbol{C}(X;\Theta) = \{h(X;\theta);\theta \in \Theta\}. \tag{2.4.29}$$

For any $\theta_1, \theta_2 \in \Theta$,

$$\theta_1 = \theta_2 \tag{2.4.30}$$

if and only if

$$h(X(\omega);\theta_1) = h(X(\omega);\theta_2), \quad a.e.on \Omega. \tag{2.4.31}$$

Under Assumption 2.4.1, the problem of finding a possibly infinite-dimensional vector $\{P_{\mathbf{C}} m(x)\}_{x \in \mathscr{X}}$ reduces to a finite-dimensional problem with a parameter $\theta \in \Theta$ and its solution θ_0 is unique.

Lemma 2.4.5 *Suppose the set* $\boldsymbol{C} := \boldsymbol{C}(X;\Theta) \subset \mathscr{M}(X)$ *satisfies Assumption 2.4.1, and a projection of* Y *onto* $\mathbf{C}$ *exists and is unique. Then, there exists a unique element* $\theta_0 \in \Theta$ *(the projection parameter) such that*

$$\|Y - h(X;\theta_0)\| < \|Y - h(X;\theta)\|, \quad \forall \theta \neq \theta_0 \tag{2.4.32}$$

and

$$\|m(X) - h(X;\theta_0)\| < \|m(X) - h(X;\theta)\|, \quad \forall \theta \neq \theta_0. \tag{2.4.33}$$

Lemma 2.4.5 implies that $h(X;\theta_0)$, the best predictor of Y in the class of functions $\boldsymbol{C}(X;\Theta)$ in terms of the norm $\|\cdot\|$, is also the best approximate model of the conditional expectation $m(X)$ inby the same criteria.

Example 2.4.1 $\boldsymbol{C}(X;\Theta)$ is a subspace spanned by a set of basis functions

$$p^k(x) = (p_{1k}(x), \cdots, p_{kk}(x))' \tag{2.4.34}$$

and its element is indexed by θ defined on a nonempty convex subspace Θ of $\mathbb{R}^m$. Every element of $\boldsymbol{C}(X;\Theta)$ can be written as

$$p^k(x)'\theta \tag{2.4.35}$$

for some $\theta \in \Theta$. Convexity and closedness of $\boldsymbol{C}(X;\Theta)$ are automatic. Note that basis

functions are not necessarily continuous in x, for example when splines are employed.θ_0 is identified if and only if

$$\mathbb{E}\left[p^k(X)' p^k(X)\right] \tag{2.4.36}$$

is full-ranked so that

$$\theta_0 = \left(\mathbb{E}\left[p^k(X)' p^k(X)\right]\right)^{-1} \mathbb{E}\left[p^k(X)' Y\right]. \tag{2.4.37}$$

Example 2.4.2 Set $\boldsymbol{C} := \boldsymbol{C}(X;\Theta)$ as

$$\left\{\exp\left(X'\theta\right) : \theta \in \Theta\right\} \tag{2.4.38}$$

where Θ is a compact convex subset of $\mathbb{R}^m$. Then, if there exists some $\theta_0 \in \Theta$ such that

$$\langle m(X) - \exp\left(X'\theta_0\right), \exp\left(X'\theta\right)\rangle = 0, \tag{2.4.39}$$

then projection $P_{\mathbf{C}}(X)$ is unique and it is associated with a unique element $\theta_0 \in \Theta$ so that

$$P_{\mathbf{C}}(X) = \exp\left(X'\theta_0\right). \tag{2.4.40}$$

Contrary to linear models, where identification of the projection parameter θ_0 can be translated into the rank condition, the uniqueness and existence of θ_0 is often difficult to check. Such potential issue may favor the use of linear models as approximate models.

2.5. Bounds for conditional expectation and nonlinear regression

2.5.1. Upper bounds for the unobserved approximation error

This section establishes approximation bounds for nonlinear regression, a device which plays a pivotal role in our approach by stochastically controlling the approximation error in the conditional expectation. Under general conditions, as illustrated in Section 2.4.2, for $Y \in L_2$ and $\boldsymbol{C} := \boldsymbol{C}(X) \subset \mathscr{M}(X)$, there exists a unique element $P_{\mathbf{C}} m(X) \in \boldsymbol{C}$, the

projection of Y onto $\mathbf{C}$, such that

$$\mathrm{Var}\left(Y-m\left(X\right)\right)\leq\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)<\mathrm{Var}\left(Y-h\left(X\right)\right),\quad\forall h\in\boldsymbol{C}\backslash\left\{P_{\mathbf{C}}m\left(X\right)\right\}. \tag{2.5.1}$$

Further, by Lemma 2.4.4

$$P_{\mathbf{C}}Y=P_{\mathbf{C}}m\left(X\right) \tag{2.5.2}$$

so that

$$\mathrm{Var}\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right)<\mathrm{Var}\left(m\left(X\right)-h\left(X\right)\right),\quad\forall h\in\boldsymbol{C}\backslash\left\{P_{\mathbf{C}}m\left(X\right)\right\}. \tag{2.5.3}$$

This entails that $P_{\mathbf{C}}Y$ is the best approximate of the conditional expectation $m\left(X\right)$ in terms of the MSE in the given family of square-integrable functions of X. There still, however, remains the question of how far $P_{\mathbf{C}}Y$ is from $m\left(X\right)$, or equivalently how large the approximation error $P_{\mathbf{C}}Y-m\left(X\right)$ is. We approach this question by bounding the approximation error in a stochastic sense. To elaborate this notion, suppose for a constant $\alpha\in\left(0,1\right)$, some positive constant c_{α} such that

$$\Pr\left(\left|m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right|>c_{\alpha}\right)\leq\alpha \tag{2.5.4}$$

is available. Then,

$$\left[P_{\mathbf{C}}m\left(X\right)-c_{\alpha},P_{\mathbf{C}}m\left(X\right)+c_{\alpha}\right] \tag{2.5.5}$$

is a confidence set for $m\left(X\right)$ with level $1-\alpha$. This section is concerned with construction of such constants c_{α}'s. We show that they can be constructed only with easily accessible characteristics of the distribution of $\left(Y,X\right)$, such as the moments. Then, since the set $\mathbf{C}$ is specified by the practitioner, the pair $\left(P_{\mathbf{C}}m\left(X\right),c_{\alpha}\right)$ does not involve any nuisance parameter, in particular, $m(\cdot)$. In practice, $\left(P_{\mathbf{C}},c_{\alpha}\right)$ has to be estimated. We defer the discussion of the impact on inference until until Section 2.8.

In Section 2.5.2, we pose a general problem of finding bounds from which an unobserved random variable deviates with at most a pre-specified probability. Our particular interest of bounding the approximation error $\left|m\left(X\right)-P_{\mathbf{C}}Y\right|$ is seen as a special case of such a problem where $\left|m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right|$ is treated as an unknown object. We propose in Section 2.5.3 a construction of nontrivial bounds based only on the knowledge of some

moments of observable variables (Y,X). We note that our approach relies on concentration inequalities; one only requires knowledge or estimatbility of a certain moment; this in turn implies that any resulting bound is guaranteed to yield a valid confidence set, but it can be much wider than the best bound obtained when the distribution of $|m(X) - P_{\mathbf{C}}m(X)|$ is fully known. We circumvent this difficulty pertinent to this type of inequalities by combining multiple (possibly an uncountable set of) bounds to yield a single refined bound. Section 2.5.5 serves as display of one-sided bounds as opposed to two-sided ones.

2.5.2. Generic concentration bounds for an unobserved random variable

Chebyshev's inequality states that for any random variable Z with finite second moment and a positive constant $c > 0$,

$$\Pr\left(|Z - \mathbb{E}[Z]| \geq c\right) \leq \frac{\mathrm{Var}(Z)}{c^2}. \tag{2.5.6}$$

This moment-based inequality has an appealing feature that the probability of deviation from a central measure $\mathbb{E}[Z]$ larger in magnitude than any constant $c > 0$ can be bounded without completely specifying the distribution of Z. The moments of Z could be consistently estimated by the sample analogues under general conditions, such as ergodicity and existence of appropriate moments. Bounds based on higher order moments and the moment generating function (exponential bounds by Chernoff (1952)) have been considered as variations of Chebyshev's inequality and each of them has its own merits (see for example Boucheron, Lugosi and Massart (2013) for discussion).

The main takeaway here is that even if Z is not observed or difficult to estimate, an analogous type of bound can be obtained as long as we are able to bound the second moment$\mathrm{Var}(Z)$ in (2.5.6). To motivate the succeeding discussion, let us consider the following simple but insightful problem and its solutions observed by Meyer (1974). He pointed out that for a pair of univariate standardized random variables (X,Y) with correlation ρ,

$$\Pr\left(|m(X) - \mathbb{E}[Y] - \rho(X - \mathbb{E}[X])| \geq c\right) \leq \frac{1-\rho^2}{c^2} = \frac{\mathrm{Var}(Y - \rho X)}{c^2} \quad ,\forall c > 0. \tag{2.5.7}$$

This inequality states that regardless of the functional form of m, the probability that the deviation from linearity $|m(X) - \rho X|$ exceeds a threshold c can be bounded from above using variation of observables: $\mathrm{Var}(Y - \rho X)$. It can be derived by applying Chebyshev inequality and then observing that

$$\mathrm{Var}(m(X) - \rho X) \leq \mathrm{Var}(Y - \rho X). \tag{2.5.8}$$

Our results are more general than Meyer (1974)'s in at least three ways: (i) our result is concerned with an approximation error with respect to the conditional expectation of a general class of projection models with multiple regressors, not just a simple linear model, (ii) our bounds could exploit not only the second moment, but other characteristics of the underlying distribution, such as higher order moments and they are combined to yield more refined bounds. (iii) we also provide single-sided bounds (Cantelli-type bounds) along with two-sided bounds. To this end, we first need the following lemma.

Lemma 2.5.1 GENERIC MOMENT BASED BOUNDS AND IMPLIED CONFIDENCE SET. *Suppose f is a nondecreasing nonnegative function on $\mathbb{R}_+$, Z is a univariate real-valued (possibly latent) random variable, $d \geq 0$ and $f(d) > 0$. If there exists some positive constant $\bar{f}$ such that*

$$\mathbb{E} f(|Z|) \leq \bar{f} \tag{2.5.9}$$

then

$$Pr(|Z| \geq d) \leq \frac{\bar{f}}{f(d)}. \tag{2.5.10}$$

Furthermore, for every $\alpha \in (0,1)$,

$$Pr(Z \in [-d_\alpha, d_\alpha]) \geq 1 - \alpha \tag{2.5.11}$$

where d_α is defined on the extended real line as

$$d_\alpha \equiv d_\alpha\left(\bar{f}\right) = \inf\left\{d \in \mathbb{R}_{++} : f(d) \geq \frac{\bar{f}}{\alpha}\right\}. \tag{2.5.12}$$

As stated earlier, even if Z is unobserved or difficult to estimate, if we can find an upper bound $\bar{f}$ for $\mathbb{E} f(|Z|)$, we can construct a conservative confidence set for Z with any level.

In practice, the upper bound $\bar{f}$ has to be known or estimable. In general, if we denote the distribution of observables by F_W, possible construction of $\bar{f}$ would be to define

$$\bar{f} = \phi_f(F_W) \tag{2.5.13}$$

where $\phi(F_W)$ maps from any probability distribution to a positive real line. Note that Meyer (1974)'s bound can be obtained as a special case by setting for $W = (Y,X)$ with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$,

$$\begin{aligned} Z &= m(X) - \rho X, && (2.5.14)\\ f(x) &= x^2, && (2.5.15)\\ \bar{f} &= \mathrm{Var}(Y - \rho X). && (2.5.16) \end{aligned}$$

2.5.3. Unconditional bounds for nonlinear regression

Now, we move on to our particular problem where

$$Z = m(X) - P_{\boldsymbol{C}} m(X) \tag{2.5.17}$$

where $P_{\boldsymbol{C}} Y$ is defined in Definition 2.4.2 with $\boldsymbol{C} := \boldsymbol{C}(X)$. Due to the identity:

$$P_{\boldsymbol{C}} m(X) = P_{\boldsymbol{C}} Y, \tag{2.5.18}$$

we assume in this section that $P_{\boldsymbol{C}} m(x)$ for each $x \in \mathscr{X}$ is observed and defer inference on this object to a later section. The following proposition is pivotal in finding the upper $\bar{f}$ in Lemma 2.5.1

Proposition 2.5.2 BOUNDS FOR THE MOMENTS OF A CONVEX TRANSFORMATION OF THE APPROXIMATION ERROR. *Suppose $f:\ [0,\infty) \to [0,\infty)$ is a convex nonnegative function with $\mathbb{E} f(m(X) - P_{\boldsymbol{C}}(X)) < \infty$. Then,*

$$\mathbb{E} f(m(X) - P_{\boldsymbol{C}} m(X)) \leq \mathbb{E} f(Y - P_{\boldsymbol{C}} m(X)) \tag{2.5.19}$$

The following two corollary is immediate from Proposition 2.5.2

Corollary 2.5.3 CENTRAL MOMENT AND EXPONENTIAL INEQUALITIES.

1. *Suppose* $\mathbb{E}\left|m(X)-P_{\mathbf{C}}m(X)\right|^{\gamma}<\infty$ *for some* $\gamma\geq 1$. *Then,*

$$\mathbb{E}\left|m(X)-P_{\mathbf{C}}m(X)\right|^{\beta}\leq\mathbb{E}\left|Y-P_{\mathbf{C}}m(X)\right|^{\beta} \tag{2.5.20}$$

for any $\beta\in[1,\gamma]$.

2. *For a given* $t\in\mathbb{R}$, *suppose* $\mathbb{E}\left[\exp\left(t\left(m(X)-P_{\mathbf{C}}m(X)\right)\right)\right]<\infty$. *Then,*

$$\mathbb{E}\left[\exp\left(t\left(m(X)-P_{\mathbf{C}}m(X)\right)\right)\right]\leq\mathbb{E}\left[\exp\left(t\left(Y-P_{\mathbf{C}}m(X)\right)\right)\right]. \tag{2.5.21}$$

The following results follows from the properties of the conditional expectation.

Corollary 2.5.4 DECOMPOSITION OF VARIANCES. *Let* $P_{\mathbf{C}}m(X)$ *be an unbiased projection, that is*

$$\mathbb{E}\left[m(X)\right]=\mathbb{E}\left[P_{\mathbf{C}}m(X)\right]. \tag{2.5.22}$$

Then, the following decomposition of $\left(Var(Y),Var(m(X)),Var(Y-P_{\mathbf{C}}m(X))\right)$ *holds:*

$$\begin{aligned} Var(Y) &= Var(\varepsilon)+Var(m(X)) && (2.5.23)\\ Var(m(X)) &= Var(P_{\mathbf{C}}m(X))+Var(m(X)-P_{\mathbf{C}}m(X)) && (2.5.24)\\ Var(Y-P_{\mathbf{C}}m(X)) &= Var(\varepsilon)+Var(m(X)-P_{\mathbf{C}}m(X)). && (2.5.25)\end{aligned}$$

For example, if $\mathbf{C}$ is a subspace and includes a nonzero constant as a base function, then $P_{\mathbf{C}}m(X)$ is unbiased. Corollary 2.5.4 is well-known, but 1. of Corollay 2.5.3 is more general and states that the moments of the approximation error of any order greater than 1 are bounded by the moments of the observed difference $Y-P_{\mathbf{C}}$ of the same order. According to 2., such inequality also holds for an exponential bound.

Lemma 2.5.1 combined with Proposition 2.5.2 implies a simple method of constructing a confidence set for $m(X)$.

Proposition 2.5.5 UNCONDITIONAL BOUND FOR NONLINEAR REGRESSION. *(i) Suppose* $f:\ [0,\infty)\to(0,\infty)$ *is convex and nondecreasing with* $\mathbb{E}f\left(\left|m(X)-P_{C(X)}Y\right|\right)<\infty$.

Then, for any $c > 0$ such that $f(c) > 0$,

$$Pr(|m(X) - P_{\mathbf{C}}m(X)| \geq c) \leq \frac{\mathbb{E}f(|Y - P_{\mathbf{C}}m(X)|)}{f(c)}. \tag{2.5.26}$$

Furthermore, for any $\alpha \in (0,1)$

$$m(X) \in \left[P_{\mathbf{C}(X)}Y - c_\alpha, P_{\mathbf{C}(X)}Y + c_\alpha\right] \tag{2.5.27}$$

where c_α is defined on the extended real line as

$$c_\alpha = \inf\left\{c \in \mathbb{R}_{++} : f(c) = \frac{\mathbb{E}f(|Y - P_{\mathbf{C}}m(X)|)}{\alpha}\right\} \tag{2.5.28}$$

with probability at least $1 - \alpha$.

Note that if f is strictly increasing and unbounded, then

$$c_\alpha = f^{-1}\left(\frac{\mathbb{E}f(|Y - P_{\mathbf{C}}m(X)|)}{\alpha}\right). \tag{2.5.29}$$

In particular, if for $\beta \geq 1$, $\mathbb{E}|m(X) - P_{\mathbf{C}}m(X)|^\beta < \infty$ then

$$\Pr(|m(X) - P_{\mathbf{C}}m(X)| \geq c) \leq \frac{\mathbb{E}|Y - P_{\mathbf{C}}m(X)|^\beta}{c^\beta} \tag{2.5.30}$$

so that to achieve a desired coverage probability $1 - \alpha$ ($\alpha \in (0,1)$), we set

$$c_\alpha = \frac{E|Y - P_{\mathbf{C}}m(X)|^\beta}{\alpha}^{1/\beta}. \tag{2.5.31}$$

Similarly, since the finite sum of convex functions is also convex, it is also possible to construct a bound based on a polynomial equation of the absolute moments by considering

$$f(|x|) = \sum_{i=1}^{k} c_i |x|^i \quad , c_i \geq 0 \ \forall i, \tag{2.5.32}$$

Observe further that this bound and the induced confidence set are valid for any convex,

nondecreasing, and nonnegative function f . However, the choice of the function f matters for the volume of the confidence set. For example, as pointed by Dufour and Hallin (1992), the usual Cherbyshev's inequality based on the second moment can be significantly improved by employing higher order moments. Then, it is sensible to choose the best f, i.e. the one that implies the tightest bound. The next theorem formalizes the intuition.

Proposition 2.5.6 UNIFIED UNCONDITIONAL BOUND FOR NONLINEAR REGRESSION. *Let $\mathscr{H}_{\wedge}^{+}$ be the class of all convex, nondecreasing, and nonnegative functions from $\mathbb{R}_{+}$ to $\mathbb{R}_{+}$. For any subclass $\mathscr{K}$ of $\mathscr{H}_{\wedge}^{+}$ and any constant $c>0$,*

$$Pr\left(\left|m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right|\geq c\right)\leq\inf_{f\in\mathscr{K}}\frac{\mathbb{E}f\left(\left|Y-P_{\mathbf{C}}m\left(X\right)\right|\right)}{f\left(c\right)} \tag{2.5.33}$$

Hence, for any $\alpha\in(0,1)$,

$$m\left(X\right)\in\left[P_{\mathbf{C}}m\left(X\right)-c_{\alpha}^{(\mathscr{K})},P_{\mathbf{C}}m\left(X\right)+c_{\alpha}^{(\mathscr{K})}\right] \tag{2.5.34}$$

where

$$c_{\alpha}^{(\mathscr{K})}=\inf_{f\in\mathscr{K}}c_{\alpha,f} \tag{2.5.35}$$

with $c_{\alpha,f}$ is defined on the extended real line as

$$c_{\alpha,f}=\inf\left\{c\in\mathbb{R}_{++}:f\left(c\right)=\frac{\mathbb{E}f\left(\left|Y-P_{\mathbf{C}}m\left(X\right)\right|\right)}{\alpha}\right\}. \tag{2.5.36}$$

For example, setting $\mathscr{K}$ as

$$\mathscr{K}_{d}\equiv\left\{f_{\beta}:\mathbb{R}_{+}\rightarrow\mathbb{R}_{+}|f_{\beta}\left(x\right)=\left(x-d\right)^{\beta},\beta\geq1\right\} \tag{2.5.37}$$

for a given constant $d\in\mathbb{R}$ yields the shifted moment bound with shift parameter d. The "moment bound" is obtained as a special case where $d=0$ and

$$c_{\alpha}^{(\mathscr{K}_{0})}=\inf_{\beta\geq1}\left(\frac{\mathbb{E}\left|Y-P_{C(X)}Y\right|^{\beta}}{\alpha}\right)^{1/\beta} \tag{2.5.38}$$

We note that Philips and Nelson (1995) showed that the moment-based bound is tighter

than the exponential bound and thus is preferable in practice.

2.5.4. Optimality properties of the approximation bounds: sharpness and honesty

In this section, we examine optimality and uniformity properties of the approximation bounds and associated confidence sets for $m(X)$ proposed in Section 2.5.3 under the criterion of tightness and sharpness. A bound c_α for given level $\alpha \in (0,1)$ is said to be sharp if it cannot be improved, i.e. c_α is a constant such that

$$\sup_{F \in \mathscr{F}} \Pr\left(|m(X) - P_{\mathbf{C}} m(X)| \geq c_\alpha\right) = \alpha. \tag{2.5.39}$$

where $\mathscr{F}$ is a family of probability distributions of (Y,X). As a consequence, the confidence set

$$[P_{\mathbf{C}} m(X) - c_\alpha, P_{\mathbf{C}} m(X) + c_\alpha] \tag{2.5.40}$$

with the tight bound c_α has size $1-\alpha$. In Proposition 2.5.7, we show that the central moments based bound in (2.5.38). satisfies the tightness condition (2.5.39) under

$$\mathscr{F} = \{F_{Y,X} : \ \mathbb{E}|Y| < \infty\}. \tag{2.5.41}$$

Proposition 2.5.7 TIGHTNESS OF THE CENTRAL MOMENTS BASED BOUNDS. *Let $c_\alpha^{(\mathscr{K}_0)}$) be the central moment based bound in (*2.5.38*). Then, $c_\alpha^{(\mathscr{K}_0)}$ is tight, i.e.*

$$\sup_{F \in \mathscr{F}} Pr_F\left(|m(X) - P_{\mathbf{C}} m(X)| \geq c_\alpha^{(\mathscr{K}_0)}\right) = \alpha. \tag{2.5.42}$$

This result states that it is not possible to improve the bound $c_\alpha^{(\mathscr{K}_0)}$ without imposing restrictions on the data generating process. In this sense, $c_\alpha^{(\mathscr{K}_0)}$ is the best approximation bound for the approximate model $P_{\mathbf{C}} m(X)$ for $m(X)$ under the family of distributions $\mathscr{F}$. Another desirable property of a bound is uniformity; the bound is uniform over some family of regression functions. We employ the notion of honesty, first posed by Li (1989), which requires that the coverage probability holds uniformly over some class $\mathscr{M}$ of regression

functions m. To this end, consider the space $\mathscr{F}(F_X, F_Y)$ of $F_{Y,X}$ given F_X and F_Y:

$$\mathscr{F}(F_X, F_Y) = \{F_{Y,X} : \mathbb{E}|Y| < \infty, F_X \text{ and } F_Y \text{ are distributions of } X \text{ and } Y, \text{ respectively}\} \subset \mathscr{F} \tag{2.5.43}$$

and then define a family $\mathscr{M}(F_X, F_Y)$ of regression functions consistent with the marginal distributions (F_X, F_Y):

$$\mathscr{M}(F_X, F_Y) = \{m(\cdot\,; F_{Y,X}) : m(\cdot\,; F_{Y,X}) \text{ is the regression function of } Y \text{ given } X \text{ under } F_{Y,X} \in \mathscr{F}(F_X, F_Y)\}. \tag{2.5.44}$$

Note that $\mathscr{M}(F_X, F_Y)$ is considerably large since the marginal distributions F_Y, F_X put no restriction on the dependence structure of Y and X. The following result states that any approximation bound based on the observed difference $Y - P_{\mathbf{C}} m(X)$ is uniform over $\mathscr{M}(F_X, F_Y)$ and thus honest.

Proposition 2.5.8 HONESTY OF THE APPROXIMATION BOUNDS BASED ON THE OBSERVED DIFFERENCE. *Consider the approximation bound $c_{\alpha,f}$ constructed as in Proposition 2.5.5 given a function f which satisfies the requirements in the proposition. Then,*

$$\sup_{m \in \mathscr{M}(F_X, F_Y)} Pr\left(|m(X) - P_{\mathbf{C}} m(X)| \geq c_{\alpha,f}\right) \leq \alpha. \tag{2.5.45}$$

Armstrong and Kolesár (2020) propose a honest confidence interval for a scalar parameter of interest, such as a regression function evaluated at a given point. Their results are restricted to some class of smooth regression functions (e.g. a class of Hölder functions) while ours are applicable without any smoothness conditions to a general class of regression functions.

2.5.5. *Unconditional bounds: Single Tailed

One may be interested in obtaining only an upper or lower bound for $m(X)$. There is a single-tailed version of Chebyshev's inequality available known as Cantelli's inequality (Rao (1973)), whose generalization is considered in the following lemma.

Lemma 2.5.9 GENERALIZED CANTELLI'S INEQUALITY. *Let Z be a real-valued ran-*

dom variable. Given $c \in \mathbb{R}$, suppose $\mathscr{H}_{ND(\geq c)}$ is some family of functions such that

$$\mathscr{H}_{ND(\geq c)} = \{h : \mathbb{R} \to \mathbb{R}_+ \,|\, h \text{ is nondecreasing for } \forall x \geq c \text{ and } h(c) > 0\}. \tag{2.5.46}$$

Then,

$$Pr(Z \geq c) \leq \inf_{h \in \mathscr{H}_{ND(\geq c)}} \frac{\mathbb{E}h(Z)}{h(c)}. \tag{2.5.47}$$

Similarly, suppose $\mathscr{H}_{NI(\leq c)}$ is some family of functions such that

$$\mathscr{H}_{NI(\leq c)} = \{h : \mathbb{R} \to \mathbb{R}_+ \,|h \text{ is nonincreasing for } \forall x \leq c \text{ and } h(c) > 0\}. \tag{2.5.48}$$

Then,

$$Pr(Z \geq c) \geq 1 - \sup_{h \in \mathscr{H}_{ND(\geq c)}} \frac{\mathbb{E}h(Z)}{h(c)}. \tag{2.5.49}$$

The original Cantelli's inequality can be obtained by letting $\mathbb{E}Z = 0$ and setting $\mathscr{H}_{ND(\geq c)}$ for $c > 0$ as

$$\left\{h_\lambda : \mathbb{R} \to \mathbb{R}_+ \,|\, h_\lambda(z) = (z+\lambda)^2, \lambda \geq 0\right\} \tag{2.5.50}$$

and $\mathscr{H}_{NI(\leq c)}$ for $c < 0$ as

$$\left\{h_\lambda : \mathbb{R} \to \mathbb{R}_+ \,|\, h_\lambda(z) = (-z-\lambda)^2, \lambda \geq 0\right\}. \tag{2.5.51}$$

Similarly, the Chernoff bounds are obtained by setting $\mathscr{H}_{ND(\geq c)}$ as

$$\{h_t : \mathbb{R} \to \mathbb{R}_+ | h_t(z) = \exp(t\lambda), \lambda > 0\} \tag{2.5.52}$$

and $\mathscr{H}_{NI(\leq c)}$ as

$$\{h_t : \mathbb{R} \to \mathbb{R}_+ | h_t(z) = \exp(-t\lambda), \lambda > 0\}. \tag{2.5.53}$$

Theorem 2.5.10 UNCONDITIONAL BOUND FOR NONLINEAR REGRESSION: SINGLED TAILED. *Given $c \in \mathbb{R}$, suppose $\mathscr{H}_{ND(\geq c),\vee}$ is some family of functions such that*

$$\mathscr{H}_{ND(\geq c),\vee} = \{h : \mathbb{R} \to \mathbb{R}_+ \,|\, h \text{ is convex and is nondecreasing for any } x \geq c \text{ with } h(c) > 0\}. \tag{2.5.54}$$

For any subclass $\mathscr{K}$ of $\mathscr{H}_{ND(\geq c),\vee}$,

$$Pr\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\geq c\right)\leq\inf_{h\in\mathscr{K}}\frac{\mathbb{E}h\left(\left|Y-P_{\mathbf{C}}m\left(X\right)\right|\right)}{h\left(c\right)} \tag{2.5.55}$$

Hence, for any $\alpha\in(0,1)$, a confidence interval for $m(X)$ is given by

$$\left[-\infty,P_{\mathbf{C}}m\left(X\right)+c_{\alpha}^{(\mathscr{K})}\right] \tag{2.5.56}$$

where

$$c_{\alpha}^{(\mathscr{K})}=\inf_{h\in\mathscr{K}}c_{\alpha,h} \tag{2.5.57}$$

with $c_{\alpha,f}$ is defined on the extended real line as

$$c_{\alpha,h}=\inf\left\{c\in\mathbb{R}_{++}:h\left(c\right)=\frac{\mathbb{E}h\left(\left|Y-P_{\mathbf{C}}m\left(X\right)\right|\right)}{\alpha}\right\}. \tag{2.5.58}$$

Theorem 2.5.10 implies a usual Cantelli-type inequality:

$$\Pr\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\geq c\right)\leq\frac{\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)}{\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)+c^{2}}. \tag{2.5.59}$$

2.6. Conditional bounds for nonlinear regression

In Section 2.5, the approach is unconditional in the sense that our object of interest $\mathbb{E}\left[Y|X\right]$ is investigated under no restriction on a random variable X. It is also of particular interest to make inference conditionally on X being in an event A or to compare the mean effects of X on Y conditional on X belonging in two different events A and B due to the following observation: identification of and inference on $m\left(x_0\right)$ at a fixed point $x_0\in\mathscr{X}$ are feasible only under fairly strong restrictions on the functional form of $m\left(X\right)$ and the distribution of (Y,X). On the other hand,suppose A is a set including x_0 as its element. If A is sufficiently small so that $g\left(x\right)$ does not vary so much on A, conditioning on the set A delivers useful information on $m\left(x_0\right)$, $x_0\in A$. Hence, set-conditioning bridges the gap between the unconditional and fully-conditional (at a point) approaches. Then, it is natural to introduce truncated random variables whose domains are restricted to a set of interest. It turns out that the framework proposed in Section 2.5 can be extended to achieve counterparts of bounds

proposed in Proposition 2.5.6 by deliberately redefining the sample space.

Definition 2.6.1 TRUNCATED RANDOM VARIABLES. *For a Borel measurable set $A \subset \mathscr{X}$, define a pair of random variables $(Y_A,\, X_A)$, truncated random variables of $(Y,\, X)$ with the domain of X restricted to A. Hence, the distribution of $(Y_A,\, X_A)$ is the conditional distribution of $(Y,\, X)$ given $X \in A$: i.e. $Pr((Y_A,X_A) \leq (y,x)) = Pr((Y,X) \leq (y,x)\,|X \in A)$.*

Lemma 2.6.1 *Suppose $A \in \mathscr{B}(\mathscr{X})$ and $(Y_A,\, X_A)$ is defined as in Definition 2.6.1. Define the conditional expectation of Y_A given $X_A = x$ $(x \in A)$ as*

$$m_A(x) \equiv \mathbb{E}[Y_A\,|X_A = x]. \tag{2.6.1}$$

Then,

$$Pr(\omega \in X^{-1}(A)\,|m_A(X_A(\omega)) \neq m(X_A(\omega))) = 0 \tag{2.6.2}$$

Intuitively, Lemma 2.6.1 states that the functional form of the conditional expectation remains invariant to restriction of the domain of X. This means that we can write $m_A(X_A) = m(X_A)$ for a random variable X_A and $m_A(x) = m(x)$ for any $x \in A$. Hence, the local properties of $m(X)$ on the set A can be studied though (Y_A,X_A) due to this functional invariance of the conditional expectation with respect to a set-conditioning. Further, define (Y_B,X_B) and m_B similarly and suppose that m_A and m_B do not change "too much" within their domains. Then, $m_A(X_A) - m_B(X_B)$ should be a good predictor of $m_A(x_a) - m_B(x_b)$ where x_a, x_b are some fixed (instead of random) elements of A and B, respectively. The implication here is that our conditional approach makes it possible to attain good approximation of $\mathbb{E}[Y|X = x_a] - \mathbb{E}[Y|X = x_b]$ for any given x_a, $x_b \in \mathscr{X}$ by considering two sets A, B, each containing x_a, x_b. The projection operator $P_{\mathscr{L}}$ in the unconditional setting (Definition 2.4.2) was characterized by its orthogonality to the residual in the covariance sense. We introduce the notion of a local projection operator, which satisfies the same orthogonality conditions but locally on A.

Definition 2.6.2 LOCAL PROJECTION OPERATOR. *Given a set $A \subset \mathscr{X}$, $P_{\mathscr{L}_A}(X)$ is said to be a local projection of Y_A onto a subspace $\mathscr{L}_A$ generated by the transformation of X_A if $P_{\mathscr{L}_A}$ satisfies*

$$Cov(W, P_{\mathbf{C}} m(X_A)) = 0 \quad ,\forall W \in \mathscr{L}_A \tag{2.6.3}$$

We say $P_{C(X_A)}Y_A$ is unbiased if

$$\mathbb{E}\left[P_{\mathbf{C}}m\left(X\right)\right]=\mathbb{E}\left[m\left(X\right)|X\in A\right] \tag{2.6.4}$$

Now, it is clear that the results in Section 2.5 remain valid when (Y,X), $P_{C(X)}$ are replaced by (Y_A,X_Y) and $P_{C(X_A)}$. To avoid repetition, we only show our conditional bounds for nonlinear regression; those are direct applications of Proposition 2.5.6.

Proposition 2.6.2 CONDITIONAL BOUND FOR NONLINEAR REGRESSION. *Let $\mathscr{H}_{\wedge}^{+}$ be the class of all convex, nondecreasing, and nonnegative functions from $\mathbb{R}_+$ to $\mathbb{R}_+$. For any subclass $\mathscr{K}$ of $\mathscr{H}_{\wedge}^{+}$ and any constant $c>0$,*

$$Pr\left(\left|m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right|\geq c\,|\,X\in A\right)\leq\inf_{f\in\mathscr{K}}\frac{\mathbb{E}f\left(\left|Y_A-P_{\mathbf{C}}m\left(X_A\right)\right|\right)}{f\left(c\right)} \tag{2.6.5}$$

Hence, for any $\alpha\in(0,1)$, conditional on $[X\in A]$, a confidence interval for $m(X)$ is given by

$$\left[-\infty,P_{\mathbf{C}}m\left(X_A\right)+c_{\alpha,A}^{(\mathscr{K})}\right] \tag{2.6.6}$$

where

$$c_{\alpha,A}^{(\mathscr{K})}=\inf_{f\in\mathscr{K}}c_{\alpha,f} \tag{2.6.7}$$

with $c_{\alpha,f}$ defined on the extended real line as

$$c_{\alpha,f}=\inf\left\{c\in\mathbb{R}_{++}\,:\,f\left(c\right)\geq\frac{\mathbb{E}f\left(\left|Y_A-P_{\mathbf{C}}m\left(X_A\right)\right|\right)}{\alpha}\right\}. \tag{2.6.8}$$

The following result provides an alternative conditional bound by exploiting the maximum and minimum values of the approximate model $P_{\mathbf{C}}m(\cdot)$ in the set A.

Theorem 2.6.3 *Conditional on $[X\in A]$, a confidence interval of $m\left(X\right)$ with level $1-\alpha$ is given by*

$$\left[\inf_{x\in A}P_{\mathbf{C}}m\left(x\right)-c_{\alpha,A},\sup_{x\in A}P_{\mathbf{C}}m\left(x\right)+c_{\alpha,A}\right] \tag{2.6.9}$$

where

$$c_{\alpha,A}=\sqrt{\frac{Var\left(Y-P_{\mathbf{C}}m\left(X_A\right)\,|\,X\in A\right)}{\alpha}} \tag{2.6.10}$$

Theorem 2.6.3 states that even if point-wise approximation errors $m(x) - P_{\mathbf{C}(X_A)}Y_A$ for each point $x \in A$ are not known, the maximum and minimum values of $P_A(x)$ combined with the variation $\text{Var}(Y - P_{\mathbf{C}}m(X_A)|X \in A)$ are useful in predicting $m(X)$ conditional on a random variable X being in A. In practice, the confidence intervals proposed in Theorem 2.6.2 and 2.6.3 are infeasible and the local projection $P_{\mathbf{C}}m(X_A)$ and the constant $c_{\alpha,A}$ have to be estimated. We will discuss these matters in Section 2.8.

2.7. Alternative bounds under continuity

2.7.1. Implications of continuity of the regression function

In this section, we discuss how shape restrictions on the regression function $m(\cdot)$ can be utilized in our framework. While approximation bounds proposed in Section 2.5 and 2.6 do not impose any restrictions on $m(\cdot)$, we show that additional information which restricts the class of functions that $m(\cdot)$ belongs to can be readily incorporated in our framework and often leads to sharper bounds than those based on the observed variation $Y - P_{\mathscr{C}}(X)$. Shape restrictions can be also imposed on approximation error $(m - P_C)(\cdot)$, for example, when discontinuity in $m(\cdot)$ is taken into account in the approximate model $P_C(\cdot)$ chosen by the practitioner so that the approximation error possesses a certain continuity property even if $m(\cdot)$ doesn't. To motivate further discussions, consider the Chebyshev-type approximation bound:

$$\Pr(|m(X) - P_{\mathbf{C}}m(X)| \geq c) \leq \frac{\text{Var}(m(X) - P_{\mathbf{C}}m(X))}{f(c)} \leq \frac{\text{Var}(Y - P_{\mathbf{C}}m(X))}{f(c)} \tag{2.7.11}$$

Then, Proposition 2.5.7 implies that without any restrictions on $m(\cdot)$, the second inequality in (2.7.11) is sharp, i.e. there exists some data generating process such that

$$\text{Var}(m(X) - P_{\mathbf{C}}m(X)) = \text{Var}(Y - P_{\mathbf{C}}m(X)). \tag{2.7.12}$$

On the other hand, if ε is large relative to $m(X) - P_{\mathscr{L}}(X)$ in the true data generating process, then the variance of the observed difference $\text{Var}(Y - P_{\mathbf{C}}m(X))$ can be much larger than the variance of the approximation error $\text{Var}(m(X) - P_{\mathbf{C}}m(X))$ so that the resulting confidence set can be too wide for practical purposes. Economics theory often does not fully pin

down $m(\cdot)$ but implies informative restrictions on the shape of $m(\cdot)$, for example continuity, monotonicity, and concavity. In particular, we will see that under appropriate smoothness conditions on $m(\cdot)$, a sharper bound than the second inequality in (2.7.11) is available. In the next section, we study an approximation bound implied by continuity of the regression function $m(\cdot)$.

2.7.2. Chebyshev's-type bound under continuity of m

In Corollary 2.5.4, we observed the decomposition:

$$\mathrm{Var}\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right)=\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)-\mathrm{Var}\left(\varepsilon\right) \tag{2.7.13}$$

if $P_{\mathscr{C}}\left(X\right)$ is unbiased for Y, i.e. $\mathbb{E}\left[P_{\mathbf{C}}m\left(X\right)\right]=E\left[Y\right]$. Suppose the variance $\mathrm{Var}\left(\varepsilon\right)$ of expectation error is known or can be consistently estimated. Since $\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)$ only involves the moments of (Y,X) and does not depend on the unknown infinite-dimensional parameter m, it is identifiable and may be easily estimated. Then, the variance $\mathrm{Var}\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right)$ of approximation error can be identified from the obvious relation:

$$\mathrm{Var}\left(m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right)=\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)-\mathrm{Var}\left(\varepsilon\right). \tag{2.7.14}$$

In such a case, one can simply apply Chebyshev inequality:

$$\Pr\left(\left|m\left(X\right)-P_{\mathbf{C}}m\left(X\right)\right|>c\right)\leq\frac{\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)-\mathrm{Var}\left(\varepsilon\right)}{c^{2}} \tag{2.7.15}$$

and obtain an approximation bound for level $\alpha\in(0,1)$:

$$\tilde{c}_{\alpha}^{(2)}=\sqrt{\frac{\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)-\mathrm{Var}\left(\varepsilon\right)}{\alpha}}. \tag{2.7.16}$$

This bound improves the Chebyshev's-type bound based on the observed difference $Y-P_{\mathbf{C}}m\left(X\right)$:

$$c_{\alpha}^{(2)}=\sqrt{\frac{\mathrm{Var}\left(Y-P_{\mathbf{C}}m\left(X\right)\right)}{\alpha}} \tag{2.7.17}$$

as $\tilde{c}_\alpha^{(2)} < c_\alpha^{(2)}$ unless $Y = P_{\mathbf{C}} m(X)$ almost everywhere. Note that when $\mathrm{Var}(\varepsilon)$ is identified, $c_\alpha^{(2)}$ is not tight anymore in the sense of Proposition 2.5.7 and improvement is possible. We show below that by restricting the class of regression functions to be a continuous function class, this additional information regarding the moment of the expectation error ε can be obtained. We also require the following regularity conditions.

Assumption 2.7.1 CONDITIONAL DENSITY OF ε. *The conditional density $f_{Y|X=x}$ of Y given $X = x$ with respect to the Lebesgue measure exists and $f_{Y|X=x}$ is continuous with respect to x uniformly on $\mathscr{X}$. Furthermore, the density f_X of X exists.*

Assumption 2.7.2 BOUNDED CONDITIONAL VARIANCE OF ε. *There exists some positive constant $\bar{C} < \infty$ such that $Var(\varepsilon|X = x) < \bar{C}$ for any $x \in \mathscr{X}$.*

Given these assumptions, we have the following decoupling representation of $\mathrm{Var}(\varepsilon)$.

Lemma 2.7.1 DECOUPLING REPRESENTATION OF THE ERROR VARIANCE. *Suppose Assumptions 2.7.1 and 2.7.2 hold and the true model $m(x)$ belongs to a family $\mathscr{M}_C$ of continuous regression functions defined as*

$$\mathscr{M}_C = \{\tilde{m} : \mathscr{X} \to \mathscr{Y} \,|\, \tilde{m} \text{ is continuous and } \mathbb{E}[Y - \tilde{m}(x) \,|\, X = x] = 0, \quad \forall x \in \mathscr{X}\}. \tag{2.7.18}$$

Then,

$$\begin{aligned} Var(\varepsilon) &= \frac{1}{2}\int_{\mathscr{X}}\int_{\mathscr{Y}}\int_{\mathscr{Y}} (y_1 - y_2)^2 \, dF_{Y|X=x_0}(y_1)\, dF_{Y|X=x_0}(y_2)\, dF_X(x_0) \\ &= \frac{1}{2}\lim_{\delta\to 0}\int_{\mathscr{X}}\int_{\mathscr{Y}}\int_{\mathscr{Y}} (y_1 - y_2)^2 \, dF_{Y|X=x_0+\delta}(y_1)\, dF_{Y|X=x_0}(y_2)\, dF_X(x_0) \end{aligned} \tag{2.7.19}$$

In Section 2.8, we consider the estimation of the bound $\tilde{c}_\alpha^{(2)}$ in (2.7.16) under continuity. When X is continuous, any observation of X takes a distinct value with probability one. The second representation of $\mathrm{Var}(\varepsilon)$ implies useful approximation of the first representation and leads to a difference-based estimation method for $\mathrm{Var}(\varepsilon)$.

2.8. Inference

This section discusses inference for confidence sets for nonparametric regression $m(X)$ considered in Section 2.5. The overview of the inference framework is provided in Section 2.8.1. Section 2.8.2 considers inference for a parametric approximate model for $m(\cdot)$ in the extremum estimation framework. In Section 2.8.3, estimators of approximation bounds proposed in Section 2.5 and 2.7.2 are provided and conditions for consistency are given. Section 2.8.4 provides a feasible confidence set for $m(X)$ given results in Section 2.8.2 and 2.8.3.

2.8.1. Overview: estimation of the proposed confidence sets

We provide here an overview of succeeding results. Recall that the proposed (two-sided) confidence set for nonparametric regression $\mathbb{E}[Y|X]$ with level $1-\alpha$ ($\alpha \in (0,1)$) is of the form:

$$CS_{1-\alpha}(X; P_{\mathbf{C}}, c_\alpha) := [P_{\mathbf{C}}(X) - c_\alpha, P_{\mathbf{C}}(X) + c_\alpha]. \tag{2.8.20}$$

where $P_{\mathbf{C}}(\cdot)$ is any approximate model of m and c_α is a positive constant which depends on the distribution of (Y,X) such that

$$\Pr(m(X) \in CS_{1-\alpha}(X; P_{\mathbf{C}}, c_\alpha)) \geq 1-\alpha. \tag{2.8.21}$$

We assume that the model $P_{\mathbf{C}}$ belongs to some function class $\mathscr{C}$ indexed by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^q$ where Θ is a subset of $\mathbb{R}^q$:

$$\mathscr{C} := \mathscr{C}(\Theta) = \{h(\cdot;\theta) : \theta \in \Theta\} \tag{2.8.22}$$

and there is some unique $\theta_0 \in \Theta$ such that $P_{\mathbf{C}} = h(\cdot;\theta_0)$ and

$$\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}\left[(m(X) - h(X;\,\theta))^2\right]. \tag{2.8.23}$$

In this framework, the inference problem involving the approximate model $P_{\mathbf{C}}$ reduces to inference on a finite-dimensional parameter θ_0. Since $P_{\mathbf{C}}(\cdot)$ is not assumed to be correctly specified, θ_0 can be interpreted as a "pseudo true value" of θ which minimizes the mean square error with respect to $m(X)$ (or equivalently minimizes the Kullback–Leibler diver-

gence when a Gaussian density with fixed variance is employed as a pseudo density (White (1982)). We consider inference for θ_0 in the extremum estimation framework (Amemiya (1985), Newey and McFadden (1994), van der Vaart (1998)).

Under some regularity conditions that for a fixed value $x \in \mathscr{X}$, we have

$$\sqrt{n}\left(h\left(x;\hat{\theta}_n\right) - h\left(x;\theta_0\right)\right) \xrightarrow{d} \mathscr{N}\left(0, \sigma_h^2(x)\right) \tag{2.8.24}$$

for some function $\sigma_h(x) : \mathscr{X} \to \mathbb{R}_{++}$. Then, conditional on X ,

$$\sqrt{n}\left(h\left(X;\hat{\theta}_n\right) - h\left(X;\theta_0\right)\right) \xrightarrow{d} \mathscr{N}\left(0, \sigma_h^2(X)\right) \tag{2.8.25}$$

as long as the dependence between $\hat{\theta}_n$ and X is negligible asymptotically. Thus, a confidence set for the approximate model $h(X;\theta_0)$ with asymptotic size $1-\alpha_1$ ($\alpha_1 \in (0,1)$) is given by

$$CS_{h,n} := \left[h\left(X;\hat{\theta}_n\right) - d_{\alpha_1,n}(X), h\left(X;\hat{\theta}_n\right) + d_{\alpha_1,n}(X)\right] \tag{2.8.26}$$

where

$$d_{\alpha_1,n}(x) := \frac{\hat{\sigma}_h(x)}{\sqrt{n}} q_{1-\alpha_1/2} \tag{2.8.27}$$

and $\hat{\sigma}_h(x)$ is a consistent estimator of $\sigma_h(x)$,and $q_{1-\alpha_1/2}$ is the $100(1-\alpha_1/2)$ percentile of the standard normal distribution. Then, we show that for any of the bounds considered in Section 2.5. and Section 2.7.2, there exists a estimator $\hat{c}_{\alpha_2}$ of c_{α_2} for $\alpha_2 \in (0,1)$:

$$\hat{c}_{\alpha_2} \xrightarrow{p} c_{\alpha_2}, \tag{2.8.28}$$

Then, if (α_1, α_2) are chosen so that $\alpha = 1-(1-\alpha_1)(1-\alpha_2)$, a confidence set for $m(X)$ with asymptotic level $1-\alpha$ is given by

$$CS_{1-\alpha,n}(X;\alpha_1,\alpha_2) := \left[h\left(X;\hat{\theta}_n\right) - \hat{D}(X;\alpha_1,\alpha_2), h\left(X;\hat{\theta}_n\right) + \hat{D}(X;\alpha_1,\alpha_2)\right]. \tag{2.8.29}$$

where

$$\hat{D}(x;\alpha_1,\alpha_2) = d_{\alpha_1,n}(X) + \hat{c}_{\alpha_2}. \tag{2.8.30}$$

Note that inference for conditional confidence sets proposed in Section 2.6 is analogous

to the unconditional case considered here. Hence, results derived for the unconditional case are readily applicable once the estimation procedure is adapted for a truncated random vector (X_A, Y_A) defined as in Definition 2.6.1 for a set of interest, $A \subset \mathscr{X}$. Then, given a sample set $\{Y_i,\, X_i\}_{i=1}^n$, we generate a purposive sample $\left\{\left(Y_{A,,j}, X_{A,j}\right)\right\}_{j=1}^{n_A}$ by selecting elements of $\{Y_i,\, X_i\}_{i=1}^n$ if and only if $X_i \in A$ to estimate an local approximate model and bounds.

2.8.2. Inference on approximate models

The previous section considers inference for the approximate model $P_{\mathbf{C}}(\cdot) = h(\cdot;\theta_0)$ where θ_0 is identified as an unique element of Θ such that

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathbb{E}\left[(m(X) - h(X;\,\theta))^2\right]. \tag{2.8.31}$$

More generally, it can be assumed that θ_0 is a unique maximizer of some objective function of θ as follows:

Assumption 2.8.1 PARAMETER IDENTIFIED AS A UNIQUE MAXIMIZER OF AN OBJECTIVE FUNCTION. *For a family of parametric models $\boldsymbol{C}_{\Theta} := \{h(\cdot;\theta) : \theta \in \Theta\}$, define an objective function $Q : \Theta \to \mathbb{R}$. Then, there exists a unique element $\theta_0 \in \Theta$ such that*

$$\theta_0 = \arg\max_{\theta\in\Theta} Q(\theta) \tag{2.8.32}$$

A natural choice in our case is

$$Q(\theta) = -\mathbb{E}\left[(Y - h(X;\theta))^2\right]. \tag{2.8.33}$$

Clearly, the parameter θ_0 is identified if

$$\mathbb{E}\left[(Y - h(X;\theta_0)^2\right] < \inf_{\Theta\setminus\{\theta_0\}} \mathbb{E}\left[(Y - h(X;\theta))^2\right]. \tag{2.8.34}$$

We emphasize that identification of θ_0 is in general a separate issue from identification of $m(\cdot)$ and, in particular, the latter is not necessary for the former. Given some sample objective function Q_n (typically sample analogue of Q), we consider an estimator $\hat{\theta}_n$ of θ_0

as the minimizer of Q_n

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta). \tag{2.8.35}$$

Consistency and asymptotic normality results for $\hat{\theta}_n$ are widely available in the literature (Amemiya (1985), Newey and McFadden (1994), van der Vaart (1998)); we reproduce the results by Hayashi (2011) in the appendix for completeness.

Then, under the assumptions in Lemma 2.B.1, we have

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \Lambda) \tag{2.8.36}$$

where Λ is defined in 2.B.22.

The next lemma provides an asymptotically valid confidence set for $h(x;\hat{\theta}_n)$ where x is a fixed point of $x \in \mathcal{X}$.

Lemma 2.8.1 *For a given $x \in \mathcal{X}$, suppose $h(x;\theta)$ is continuously differentiable with respect to θ in an open neighborhood of θ_0 and the derivative $\partial_\theta h(x;\theta_0)$ is nonzero. Further, suppose that there is a consistent estimator $\hat{\Lambda}_n$ of Λ. Then, for $\alpha \in (0,1)$, the random set*

$$CS_{h,n} := \left[h(x;\hat{\theta}_n) - d_{\alpha,n}(x), h(x;\hat{\theta}_n) + d_{\alpha,n}(x)\right] \tag{2.8.37}$$

where

$$d_{\alpha,n}(x) = \frac{\hat{\sigma}_h(x)}{\sqrt{n}} q_{1-\alpha_1/2} \tag{2.8.38}$$

and

$$\hat{\sigma}^2_{h,x} = (\nabla_\theta h(x;\hat{\theta}_n))' \hat{\Lambda}_n (\nabla_\theta h(x;\hat{\theta}_n)) \tag{2.8.39}$$

is a confidence interval for $h(x;\theta_0)$ with asymptotic size $1-\alpha$.

However, in order to estimate a confidence set for $m(X)$ where X is random, we need to establish a confidence set for $h(X;\hat{\theta}_n)$. We accommodate the case where $\hat{\theta}_n$ and X are not independent, e.g. when the same data are used to $\hat{\theta}_n$ and to evaluate $h(\cdot;\hat{\theta}_n)$ as long as the influence of X on $\hat{\theta}_n$ is asymptotically negligible. Such a condition is stated in terms of the notion of asymptotic independence.

Definition 2.8.1 ASYMPTOTIC INDEPENDENCE. *Suppose $\{X_n\}$ and $\{Y_n\}$ are convergent sequences of random elements defined on measurable spaces $(\mathcal{X}, \mathcal{G}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{G}_\mathcal{Y})$,*

respectively. Then, $\{X_n\}$ and $\{Y_n\}$ are asymptotically independent if

$$|Pr(\{X_n \in A\} \cap \{Y_n \in B\}) - Pr(\{X_n \in A\})Pr(\{Y_n \in B\})| \to 0, \qquad \forall A \in \mathscr{G}_{\mathscr{X}}, B \in \mathscr{G}_{\mathscr{Y}} \tag{2.8.40}$$

as $n \to \infty$. Let X be a random element in $(\mathscr{X}, \mathscr{G}_{\mathscr{X}})$. Then, X and $\{Y_n\}$ are asymptotically independent if it holds for $\{X_n\}$ where $X_n = X$, $\forall n$ and $\{Y_n\}$ are asymptotically independent.

Suppose $\hat{\theta}_n$ is estimated from an i.i.d. data $\{(y_i, x_i)\}_{i=1}^n$ and (y_s, x_s) is picked randomly from this set of observations. Let $\hat{\theta}_{n,s}$ be an estimator based on $\{(y_i, x_i)\}_{i=1}^n \backslash (y_s, x_s)$.Then, (y_s, x_s) and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ are asymptotically independent if the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is identical to that of $\sqrt{n}(\hat{\theta}_{n,s} - \theta_0)$. This holds for a wide range of finite-dimensional parameter estimators that treat all the observations equally and such a condition can be examined via influence functions (Jann (2019)). The following proposition attains the desired result under asymptotic independence of X and $\{\sqrt{n}(\theta_n - \theta_0)\}$.

Proposition 2.8.2 CONFIDENCE SET FOR $h(X)$. *Suppose X and $\{\sqrt{n}(\theta_n - \theta_0)\}$ are asymptotically independent. Further, suppose that $h(x;\theta)$ is continuously differentiable with respect to θ in some open neighborhood of θ_0 and the derivative $\partial_\theta h(x;\theta_0)$ is nonzero for any $x \in \mathscr{X}$. Finally, let $\hat{\Lambda}_n$ be a consistent estimator of Λ. Then, for $\alpha \in (0,1)$,*

$$CS_{h,n} := \left[h(X;\hat{\theta}_n) - d_{\alpha,n}(X), h(X;\hat{\theta}_n) + d_{\alpha,n}(X)\right] \tag{2.8.41}$$

where $d_{\alpha,n}(x)$ is defined as (2.8.38) *is a confidence interval for $h(X;\theta_0)$ with asymptotic size $1-\alpha$.*

Given the results presented here, inference for approximation bounds is studied in the next section.

2.8.3. Estimation of approximation bounds

We consider estimation of the bounds based on the observed difference $Y - h(x;\theta_0)$ considered in Section 2.5 as well as the Chebyshev bound obtained under continuity (Section 2.7.2).

2.8.3.1. Approximation bounds based on observed difference

Let $h_0(\cdot)$ be an approximate model of $m(\cdot)$ and suppose there exists a consistent estimator $\hat{\theta}_n \xrightarrow{p} \theta_0$ where θ_0 is such that $h_0(\cdot) = h(;\theta_0)$. For any nondecreasing, convex function $f : \mathbb{R}_+ \to \mathbb{R}_{++}$, define the constant $c_{\alpha,f}$

$$c_{\alpha,f} = \inf\left\{c \in \mathbb{R}_{++} : f(c) = \frac{\mathbb{E}f(|Y - h_0(X)|)}{\alpha}\right\}. \tag{2.8.42}$$

Then, according to Proposition 2.5.5,

$$[h_0(X) - c_{\alpha,f}, h_0(X) + c_{\alpha,f}] \tag{2.8.43}$$

is a confidence set for $m(X)$ with level $1-\alpha$ ($\alpha \in (0,1)$). Furthermore, consider combined bounds $\left\{c_{\alpha,f}\right\}_{f\in\mathscr{F}}$ implied by a class $\mathscr{F}$ of nondecreasing and convex functions from $\mathbb{R}_+$ to $\mathbb{R}_{++}$ by considering

$$c_\alpha^{(\mathscr{F})} = \inf_{f\in\mathscr{F}} c_{\alpha,f}. \tag{2.8.44}$$

First, we are going to construct an estimator $\hat{c}_{\alpha,f}$ of $c_{\alpha,f}$ in (2.8.42) for an individual function f and show its consistency. Then, we show that an estimator of the unified bound (2.8.44) defined as

$$\hat{c}_\alpha^{(\mathscr{F})} = \inf_{f\in\mathscr{F}} \hat{c}_{\alpha,f} \tag{2.8.45}$$

is consistent under the assumptions presented below. Given the results in the previous section, we posit the existence of a consistent estimator $\hat{\theta}_n$ of θ_0.

Assumption 2.8.2 CONSISTENCY OF THE APPROXIMATE MODEL PARAMETER ESTIMATOR. *The estimator $\hat{\theta}_n$ of θ_0 is consistent:*

$$\hat{\theta}_n \xrightarrow{p} \theta_0. \tag{2.8.46}$$

We assume that an available set of observations: $\{w_i\} := \{y_i, x_i\}_{i=1}^n$ is stationary and β-mixing, the definition of which is given below:

Definition 2.8.2 β-MIXING PROCESS. *For a stationary sequence $\{w_t\}_{t\in\mathbb{N}_+}$ of random elements defined on a probability space $(\Omega, \mathscr{A}, \mathsf{P})$, let $\mathscr{M}_a^b$ be the σ-algebra generated by*

$w_a, \ldots, w_b$ ($a, b \in \mathbb{Z}$ ($a \leq b$)). The sequence $\{w_t\}_{t\in\mathbb{Z}}$ is said to be β-mixing (or absolutely regular) with coefficients $\{\beta(s)\}_{s=1}^{\infty}$ where

$$\beta(s) = \mathbb{E} \sup_{m\geq 1} \left\{ \mathsf{P}\left(B|\mathscr{M}_1^m\right) - \mathsf{P}\left(B\right) : B \in \mathscr{M}_{m+s}^{\infty} \right\} \tag{2.8.47}$$

if $\beta(s) \to 0$ as $s \to \infty$.

Assumption 2.8.3 β-MIXING SEQUENCE. *The sequence $\{w_t\}_{t=1}^n = \{(x_t,\, y_t)\}_{t=1}^n$ of observations is stationary and β-mixing.*

We start with inference for the bound $c_{\alpha,f}$ in (2.8.42) for a given f and then extend the results to the unified bound $\hat{c}_{\alpha}^{(\mathscr{F})}$ in (2.8.52) . In addition, we propose a consistent estimator of the Chebyshev bound under continuity.

Single bound. Given $\{w_i\}_{i=1}^n$, we define an estimator $\hat{c}_{\alpha,f}$ of $c_{\alpha,f}$ as

$$\hat{c}_{\alpha,f} = \inf \left\{ c \in \mathbb{R}_{++} : f\left(c\right) = \frac{\mathbb{E}_n f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right)}{\alpha} \right\}. \tag{2.8.48}$$

where $\mathbb{E}_n[\cdot]$ is the expectation operator given the empirical distribution of $\{w_i\}_{i=1}^n$.

We impose the following regularity assumptions to achieve consistency of $\hat{c}_{\alpha,f}$.

Assumption 2.8.4 EXISTENCE OF THE CONTINUOUS INVERSE. *For $c^* \in \mathbb{R}$ defined as*

$$f\left(c^*\right) = \frac{\mathbb{E} f\left(|Y - h_0(X)|\right)}{\alpha}, \tag{2.8.49}$$

the function $f : \mathbb{R}_+ \to \mathbb{R}_{++}$ is strictly increasing in some open neighborhood of c^*.

If Assumption 2.8.4 is satisfied, the inverse f^{-1} of f at the point of interest is well-defined and is continuous in its neighborhood.

Assumption 2.8.5 ENVELOPE FUNCTIONS. *There exists an open neighborhood $\mathscr{N}^{(E)}(\theta_0)$ such that*

$$\mathbb{E} \sup_{\theta\in\mathscr{N}^{(E)}(\theta_0)} f\left(|Y - h(X;\theta)|\right) < \infty. \tag{2.8.50}$$

Assumption 2.8.6 REGULARITY: CONTINUITY OF THE APPROXIMATE MODEL WITH RESPECT TO THE PARAMETER. *$h(X;\theta)$ is continuous with respect to θ in some open neighborhood $\mathscr{N}^{(h)}(\theta_0)$ with probability one.*

Assumption 2.8.7 REGULARITY: CONTINUITY OF THE EXPECTATION. *The function*

$$k(\theta) = \mathbb{E} f(|Y - h(X;\theta)|) \tag{2.8.51}$$

is continuous in some open neighborhood $\mathscr{N}^{(Eh)}(\theta_0)$.

If (Y,X) is continuous, Assumptions 2.8.5 and 2.8.6 imply this assumption for $\mathscr{N}^{(\mathrm{Eh})}(\theta_0) = \mathscr{N}^{(\mathrm{E})}(\theta_0) \cap \mathscr{N}^{(\mathrm{h})}(\theta_0)$ by application of the dominated convergence theorem. Under these assumptions, we establish consistency of $\hat{c}_{\alpha,f}$.

Lemma 2.8.3 *Suppose Assumption 2.8.2-2.8.6 hold. Then, the estimator $\hat{c}_{\alpha,f}$ defined as (*2.8.48*) converges to $c_{\alpha,f}$ of (*2.8.42*) in probability.*

In the next section, inference for the unified bound is considered.

Unified bound. We provide below regularity conditions under which we achieve consistency of an estimator of the unified bound defined as $\hat{c}_\alpha^{(\mathscr{F})}$

$$\hat{c}_\alpha^{(\mathscr{F})} = \inf_{f \in \mathscr{F}} \hat{c}_{\alpha,f} \tag{2.8.52}$$

where $\mathscr{F}$ is a class of nondecreasing and convex functions from $\mathbb{R}_+$ to $\mathbb{R}_{++}$ and $\hat{c}_{\alpha,f}, f \in \mathscr{F}$ is given in (2.8.48).

Assumption 2.8.8 EXISTENCE OF CONTINUOUS INVERSE. *For $c_f^* \in \mathbb{R}$ defined as*

$$f(c_f^*) = \frac{\mathbb{E} f\left(|Y - h_0(X)|\right)}{\alpha} \quad , f \in \mathscr{F}, \tag{2.8.53}$$

there exists some $\varepsilon > 0$ such that for f is strictly increasing in an ε-open neighborhood $\mathscr{N}_\varepsilon(c_{f^*})$. Furthermore, the inverse f^{-1} of f defined on the restricted domain $\mathscr{N}_\varepsilon(c_f^*)$ satisfies the following uniform continuity condition: for any $\gamma > 0$, there exists some $\delta > 0$ such

that for

$$\sup_{f\in\mathscr{F}}\left|f^{-1}(y_1)-f^{-1}(y_2)\right|<\gamma \tag{2.8.54}$$

whenever $|y_1-y_2|<\delta$.

Assumption 2.8.9 UNIFORM ENVELOPE FUNCTION. *There exists an open neighborhood* $\mathscr{N}^{(E,\mathscr{F})}(\theta_0)$ *of* θ_0 *such that*

$$\mathbb{E}\sup_{\theta\in\mathscr{N}^{(E,\mathscr{F})}(\theta_0)}\sup_{f\in\mathscr{F}}f\left(|Y-h(X;\theta)|\right)<\infty. \tag{2.8.55}$$

Assumption 2.8.10 REGULARITY: UNIFORM CONTINUITY OF THE APPROXIMATE MODEL. *Define* $\mathscr{K}=\left\{k_f(\theta)\right\}_{f\in\mathscr{F}}$ *where*

$$k_f(\theta)=\mathbb{E}f\left(|Y-h(X;\theta)|\right). \tag{2.8.56}$$

Then, there exists an open neighborhood $\mathscr{N}^{(\mathscr{K})}(\theta_0)$ *of* θ_0 *such that* $\left\{k_f(\theta)\right\}_{f\in\mathscr{F}}$ *is uniformly continuous, i.e. for any* $\gamma>0$*, there exists some* $\delta>0$ *such that*

$$\sup_{f\in\mathscr{F}}\left|k_f(\theta_1)-k_f(\theta_2)\right|\leq\gamma \tag{2.8.57}$$

whenever

$$|\theta_1-\theta_2|<\delta. \tag{2.8.58}$$

Note Assumption 2.8.10 holds, for example, if the probability measure of (Y,X) is absolutely continuous and $\mathscr{F}$ is a class of uniformly continuous functions. Finally, the following assumption controls the entropy of the function class.

Assumption 2.8.11 ENTROPY CONDITION FOR THE FUNCTION CLASS. *Let*

$$g_f(w;\theta)=f\left(|y-h(x;\theta)|\right) \tag{2.8.59}$$

and define a function class $\mathscr{G}_\varepsilon$ *indexed by* (θ,f) *for given* $\varepsilon>0$*,*

$$\mathscr{G}_\varepsilon=\cup_{f\in\mathscr{F}}\left\{g_f(\cdot;\theta)|\theta:\|\theta-\theta_0\|\leq\varepsilon\right\}. \tag{2.8.60}$$

Then, there exists some $\varepsilon > 0$ such that $\mathscr{G}_\varepsilon$ is a Glivenko-Cantelli class with respect to the probability measure $\mathbb{P}^* = \Pi_{i=1}^{\infty} P_i$ where P_i is the marginal distribution of w_i.

Assumption 2.8.11 only requires $\mathscr{G}_\varepsilon$ be a Glivenko-Cantelli class with respect to the product measure of the marginal distributions of w_i's, $i = 1, \ldots$, not with respect to the joint distribution of the β-mixing sequence $\{w_i\}_{i=1}^{\infty}$. Intuitively, it states the dependent sequence $\{w_i\}_{i=1}^{\infty}$ may be treated as if it were an i.i.d. to examine whether the condition of Assumption 2.8.11 is satisfied. Checking such condition for a dependent sequence is in general significantly more challenging since much of the theory of empirical processes are concerned with i.i.d sequences. Given this set of assumptions, we establish consistency of an estimator of the unified bound.

Lemma 2.8.4 *Suppose Assumption 2.8.2, 2.8.3, 2.8.6, 2.8.8-2.8.11 hold. Then, the estimator $\hat{c}_\alpha^{(\mathscr{F})}$ in* (2.8.52) *converges to $c_\alpha^{(\mathscr{F})}$ in (*2.8.44*) in probability.*

Consider

$$\mathscr{F}_\beta = \left\{ f_\gamma(x) = x^\beta \mid 1 \leq \beta \leq \bar{\beta} \right\} \tag{2.8.61}$$

Observe that any $f_\beta \in \mathscr{F}_\beta$ is invertible and the inverse function is given by

$$f_\beta^{-1}(y) = y^{1/\beta}. \tag{2.8.62}$$

Furthermore, for any $\delta > 0$ pick any $y_1, y_2 \in \mathbb{R}_=$ such that $|y_1 - y_2| < \delta^{1/\bar{\gamma}}$. Then, for any $\beta \in [1, \bar{\beta}]$,

$$\left| f_\beta^{-1}(y_1) - f_\beta^{-1}(y_2) \right| < \delta. \tag{2.8.63}$$

Thus, Assumption 2.8.8 is satisfied. Assumption 2.8.9 is simplified to the following condition; there exists an open neighborhood $\mathscr{N}^{(\mathrm{E},\mathscr{F})}(\theta_0)$ of θ_0 such that

$$\mathbb{E} \sup_{\theta \in \mathscr{N}^{(\mathrm{E},\mathscr{F})}(\theta_0)} |Y - h(X;\theta)|^{\bar{\gamma}} < \infty. \tag{2.8.64}$$

Assumption 2.8.10 holds if

$$\mathbb{E}|Y - h(X;\theta)|^{\bar{\beta}} \tag{2.8.65}$$

is continuous in some open neighborhood of θ_0. Finally, Assumption 2.8.11 holds since $\mathscr{F}_\gamma$ is indexed by a finite-dimensional parameter in a compact space and thus $\mathscr{G}_\varepsilon$ is indexed by

a pair of finite-dimensional parameters (β, γ) in a compact set $[1, \bar{\beta}] \times \{\theta : \|\theta - \theta_0\| \leq \varepsilon\}$ (Vaart and Wellner (2000)). This shows that under the conditions described above, the unified central moment based bound given in (2.5.38) can be consistently estimated.

2.8.3.2. Bounds under continuity

This section is concerned with inference for the approximation bound under continuity considered in Section 2.7.2 where

$$\tilde{c}_{\alpha}^{(2)} = \sqrt{\frac{\mathrm{Var}\left(Y - h(X;\theta_0)\right) - \mathrm{Var}\left(\varepsilon\right)}{\alpha}}. \tag{2.8.66}$$

In the literature of nonparametric regression, inference for $\mathrm{Var}(\varepsilon)$ without estimating the regression function $m(\cdot)$ has been an active topic of research since Rice (1984), who revives a difference based estimator of $\mathrm{Var}(\varepsilon)$ by Von Neumann (1941). He assumes that X is univariate, Following Rice (1984), we assume here X is univariate for brevity, however multivariate extensions have been proposed in the literature (Cai, Levine and Wang (2009)) and may be employed in our framework. Let $\{(y_i,\, x_i)\}_{i=1}^n$ be a set of observations, consider reordering $\{x_{[j]}\}_{j=1}^n$ of $\{x_i\}_{i=1}^n$ so that $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n]}$. and define $\{y_{[j]}\}_{j=1}^n$ based on the same indices, i.e. for each $i \in \{1, \ldots, n\}$, there exists some unique $[j'] \in \{1, \ldots, n\}$ such that $(x_i, y_i) = (x_{[j']}, y_{[j']})$. Then, the Rice estimator is defined as

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{2\,(n-1)} \sum_{j=2}^{n} \left(y_{[j]} - y_{[j-1]}\right)^2. \tag{2.8.67}$$

Rice (1984) proves consistency of $\hat{\sigma}_{\varepsilon}^2$ in a fixed design, in which $\{x_i\}$ is fixed. Lemma 2.8.5 extends to a random sampling design as in Assumption 2.8.12 under the following additional assumptions.

Assumption 2.8.12 I.I.D. OBSERVATIONS. *The sequence $\{(y_i,\, x_i)\}_{i=1}^n$ is independent and identically distributed with $\mathbb{E}\,|x_i|^2 < \infty$ and $\mathbb{E}\,|y_i|^2 < \infty$.*

Assumption 2.8.13 HÖLDER CONTINUITY. *There exist some constants C and γ such*

that $C > 0$, $\gamma > 0$ and

$$|m(\bar{x}_1) - m(\bar{x}_2)| \leq C\,|\bar{x}_1 - \bar{x}_2|^{\gamma}, \qquad \forall \bar{x}_1, \bar{x}_2 \in \mathscr{X}. \tag{2.8.68}$$

Assumption 2.8.14 CONTINUOUS SUPPORT OF X. *The support $\mathscr{X}$ of X is an interval in $\mathbb{R}$ and for given γ such that it satisfies Assumption 2.8.13 for some $C > 0$,*

$$\max_{2 \leq i \leq n} \left| x_{[i]} - x_{[i-1]} \right|^{2\gamma} = o_p\left(n^{-1}\right). \tag{2.8.69}$$

Lemma 2.8.5 *Under Assumption 2.8.12-2.8.14, the estimator $\hat{\sigma}_{\varepsilon}^2$ in* (2.8.67) *is a consistent estimator of σ_{ε}^2.*

Given $\hat{\sigma}_{\varepsilon}^2$, we may consider an estimator of $\tilde{c}_{\alpha}^{(2)}$ in (2.8.66) defined as

$$\hat{c}_{\alpha}^{(2)} = \sqrt{\frac{\max(\hat{\sigma}_{h,n}^2 - \gamma_n \hat{\sigma}_{\varepsilon}^2, 0\}}{\alpha}} \tag{2.8.70}$$

where

$$\hat{\sigma}_{h,n}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i; \hat{\theta}_n))^2 \tag{2.8.71}$$

and $\{\gamma_n\}$ is a nondecreasing sequence of random values in $[0,1]$ such that $\gamma_n \to 1$. Note that $\hat{\sigma}_{h,n}^2 - \hat{\sigma}_{\varepsilon}^2$ may be negative in practice especially under a small sample size and thus the scaling term γ_n may be chosen so that $\hat{\sigma}_{h,n}^2 < \gamma_n \hat{\sigma}_{\varepsilon}^2$ will happen with negligible probability. The assumption that $\gamma_n \to 1$ is required for consistency of $\hat{c}_{\alpha}^{(2)}$, however it is not essential for the approximate bound to sustain the level $1 - \alpha$ since

$$\sqrt{\frac{\sigma_{h,n}^2 - \gamma_n \hat{\sigma}_{\varepsilon}^2}{\alpha}} \geq \sqrt{\frac{\sigma_{h,n}^2 - \hat{\sigma}_{\varepsilon}^2}{\alpha}}, \qquad \forall \gamma_n \in [0,1]. \tag{2.8.72}$$

2.8.4. Feasible confidence set for nonlinear regression

Given the succeeding discussion, we are going to establish the asymptotic validity of the feasible confidence set for $m(X)$ defined in (2.8.29) . To this end, we impose the following assumption.

Assumption 2.8.15 NO MASS ON THE BOUNDARY. *For given $\alpha \in (0,1)$, consider an approximation bound c_α of level $1-\alpha$. Then,*

$$Pr(|m(X) - h(X;\theta_0)| = c_\alpha) = 0. \tag{2.8.73}$$

Under Assumption 2.8.15, it can be shown (in the proof of Proposition 2.8.6) that any consistent estimator $\hat{c}_\alpha$ of c_α is an approximation bound of asymptotic level $1-\alpha$. Combining with Proposition 2.8.2, we show that the feasible confidence set $CS_{1-\alpha,n}(X;\alpha_1,\alpha_2)$ in (2.8.29) has asymptotic level $1-\alpha$.

Proposition 2.8.6 FEASIBLE CONFIDENCE SET FOR THE CONDITIONAL EXPECTATION. *For any $\alpha \in (0,1)$, choose a pair $(\alpha_1,\alpha_2) \in (0,\alpha)^2$ such that $\alpha_1+\alpha_2=\alpha$. Suppose Assumption 2.8.15 holds for the approximation bound c_{α_2} of level $1-\alpha_2$ and further maintain assumptions in Proposition 2.8.2. Let $\hat{c}_{\alpha_2}$ be a consistent estimator of c_{α_2}. Define a random set*

$$CS_{1-\alpha,n}(X;\alpha_1,\alpha_2) = \left[h(X;\hat{\theta}_n) - \hat{D}(X;\alpha_1,\alpha_2), h(X;\hat{\theta}_n) + \hat{D}(X;\alpha_1,\alpha_2)\right] \tag{2.8.74}$$

where

$$\hat{D}(x;\alpha_1,\alpha_2) = d_{\alpha_1,n}(x) + \hat{c}_{\alpha_2} \tag{2.8.75}$$

with $d_{\alpha_1,n}(x)$ is defined in (2.8.27). Then, $CS_{1-\alpha,n}(X;\alpha_1,\alpha_2)$ is a confidence interval for $m(X)$ with asymptotic level $1-\alpha$.

Assumption 2.8.15 is generally justified in practice, in particular, when X is continuous. However, one may find data generating processes such that this assumption is violated. For example, suppose c_α is tight and the distribution of X has mass points at x's such that

$$m(x) = h(x;\theta_0) + c_\alpha \text{ or } m(x) = h(x;\theta_0) - c_\alpha. \tag{2.8.76}$$

In order to accommodate cases where the underlying generating process may violate Assumption 2.8.15, we consider an alternative assumption imposing strong monotonicity of c_α with respect to α that imposes a different requirement on (α_1,α_2): $\alpha_1+\alpha_2<\alpha$ instead of $\alpha_1+\alpha_2=\alpha$.

Assumption 2.8.16 STRONG MONOTONICITY OF THE BOUND. *The bound c_α as a function of α is strictly decreasing in α.*

The following corollary shows that the conclusion of Proposition 2.8.6 holds under Assumption 2.8.16 in place of Assumption 2.8.15.

Corollary 2.8.7 FEASIBLE CONFIDENCE SET WITH MASSES. *Choose $(\alpha_1, \alpha_2) \in (0, \alpha)^2$ so that $\alpha_1 + \alpha_2 < \alpha$ and replace Assumption 2.8.15 by Assumption 2.8.16 in Proposition 2.8.6. Then, the feasible confidence set in (*2.8.74*) has asymptotic level $1 - \alpha$.*

We investigate finite-sample properties of the estimated confidence set through Monte Carlo experiments in Section 2.9.

2.9. Monte Carlo simulation

2.9.1. Simulation design

This section provides Monte Carlo evidence on the finite sample properties of the bound approach in comparison with alternative nonparametric methods: kernel regression, sieve method, random forest, LASSO (least absolute shrinkage and selection operator), and neural network. Comparisons will be made in terms of the size and average width of confidence sets for $m(X)$ and the mean squared error of point estimates associated with each method with respect to $m(X)$.

We consider various data generating processes based on 4 different specifications of the regression function and the distributions of the conditioning variable X (and W) as described in Table 2.1. For all cases, it is assumed that the expectation error $\varepsilon = Y - \mathbb{E}[Y \mid X]$ is independent of X and is normally distributed. The ratio of the unconditional variance $\text{Var}(\varepsilon)$ of ε relative to the total variance $\text{Var}(Y)$ ranges from $\{.01, .1, .2\}$ in order to examine the effect of different signal-to-noise ratios. We also alter the sample size n of the training set: $n \in \{50, 200, 500\}$.

For each data generating process, we simulate M ($M = 500$) independent training sets, the l-th set of which is denoted by $\left\{ \left(X_i^{(l)}, Y_i^{(l)} \right) \right\}_{i=1}^{n}$. Given each training set, confidence sets, if available, and point predictions of $m(X)$ are estimated for each method. To conduct out-of-sample evaluation of the estimated confidence set for $m(X)$, we simulate a test

set $\left\{\left(X_j^{(\text{test})}, m\left(X_j^{(\text{test})}\right)\right)\right\}_{j=1}^{N}$, where $N = 10,000$, independently from the training sets. Then, we compute the empirical coverage and average widths of the confidence sets and the mean squared error of the point estimates. A more detail account of the simulation procedure is given as follows:

Simulation procedure

1. For each l, draw a set of i.i.d. observations $\left\{\left(X_i^{(l)}\right)\right\}_{i=1}^{N}$ according to a data generating process in Table 2.1. Then, draw a set of i.i.d. observations $\left\{\left(\varepsilon_i^{(l)}\right)\right\}_{i=1}^{N}$, independently of $\left\{\left(X_i^{(l)}\right)\right\}_{i=1}^{N}$, from a normal distribution with mean 0 and variance $\text{Var}(\varepsilon) = c_\varepsilon \text{Var}(Y)$ where $c_\varepsilon \in \{.01, .1, .2\}$.

2. Compute the dependent variable $Y_i^{(l)}$ as

$$Y_i^{(l)} = m\left(X_i^{(l)}\right) + \varepsilon_i^{(l)}, i = 1, \cdots, N. \tag{2.9.1}$$

3. For each method, using $\left\{\left(X_i^{(l)}, Y_i^{(l)}\right)\right\}_{i=1}^{N}$, estimate a confidence set for $m(X)$ with asymptotic level 95% and construct point estimates for $\{m(x)\}_{x\in\mathscr{X}}$, each of which is denoted by $\hat{C}_l(X)$ and $\{\hat{m}_l(x)\}_{x\in\mathscr{X}}$.

4. Given the test set $\left\{\left(X_j^{(\text{test})}, m\left(X_j^{(\text{test})}\right)\right)\right\}_{j=1}^{N}$, compute the empirical coverage level and average width of $\hat{C}_l(X)$:

$$ECP_m = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\left\{m\left(X_j^{(\text{test})}\right) \in \hat{C}_m\left(X_j^{(\text{test})}\right)\right\}$$

and

$$WTH_m = \frac{1}{N}\sum_{j=1}^{N} \lambda\left(C_l\left(X_j^{(\text{test})}\right)\right)$$

where $\lambda(\cdot)$ is the Lebesgue measure on $\mathbb{R}$. The empirical relative MSE is given by

$$RMSE_m = \frac{\frac{1}{N}\sum_{j=1}^{N}\left(m\left(X_j^{(\text{test})}\right) - \hat{m}_l\left(X_j^{(\text{test})}\right)\right)^2}{\text{Var}(m(X))}.$$

We report the average of each of $\{ECP_l, WTH_l, MSE_l\}_{l=1}^{M}$:

$$ECP = \frac{1}{M}\sum_{l=1}^{M} ECP_l, WTH = \frac{1}{M}\sum_{l=1}^{M} WTH_l, RMSE = \frac{1}{M}\sum_{l=1}^{M} RMSE_l.$$

2.9.2. Data generating processes

We generate observations according to data generating processes provided in Table 2.1. Note that for each model, we consider four specifications of the noise ratio: $\text{Var}(\varepsilon)/\text{Var}(Y) \in \{.01, .05, .1, .2\}$. Each model embodies certain limitations that (some of) standard standard nonparametric and machine learning methods face.

In (1), the regression function is specified as a step function. The sieve method (LASSO) can incorporate such a feature of the function $m(x)$ into the sieve space only if the points of discontinuity (in this case, $x = 1, \ldots, 10$) are known, which is often not the case in practice. Kernel methods require $m(x)$ to be continuous. Thus, in this setup, we can examine the effect of this violation of the key assumptions on Nadaraya-Watson and local linear methods.

We investigate through Model (2) the impact of the curse of dimensionality due to the presence of the irrelevant regressors.

The regression function $m(x)$ in Model (3) is specified a continuous but non-differentiable periodic function. We assume that $m(x)$ is only known to be periodic and continuous. However, its functional form and its period are considered unknown. It is challenging for methods, such as regression forest and kernel methods, to incorporate such prior information efficiently while it allows for series based methods (the bound approach, the sieve method, and LASSO) to employ a suitable class of base functions, such as a trigonometric series.

On the other hand, the function $m(x)$ in Model (4) is a function that is Hölder continuous everywhere but differentiable nowhere, known as a Weierstrass function. We examine each method under such a moderate degree of smoothness.

2.9.3. Implementation of each method

2.9.3.1. Approximation bound approach

For all DGPs, we consider a class of approximate model of the form:

$$P_{\mathrm{C}}(x\ ;\ \theta) = p(x)'\theta_0 \tag{2.9.2}$$

where $p(x)$ is a k-dimensional vector specified below for each model and θ_0 is a k-dimensional parameter such that

$$\theta_0 = \arg\inf_{\theta\in\mathbb{R}^k} \mathbb{E}\left[(Y - p(X)'\theta)^2\right] \tag{2.9.3}$$

We consider (i) linear (**DT(Linear)**) and (ii) finite-series specifications (**DT(Sieve)**) defined as follows. In **DT(Linear)**, $p(x)$ is specified as power series of order up to 1 for Model (1)-(2) and a trigonometric series of order up to 1 for Model (3)-(4), i.e.

$$p(x) = \begin{cases} (1,x) & \text{(Model (1)),} \\ (1,x_1,x_2,\ldots,x_8)' & \text{(Model (2)),} \\ (1,\cos(x),\sin(x)) & \text{(Model (3)-(4)).} \end{cases} \tag{2.9.4}$$

In **DT(Sieve)**, we consider a finite set of candidate models $S = \{h_1,\ldots,h_L\}$ for each model and then pick a model $h^* \in S$ which minimizes TIC (Takeuchi information criterion). TIC is a generalization of AIC and is known to be robust to model misspecification (Takeuchi (1976)). In Model (1), an element of S is of the form (2.9.2) where $p(x)$ is a subvector of $(1,x,x^2,\ldots,x^6)'$, i.e. the base functions of a power series of order up to 6 but assuming $p(x)$ always includes a constant term. Then, the cardinality $|S|$ of S here is $64(=2^6)$. Similarly in Model (3)-(4), S consists of a model of the form (2.9.2) where $p(x)$ is a subvector of $(1,\cos(x),\sin(x),\ldots,\cos(6x),\sin(6x))'$, i.e. the base functions of a trigonometric series of order up to 6. In Model (2), the base functions $p(x)$ of a candidate model in S is a subvector of the base functions of a power series of order up to 2, including the cross terms, e.g. x_1x_2. Then, the set S includes 2^{42} possible models. Since evaluating TIC for all models is computationally prohibitive, we instead employ the forward stepwise approach with bidirectional elimination.

In Model (1)-(2), for both **DT(Sieve)** and **DT(Sieve)**, we consider an approximation bound based on the central-based moment bounds up to 8, given a given approximate model $P_{\mathbf{C}}$:

$$c_{\alpha}^{(1),(2)} = \inf_{1 \leq \beta \leq 8} \left(\frac{\mathbb{E}\left|Y - P_{\mathbf{C}} Y\right|^{\beta}}{\alpha} \right)^{1/\beta}. \tag{2.9.5}$$

where $\alpha = .95$. For Model (3)-(4), we also incorporate the Chebyshev bound under continuity:

$$c_{\alpha}^{(\mathbf{con})} = \sqrt{\frac{\mathrm{Var}\left(Y - P_{\mathbf{C}} m\left(X\right)\right)}{\alpha}} \tag{2.9.6}$$

and combine such bound with (2.9.5) by defining the approximation bound as

$$c_{\alpha}^{(3),(4)} = \begin{cases} c_{\alpha}^{(\mathbf{con}))} & \text{if } 0 < c_{\alpha}^{(\mathbf{con}} \\ c_{\alpha}^{(1),(2)} & \text{otherwise.} \end{cases} \tag{2.9.7}$$

Estimation of the approximate model $P_{\mathbf{C}}$ and the approximation bound are conducted in the framework of Section 2.8.

2.9.3.2. Alternative methods

For the sieve method, as in **DT(Sieve)**, we pick the best model among a set of possible models S defined as in the case of **DT(Sieve)** but AIC (Akaike information criteion (Akaike (1974))) is employed for Model (1) and (3)-(4). The resulting confidence sets only account for estimation error but not approximation error, which is assumed to be negligible asymptotically, and thus we call this implementation "**Naive Sieve**".

As kernel regression methods, we consider Nadaraya-Watson (**NW**) and local linear estimators. For each method, we choose the bandwidth h_{CV} by least-squares cross-validation. It is known that under such choice of bandwidth, the estimate of $m(X)$ is biased and thus the confidence set fails to have correct size (Hall (1992)). Thus, we also present results under undersmoothing, in which the bandwidth h_{US} is set to be $h_{CV}/2$. Thus, we consider construction of confidence sets based on NW estimators with the cross-validation bandwidth h_{CV} (referred to as "**NW**"), with the bandwidth h_{US} ("**NW(US)**"), and local linear estimators with h_{CV} ("**Local Linear**") and with h_{US} ("Local Linear (**US**)").

Consistency and asymptotic normality of a random forest estimator is known only when subsampling is used to generate trees (Wager (2014), Wager and Athey (2018)). We set the subsample size$\approx n/3$. Then, following Wager, Hastie and Efron (2014), we construct confidence sets based on random forest (**RF**) estimates by computing the infinitesimal jackknife standard error. For **LASSO** and neural network (referred to as "**NN**"), there is yet no asymptotically valid method proposed for construction of confidence sets for $m(X)$ based on **LASSO** and neural network and thus only point estimates are provided.

2.9.4. Simulation results

Simulation results are provided in Table 2.2-2.13 for all models (1)-(4) with three different levels of the noise ratio: $\mathrm{Var}(\varepsilon)/\mathrm{Var}(Y)$.

In Model (1), we see that confidence sets based on both **AB(Linear)** and**AB(Sieve)** have correct coverage probabilities even though both specifications are clearly misspecified. Given the relatively small MSE, the impact of the specification error is fairly small and a simpler model (**AB(Linear)**) appears to provide better approximation especially in a small sample size: $N = 50$. As $\mathrm{Var}(\varepsilon)/\mathrm{Var}(Y)$ increases, the average width of the confidence set gets larger for both methods. However, as discussed in Section 2.5.4, this is an unavoidable feature of a confidence set for $\mathbb{E}[Y \mid X]$ with valid size when the regression function is weakly identified (and possibly nonsmooth). While the regression function $m(\cdot)$ in Model (1) exhibits a somewhat limited pattern of discontinuity characterized by a finite number of equispaced jumps, such a structure is in practice not known to the practitioner. The bound-based confidence set remains can be constructed without prior knowledge on $m(\cdot)$ and remains valid even when the number of jumps are uncountably many and the points of discontinuities are unknown. In the same model, a confidence set based on any of the alternative methods is undersized even for a relatively large sample size. In particular, for **Naive Sieve**, the coverage probability is close to zero. For this method, the estimated model is misspecified even asymptotically, however the associated confidence set only incorporates estimation error and thus shrinks to a point as the sample size increases, ignoring misspecification error. Analogous observation can be made in Model (2)-(4).

The four kernel methods and **RF** appear to be affected by the lack of smoothness of $m(\cdot)$ and produce undersized confidence sets. Note that the average width of the confidence set

associated with a kernel method is astronomically large for some cases for $n = 50, 200$. To see why this is the case, note that the estimated standard error of the kernel estimator is reciprocal to an estimated density $\hat{f}(x)$. When $\hat{f}(x)$ is evaluated at the tails of the distribution of X, it may be close to zero, especially in a small or moderate sample, which results in an extremely large value of the standard error and thus the width of the confidence set. However, such tail behavior does not contribute to the size of the estimated confidence set.

Model (2) shows all methods, except for **AB(Linear)** are severely impacted by the curse of dimensionality, especially when $\mathrm{Var}(\varepsilon)/\mathrm{Var}(Y)$ is large and yield undersized confidence sets. For **DT(Sieve)**, a chosen model is subject to overfitting with a high probability and thus the approximation error is underestimated. On the other hand, employing a misspecified linear model and then bounding the approximation error (**DT(Linear)**) does not suffer from such issue of over-fitting. We also note that **AB(Linear)** often provide the smallest MSE. These points confirm the advantage of the parsimony principle, in particular in a small to moderate sample size (relative to the number of variables).

In Model (3)-(4), the issue of undersized confidence sets is present for any of the alternative models while **AB(Linear)** and **AB(Sieve)** maintain to produce a valid confidence set even in a small sample size. For these models, the approximation bound for both **AB(Linear)** and **AB(Sieve)** take continuity of $m(\cdot)$ into account through the Chebyshev bound under continuity. As a consequence, the width of the confidence sets is only minimally affected by the increase in the noise ratio $\mathrm{Var}(\varepsilon)/\mathrm{Var}(Y)$. These results show that continuity of $m(\cdot)$ can improve the confidence set significantly without affecting its size.

In summary, the proposed bound methods (**AB(Linear)**, **AB(Sieve)**) successfully deliver a valid confidence set under the various data generating processes considered. In addition, the MSE of point estimates is often comparable to the lowest one among the alternative methods.

Table 2.1: Data generating processes in the Monte Carlo experiments

Model	Functional form of $m(x)$	Distribution of X
(1)	$\lfloor x \rfloor$ (the largest integer smaller than x)	$X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,10)$
(2)	$10 + 2x_1 + x_2 + .3x_1^2 + .2x_1x_2$	$(X_{1i}, X_{2i}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{matrix} 1 \\ 1 \end{matrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} \right)$ $X_{3:8,i} \sim \mathcal{N}(\mathbf{0}, I_6)$ independent of (X_{1i}, X_{2i})
(3)	$m(x) = \lvert x \rvert$ for $-\frac{\pi}{4} < x < \frac{\pi}{4}$ $m(x + 2\pi) = m(x)$	$X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(-6,6)$
(4)	$m(x) = \sum_{i=0}^{\infty} \frac{1}{2^i} \sin\left(2^i x\right)$	$X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,5)$

Table 2.2: Model (1) Step function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .01$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.000	0.999	0.004	0.757	0.713	0.940	0.900	0.855	NA	NA
N= 200	1.000	1.000	0.003	0.841	0.783	0.776	0.733	0.891	NA	NA
N= 500	1.000	1.000	0.002	0.874	0.838	0.864	0.830	0.902	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.917	1.866	0.004	30.827	2.5E+18	1.202	1.764	0.928	NA	NA
N= 200	1.971	1.950	0.002	1.7E+11	1.3E+60	0.534	8.5E+02	0.616	NA	NA
N= 500	1.986	1.971	0.002	0.406	0.516	0.395	0.495	0.556	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0109	0.0119	0.0124	0.0143	0.0153	0.0114	9.6209	0.0135	0.0112	0.0114
N= 200	0.0102	0.0102	0.0101	0.0052	0.0064	0.0067	0.0130	0.0050	0.0103	0.0102
N= 500	0.0101	0.0098	0.0097	0.0029	0.0038	0.0029	0.0045	0.0032	0.0102	0.0100

Table 2.3: Model (1) Step function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .1$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.000	0.998	0.008	0.817	0.815	0.924	0.897	0.865	NA	NA
N= 200	1.000	1.000	0.005	0.714	0.815	0.704	0.701	0.891	NA	NA
N= 500	1.000	1.000	0.004	0.783	0.861	0.588	0.619	0.905	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	4.859	4.706	0.009	1.338	2.175	1.383	1.401	1.905	NA	NA
N= 200	5.057	5.004	0.005	0.810	1.198	0.732	0.776	1.741	NA	NA
N= 500	5.126	5.104	0.003	0.701	0.945	0.555	0.661	1.711	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0139	0.0261	0.0297	0.0280	0.0355	0.0194	0.0448	0.0351	0.0176	0.0187
N= 200	0.0110	0.0126	0.0129	0.0148	0.0172	0.0124	0.0142	0.0257	0.0117	0.0119
N= 500	0.0105	0.0109	0.0109	0.0092	0.0113	0.0103	0.0116	0.0232	0.0107	0.0107

Table 2.4: Model (1) Step function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .2$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.000	0.999	0.009	0.829	0.817	0.898	0.873	0.876	NA	NA
N= 200	1.000	1.000	0.007	0.760	0.853	0.724	0.722	0.907	NA	NA
N= 500	1.000	1.000	0.005	0.704	0.830	0.527	0.558	0.911	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	7.122	6.814	0.013	2.042	2.1E+07	1.604	1.687	2.718	NA	NA
N= 200	7.477	7.417	0.007	0.985	1.285	0.807	0.862	2.606	NA	NA
N= 500	7.516	7.492	0.005	0.764	1.031	0.538	0.604	2.574	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0218	0.0439	0.0458	0.0504	0.0693	0.0358	0.0541	0.0663	0.0294	0.0321
N= 200	0.0130	0.0155	0.0166	0.0192	0.0232	0.0149	0.0175	0.0530	0.0143	0.0146
N= 500	0.0111	0.0122	0.0127	0.0133	0.0154	0.0121	0.0131	0.0506	0.0117	0.0118

Table 2.5: Model (2) Many regressors: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .01$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.974	0.086	0.000	0.626	0.620	0.841	0.829	0.493	NA	NA
N= 200	0.990	0.218	0.000	0.608	0.648	0.819	0.841	0.547	NA	NA
N= 500	0.993	0.255	0.000	0.581	0.625	0.792	0.837	0.591	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	3.961	1.237	0.003	1.9E+146	Inf	4.178	5.1E+11	3.890	NA	NA
N= 200	5.058	2.906	0.002	1.3E+52	7.5E+144	0.768	8.2E+05	2.441	NA	NA
N= 500	5.344	3.086	0.001	1.5E+19	9.8E+93	0.461	1.3E+05	1.799	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0138	0.8847	0.8847	0.0607	0.1600	0.0158	5.9E+27	0.3254	0.8854	0.1246
N= 200	0.0104	0.7459	0.7459	0.0127	0.0185	0.2301	2.1E+27	0.1441	0.7486	0.0072
N= 500	0.0097	0.6071	0.6071	0.0058	0.0090	0.0010	0.1603	0.0803	0.6096	0.0014

Table 2.6: Model (2) Many regressors: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .1$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.998	0.246	0.001	0.670	0.671	0.449	0.499	0.520	NA	NA
N= 200	1.000	0.645	0.000	0.587	0.676	0.411	0.482	0.609	NA	NA
N= 500	1.000	0.723	0.000	0.550	0.656	0.386	0.443	0.664	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	9.601	3.786	0.011	1.7E+148	Inf	2.5E+02	1.8E+20	4.262	NA	NA
N= 200	10.802	9.756	0.006	5.2E+19	1.8E+96	0.829	9.8E+06	2.945	NA	NA
N= 500	11.082	10.314	0.004	7.8E+07	1.7E+48	0.454	1.022	2.315	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0369	1.0039	1.0039	0.1629	0.2917	0.1170	1.4E+04	0.3434	0.8996	0.4855
N= 200	0.0156	0.7487	0.7487	0.0328	0.0729	0.0190	9.6E+02	0.1485	0.7574	0.1270
N= 500	0.0116	0.6086	0.6086	0.0142	0.0371	0.0067	0.0238	0.0834	0.6170	0.0267

Table 2.7: Model (2) Many regressors: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .2$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.999	0.337	0.001	0.658	0.679	0.342	0.405	0.578	NA	NA
N= 200	1.000	0.824	0.001	0.578	0.680	0.315	0.372	0.664	NA	NA
N= 500	1.000	0.891	0.000	0.536	0.655	0.301	0.362	0.722	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	13.894	5.631	0.017	3.4E+145	1.0E+148	2.4E+03	7.8E+29	4.911	NA	NA
N= 200	15.405	14.239	0.009	5.1E+14	2.3E+76	0.842	5.9E+03	3.488	NA	NA
N= 500	15.786	15.181	0.006	7.8E+04	1.6E+35	0.508	3.664	2.856	NA	NA

MSE

	Ours(Linear)	Ours(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.0671	1.1653	1.1653	0.2553	0.3839	51.2306	2.5E+28	0.3352	0.9069	0.9106
N= 200	0.0224	0.7454	0.7454	0.0517	0.1330	0.0370	0.4507	0.1452	0.7597	0.0945
N= 500	0.0144	0.6080	0.6080	0.0226	0.0728	0.0142	0.0514	0.0841	0.6207	0.1003

Table 2.8: Model (3) Periodic: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .01$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.963	0.985	0.004	0.285	0.259	0.424	0.286	0.742	NA	NA
N= 200	0.999	0.999	0.002	0.847	0.690	0.926	0.792	0.902	NA	NA
N= 500	1.000	1.000	0.001	0.911	0.800	0.959	0.872	0.919	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.856	0.164	0.0004	5.6E+14	1.5E+97	1.291	6.3E+03	0.395	NA	NA
N= 200	0.893	0.177	0.0002	9.523	1.5E+19	0.304	3.5E+07	0.158	NA	NA
N= 500	0.889	0.178	0.0001	0.074	2.060	0.076	0.092	0.075	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.041	0.018	0.019	0.850	0.857	4.271	1.7E+02	0.629	0.018	0.936
N= 200	1.006	0.015	0.015	0.029	0.034	0.016	0.247	0.048	0.015	0.223
N= 500	0.998	0.015	0.015	0.006	0.008	0.003	0.010	0.007	0.015	0.177

Table 2.9: Model (3) Periodic: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .1$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.973	0.996	0.008	0.319	0.294	0.423	0.374	0.741	NA	NA
N= 200	1.000	1.000	0.005	0.861	0.758	0.901	0.794	0.878	NA	NA
N= 500	1.000	1.000	0.003	0.909	0.867	0.929	0.883	0.905	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.919	0.364	0.0008	5.6E+148	2.9E+149	33.542	1.2E+19	0.418	NA	NA
N= 200	1.001	0.402	0.0004	0.192	5.7E+03	0.189	0.308	0.205	NA	NA
N= 500	0.995	0.408	0.0003	0.121	0.138	0.120	0.133	0.148	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.040	0.032	0.037	0.867	0.882	1.002	8.145	0.663	0.029	1.019
N= 200	1.004	0.018	0.019	0.058	0.077	0.045	1.210	0.071	0.017	0.245
N= 500	0.997	0.016	0.016	0.021	0.031	0.018	0.032	0.028	0.016	0.153

Table 2.10: Model (3) Periodic: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .2$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.956	0.994	0.008	0.310	0.295	0.386	0.353	0.747	NA	NA
N= 200	1.000	1.000	0.006	0.875	0.809	0.894	0.822	0.888	NA	NA
N= 500	1.000	1.000	0.004	0.906	0.888	0.915	0.897	0.911	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.960	0.533	0.001	1.6E+04	1.4E+32	0.245	1.1E+04	0.447	NA	NA
N= 200	1.112	0.584	0.0006	0.226	1.363	0.224	0.285	0.255	NA	NA
N= 500	1.111	0.588	0.0004	0.151	0.179	0.148	0.174	0.211	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.039	0.048	0.057	0.909	0.927	0.922	1.694	0.677	0.040	1.099
N= 200	1.004	0.022	0.024	0.085	0.115	0.074	1.014	0.095	0.021	0.270
N= 500	0.998	0.018	0.018	0.035	0.052	0.031	0.054	0.055	0.017	0.177

Table 2.11: Model (4) Weierstrass function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .01$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.996	0.937	0.004	0.671	0.546	0.763	0.662	0.820	NA	NA
N= 200	0.998	1.000	0.002	0.816	0.702	0.854	0.745	0.855	NA	NA
N= 500	1.000	1.000	0.001	0.853	0.810	0.870	0.827	0.878	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.505	0.421	0.001	1.5E+149	Inf	1.5E+04	1.1E+25	0.282	NA	NA
N= 200	1.487	0.451	0.0008	3.644	6.7E+15	0.119	5.4E+04	0.118	NA	NA
N= 500	1.478	0.452	0.0005	0.071	0.083	0.070	0.078	0.083	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.96725	0.11399	0.11487	0.05453	0.06047	0.18238	0.30703	0.07752	0.08298	0.34459
N= 200	0.94167	0.07024	0.06968	0.00924	0.01118	0.00767	0.02718	0.01069	0.07408	0.04841
N= 500	0.93892	0.06762	0.06762	0.00349	0.00450	0.00277	0.00494	0.00379	0.07235	0.03953

Table 2.12: Model (4) Weierstrass function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .1$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.998	0.941	0.004	0.719	0.620	0.771	0.682	0.813	NA	NA
N= 200	1.000	1.000	0.003	0.841	0.793	0.849	0.806	0.868	NA	NA
N= 500	1.000	1.000	0.003	0.866	0.860	0.868	0.856	0.896	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.586	0.531	0.002	1.6E+15	6.7E+75	0.520	3.4E+10	0.317	NA	NA
N= 200	1.572	0.619	0.0009	0.159	26.761	0.154	0.193	0.185	NA	NA
N= 500	1.569	0.613	0.0006	0.111	0.140	0.109	0.134	0.158	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.96805	0.09917	0.10596	0.07723	0.08645	0.13016	1.83311	0.08607	0.09245	64.14131
N= 200	0.94161	0.07211	0.07136	0.01846	0.02481	0.01640	0.03263	0.01957	0.07513	0.05378
N= 500	0.93576	0.06749	0.06749	0.00844	0.01248	0.00773	0.01309	0.01241	0.07249	0.03235

Table 2.13: Model (4) Weierstrass function: $\sigma_\varepsilon^2/\mathrm{Var}(Y) = .2$

Coverage Probability of Confidence Sets with Asymptotic Level 95%

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.997	0.970	0.005	0.752	0.663	0.773	0.697	0.821	NA	NA
N= 200	1.000	1.000	0.004	0.855	0.843	0.858	0.845	0.889	NA	NA
N= 500	1.000	1.000	0.002	0.879	0.911	0.880	0.918	0.916	NA	NA

Average Width of Confidence Sets

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	1.675	0.656	0.002	0.550	6.0E+09	0.322	55.124	0.364	NA	NA
N= 200	1.663	0.773	0.001	0.193	0.314	0.188	0.225	0.250	NA	NA
N= 500	1.646	0.827	0.0008	0.137	0.176	0.132	0.169	0.238	NA	NA

MSE

	AB(Linear)	AB(Sieve)	Naive Sieve	NW	NW(US)	Local Linear	Local Linear(US)	RF	LASSO	NN
N= 50	0.97519	0.09743	0.10677	0.09675	0.11137	0.13021	0.61484	0.10431	0.09591	1.07405
N= 200	0.94289	0.07486	0.07399	0.02603	0.03486	0.02398	0.12853	0.03021	0.07680	0.10154
N= 500	0.93892	0.07065	0.07002	0.01289	0.01578	0.01149	0.01454	0.02309	0.07403	0.03928

2.10. Empirical illustrations

We consider empirical applications of the bound approach. The prediction problem of the MPG given car attributes is studied in Section 2.10.1. Section 2.10.2 applies the bound approach for inference for the shape of an Engle curve.

2.10.1. Prediction of Auto miles-per-gallon

In this section, we consider a regression exercise using Auto MPG Data Set from from UCI Machine Learning Repository (Dua and Graff (2017)) in which the purpose is to construct a confidence set for the miles-per-gallon (MPG) fuel consumption of a car model given 7 attributes, 3 of which are multi-valued discrete. This problem is also considered in Wager et al. (2014), who propose a confidence set based on random forest estimates with the infinitesimal Jackknife variance. The dataset contains 392 observations and we split the dataset into a training and test sets of equal sizes 196. Since the outcome can differ considerably according to how the dataset is divided, we generated 100 pairs of training and test sets by performing random splits 100 times.

As in Section 2.9.3.1, we select an approximate model from the set $S$of candidate models where $p(x)$ of an element of S is specified as follows: let $x_c = (x_1, x_2, x_3)$ be the three-dimensional continuous covariate and define the polynomial series $x_{c,2}$ of x_c of order up to 2 without cross terms:

$$x_{c,2} = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2) \tag{2.10.8}$$

Then, $p(x)$ of an element of S is a subvector of the vector $x_{c,2}$ augmented by dummy variables based on the three discrete variables. Then, given a chosen model, we construct an approximation bound in 2.9.5 .

Our method is compared with the Nadaraya-Watson (NW) and local linear estimators with the bandwidth determined by either the least squares cross validation (**CV**) or by undersmoothing (**UC**), and random forests . For random forests, we consider two terminal sizes where the maximum terminal size is 5 with bootstrap resampling (**bts**) and the maximum terminal size is 1 with subsampling (**ss**). The case (i) is the default choice in the randomForest function in the R package *randomForest* while (ii) is suggested in the literature, such as Wager (2014) in the context of construction of a confidence set based on

random forest estimates. For each method, we compute (1) coverage probability of Y, (2) average width of 95% confidence set and (3) MSE (mean squared error). Note that in (1), $m(X)$ is not observed, we instead construct a confidence set for the dependent variable, the MPG fuel consumption as a proxy variable. In (2), for each pair of training/test samples, we estimate each model using the training set and then construct a confidence set with nominal size .95 for the dependent variable in the test sample given the estimated model and the corresponding attributes. The MSE is computed with respect to Y, however recall that the minimizer of the MSE with respect to Y also minimizes the MSE with respect to $m(X)$ and thus comparisons of the MSE across different methods provide properties of point estimates as predictor of $m(X)$.

Table 2.14 report the averages of statistic (1)-(3) for each method. We see that confidence sets based on kernel estimators are undersized regardless of the choice of the bandwidth parameter. The approximation bound and random forest-based confidence sets have valid coverage probabilities. However, among the three methods, the approximation bound approach leads to a confidence set with the shortest width. This shows that the bound approach yields the most informative confidence set among the methods in consideration. In addition, point estimates based on our method yields the smallest MSE. This suggests that while our point estimates are only predictors from an approximate and likely misspecified model and thus the approximation errors are present even asymptotically, such models could also provide better point predictors of the underlying model $m(X)$ than the alternative methods in practice.

Table 2.14: Confidence sets for the MPG

	(1) Coverage probability of 95% CS	(2) Width of the confidence set	(3) MSE
Approximation Bound	**.9746**	**15.31**	**.135**
NW (**CV**)	.8959	9322.67	.164
NW (**UC**)	.8863	9322.44	.164
Local Linear (**CV**)	.8850	11.98	.199
Local Linear (**UC**)	.8823	11.77	.199
Random Forest (**bts**)	**.9795**	15.70	.140
Random Forest (**ss**)	**.9857**	17.34	.139

2.10.2. Shape of Engle curve

The Engle curve describes a collective relationship between household income level and expenditure on a certain good. Understanding the shape of the Engle curve is a key element of welfare and commodity tax policy evaluation. Estimation of the curve has been extensively studied in the literature (Banks, Blundell and Lewbel (1997), Blundell, Duncan and Pendakur (1998), Lewbel and Pendakur (2008), Blundell, Browning and Crawford (2003), Blundell, Browning and Crawford (2008)) and nonlinearity and non-monotonicity of the curve has been reported. In this section, we consider application of the bound method to this problem, in particular an inference problem for the demand response for alcohol using Family Expenditure Survey (FES) from 2000-2001 collected by Office for National Statistics (2002). Alcohol consumption often exhibits a non-monotonic relationship with income level and is also of interest due to its implications to consumers' health (Banks et al. (1997), Andrienko, Nemtsov et al. (2005), Yakovlev (2018)).

As a set of approximate models, we consider polynomial models of order to up to 6. We assume continuity of the underlying regression function and thus the Chebyshev bound in (2.7.16) applies. Figure 2.1 presents fitting of each model along with 95% confidence sets for 20 out-of-sample data points. We see that as the model complexity increases, the width of confidence sets gets smaller up to the quadratic model. Employing an additional term higher than 4th-order does not alter the shape of the fitting much as well as the value of the associated approximation bound. This can be seen from Table 2.15, which reports regression results for each model. Note that the covariates are orthogonalized by the Gram-Schmidt process in order to isolate the contribution of additional terms. It is worth noting that the quintic (6-th order) term is not significant and the value of the approximation bound almost does not change from the quartic model to the quintic model. Including further higher terms does not improve the approximation bound. It indicates that a polynomial models of any order is likely misspecified. Even in the presence of the effect of specification error by the approximation bound incorporates, confidence sets based on the quintic models provide meaningful information on the shape of the curve.

In Figure 2.2, model fitting by kernel regression and random forest is reported. Random forest estimates exhibit extreme non-smoothness and large variation and the wide range of the confidence sets reflect poor prediction of the estimates. On the other hand, kernel regression estimates show a relatively similar curve to that from fitting by the quartic model.

The out-of-sample MSE of the quartic model (90.0%) is smaller than the smallest one from the four kernel-based models (93.6%) while the average width of confidence sets based on any of the four kernel-based models is shorter than the one from the quartic model (approximately 80% shorter). Thus, whether these kernel regression-based confidence sets have the right coverage probabilities is in question. Note that as we observe in the Monte Carlo simulation in Section 2.9, confidence sets based on kernel regression estimates are often undersized.

Thus, a simple polynomial approximate model provides superior fitting to the alternative nonparametric methods and the use of the approximation bound allows to make inference for the shape of the curve in a robust manner.

Figure 2.1: Polynomial fitting and 95% confidence sets

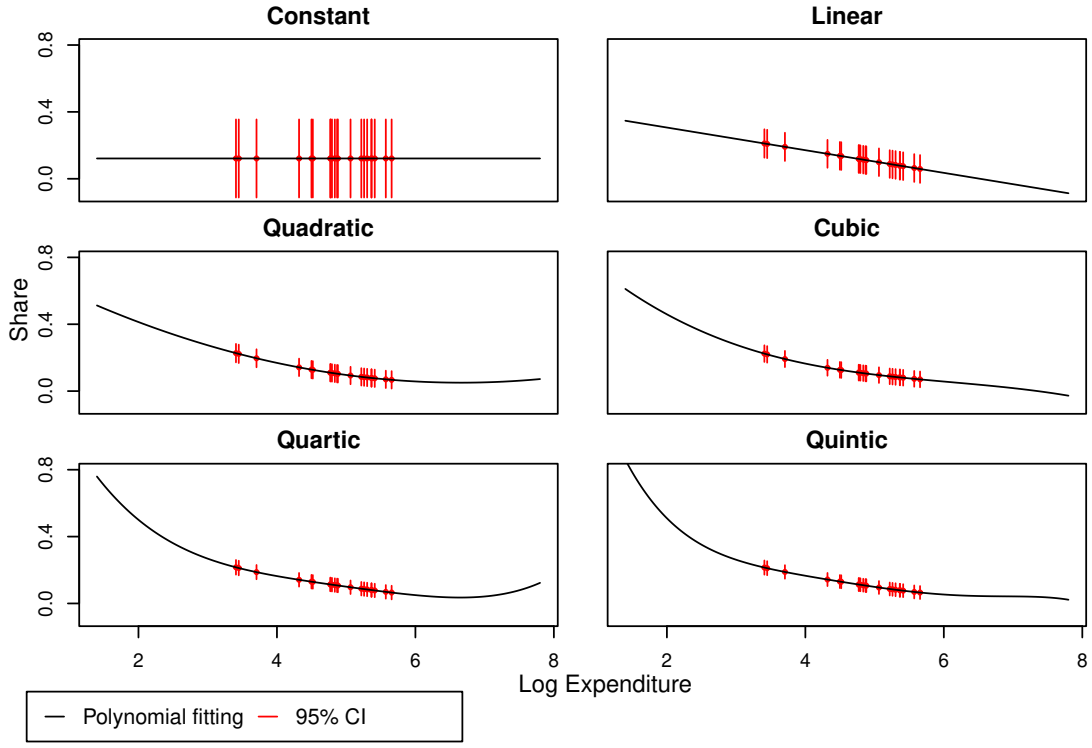


Figure 2.2: Fitting by nonparametric methods with 95% confidence sets

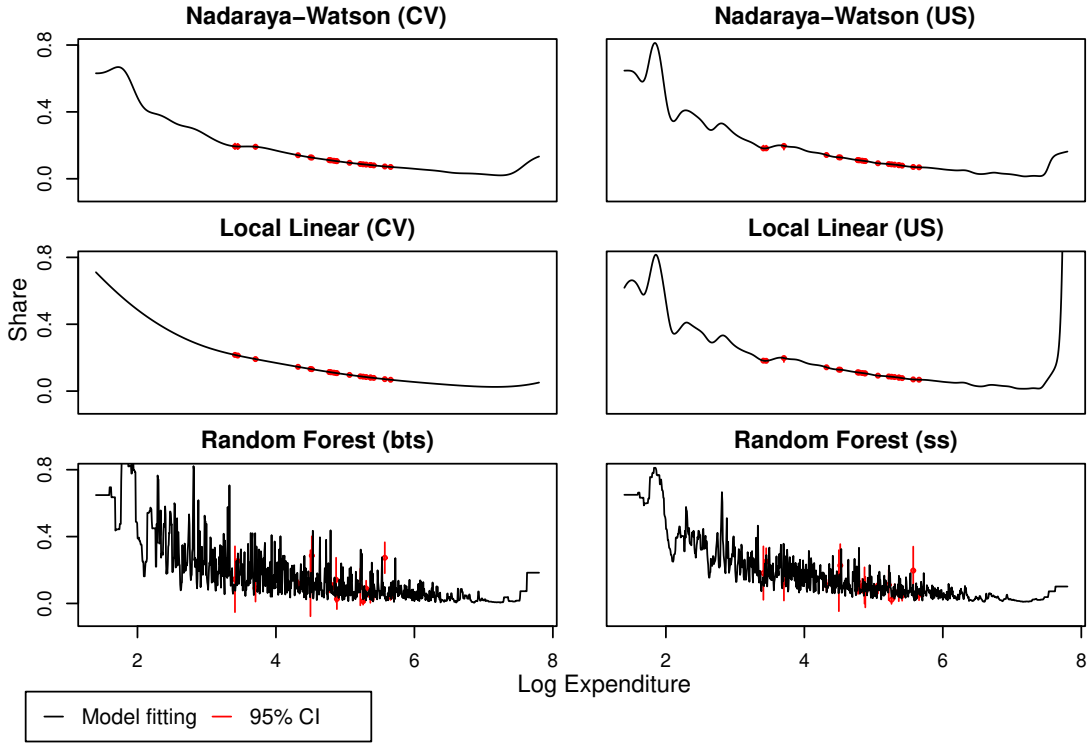


Table 2.15. Regression results of polynomial models of orders up to 6

	Dependent variable:						
	Expenditure share for alcohol						
	Constant	Linear	Quadratic	Cubic	Quartic	Quntic	Sextic
I := log(expenditure)		−0.065***	−0.065***	−0.065***	−0.065***	−0.065***	−0.065***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I_o^2			0.014***	0.014***	0.014***	0.014***	0.014***
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I_o^3				−0.002*	−0.002*	−0.002*	−0.002*
				(0.001)	(0.001)	(0.001)	(0.001)
I_o^4					0.002**	0.002**	0.002**
					(0.001)	(0.001)	(0.001)
I_o^5						−0.001	−0.001
						(0.0004)	(0.0004)
I_o^6							0.0004
							(0.0003)
Constant	0.121***	0.121***	0.121***	0.121***	0.121***	0.121***	0.121***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	6,790	6,790	6,790	6,790	6,790	6,790	6,790
R^2	0.000	0.131	0.139	0.140	0.141	0.141	0.141
Adjusted R^2	0.000	0.131	0.139	0.139	0.140	0.140	0.140
TIC	-27790.52	-28742.96	-28803.43	-28802.21	-28805.58	-28804.35	-28803.3
Residual Std. Error	0.129 (df = 6789)	0.120 (df = 6788)	0.120 (df = 6787)	0.120 (df = 6786)	0.120 (df = 6785)	0.120 (df = 6784)	0.120 (df = 6783)
Out of Sample MSE	0.9853731	0.9146866	0.9003884	0.897451	0.9007055	0.9007055	0.9034706
F Statistic		1,025.752*** (df = 1; 6788)	549.815*** (df = 2; 6787)	367.602*** (df = 3; 6786)	277.522*** (df = 4; 6785)	222.566*** (df = 5; 6784)	185.773*** (df = 6; 6783)
Approximation bound	0.2324971	0.0817974	0.0509312	0.0457895	0.0393426	0.0374863	0.0368927

Note: *p<0.1; **p<0.05; ***p<0.01

2.11. Conclusion

This paper establishes approximation bounds for regression models of arbitrary form and proposes inference based on confidence sets. Our framework is valid for any generating processes including those where identification of the regression function fails. The simulation studies support validity of our approach with fairly small sample sizes and are in favor of our approach over alternative methods in the literature, such as Kernel regression, the method of sieve and random forest.

In future work, it is of interest to further develop the conditional approach discussed in Section 2.6. As mentioned there, inference based on set conditioning can be conducted as in the unconditional case primarily considered in this paper by restricting the support $\mathscr{X}$ of the conditioning variable X as long as the samples are large enough that observations in such restricted support set are available. When $m(x)$ behaves in a smooth manner in a certain sense on a set $A \subset \mathscr{X}$, the bound approach can be employed to derive meaningful information on $m(x_A)$ evaluated at any point in $x_A \in A$ even when $m(\cdot)$ is weakly identified. Additionally, by conditioning on two disjoint sets A, B, such approach can be extended to study the difference $m(x_A) - m(x_B)$ for $x_A \in A$ and $x_B \in B$.

2.A. Nonparametric identification of functions

We review the concept of nonparametric identification of functions, following Matzkin (2007) and Lewbel (2019), among others. We primarily adapt the notation in Matzkin (2007).

Denote the set of all functions and distributions that satisfy the restrictions imposed by a model, by S. We denote any element of S by ξ and the true value of ξ by ξ_0. For any element $\xi \in S$, we denote by $F_{Y,X}(\cdot,\cdot;\xi)$ the distribution of the observable variables (Y,X) generated by ξ. When $\xi = \xi_0$, $F_{Y,X}$.

We call $\eta = \Psi(\xi)$ a feature of ξ for any function $\Psi : S \to \Lambda$. We denote by $\eta_0 = \Psi(\xi_0)$ the true value of a feature of ξ_0. Given $\psi \in \Psi(S)$, we define $\Gamma_{Y,X}(\psi, S)$ to be the set of all probability distributions of (Y,X) that are consistent with ψ and S. Formally,

$$\Gamma_{Y,X}(\psi, S) = \{F_{Y,X}(\cdot,\cdot;\xi) \mid \xi \in S, \Psi(\xi) = \psi\}. \tag{2.A.9}$$

The following notion plays a key role in identification of the feature ψ^*.

Definition 2.A.1 OBSERVATIONAL EQUIVALENCE. *$\psi, \psi^{'} \in \Lambda$ are observationally equivalent in the model S if*

$$\left[\Gamma_{Y,X}(\psi, S) \cap \Gamma_{Y,X}(\psi^{'}, S)\right] \neq \emptyset. \tag{2.A.10}$$

It states that for $\psi, \psi' \in \Lambda$, if there exists elements $\xi, \xi' \in S$ such that $\Psi(\xi) = \psi$, $\Psi(\xi') = \psi'$ and the observable distributions $F_{Y,X}$ generated by ξ and ξ' are equivalent:

$$F_{Y,X}(\cdot,\cdot;\xi) = F_{Y,X}(\cdot,\cdot;\xi'), \tag{2.A.11}$$

then ψ and ψ' are observationally equivalent.

Then, the feature ψ_0 is identified if there does not exist $\psi \in \Lambda$ such that it is observationally equivalent to ψ_0 as defined below.

Definition 2.A.2 IDENTIFICATION. *$\psi_0 \in \Lambda$ is identified in model S if for any $\psi \in \Lambda$ such that $\psi \neq \psi_0$*

$$[\Gamma_{Y,X}(\psi, S) \cap \Gamma_{Y,X}(\psi_0, S)] = \emptyset. \tag{2.A.12}$$

In the nonparametric regression setup, S corresponds to a collection of pairs $(m, F_{\varepsilon,X})$ where m is an integrable function on $\mathscr{X}$ and $F_{\varepsilon,X}$ is the joint distribution of the expectation error $\varepsilon := Y - \mathbb{E}[Y \mid X]$ and the conditioning variable X.

2.B. Consistency and asymptotic normality of an extremum estimator

Results on Consistency and asymptotic normality of an extremum estimator are reviewed.

Assumption 2.B.1 . *$\{w_i\}_{i=1}^n = \{y_i, x_i\}_{i=1}^n$ is ergodic stationary.*

Assumption 2.B.2 . *The objective function $Q(\theta)$ is of the form*

$$Q(\theta) = \mathbb{E}[g(w_i; \theta)] \tag{2.B.13}$$

and the sample objective function $Q_n(\theta)$ is given by

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta) \tag{2.B.14}$$

where g is a real-valued function of (w_i, θ) and the following conditions hold:

1. *θ_0 is in the interior of a convex parameter space Θ.*
2. *$m(w_i;\theta)$ is measurable in w_i for any $\theta \in \Theta$.*
3. *$\mathbb{E}[|m(w_i;\theta)|] < \infty$ for any $\theta \in \Theta$.*

Assumption 2.B.3 . *The objective function $Q(\theta)$ is of the form*

$$Q(\theta) = \mathbb{E}[g(w_i;\theta)] \tag{2.B.15}$$

and the sample objective function $Q_n(\theta)$ is given by

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta) \tag{2.B.16}$$

where g is a real-valued function of (w_i, θ) and the following conditions hold:

1. *θ_0 is in the interior of a convex parameter space Θ.*
2. *$m(w_i;\theta)$ is measurable in w_i for any $\theta \in \Theta$.*
3. *$\mathbb{E}[|m(w_i;\theta)|] < \infty$ for any $\theta \in \Theta$.*
4. *$m(w_i;\theta)$ is twice continuous differentiable in θ for any w_i.*
5. *For some $q \times q$ positive-definite matrix Σ,*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} s(w_i\,;\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma). \tag{2.B.17}$$

where $s(w_i;\theta)$ is defined as

$$s(w_i;\theta) = \frac{\partial m(w_i;\theta)}{\partial \theta} \tag{2.B.18}$$

6. *For some neighborhood $\mathscr{N}$ of θ_0,*

$$\mathbb{E}[\sup_{\theta \in \mathscr{N}} \|\mathbf{H}(w_i;\theta)\|] < \infty \tag{2.B.19}$$

where $\mathbf{H}(w_i;\theta)$ is defined as

$$\mathbf{H}(w_i;\theta) = \frac{\partial s(w_i;\theta)}{\partial \theta'}. \tag{2.B.20}$$

Under the assumptions above, we have the following result.

Lemma 2.B.1 HAYASHI (2011), PROPOSITION 7.8. *Suppose Assumption 2.8.1, 2.B.1-2.B.3 hold. Then,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathscr{N}(0, \Lambda) \tag{2.B.21}$$

where

$$\Lambda = \left(\mathbb{E}[\mathbf{H}(w_i;\theta)_0]\right)^{-1} \Sigma \left(\mathbb{E}[\mathbf{H}(w_i;\theta)_0]\right)^{-1} \tag{2.B.22}$$

2.C. Proofs

PROOF OF PROPOSITION 2.3.1 Using the notation in Appendix, 2.A, define S as

$$\begin{aligned} S = \{(m^*, F^*_{\varepsilon,X}) \,|m^* \in \mathscr{M}\ F^*_{\varepsilon,X} \text{ is a distribution of } (\varepsilon, X) \\ \text{such that the density } f^*_X \text{ of } X \text{exist s and is continuous,} \\ f_X(x) > 0 \text{ for any } x \in \mathscr{X}, \text{ and } \mathbb{E}[\varepsilon \,|\, X = x; F^*_{\varepsilon,X}] = 0, \forall x \in \mathscr{X}\} \end{aligned} \tag{2.C.1}$$

where $\mathbb{E}[\varepsilon \,|\, X = x; F^*_{\varepsilon,X}]$ is the expectation of ε given $X = x$ calculated using $F^*_{\varepsilon,X}$. Pick any $x_0 \in \mathscr{X}$ and let $\xi = (m, F^{'}_{\varepsilon,X}) \in S$. For an arbitrary $c \neq 0$, define a function $\tilde{m}$ as

$$m^*(x_0) = \begin{cases} m(x_0) + c & \text{if } x = \bar{x} \\ m(x_0) & \text{if } x \neq \bar{x}. \end{cases} \tag{2.C.2}$$

and let $\tilde{\xi} = (m^*, F^{'}_{\varepsilon,X}) \in S$. Then, we have

$$F_{Y|X=x}(y\,;\,\xi) = F_{Y|X=x}(y\,;\,\tilde{\xi}) \tag{2.C.3}$$

except at $x = x_0$. Since $\Pr(X = x_0) = 0$, we have

$$F_{Y,X}(\cdot,\cdot;\xi) = F_{Y,X}(\cdot,\cdot;\tilde{\xi}). \tag{2.C.4}$$

Thus, ξ and $\tilde{\xi}$ are observationally equivalent.

□

PROOF OF LEMMA 2.3.2

In this case, we have

$$S = \left\{(m^*, F^*_{\varepsilon,X}) \,|\, F^*_{\varepsilon,X} \text{ is a distribution of } (\varepsilon, X) \text{ such that } \Pr(X = x_0) > 0 \text{ and } \mathbb{E}[\varepsilon \,|\, X = x\,;\, F^*_{\varepsilon,X}] = 0\right\}. \tag{2.C.5}$$

Let $(m, F'_{\varepsilon,X})$, $(\tilde{m}, F^*_{\varepsilon,X}) \in S$ such that $\tilde{m}(\bar{x}) \neq m(\bar{x})$. Then,

$$\begin{aligned} \mathbb{E}\left[Y \mid X = x_0; m, F'_{\varepsilon,X}\right] &= \tilde{m}(\bar{x}) + \mathbb{E}\left[\varepsilon \mid X = \bar{x}; m, F'_{\varepsilon,X}\right] = \tilde{m}(\bar{x}) \quad (2.C.6) \\ \mathbb{E}\left[Y \mid X = x_0; \tilde{m}, F^*_{\varepsilon,X}\right] &= m_0(\bar{x}) + \mathbb{E}\left[\varepsilon \mid X = \bar{x}; \tilde{m}, F^*_{\varepsilon,X}\right] = m(\bar{x}) \quad (2.C.7) \end{aligned}$$

so that $F'_{\varepsilon|X=\bar{x}} \neq F^*_{\varepsilon|X=\bar{x}}$. Then, since $\Pr(X = \bar{x}) > 0$, $\tilde{F}_{\varepsilon,X} \neq F^*_{\varepsilon,X}$ and thus

$$F_{Y,X}(\cdot,\cdot; m, F'_{\varepsilon,X}) \neq F_{Y,X}(\cdot,\cdot; \tilde{m}, F^*_{\varepsilon,X}). \quad (2.C.8)$$

Thus, $m(x_0)$ is identified.

□

PROOF OF PROPOSITION 2.3.3

Consider first the case where for any n, $x_i \neq x_0$ for any $i = 1, \ldots, n$. Then, since there is no restriction on F_n, the set of observations $\{(y_i, x_i)\}_{i=1}^n$ does not convey any information on $m(x_0)$ and thus an optimal test is to draw $U \sim \text{Unif}(0,1)$ and then reject $H_0(\mu_0; x_0)$ if and only if $U \leq \alpha$. Suppose now the case where there exists some $i = 1, \ldots, n$ such that $x_i = x_0$. Then, construct a purposive sample $\left\{y_j^*\right\}_{j=1}^l$ from $\{(y_i, x_i)\}_{i=1}^n$ by only keeping y_i such that $x_i = x_0$. Note $l = \sum_{i=1}^n \mathbf{1}\{x_i = x_0\}$. Then, $\left\{y_j^*\right\}_{j=1}^l$ is a set of i.i.d. observations from the conditional distribution of Y given $X = x_0$. Then, Theorem 1 of Bahadur and Savage (1956) applies and the conclusion □

PROOF OF LEMMA 2.4.1 In a finite-dimensional inner-product space, any Chebyshev set is closed and convex. The other direction follows from the projection theorem. See Proposition 12.7 p.306 in Deutsch (2012). □

PROOF OF LEMMA 2.4.2 For any $c \in \boldsymbol{C}$

$$\|Y - c\|^2 = \|Y - q\|^2 + \|q - c\|^2 \quad (2.C.9)$$

so that

$$\|Y - q\|^2 \leq \|Y - c\|^2 \quad , \forall c \in \boldsymbol{C} \quad (2.C.10)$$

and the equality holds if and only if $q = c$, $P-a.e.$ □

PROOF OF LEMMA 2.4.3 They are elementary results in functional analysis. See, for example, Lemma 6.54 in Charalambos and Border (2006). □

PROOF OF LEMMA 2.4.4 By assumption, there exists a unique element $P_{\mathbf{C}}Y$ of $\boldsymbol{C}(X)$ such that

$$\|Y - P_{\mathbf{C}}Y\| < \|Y - h(X)\|, \quad \forall h(X) \in \boldsymbol{C}(X) \setminus \{P_{\mathbf{C}}Y\}. \tag{2.C.11}$$

Equivalently,

$$\|m(X) - P_{\mathbf{C}}Y\| + 2\langle m(X) - P_{\mathbf{C}}Y, \varepsilon\rangle < \|m(X) - h(X)\| + 2\langle m(X) - h(X), \varepsilon\rangle. \tag{2.C.12}$$

But, $m(X) - P_{\mathbf{C}}Y, m(X) - h(X) \in \mathscr{M}(X)$ and $\varepsilon \perp \mathscr{M}(X)$ so that

$$\langle m(X) - P_{\mathbf{C}}Y, \varepsilon\rangle = \langle m(X) - h(X), \varepsilon\rangle = 0. \tag{2.C.13}$$

Hence,

$$\|m(X) - P_{\mathbf{C}}Y\| < \|m(X) - h(X)\|, \quad \forall h(X) \in \boldsymbol{C}(X) \setminus \{P_{\mathbf{C}}Y\}. \tag{2.C.14}$$

This in turn implies

$$P_{\mathbf{C}}Y = P_{\mathbf{C}}m(X). \tag{2.C.15}$$

□

PROOF OF LEMMA 2.4.5 By assumption, $P_{\mathbf{C}}Y$ is unique and there exists a unique element $\theta_0 \in \Theta$ such that

$$P_{\mathbf{C}}Y = h(X;\theta_0). \tag{2.C.16}$$

Then, by Lemma 2.4.4, $P_{\mathbf{C}}m(X) = h(X;\theta_0)$ so that the assertions hold. □

PROOF OF LEMMA 2.5.1 By Markov's inequality for a nondecreasing nonnegative func-

tion,

$$\Pr(|Z| \geq d) \leq \frac{Ef(|Z|)}{f(d)} \tag{2.C.17}$$

and the results follow by observing

$$Ef(|Z|) \leq \bar{f}. \tag{2.C.18}$$

□

PROOF OF LEMMA 2.5.2 By conditional Jensen's inequality

$$\mathbb{E}[f(Y-h(X))|X] \geq f(\mathbb{E}[Y-h(X)|X]) = f(m(X)-h(X)) \tag{2.C.19}$$

and thus by the law of iterated expectation,

$$\mathbb{E}[\mathbb{E}[f(Y-h(X))|X]] \geq \mathbb{E}[f(m(X)-h(X))] \tag{2.C.20}$$

or

$$\mathbb{E}[f(Y-h(X))] \geq \mathbb{E}[f(m(X)-h(X))]. \tag{2.C.21}$$

□

PROOF OF COROLLARY 2.5.3 Observe $|x|^{\alpha}$ $(\alpha \geq 1)$ and $\exp(tx)(t \in \mathbb{R})$ are convex in x and apply Lemma 2.5.2. □

PROOF OF COROLLARY 2.5.4 Noting $\mathrm{Cov}(P_{\mathscr{L}}(X), m(X)-P_{\mathscr{L}}(X)) = 0$ and $\mathbb{E}[P_{\mathscr{L}}(X)] = E[m(X)]$,

$$\begin{aligned}
\mathrm{Var}(m(X)) &= E\left[(m(X)-E[m(X)])^2\right] && (2.C.22)\\
&= E\left[\{(m(X)-P_{\mathscr{L}}(X))+(P_{\mathscr{L}}(X)-E[m(X)])\}^2\right] && (2.C.23)\\
&= E\left[(m(X)-P_{\mathscr{L}}(X))^2\right]+E\left[(P_{\mathscr{L}}(X)-E[P_{\mathscr{L}}(X)])^2\right] && (2.C.24)\\
&\quad +2E[(m(X)-P_{\mathscr{L}}(X))(P_{\mathscr{L}}(X)-E[P_{\mathscr{L}}(X)])] && (2.C.25)\\
&= \mathrm{Var}(m(X)-P_{\mathscr{L}}(X))+\mathrm{Var}(P_{\mathscr{L}}(X)) && (2.C.26)
\end{aligned}$$

$$Y - P_{\mathscr{L}}(X) = [Y - m(X)] + [m(X) - P_{\mathscr{L}}(X)] \tag{2.C.27}$$

and

$$\begin{aligned} \operatorname{Var}(Y - P_{\mathscr{L}}(X)) &= \operatorname{Var}(Y - m(X)) + \operatorname{Var}(m(X) - P_{\mathscr{L}}(X)) && (2.C.28)\\ &\quad + \operatorname{Cov}(Y - m(X), m(X) - P_{\mathscr{L}}(X)) && (2.C.29)\\ &= \operatorname{Var}(Y - m(X)) + \operatorname{Var}(m(X) - P_{\mathscr{L}}(X)). && (2.C.30) \end{aligned}$$

□

PROOF OF PROPOSITION 2.5.5 By Markov's inequality for a nondecreasing nonnegative function,

$$\Pr(|m(X) - P_{\mathscr{L}}(X)| \geq c) \leq \frac{Ef(|m(X) - P_{\mathscr{L}}(X)|)}{f(c)}. \tag{2.C.31}$$

Since f is convex and nondecreasing and $h(x) \equiv |x|$ is convex, the composition $f(h(x))$ is also convex. Hence, by the law of total expectation and conditional Jensen's inequality

$$\begin{aligned} Ef(|Y - P_{\mathscr{L}}(X)|) &= E[E[f(|Y - P_{\mathscr{L}}(X)|)|X]] && (2.C.32)\\ &\geq E[f(|E[Y - P_{\mathscr{L}}(X)|X]|)] && (2.C.33)\\ &= E[f(|m(X) - P_{\mathscr{L}}(X)|)]. && (2.C.34) \end{aligned}$$

Combined with (2.C.31),

$$\Pr(|m(X) - P_{\mathscr{L}}(X)| \geq c) \leq \frac{Ef(|Y - P_{\mathscr{L}}(X)|)}{f(c)}. \tag{2.C.35}$$

If $c_\alpha = \infty$, the second assertion is trivially true. Now, assume $c_\alpha < \infty$. Then, by definition of c_α

$$\frac{Ef(|Y - P_{\mathscr{L}}(X)|)}{f(c_\alpha)} \leq \alpha. \tag{2.C.36}$$

□

PROOF OF PROPOSITION 2.5.6 Note $c_{\alpha,f_1} \leq c_{\alpha,f_2}$ $(f_1, f_2 \in \mathscr{K})$ implies

$$\left[P_{\mathbf{C}}m(X) - c_{\alpha,f_1}, P_{\mathbf{C}}m(X) + c_{\alpha,f_1}\right] \subset \left[P_{\mathbf{C}}m(X) - c_{\alpha,f_2}, P_{\mathbf{C}}m(X) + c_{\alpha,f_2}\right], \tag{2.C.37}$$

so that

$$\left[P_{\mathbf{C}}m(X) - c_{\alpha}^{(\mathscr{K})}, P_{\mathbf{C}}m(X) + c_{\alpha}^{(\mathscr{K})}\right] = \cap_{f\in\mathscr{K}}\left[P_{\mathbf{C}}m(X) - c_{\alpha,f}, P_{\mathbf{C}}m(X) + c_{\alpha,f}\right]. \tag{2.C.38}$$

Observe by Proposition 2.5.5

$$\inf_{f\in\mathscr{K}} \Pr\left(m(X) \in \left[P_{\mathbf{C}}m(X) - c_{\alpha,f}, P_{\mathbf{C}}m(X) + c_{\alpha,f}\right]\right) \geq 1 - \alpha \tag{2.C.39}$$

and by the monotone continuity property,

$$\inf_{f\in\mathscr{K}} \Pr\left(m(X) \in \left[P_{\mathbf{C}}m(X) - c_{\alpha,f}, P_{\mathbf{C}}m(X) + c_{\alpha,f}\right]\right) = \Pr\left(m(X) \in \cap_{f\in\mathscr{K}}\left[P_{\mathbf{C}}m(X) - c_{\alpha,f}, P_{\mathbf{C}}m(X) + c_{\alpha,f}\right]\right). \tag{2.C.40}$$

□

PROOF OF PROPOSITION 2.5.7 Fix the distribution F_X^* of X. Consider any integrable function $m(\cdot)$ on $\mathscr{X}$ such that

$$\mathbb{E}\,|m(X)| < \infty. \tag{2.C.41}$$

Let $Y = m(X)$. The joint distribution $F_{Y,X}(F_X^*, m)$ given F_X and $m(\cdot)$ is equal to $F_{m(X),X}$ so that

$$Y - P_{\mathbf{C}}m(X)) = m(X) - P_{\mathbf{C}}m(X) \tag{2.C.42}$$

with probability one. Now, for any $\beta \geq 1$, there exists a random variable V such that

$$\Pr\left(|V_\beta| \geq c_\alpha^{(\beta)}\right) = \alpha. \tag{2.C.43}$$

where

$$c_\alpha^{(\beta)} = \frac{\mathbb{E}|V|^\beta}{\alpha}^{1/\beta}. \tag{2.C.44}$$

For example, for $\beta = 2$, V_2 defined as

$$V_2 = \begin{cases} -c_\alpha^{(2)} & \text{with probability } \dfrac{1}{2\left(c_\alpha^{(2)}\right)^2}, \\ 0 & \text{with probability } 1 - \dfrac{1}{\left(c_\alpha^{(2)}\right)^2}, \\ c_\alpha^{(2)} & \text{with probability } \dfrac{1}{2\left(c_\alpha^{(2)}\right)^2}. \end{cases} \tag{2.C.45}$$

satisfies 2.C.43. Then, given the distribution F_X^* of X such that $m(X) - P_{\mathbf{C}} m(X) \overset{d}{\sim} V_\beta$ for given β,

$$\begin{aligned} \Pr_{F_{Y,X}(F_X^*,m)}(|Y - P_{\mathbf{C}} m(X)| \geq c_\alpha^{(\mathscr{K}_0)}) &= \Pr(|Y - P_{\mathbf{C}} m(X))| \geq c_\alpha^{(\beta)}) && \text{(2.C.46)} \\ &= \Pr(|m(X) - P_{\mathbf{C}} m(X))| \geq c_\alpha^{(\beta)}) && \text{(2.C.47)} \\ &= \Pr(|V_\beta| \geq c_\alpha^{(\beta)}) = \alpha. && \text{(2.C.48)} \end{aligned}$$

Since

$$\sup_{f \in \mathscr{F}} \Pr(|Y - P_{\mathbf{C}} m(X))| \geq c_\alpha^{(\mathscr{K}_0)}) \geq \Pr_{F_{Y,X}(F_X^*,m)}(|Y - P_{\mathbf{C}} m(X)| \geq c_\alpha^{(\mathscr{K}_0)}), \tag{2.C.49}$$

the assertion follows.

□

PROOF OF PROPOSITION 2.5.8 Define

$$c_{\alpha,f}^*(m) = \inf\left\{ c \in \mathbb{R}_{++} : f(c) = \frac{\mathbb{E} f(|m(X) - P_{\mathbf{C}} m(X)|)}{\alpha} \right\}, \quad m \in \mathscr{M}(X; F_Y, F_X). \tag{2.C.50}$$

Then, by Proposition 2.5.2,

$$c_{\alpha,f}^*(m) \leq c_{\alpha,f} \quad \forall m \in \mathscr{M}(X; F_Y, F_X). \tag{2.C.51}$$

so that

$$\begin{aligned}\sup_{m\in\mathscr{M}(F_Y,F_X)} \Pr_F\left(|m(X)-P_{\mathbf{C}}m(X)|\geq c_{\alpha,f}\right) &\leq \sup_{m\in\mathscr{M}(F_Y,F_X)} \Pr_F\left(|m(X)-P_{\mathbf{C}}m(X)|\geq c^*_{\alpha,f}(m)\right)\\ &\leq \alpha. \end{aligned} \tag{2.C.52}$$

□

PROOF OF LEMMA 2.5.9 For any $h\in\mathscr{H}_{ND(\geq c)}$,

$$\begin{aligned} Eh(Z) &= E\left[h(Z)\mathbf{1}\{Z\geq c\}\right]+E\left[h(Z)\mathbf{1}\{Z<c\}\right] && \text{(2.C.53)}\\ &\geq E\left[h(Z)\mathbf{1}\{Z\geq c\}\right] && \text{(2.C.54)}\\ &\geq \inf_{z\geq c} h(z)\,E\left[\mathbf{1}\{Z\geq c\}\right]=h(c)\Pr(Z\geq c) && \text{(2.C.55)}\end{aligned}$$

and hence

$$\Pr(Z\geq c)\leq \frac{Eh(Z)}{h(c)} \quad \forall h\in\mathscr{H}_{ND(\geq c)}. \tag{2.C.56}$$

For any $h\in\mathscr{H}_{NI(\leq c)}$,

$$\begin{aligned} Eh(Z) &= E\left[h(Z)\mathbf{1}\{Z\geq c\}\right]+E\left[h(Z)\mathbf{1}\{Z<c\}\right] && \text{(2.C.57)}\\ &\geq E\left[h(Z)\mathbf{1}\{Z<c\}\right] && \text{(2.C.58)}\\ &\geq \inf_{z<c} h(z)\,E\left[\mathbf{1}\{Z<c\}\right]\geq h(c)\Pr(Z<c) && \text{(2.C.59)}\end{aligned}$$

and by taking the complement,

$$\Pr(Z\geq c)\geq 1-\frac{Eh(Z)}{h(c)} \quad \forall h\in\mathscr{H}_{NI(\leq c)}. \tag{2.C.60}$$

□

PROOF OF THEOREM 2.5.10 By Lemma 2.5.9,

$$\Pr(Z\geq c)\leq \inf_{h\in\mathscr{H}_{ND(\geq c),\vee}} \frac{Eh(m(X)-P_{\mathscr{L}}(X))}{h(c)} \tag{2.C.61}$$

and for any $h \in \mathscr{H}_{ND(\geq c),\vee}$,

$$Eh\left(m\left(X\right)-P_{\mathscr{L}}\left(X\right)\right) \leq Eh\left(Y-P_{\mathscr{L}}\left(X\right)\right) \tag{2.C.62}$$

by Proposition 2.5.2. □

PROOF OF LEMMA 2.6.1 For any function h,

$$E\left[\left(Y-E\left[Y|X\right]\right)^2\right] \leq E\left[\left(Y-h\left(X\right)\right)^2\right] \tag{2.C.63}$$

and equality holds if and only if $\mathbb{E}\left[Y|X\right]=h\left(X\right)$ with probability one. Similarly, for any function h_A,

$$E\left[\left(Y_A-E\left[Y_A|X_A\right]\right)^2\right] \leq E\left[\left(Y_A-h_A\left(X_A\right)\right)^2\right] \tag{2.C.64}$$

or equivalently,

$$E\left[\left(Y-m_A\left(X\right)\right)^2|X\in A\right] \leq E\left[\left(Y-h_A\left(X\right)\right)^2 \mid X\in A\right] \tag{2.C.65}$$

Note that

$$\begin{aligned} E\left[\left(Y-E\left[Y|X\right]\right)^2\right] &= \Pr\left(X\in A\right)E\left[\left(Y-E\left[Y|X\right]\right)^2|X\in A\right] && (2.C.66)\\ &\quad +\Pr\left(X\notin A\right)E\left[\left(Y-E\left[Y|X\right]\right)^2|X\notin A\right] && (2.C.67)\\ &= \Pr\left(X\in A\right)E\left[\left(Y-m\left(X_A\right)\right)^2|X\in A\right] && (2.C.68)\\ &\quad +\Pr\left(X\notin A\right)E\left[\left(Y-E\left[Y|X\right]\right)^2|X\notin A\right] && (2.C.69) \end{aligned}$$

Suppose by contradiction that $\Pr(\omega\in X^{-1}\left(A\right)|m_A\left(X_A\left(\omega\right)\right)\neq m\left(X_A\left(\omega\right)\right))>0$. Then, it must be either

$$E\left[\left(Y-m_A\left(X\right)\right)^2|X\in A\right] > E\left[\left(Y-m\left(X\right)\right)^2|X\in A\right] \tag{2.C.70}$$

or

$$E\left[\left(Y-m_A\left(X\right)\right)^2|X\in A\right] < E\left[\left(Y-m\left(X\right)\right)^2|X\in A\right]. \tag{2.C.71}$$

Suppose $\mathbb{E}\left[\left(Y-m_A(X)\right)^2|X\in A\right] > E\left[\left(Y-m(X)\right)^2|X\in A\right]$. Define

$$(x) = \begin{cases} m(x) & ,x\notin A \\ m_A(x) & ,x\in A \end{cases}. \tag{2.C.72}$$

Then,

$$E\left[\left(Y-E[Y|X]\right)^2\right] > E\left[\left(Y-h(X)\right)^2\right] \tag{2.C.73}$$

a contradiction to (2.C.63). If $\mathbb{E}\left[\left(Y-m_A(X)\right)^2|X\in A\right] < E\left[\left(Y-m(X)\right)^2|X\in A\right]$, a contradiction to (2.C.65). □

PROOF OF THEOREM 2.6.3 First, note that

$$\Pr\left(|m(X)-P_A(X)|\geq c|X\in A\right) \leq \frac{\mathrm{Var}\left(Y-P_A(X)|X\in A\right)}{c^2} \tag{2.C.74}$$

For a fixed $\alpha\in[0,1]$, define $c_\alpha = \sqrt{\dfrac{\mathrm{Var}\left(Y-P_A(X)|X\in A\right)}{\alpha}}$ so that

$$\Pr\left(|m(X)-P_A(X)|\geq c_\alpha|X\in A\right) \leq \alpha \tag{2.C.75}$$

This means that there exists a set $W\subset\Omega$ such that $P(\omega\in W\,|X(\omega)\in A)\geq 1-\alpha$ and for $\forall\omega\in W$,

$$|m(X(\omega))-P_A(X(\omega))| \leq c_\alpha \tag{2.C.76}$$

or

$$P_A(X(\omega))-c_\alpha \leq m(X(\omega)) \leq P_A(X(\omega))+c_\alpha. \tag{2.C.77}$$

Now, since $\sup_{x\in A}P_A(x)\geq P_A(X(\omega))$ and $P_A(X(\omega))\leq\inf_{x\in A}P_A(x)$ for any ω,

$$\inf_{x\in A}P_A(x)-c_\alpha \leq m(X(\omega)) \leq \sup_{x\in A}P_A(x)+c_\alpha. \tag{2.C.78}$$

It follows that

$$\Pr\left(\left\{\omega : m\left(X\left(\omega\right)\right) \in \left[\inf_{x\in A} P_A\left(x\right) - c_\alpha, \sup_{x\in A} P_A\left(x\right) + c_\alpha\right]\right\} | X\left(\omega\right) \in A\right) \geq 1-\alpha. \tag{2.C.79}$$

□

PROOF OF LEMMA 2.7.1 Pick any $\tilde{m} \in \mathscr{M}_{\mathrm{C}}$. Then, the conditional density of ε given $X = x$ and $\tilde{m}$ is well-defined as

$$f_{\varepsilon|X=x}(\varepsilon|X=x;\tilde{m}) = f_{Y|X=x}(y-\tilde{m}(x)) \tag{2.C.80}$$

and is continuous with respect to x. Thus,

$$\begin{aligned}
& \frac{1}{2}\int\int\int \left(y_1-y_2\right)^2 dF_{Y|X=x_0}\left(y_1\right) dF_{Y|X=x_0}\left(y_2\right) dF_X\left(x_0\right) && (2.C.81)\\
= & \frac{1}{2}\int\int\int (\varepsilon_1^2+\varepsilon_2^2) dF_{\varepsilon|X=x_0}\left(\varepsilon_1\right) dF_{\varepsilon|X=x_0}\left(\varepsilon_2\right) dF_X\left(x_0\right) && (2.C.82)\\
= & \int\int \varepsilon_1^2 dF_{\varepsilon|X=x_0}\left(\varepsilon_1\right) dF_X\left(x_0\right) && (2.C.83)\\
= & \int \varepsilon_1^2 dF_\varepsilon\left(\varepsilon_1\right) = \mathrm{Var}\left(\varepsilon\right) && (2.C.84)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
& \frac{1}{2}\int\int\int \left(y_1-y_2\right)^2 dF_{Y|X=x_0+\delta}\left(y_1\right) dF_{Y|X=x_0}\left(y_2\right) dF_X\left(x_0\right) && (2.C.85)\\
= & \frac{1}{2}\int\int\int (\varepsilon_1^2+\varepsilon_2^2) dF_{\varepsilon|X=x_0}\left(\varepsilon_1\right) dF_{\varepsilon|X=x_0}\left(\varepsilon_2\right) dF_X\left(x_0\right) && (2.C.86)\\
& + \frac{1}{2}\left\{\left(\tilde{m}(x_0+\delta)\right)^2 - \left(\tilde{m}(x_0)\right)^2\right\} dF_X\left(x_0\right) && (2.C.87)\\
& + \tilde{m}(x_0)\int \varepsilon_1 dF_{\varepsilon|X=x_0}\left(\varepsilon_1\right) && (2.C.88)\\
& + \tilde{m}(x_0+\delta)\int \varepsilon_2 dF_{\varepsilon|X=x_0+\delta}\left(\varepsilon_2\right) && (2.C.89)\\
& = \mathrm{Var}\left(\varepsilon\right) && (2.C.90)\\
& + \frac{1}{2}\int\left\{\left(\tilde{m}(x_0+\delta)\right)^2 - \left(\tilde{m}(x_0)\right)^2\right\} dF_X\left(x_0\right) && (2.C.91)
\end{aligned}$$

Then, since $\tilde{m}$ is continuous, for each x_0,

$$\lim_{\delta \to 0} \left(\tilde{m}(x_0+\delta) f_X(x_0)\right)^2 = \tilde{m}(x_0) f_X(x_0) \tag{2.C.92}$$

and thus

$$\frac{1}{2} \lim_{\delta \to 0} \int \left\{ \left(\tilde{m}(x_0+\delta)\right)^2 - \left(\tilde{m}(x_0)\right)^2 \right\} dF_X(x_0) = 0. \tag{2.C.93}$$

□

PROOF OF PROPOSITION 2.8.2

$$\sqrt{n}(h(x;\hat{\theta}_n) - h(x;\theta_0)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{h,x}) \tag{2.C.94}$$

where $\sigma^2_{h,x} = (\nabla_\theta h(x;\theta_0))' \Lambda (\nabla_\theta h(x;\theta_0))$. Furthermore, given consistency of $\hat{\Lambda}_n$,

$$\hat{\sigma}^2_{h,x} = (\nabla_\theta h(x;\hat{\theta}_n))' \hat{\Lambda}_n (\nabla_\theta h(x;\hat{\theta}_n)) \tag{2.C.95}$$

is a consistent estimator of $\sigma^2_{h,x}$. Thus,

$$\lim_{n \to \infty} \Pr\left(h(x;\theta_0) \in CS_{h,n}\right) = 1 - \alpha \tag{2.C.96}$$

as desired. □

PROOF OF PROPOSITION 2.8.2 Given asymptotic independence of X and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ conditional on X, we still have that the conditional distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ given X is identical to its unconditional distribution. Then, the assertion follows from the proof of Lemma 2.8.1. □

Lemma 2.C.1 NOBEL AND DEMBO (1993). *Let $\mathcal{F}$ be a permissible family of functions having an envelope function. If $\mathcal{F}$ satisfies a uniform law of numbers with respect to an i.i.d. process having distribution $\mathbb{P}_0 = \Pi_{-\infty}^{\infty} P$ then*

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_0) \to 0 \tag{2.C.97}$$

with probability $\mathbb{P}$-probability one for every stationary and β-mixing stochastic process $\{X_i\}_{i=-\infty}^{\infty}$ having distribution $\mathbb{P}$ with marginal P.

PROOF OF LEMMA 2.8.3 First, we are going to show that

$$\mathbb{E}f(\left|Y-h(X;\hat{\theta}_n)\right| \xrightarrow{p} \mathbb{E}f\left(\left|Y-h_0(X)\right|\right). \tag{2.C.98}$$

Pick any $\varepsilon > 0$ such that $\|\theta-\theta_0\| < \varepsilon$ implies $\theta \in \mathscr{N}^{(\mathrm{E})}(\theta_0) \cap \mathscr{N}^{(\mathrm{h})}(\theta_0) \cap \mathscr{N}^{(\mathrm{Eh})}(\theta_0)$. Define function classes

$$\mathscr{F}_\varepsilon = \{g(w;\theta) \,|\theta : \|\theta-\theta_0\| < \varepsilon\} \tag{2.C.99}$$

and

$$\mathscr{K}_\varepsilon = \{k(w;\theta) \,|\theta : \|\theta-\theta_0\| < \varepsilon\}. \tag{2.C.100}$$

where

$$g(w;\theta) = f\left(\left|y-h(x;\theta)\right|\right), k(w;\theta) = \left|y-h(x;\theta)\right|. \tag{2.C.101}$$

Observe that

$$\mathbb{E}\sup_{\theta\in\mathscr{N}^{(\mathrm{E})}(\theta_0)} f\left(\left|Y-h(X;\theta)\right|\right) \geq \mathbb{E}f(\sup_{\theta\in\mathscr{N}^{(\mathrm{E})}(\theta_0)} \left|Y-h(X;\theta)\right|) \tag{2.C.102}$$

$$\geq f(\mathbb{E}[\sup_{\theta\in\mathscr{N}^{(\mathrm{E})}(\theta_0)} \left|Y-h(X;\theta)\right|]) \tag{2.C.103}$$

by Jensen's inequality and thus

$$\mathbb{E}\sup_{\theta\in\mathscr{N}^{(\mathrm{E})}(\theta_0)} \left|Y-h(X;\theta)\right| < \infty \tag{2.C.104}$$

by Assumption 2.8.5. It follows that $\mathscr{K}_\varepsilon$ is a Vapnik-Chervonenkis-subgraph class and thus a Glivenko-Cantelli class with respect to the probability measure $\mathbb{P}^* = \Pi_{i=1}^{\infty} P_i$ where P_i is the marginal distribution of w_i and hence, by Lemma 2.C.1,with respect to the probability measure of the β-mixing sequence $\{w_i\}_{i=1}^{\infty}$. Since f is continuous, it follows from the preservation theorem that $\mathscr{F}_\varepsilon$ is also a Glivenko-Cantelli class with respect to the same

measure. By Assumption 2.8.2, for any $\delta \in (0,1)$, there exists some $n^* \in \mathbb{N}$ such that

$$\hat{\theta}_n \in \mathscr{N}^{(\mathrm{E})}(\theta_0) \cap \mathscr{N}^{(\mathrm{h})}(\theta_0) \cap \mathscr{N}^{(\mathrm{Eh})}(\theta_0) \tag{2.C.105}$$

so that

$$g(w;\hat{\theta}_n) \in \mathscr{F}_\varepsilon, k(w;\hat{\theta}_n) \in \mathscr{K}_\varepsilon, \forall n \geq n^*. \tag{2.C.106}$$

with probability $1-\delta$. Thus,

$$\mathbb{E}_n f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right) - \mathbb{E} f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right) \xrightarrow{p} 0. \tag{2.C.107}$$

Given Assumption 2.8.2 and 2.8.7, it follows from the continuous mapping theorem that

$$\mathbb{E} f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right) - \mathbb{E} f\left(|Y - h(X;\theta_0)|\right) \xrightarrow{p} 0. \tag{2.C.108}$$

Thus, (2.C.98) holds by the triangle inequality. Then, by Assumption 2.8.4, f is invertible at $c = \hat{c}_{\alpha,f}$ such that

$$f(\hat{c}_{\alpha,f}) = \frac{\mathbb{E}_n f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right)}{\alpha} \tag{2.C.109}$$

so that

$$\hat{c}_{\alpha,f} = f^{-1}\frac{\mathbb{E}_n f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right)}{\alpha} \tag{2.C.110}$$

for any $n \geq n^{**}$ for some $n^{**} \in \mathbb{N}$ with probability approaching to one. Finally, the continuous mapping theorem implies

$$\hat{c}_{\alpha,f} = f^{-1}\frac{\mathbb{E}_n f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right)}{\alpha} \xrightarrow{p} f^{-1}\frac{\mathbb{E} f\left(|Y - h_0(X)|\right)}{\alpha} = c_{\alpha,f}. \tag{2.C.111}$$

□

PROOF OF LEMMA 2.8.4 We are first going to show

$$\sup_{f \in \mathscr{F}} \left| f\left(\left|Y - h(X;\hat{\theta}_n)\right|\right) - \mathbb{E} f\left(|Y - h_0(X)|\right)\right| \xrightarrow{p} 0. \tag{2.C.112}$$

To this end, observe by the triangle inequality that

$$\begin{aligned}\sup_{f\in\mathscr{F}}\left|\mathbb{E}_n f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)-\mathbb{E}f\left(|Y-h_0(X)|\right)\right| \;\leq\; & \sup_{f\in\mathscr{F}}\left|\mathbb{E}_n f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)-\mathbb{E}f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)\right| \\ & +\sup_{f\in\mathscr{F}}\left|\mathbb{E}f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)-\mathbb{E}f\left(|Y-h_0(X)|\right)\right| \qquad (2.C.113)\end{aligned}$$

By Assumption 2.8.2, $g_f(w;\hat{\theta}_n)\in\mathscr{G}_\varepsilon$ with probability approaching to one and thus by Assumption 2.8.11 and Lemma 2.C.1

$$\sup_{f\in\mathscr{F}}\left|\mathbb{E}_n f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)-\mathbb{E}f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)\right| \;\leq\; \sup_{g\in\mathscr{G}}\left|\mathbb{E}_n g(W)-\mathbb{E}g(W)\right| \xrightarrow{p} 0 \qquad (2.C.114)$$

Furthermore, Assumption 2.8.8 and 2.8.9 combined with the continuous mapping theorem imply

$$\sup_{f\in\mathscr{F}}\left|\mathbb{E}f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)-\mathbb{E}f\left(|Y-h_0(X)|\right)\right| \xrightarrow{p} 0. \qquad (2.C.115)$$

Finally, by Assumption 2.8.8, for any $\varepsilon>0$,

$$\sup_{f\in\mathscr{F}}\left|f^{-1}\frac{\mathbb{E}_n f\left(\left|Y-h(X;\hat{\theta}_n)\right|\right)}{\alpha}-f^{-1}\frac{\mathbb{E}f\left(|Y-h(X;\theta_0)|\right)}{\alpha}\right|\leq\varepsilon \qquad (2.C.116)$$

and thus

$$\sup_{f\in\mathscr{F}}\left|\hat{c}_{\alpha,f}-c_{\alpha,f}\right| \xrightarrow{p} 0. \qquad (2.C.117)$$

□

PROOF OF LEMMA 2.8.5 Observe the decomposition of $\hat{\sigma}^2_\varepsilon$:

$$\begin{aligned}\hat{\sigma}^2_\varepsilon &= \frac{1}{2(n-1)}\sum_{j=2}^{n}\left(y_{[j]}-y_{[j-1]}\right)^2 \\ &= \frac{1}{2(n-1)}\sum_{j=2}^{n}\left(\varepsilon_{[j]}-\varepsilon_{[j-1]}\right)^2 \\ &\quad+\frac{1}{2(n-1)}\sum_{j=2}^{n}\left(m\left(x_{[j]}\right)-m\left(x_{[j-1]}\right)\right)^2\end{aligned}$$

$$+\frac{1}{2(n-1)}\sum_{j=2}^{n}\left(\varepsilon_{[j]}-\varepsilon_{[j-1]}\right)\left(m\left(x_{[j]}\right)-m\left(x_{[j-1]}\right)\right). \tag{2.C.118}$$

By Assumption 2.8.12,

$$\frac{1}{2(n-1)}\sum_{j=2}^{n}\left(\varepsilon_{[j]}-\varepsilon_{[j-1]}\right)^2=\frac{1}{(n-1)}\sum_{j=2}^{n}\varepsilon_j^2+o_p(1)\xrightarrow{p}\sigma_\varepsilon^2. \tag{2.C.119}$$

Furthermore, Assumption 2.8.13-2.8.14 imply

$$\begin{aligned}\frac{1}{2(n-1)}\sum_{j=2}^{n}\left(m\left(x_{[j]}\right)-m\left(x_{[j-1]}\right)\right)^2 &\leq \frac{1}{2(n-1)}C^2\sum_{j=2}^{n}\left|x_{[j]}-x_{[j-1]}\right|^{2\gamma}\\ &\leq \frac{1}{2(n-1)}C^2\max_{2\leq j\leq n}\left|x_{[j]}-x_{[j-1]}\right|^{2\gamma}\\ &= o_p(1)\xrightarrow{p}0,\end{aligned} \tag{2.C.120}$$

Finally, we have

$$\begin{aligned}\frac{1}{2(n-1)}\sum_{j=2}^{n}\left(\varepsilon_{[j]}-\varepsilon_{[j-1]}\right)\left(m\left(X_{[j]}\right)-m\left(X_{[j-1]}\right)\right) &= \frac{1}{n-1}\sum_{j=2}^{n}\varepsilon_j m\left(x_j\right)\\ &\quad+\frac{1}{2(n-1)}\sum_{i=2}^{n}\left\{\varepsilon_{[j]}m\left(x_{[j-1]}\right)-\varepsilon_{[i-1]}m\left(x_{[i]}\right)\right\}\\ &= o_p(1)+o_p(1)\xrightarrow{p}0.\end{aligned} \tag{2.C.121}$$

We conclude

$$\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2+o_p(1). \tag{2.C.122}$$

□

PROOF OF PROPOSITION 2.8.6 Define events

$$A_n=\left\{\left|m(X)-h(X;\theta_0)\right|>\hat{c}_{\alpha_2}\right\} \tag{2.C.123}$$

$$B_n=\left\{\left|h(X;\hat{\theta}_n)-h(X;\theta_0)\right|>\hat{d}\left(x;\alpha_1\right)\right\} \tag{2.C.124}$$

Then, under Assumption 2.8.15,

$$\limsup_{n\to\infty} \Pr(A_n) \leq \alpha_2. \tag{2.C.125}$$

By Proposition 2.8.2, we have

$$\limsup_{n\to\infty} \Pr(B_n) = \alpha_1. \tag{2.C.126}$$

Finally,

$$\begin{aligned} \limsup_{n\to\infty} \Pr(m(X) \in CS_{1-\alpha,n}(X;\alpha_1,\alpha_2)) \leq & \quad \limsup_{n\to\infty} \Pr(A_n) + \limsup_{n\to\infty} \Pr(B_n) \\ \leq & \quad \alpha_1 + \alpha_2 = \alpha. \end{aligned} \tag{2.C.127}$$

□

PROOF OF COROLLARY 2.8.7 Define the events (A_n, B_n) as in the proof of Proposition 2.8.6. Then, $\limsup_{n\to\infty} \Pr(B_n) = \alpha_1$. Choose any sufficiently small constant $\varepsilon > 0$ and let $\alpha_2^* = \alpha_2 - \varepsilon$. Then, since c_α is strictly decreasing in α,

$$[h(X;\theta_0) - c_{\alpha_2^*}, h(X;\theta_0) + c_{\alpha_2^*}] \subset [h(X;\theta_0) - \hat{c}_{\alpha_2}, h(X;\theta_0) + \hat{c}_{\alpha_2}] \tag{2.C.128}$$

with probability approaching to one. Thus,

$$\limsup_{n\to\infty} \Pr(B_n) \leq \alpha_2^*. \tag{2.C.129}$$

Then,

$$\limsup_{n\to\infty} \Pr(m(X) \in CS_{1-\alpha,n}(X;\alpha_1,\alpha_2)) \leq \limsup_{n\to\infty} \Pr(A_n) + \limsup_{n\to\infty} \Pr(B_n) \tag{2.C.130}$$

$$\leq \alpha_1 + \alpha_2^* = \alpha_1 + \alpha_2 - \varepsilon. \tag{2.C.131}$$

Since ε is arbitrary, the assertion follows. □

Chapter 3

Generalized C(α) tests with nonstandard convergence rates

3.1. Introduction

In this paper, we develop generalized $C(\alpha)$-type tests for linear and nonlinear hypotheses, in order to allow for nonstandard (possibly slow) convergence rates on the parameter estimates used, in the context of models specified through estimating functions or moment equations [*e.g.*, using the generalized method of moments (GMM)].

In parametric models, likelihood ratio (LR) tests [Neyman and Pearson (1928)], Wald tests [Wald (1943)] and score tests [introduced by Rao (1948)], along with various extensions, constitute the basis of statistical hypothesis testing. In likelihood models, LR tests require one to estimate the models under both the null hypothesis and without restrictions, Wald tests only require unrestricted estimators, while score tests only require restricted estimators. Under standard regularity conditions, the three tests have local asymptotic efficiency. These general testing procedures can be extended to more general setups where moments (or estimating functions) play the role as a score-type function [see, for example, Newey and West (1987*a*), Gourieroux and Monfort (1995), Dufour, Trognon and Tuvaandorj (2017)].

Optimization of a likelihood function under constraints can be computationally expensive. The $C(\alpha)$ test procedure proposed by Neyman (1954, 1959) extends Rao's score

test by allowing to replace the maximum likelihood estimator by any root-n consistent restricted estimator. As in the original score test, the $C(\alpha)$ test enjoys optimality properties and has been extended to more general setups. This provides great flexibility in the choice of estimator used, because the asymptotic distribution of the restricted estimator need not be known (or Gaussian).

The literature on $C(\alpha)$ tests and related procedures is now extensive; see Le Cam (1956), Bhat and Nagnur (1965), Bühler and Puri (1966), Bartoo and Puri (1967), Moran (1970, 1973), Chibisov (1973), Chant (1974), Ray (1974), Singh and Zhurbenko (1975), Foutz (1976), Vorob'ev and Zhurbenko (1979), Bernshtein (1976, 1978, 1980*a*, 1980*b*, 1981), Le Cam and Traxler (1978), Neyman (1979), Tarone (1979, 1985), Tarone and Gart (1980), Wang (1981, 1982), Basawa (1985), Ronchetti (1987), Smith (1987*a*, 1987*b*), Berger and Wallenstein (1989), Hall and Mathiason (1990), Paul and Barnwal (1990), Wooldridge (1990), Dagenais and Dufour (1991), Davidson and MacKinnon (1991, 1993), Kocherlakota and Kocherlakota (1991), Dufour and Dagenais (1992), Bera and Yoon (1993), Jaggia and Trivedi (1994), Rao (1996), Bera and Bilias (2001), Pal (2003), Dufour and Valéry (2009), Chaudhuri and Zivot (2011), Bontemps and Meddahi (2012), Dufour et al. (2016, 2017), Bontemps (2019).

The fundamental idea of the $C(\alpha)$ test is to orthogonalize the scores associated with the parameters of interest (which are restricted by the null hypothesis) with respect to the scores of the nuisance parameters (at least, under the null hypothesis). This reduces the sensitivity of the test statistic to the distribution of the nuisance parameter estimate, and indeed evacuates it from the asymptotic distribution of the test statistic (under appropriate regularity conditions). The $C(\alpha)$ test can be further generalized to relax orthogonality conditions. Indeed, the assumption of differentiability of the log-likelihood function with respect to model parameters (the usual score function based on likelihood function) can also be also be abandoned; see Dufour et al. (2016). Existing work typically assumes the existence of a $n^{1/2}$-consistent restricted estimator, so that these results are not applicable in cases where the restricted estimator used has a slower convergence rate.

In this paper, we *first* extend the generalized $C(\alpha)$ test proposed in Dufour et al. (2016) for testing general parameter restrictions using a vector of estimating functions. We allow for the restricted estimator to converge at a rate slower than $n^{1/2}$. When the estimating function converges at the standard rate $n^{1/2}$, our conditions entail that the convergence

rate of the estimator should be faster than $n^{1/4}$. *In some cases, it could even be slower.* Naturally, since the results presented extend those of Dufour et al. (2016), they do not require orthogonality of the moment equation with respect to the score of the log-likelihood function, nor even the existence of the latter. On the other hand, the specific form of the restriction (and their derivatives) plays a central role in creating the required asymptotic invariance.

There are many examples where slow convergence can happen. For example, the maximum score estimator of Manski (1975) in the context of discrete choice model is known to converge at the cube rate $n^{1/3}$, and an improved estimator of Horowitz (1992) utilizing kernel smoothing enjoys a faster rate of convergence, but still does not attain the parametric rate $n^{1/2}$. The problem involving cubic root asymptotics was first pointed out by Chernoff (1964) and has been studied by Kim and Pollard (1990) and Seo and Otsu (2018), among others. Caner (2006) shows that the convergence rate of an M-estimator with weakly dependent data depends on the decay rate of the mixing coefficients and smoothness of the objective function and may be slower than $n^{1/2}$. Additionally, $n^{1/2}$-consistent estimation in the presence of an infinite-dimensional parameter is not necessarily feasible [see Firpo, Fortin and Lemieux (2009)] and this appliemore generally for estimators based on nonparametric regressions.

Second, we also let the estimating functions converge to a non-degenerate limit at a more general rate than $n^{1/2}$. We only require that the restricted estimator converge faster than the estimating function to show that the asymptotic distribution of the proposed test statistic is not affected by estimation error involving the restricted estimator and is distributed according to the usual chi-square distribution. This allows one to use an estimating function and a restricted estimator based on different samples – whose size can be quite different – which can lead to different rates of convergence. This feature is easily accommodated by the asymptotic invariance of $C(\alpha)$ statistic with respect to the distribution of the estimator used.

Third, we allow for the presence of additional nuisance parameters for which an estimator with a possibly *different* rate of convergence is available, along (possibly) with additional *auxiliary* estimating functions. The primary and auxiliary estimating functions can have different rates of convergence to a limiting distribution, in order to accommodate cases where one of them involves kernel-smoothing or the two estimators are based

on very different sample sizes. For this extended setup, we propose *extended generalized* $C(\alpha)$ *statistics* [$EC(\alpha)$], and we give conditions under which the distribution is asymptotically chi-square under the null hypothesis. In particular, these include on the convergence rates of the estimators of the main parameter vector (on which restrictions are imposed) and the additional nuisance parameters.

Fourth, applications to local estimating equation models [Carroll, Ruppert and Welsh (1998), Xu (2020), Lewbel (2007), Gagliardini, Gourieroux and Renault (2011)] and problems involving data sets with asymptotically unbalanced sample sizes are discussed. Local estimating equations arise naturally when the parameter of interest is the value of a functional evaluated at a particular point of the covariate. Both the estimating functions and the restricted estimator typically converge at a rate slower than $n^{1/2}$, due to kernel smoothing. Furthermore, the sensitivity of the estimating functions with respect to each element of the parameter vector can depend on the sample size and thus may be properly measured and taken into account only when the parameter vectors are scaled by a diagonal matrix of scaling factors in the estimating equations, as we allow in our framework. We apply the test procedure to hypothesis testing on: (i) derivatives of a nonparametric regression function; (ii) average treatment effects in regression discontinuity designs; (iii) semiparametric stochastic discounting factors in the local estimating function framework. The problem of asymptotically unbalanced sample sizes is commonly observed in practice when the data consists of observations from multiple populations/sources [**?**, Jonker and Van der Vaart (2014)]. When the problem is concerned with the increase in the sample sizes of two populations at different orders, the estimating functions can be split into the primary and auxiliary ones which have different rate of convergence, to apply the extended generalized $C(\alpha)$ test. We consider hypothesis testing on homogeneity of regression models among different groups/populations under unbalanced sample sizes in this framework.

The paper is organized as follows. In Section 3.2, we describe the setup and present the general idea of the test procedures. In Section 3.3, we generalize further the generalized $C(\alpha)$ test of Dufour et al. (2016), and we relax the assumption on the convergence rate of the restricted estimator. Section 3.4 extends these results to testing problems involving nuisance parameters estimated from an auxiliary estimating function. In Section 3.5, we present applications of the test procedures in the local estimating function framework. In Section 3.6, the extended generalized $C(\alpha)$ test is applied to a hypothesis testing on ho-

mogeneity of regression models of two populations under unbalanced sample sizes. We conclude in Section 3.7. The proofs are provided in Appendix.

3.2. Generalized $C(\alpha)$ statistic under general convergence rates

Consider an $m \times 1$ vector of estimating (or moment) functions $D_n(\theta; Z_n)$ which depend on a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$ and a vector Z_n of size n with nonrandom limit $\bar{D}_\infty(\theta; \theta_0)$ which depends on the "true value" θ_0:

$$D_n(\theta; Z_n) \xrightarrow[n\to\infty]{\mathsf{p}} \bar{D}_\infty(\theta; \theta_0)\,. \tag{3.2.1}$$

The parameter θ is often estimated by minimizing a criterion function of the form

$$M_n(\theta, W_n) = D_n(\theta; Z_n)' \, W_n \, D_n(\theta; Z_n) \tag{3.2.2}$$

where W_n is a symmetric positive definite matrix. This setup comprises as special cases the method of estimating functions [Durbin (1960), Godambe (1960, 1991), Small and McLeish (1994), Basawa, Godambe and Taylor (1997), Heyde (1997)], the generalized method of moments [Hansen (1982), Hall (2004)], maximum likelihood, pseudo-maximum likelihood, M-estimation and instrumental-variable methods.

We are interested in testing hypotheses of the form

$$H_0 : \psi(\theta) = 0 \tag{3.2.3}$$

where $\psi(\theta)$ is a $p_1 \times 1$ continuously differentiable function ($p_1 \leq p$). We suppose that we have a restricted estimator $\tilde{\theta}_n^0$ which converges to θ_0 at rate n^{r_θ} under H_0:

$$n^{r_\theta}(\tilde{\theta}_n^0 - \theta_0) = O_p(1) \quad \text{with } r_\theta > 0\,. \tag{3.2.4}$$

Further,

$$n^{r_D} D_n(\theta_0; Z_n) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, I(\theta_0)] \quad \text{with } r_D \geq 0 \tag{3.2.5}$$

where $I(\theta_0)$ is an $m \times m$ nonsingular matrix. The value $r_D = 0$ means that $D_n(\theta_0; Z_n)$ is exactly normal or converges to normality without rescaling. $D_n(\theta_0; Z_n)$ and $\tilde{\theta}_n^0$ may come from different samples.

For the case where $r_\theta = r_D = 1/2$, Dufour et al. (2016) propose a generalized $C(\alpha)$ statistic based on a restricted estimator $\tilde{\theta}_n^0$ and show (under weak regularity conditions) that its asymptotic distribution is $\chi^2(p_1)$ under H_0, irrespective of the asymptotic distribution of $\sqrt{n}(\tilde{\theta}_n^0 - \theta_0)$ [indeed, even if such a distribution does not exist]. This statistic nests earlier $C(\alpha)$ tests. Further, the score (or the estimating function) need not be orthogonal to the log-likelihood score of the nuisance parameters [in contrast with Neyman (1959)]. The generalized $C(\alpha)$ statistic relies on the following transformation $D_n(\theta; Z_n)$:

$$s_n(\theta; Z_n) = \tilde{Q}_n D_n(\tilde{\theta}_n^0; Z_n) \tag{3.2.6}$$

where $\tilde{Q}_n$ is a $p_1 \times p$ matrix that converges in probability to a nonrandom limit $Q(\theta_0)$ such that $\text{rank}[Q(\theta_0)] = p_1$ and satisfies

$$\sqrt{n}\,[s_n(\tilde{\theta}_n; Z_n) - Q(\theta_0) D_n(\theta_0; Z_n)] \xrightarrow[n\to\infty]{\mathsf{p}} 0\,. \tag{3.2.7}$$

The transformation matrix $\tilde{Q}_n$ depends on the local sensitivity of $D_n(\theta; Z_n)$ to the value of θ in a neighborhood of θ_0 – though $D_n(\theta; Z_n)$ may not be differentiable – and the Jacobian of the restriction function $\psi(\theta)$. Both features interact and play a role in determining a transformation that can eliminate the distribution of $\sqrt{n}(\tilde{\theta}_n^0 - \theta_0)$ from the asymptotic distribution of the test statistic. Precise expressions for $\tilde{Q}_n$ and $Q(\theta)$ are given in Dufour et al. (2016, Proposition 3.1). In this case, it follows from (3.2.5) and Slutsky's theorem that

$$\sqrt{n}\, s_n(\tilde{\theta}_n; Z_n) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, Q(\theta_0) I(\theta_0) Q(\theta_0)']\,. \tag{3.2.8}$$

The generalized $C(\alpha)$ statistic is given by

$$PC(\tilde{\theta}_n^0; \psi, W_n) = n\, s_n(\tilde{\theta}_n^0; Z_n)' \tilde{\Sigma}_n^{-1} s(\tilde{\theta}_n^0; Z_n) \tag{3.2.9}$$

where $\tilde{\Sigma}_n$ is a consistent estimator of $\Sigma(\theta_0) := Q(\theta_0) I(\theta_0) Q(\theta_0)'$ and $\Sigma(\theta_0)$ is invertible. Under the null hypothesis, $PC(\tilde{\theta}_n^0; \psi, W_n)$ converges to a $\chi^2(p_1)$ distribution.

In this paper, we show that where $D_n(\theta_0; Z_n)$ converges in distribution at the rate $r_D = 1/2$, we require $r_\theta > 1/4$ instead of $r_\theta = 1/2$ if $D_n(\theta; Z_n)$ and the restriction function ψ admit representation akin to second-order approximation around $\theta = \theta_0$ and $\tilde{Q}_n$ converges to Q_0 not too slowly as detailed in Section 3.3. This observation can be extended to more general rates r_D by replacing the scaling constant n in (3.2.9) by n^{2r_D}. This yields the statistic:

$$PC(\tilde{\theta}_n^0;\, \psi,\, r_D W_n) = n^{2r_D} s_n(\tilde{\theta}_n^0; Z_n)' \tilde{\Sigma}_n^{-1} s(\tilde{\theta}_n^0; Z_n)\,. \tag{3.2.10}$$

We show that if

$$r_\theta > r_D/2 \tag{3.2.11}$$

then $PC(\tilde{\theta}_n^0;\, \psi, r_D W_n)$ is still asymptotically distributed as $\chi^2(p_1)$.

In Section 3.4, we allow for the score function $D_n(\theta, \eta; Z_n)$ to depend on an additional parameter $\eta \in \mathscr{E} \subset \mathbb{R}^q$ which is not tested and thus is treated as a nuisance parameter. The parameter η is assumed to be identified and also possibly estimated from a $q \times 1$ auxiliary score $G_n(\eta; X_n)$ which depends on η, but not θ and the data X_n may or may not overlap with Z_n:

$$n^{r_D} D_n(\theta_0, \eta_0; Z_n) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, I(\theta_0, \eta_0)]\,, \tag{3.2.12}$$

$$n^{r_G} G_n(\eta_0; X_n) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, \Lambda(\eta_0)]\,, \tag{3.2.13}$$

where $I(\theta_0, \eta_0)$ and $\Lambda(\eta_0)$ are singular matrices of size p_1 and q_1, respectively. First, treat η_0 as if is known and consider, as in the case without η, transformation $s_n(\theta, \eta_0; Z_n)$ of $D_n(\theta, \eta_0; Z_n)$:

$$s_n(\theta, \eta_0; Z_n) = \tilde{Q}_n D_n(\tilde{\theta}_n^0, \eta_0; Z_n) \tag{3.2.14}$$

where $\tilde{Q}_n$ is a $p_1 \times p$ matrix which converges in probability to a nonrandom limit $Q(\theta_0, \eta_0)$ such that $\mathrm{rank}[Q(\theta_0, \eta_0)] = p_1$ and satisfies

$$n^{r_D}[s_n(\theta, \eta_0; Z_n) - Q(\theta_0, \eta_0) D_n(\theta_0, \eta_0; Z_n)] \xrightarrow[n\to\infty]{\mathsf{p}} 0\,. \tag{3.2.15}$$

Let $\hat{\eta}_n$ be an estimator of η_0 such that, for some constant $r_\eta > 0$,

$$n^{r_\eta}(\hat{\eta}_n - \eta_0) = O_p(1) \tag{3.2.16}$$

at least under H_0. Given $s_n(\theta,\eta;Z_n)$, we define the score

$$s_n^*(\theta,\eta;Z_n,X_n) = s_n(\theta,\eta;Z_n) - \tilde{T}_n G_n(\eta;X_n) \tag{3.2.17}$$

where $\tilde{T}_n$ is a $p_1 \times q$ matrix defined in Section 3.4. In this case, we show that the rate condition

$$\min(r_\theta, r_G) > \max(r_D, r_G) \tag{3.2.18}$$

ensures that so that the effect of estimation error of $(\tilde{\theta}_n^0, \hat{\eta}_n)$ on the score:

$$s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n) - s_n^*(\theta_0, \eta_0; Z_n, X_n) = o_p\left(n^{-\min(r_D, r_G)}\right). \tag{3.2.19}$$

Depending on which one of (i) $r_D < r_G$, (ii) $r_D > r_G$, or (iii) $r_G = r_G$ holds, the score $s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n)$ converges at a different rate to a Gaussian limit with a different covariance matrix:

$$s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n) = s(\tilde{\theta}_n^0, \eta_0; Z_n) - \tilde{T}_n G_n(\hat{\eta}_n; X_n) \tag{3.2.20}$$

For example, in case (i), the first term $s_n(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n)$ dominates $\tilde{T}_n G_n(\hat{\eta}_n; X_n)$ in (3.2.20) asymptotically and $s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n)$ converges to a (non-degenerate) Gaussian limit at rate r_G. We show that the extended generalized $C(\alpha)$ statistic $EC(\tilde{\theta}_n^0, \hat{\eta}_n; \tilde{\Lambda}_n^{-1}\psi)$ given below is asymptotically chi-squared:

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \tilde{\Lambda}_n^{-1}\psi) = n^{2\min(r_D, r_G)} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n)' \tilde{\Lambda}_n^- s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n) \tag{3.2.21}$$

where A^- denotes the Moore-Penrose inverse of a square matrix A and $\tilde{\Lambda}_n$ is a consistent estimator of the asymptotic covariance Λ_0 of $n^{\min(r_D, r_G)} s_n^*(\theta_0, \eta_0; Z_n, X_n)$, the analytical expression is given in Section 3.4. Note that the asymptotic covariance of $s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n; Z_n, X_n)$ may be singular in case (ii).

Notation – $\|\cdot\|$ denotes Euclidean norm for vectors and a matrix norm for matrices. $\xrightarrow[n\to\infty]{\mathsf{p}}$ convergence in probability, $\xrightarrow[n\to\infty]{L}$ convergence in distribution, $X_n = o_p(R_n)$ means $X_n = Y_n R_n$ and $Y_n \xrightarrow[n\to\infty]{\mathsf{p}} 0$ and $X_n = O_p(R_n)$ means $X_n$is asymptotically bounded if $R_n = 1$and $X_n = R_n Y_n$ where $Y_n = O_p(1)$ more generally. When the term neighborhood is used, it is assumed that it is open and non-empty.

3.3. Asymptotic distribution of generalized $C(\alpha)$ statistics

Dufour et al. (2016) consider the problem of testing a general (possibly nonlinear) restriction on a finite-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ of the form:

$$H_0 : \psi(\theta) = 0 \tag{3.3.22}$$

where $\psi : \Theta \mapsto \mathbb{R}^{p_1}$: when the parameter θ is specified by an $m \times 1$ estimating equation $D_n(\theta)$, $(p_1 \leq p \leq m)$ and any restricted estimator $\tilde{\theta}_n^0$ of θ which converges in probability to the "true value" θ_0 under H_0. Assuming along with other regularity conditions that

$$\sqrt{n} D_n(\theta_0) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, I_0] \tag{3.3.23}$$

where I_o is a $m \times m$ nonsingular matrix and under H_0,

$$\tilde{\theta}_n^0 - \theta_0 = O_p(n^{-1/2}), \tag{3.3.24}$$

they show that their proposed generalized $C(\alpha)$ statistic is asymptotically distributed as $\chi^2(p_1)$.

When the score $D_n(\theta)$ is employed to construct the restricted estimator $\tilde{\theta}_n^0$, e.g. by minimizing (3.2.2) over Θ under constraint (3.3.22), the convergence rate of $\tilde{\theta}_n^0$ is closely related to that of $D_n(\theta_0)$ and thus (3.3.23) may not hold when $\tilde{\theta}_n^0$ is consistent but not does converge at the parametric rate.

In this section, we extend the generalized $C(\alpha)$ test by Dufour et al. (2016) to allow for the restricted estimator $\tilde{\theta}_n^0$ and the score $D_n(\theta_0)$ to converge at nonstandard rates. It is shown that when (3.3.23), their generalized $C(\alpha)$ statistic is still asymptotically $\chi^2(p_1)$ even when (3.3.24) is replaced by a weaker condition that $\tilde{\theta}_n^0$ converges to θ_0 at rate faster than $n^{1/4}$ under the null. (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018), Bontemps (2019)). Such relaxation is achieved by exploiting further expansions of the score $D_n(\theta_0)$ and the restriction function $\psi(\theta)$ around $\theta = \theta_0$. This result can be generalized to the case when the score $D_n(\theta_0)$ converges at a general rate n^{r_D} where $r_D > 0$: we require that $\tilde{\theta}_n^0$ converges at a rate faster than $r_D/2$. Note that many of

the regularity conditions imposed are adapted from Dufour et al. (2016).

Let r_D, r_θ, r_M be positive constants

Assumption 3.3.1 EXISTENCE OF SCORE-TYPE FUNCTIONS.

$$D_n(\theta;\, \omega) = [D_{1n}(\theta;\, \omega), \ldots, D_{mn}(\theta;\, \omega)]',\ \omega \in \mathscr{Z},\ n = 1, 2, \ldots \tag{3.3.25}$$

is a sequence of $m \times 1$ random vectors, defined on a common probability space $(\mathscr{Z}, \mathscr{A}_{\mathscr{Z}}, \mathsf{P})$, which are functions of a $p \times 1$ parameter vector θ, where $\theta \in \Theta \subseteq \mathbb{R}^p$ ($p \leq m$) and Θ is a non-empty open subset of $\mathbb{R}^p$. All the random variables considered here as well in the following assumptions are functions of ω, so the symbol ω may be dropped to simplify notations [e.g., $D_n(\theta) := D_n(\theta; \omega)$] *. There is a unique vector $\theta_0 \in \Theta$ called the "true parameter value".*

The score $D_n(\theta)$ evaluated at $\theta = \theta_0$ and the restricted estimator $\tilde{\theta}_n^0$ converge to a Gaussian limit at the rate r_D and r_θ, respectively.

Assumption 3.3.2 SCORE ASYMPTOTIC NORMALITY.

$$n^{r_D} D_n(\theta_0) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, I(\theta_0)] \tag{3.3.26}$$

where $I(\theta_0)$ is a nonsingular $m \times m$ matrix.

We note that normality of the limiting distribution of $n^{r_D} D_n(\theta_0)$ is imposed to construct an asymptotically chi-squared test statistic and does not contribute to asymptotic negligibility of estimation error of $\tilde{\theta}_n$ in the test procedure.

Assumption 3.3.3 CONVERGENCE RATE OF THE RESTRICTED ESTIMATOR. *$\tilde{\theta}_n^0$, $n \geq 1$ is a random sequence on Θ such that*

$$\|\tilde{\theta}_n^0 - \theta_0\| = O_p\left(n^{-r_\theta}\right) \tag{3.3.27}$$

under H_0 in (3.3.22).

Assumption 3.3.3 only states that the scaled estimation error

$$r^\theta(\tilde{\theta}_n^0 - \theta_0) \tag{3.3.28}$$

is stochastically bounded and neither require the asymptotic distribution of $\tilde{\theta}_n^0$ to be known nor even exist. The following two assumptions play a key role in allowing for $r_\theta < r_D$. Assumption 3.4 in Dufour et al. (2016) is strengthened by imposing the existence of second-order expansion of $D_n(\theta)$ around $\theta = \theta_0$. Assumption 3.3.4 generalizes Assumption 3.4 in Dufour et al. (2016) by allowing for the linear expansion term to be scaled by some orders of n.

Assumption 3.3.4 SCORE EXPANSION. *For some non-empty open neighborhood $\mathscr{V}_D$ of θ_0 and some $p \times 1$ dimensional nonnegative vector $\beta = \left(\beta_1, \ldots \beta_p\right)'$, we have*:

$$D_n(\theta;\, \omega) = D_n(\theta_0;\, \omega) + J_\theta(\theta_0)\mathbf{n}^{-\beta}(\theta - \theta_0) + B_n(\theta, \theta_0; \omega) \tag{3.3.29}$$

for $\omega \in \mathscr{D}_J$ where $\mathscr{D}_J$ is an event with probability one, $J_\theta(\theta_0)$ is an $m \times p$ (nonrandom) function of θ, $\mathbf{n}^{-\beta}$ is a $p \times p$ diagonal matrix with the i-th entry $n^{-\beta_i}$, $i = 1, \ldots p$, and the remainder vector

$$B_n(\theta, \theta_0; \omega) = o_p(\|\theta - \theta_0\|). \tag{3.3.30}$$

In a standard case, β is a zero matrix so that $\mathbf{n}^{-\beta}$ is an identity matrix. However, there are problems where taking $\beta = 0$ leads to rank deficiency of $J_\theta(\theta_0)$. Such a problem is studied in Section 3.5. The scaling matrix $\mathbf{n}^{-\beta}$ essentially captures the rate at which the sensitivity of $D_n(\theta)$ with respect to $\theta - \theta_0$ decays around $\theta = \theta_0$.

Assumption 3.3.5 SCORE SECOND-ORDER EXPANSION. *Assumption* 3.3.4 *holds with*

$$B_n(\theta, \theta_0; \omega) = O_p(\|\theta - \theta_0\|^2). \tag{3.3.31}$$

Assumption 3.3.5 strengthens Assumption 3.4 in Dufour et al. (2016), who only assume $B_n(\theta, \theta_0; \omega)$ to be $o_p(\|\theta - \theta_0\|)$. It is satisfied when $D_n(\theta)$ is twice differentiable with probability one, however, it also covers a certain class of nonsmooth functions (see Bontemps (2019)).

Assumption 3.3.6 RESTRICTION FUNCTION: CONTINUOUS DIFFERENTIABILITY. *The function*

$$\psi(\theta) = [\psi_1(\theta), \ldots, \psi_{p_1}(\theta)]' \tag{3.3.32}$$

is a $p_1 \times 1$ *twice differentiable vector function of* θ *with first derivative*

$$P(\theta) = [P_1(\theta)', \ldots, P_{p_1}(\theta)']' \tag{3.3.33}$$

where

$$P_l(\theta) = \frac{\partial \psi_l(\theta)}{\partial \theta'}, \quad l = 1, \ldots, p_1 . \tag{3.3.34}$$

Assumption 3.3.7 RESTRICTION FUNCTION: TWICE DIFFERENTIABILITY. *Under Assumption* 3.3.6*, there exists an open neighborhood* $\mathcal{N}_P(\theta_0)$ *such that for any* $\theta^* \in \mathcal{N}_P(\theta_0)$ *the derivative* $\mathbb{H}_l(\theta^*)$ *of* $P_l(\theta^*)$ *exists and is bounded for any* $l = 1, \ldots, p_l$, i.e.

$$\sup_{1 \le l \le p_1} \sup_{\theta^* \in \mathcal{N}_P(\theta)} \|\mathbb{H}_l(\theta^*)\| < C_{P,\theta} \tag{3.3.35}$$

for some positive constant $C_{P,\theta}$.

While Assumption 3.3.7 imposes additional smoothness on ψ to Assumption 3.3.6 originally made in Dufour et al. (2016), empirically relevant equality constraints, such as linear or more generally polynomial restrictions, satisfy Assumption 3.3.6.

Assumption 3.3.8 LIPSCHITZ CONDITION ON $J(\theta)$. *There exists a nonempty open neighborhood* $\mathcal{V}_{J,1}(\theta_0)$ *such that for any* $\theta \in \mathcal{V}_{J,1}(\theta_0)$*,*

$$\|J_\theta(\theta) - J_\theta(\theta_0)\| \le C_{J,\theta_0} \|\theta - \theta_0\| \quad \text{for some constant } C_{J,\theta_0} > 0 . \tag{3.3.36}$$

Assumption 3.3.9 CONVERGENCE RATE OF THE ESTIMATOR $\tilde{J}_n(\theta)$ OF $J_\theta(\theta)$. $\{\tilde{J}_{\theta,n}(\theta) : n \ge 1\}$ *is a sequence of* $m \times p$ *random matrices such that there exists a nonempty open neighborhood* $\mathcal{V}_{J,2}(\theta_0)$ *such that under* H_0

$$\sup_{\theta \in \mathcal{V}_{J,2}} \|\tilde{J}_{\theta,n}(\theta) - J_\theta(\theta)\| = O_P(n^{-r_M}) \quad \text{where } r_M > 0 . \tag{3.3.37}$$

Assumption 3.3.10 CONVERGENCE RATE OF THE WEIGHT MATRIX. W_n, $n \ge 1$, *is a random sequence of* $m \times m$ *symmetric nonsingular* (*random*) *matrices such that*

$$\|W_n - W_0\| = O_P(n^{-r_M}) \tag{3.3.38}$$

where W_0 is a nonsingular non-random matrix.

Assumption 3.3.11 imposes restrictions on the constants (r_D, r_θ, r_M).

Assumption 3.3.11 RESTRICTIONS ON THE CONVERGENCE RATES. *$r_\theta > r_D/2$ and $r_\theta + r_M > r_D$.*

Assumption 3.3.12 CONVERGENCE RATE OF THE ESTIMATOR OF $J(\theta)$. *$\tilde{I}_n, n \geq 1$ is a sequence of $p \times p$ random matrices such that*

$$\|\tilde{I}_n - I(\theta_0)\| = o_p(1)\,. \tag{3.3.39}$$

Assumption 3.3.13 NON-DEGENERACY OF THE INFORMATION MATRIX.

$$\text{rank}\,[I(\theta_0)] = m\,. \tag{3.3.40}$$

Assumption 3.3.14 NON-DEGENERACY OF THE JACOBIAN MATRIX. *Assumption* 3.3.4 *is satisfied with some nonnegative value of β such that*

$$\text{rank}\,[J_\theta(\theta_0)] = p\,. \tag{3.3.41}$$

Assumption 3.3.15 MATRIX RANK: $(\tilde{J}_{\theta,n}(\tilde{\theta}_n^0), \tilde{I}_n)$. *For any $n \geq 1$, the matrices $\tilde{J}_{\theta,n}(\tilde{\theta}_n^0)$ and $\tilde{I}_n$ have full rank with probability one.*

Assumption 3.3.16 NON-DEGENERACY OF RESTRICTION JACOBIAN.

$$\text{rank}\,[P(\theta)] = p_1 \tag{3.3.42}$$

for any θ such that $\psi(\theta) = 0$.

Assumption 3.3.17 RESTRICTED ESTIMATOR. *$\psi(\theta_0) = 0$ and $\psi(\check{\theta}_n) = O_p(n^{-2r_\theta})$ where $\check{\theta}_n$ is defined as*

$$\check{\theta}_n = \theta_0 + n^{-\beta}(\tilde{\theta}_n^0 - \theta_0) \tag{3.3.43}$$

for such value of β that satisfies Assumption 3.3.14.

In the standard case where β is a zero vector, Assumption 3.3.17 allows for the restricted estimator $\tilde{\theta}_n^0$ to satisfy the restriction $\psi(\theta) = 0$ only asymptotically at the rate n^{-2r_θ}. If $\psi(\tilde{\theta}_n^0) = 0$, then this assumption holds trivially since $\tilde{\theta}_n^0 = \check{\theta}_n$. If β is not a zero vector, $\check{\theta}_n$ is as an (infeasible) estimator of θ_0 which strictly improves $\tilde{\theta}_n^0$ in terms of bias, i.e. $\|\check{\theta}_n - \theta_0\| < \|\tilde{\theta}_n - \theta_0\|$ everywhere while it may not satisfy the restriction $\psi(\theta) = 0$ exactly. Assumption 3.3.17 requires that the distance $\psi(\check{\theta}_n) - \psi(\theta_0)$ is of order n^{-2r_θ}. This condition can be checked, for example, when the order of bias of $\tilde{\theta}_n^0$ is known. In addition, for a restriction fixing a subvector of θ, $\psi(\tilde{\theta}_n^0) = 0$ implies $\psi(\check{\theta}_n) = 0$. Let

$$\tilde{Q}_n := \tilde{Q}[W_n] = \tilde{P}_n[\tilde{J}'_{\theta,n} W_n \tilde{J}_{\theta,n}]^{-1} \tilde{J}'_{\theta,n} W_n \tag{3.3.44}$$

where $\tilde{J}_{\theta,n} := \tilde{J}_{\theta,n}(\tilde{\theta}_n^0)$, $\tilde{P}_n := P(\tilde{\theta}_n^0)$. Define the $p_1 \times 1$ estimating function

$$s_n(\theta) = \tilde{Q}_n D_n(\theta) \tag{3.3.45}$$

and the generalized $C(\alpha)$ statistic $PC\big(\tilde{\theta}_n^0; \psi\big)$

$$PC(\tilde{\theta}_n^0; \psi) = n^{2r_D}\, s_n(\tilde{\theta}_n^0)' \left[\tilde{Q}_n \tilde{I}_n \tilde{Q}_n'\right]^{-1} s_n(\tilde{\theta}_n^0)\,. \tag{3.3.46}$$

Under the assumptions above, we establish the asymptotic distributions of $s_n(\tilde{\theta}_n^0)$ and the test statistic $PC(\tilde{\theta}_n^0; \psi)$.

Proposition 3.3.1 ASYMPTOTIC DISTRIBUTION OF GENERALIZED $C(\alpha)$ STATISTIC. *If Assumption* 3.3.1 *to* 3.3.17 *are satisfied, under* H_0*,*

$$n^{r_D}\, s_n(\tilde{\theta}_n^0) \xrightarrow[n\to\infty]{L} \mathrm{N}\big[0,\, Q(\theta_0) I(\theta_0) Q(\theta_0)'\big] \tag{3.3.47}$$

where $s_n(\theta)$ *is defined in* (3.3.45)*, and*

$$Q(\theta_0) = P(\theta_0)\big[J_\theta(\theta_0)' W_0 J_\theta(\theta_0)\big]^{-1} J_\theta(\theta_0)' W_0\,. \tag{3.3.48}$$

Furthermore, the generalized $C(\alpha)$ *statistic* $PC(\tilde{\theta}_n^0; \psi)$ *in* (3.3.46) *is asymptotically distributed as* $\chi^2(p_1)$*. When* $r_\theta \geq r_D$*, these above assertions hold without Assumptions* 3.3.5 *and* 3.3.7*.*

Proposition 3.3.1 generalizes Proposition 3.1 of Dufour et al. (2016), which establishes the asymptotic distribution of $PC(\tilde{\theta}_n^0; \psi)$ under $r_D = r_\theta = 1/2$. When $r_D = 1/2$, our result holds if $r_\theta > 1/4$, *i.e.* $\tilde{\theta}_n^0$ converges at a rate faster than $n^{1/4}$ as long as $\tilde{Q}_n$ in (3.3.44) converges to $Q(\theta_0)$ at an appropriately fast rate, which equates to the condition

$$r_\theta + r_M > r_D \tag{3.3.49}$$

in Assumption 3.3.11. In the worse case where r_θ is arbitrarily close to $1/4$, 3.3.49 requires r_M be also larger than $1/4$. On the other hand, when $r_\theta = 1/2$, (3.3.49) always holds when $r_M > 0$ so that the convergence rate of $\tilde{Q}_n$ can be arbitrarily slow.

Note that given the asymptotic normality of $s_n(\tilde{\theta}_n^0)$ in (3.3.47), an alternative $C(\alpha)$-type statistic can be constructed by considering linear transformation of $s_n(\tilde{\theta}_n^0)$: by some $p^* \times p_1$ matrix R_S where $p^* \in \{1, \ldots, p_1\}$:

$$s_n^{(R_s)}(\tilde{\theta}_n^0) = R_s s_n(\tilde{\theta}_n^0) \tag{3.3.50}$$

Such transformation puts on different weights on the p_1 restrictions specified by each element of ψ. A test statistic based on the estimating function $s_n^{(R_s)}(\tilde{\theta}_n^0)$ may have improvement in power while it may could lose power in certain directions of the restrictions.

In Section 3.4, we allow for the presence of an nuisance parameter in the testing problem and establish a valid $C(\alpha)$ test.

3.4. Extended $C(\alpha)$ statistics with multiple convergence rates

In this section, we extend the framework in Section 3.3 to testing problems in the presence of nuisance parameters. Specifically, we are interested in testing the hypothesis of the form:

$$H_0 : \psi(\theta) = 0 \tag{3.4.51}$$

in the presence of η where $(\theta, \eta) \in \Theta \times \mathscr{E} \subseteq \mathbb{R}^p \times \mathbb{R}^q$ is a pair of the parameter of interest θ and the nuisance parameter η defined by the primary estimating function $D_n(\theta, \eta)$ which

depends on both θ and η and the auxiliary one $G_n(\eta)$ which only depends on η. Let $r_D, r_G, r_\theta, r_\eta$ be positive constants.

Assumption 3.4.1 EXISTENCE OF SCORE-TYPE FUNCTIONS.

$$D_n(\theta,\eta;\,\omega) = [D_{1n}(\theta,\eta;\,\omega),\ldots,D_{mn}(\theta,\eta;\,\omega)]',\ \omega\in\mathscr{Z},\ n=1,2,\ldots \tag{3.4.52}$$

is a sequence of $m\times 1$ random vectors, defined on a common probability space $(\mathscr{Z}, \mathscr{A}_{\mathscr{Z}}, \mathsf{P})$, which are functions of a pair of parameter vectors $(\theta,\eta)\in\Theta\times\mathscr{E}\subset\mathbb{R}^p\times\mathbb{R}^q$ where Θ is a subset of $\mathbb{R}^p$ and ε is a subset of $\mathbb{R}^q$.

$$G_n(\eta;\,\omega) = [G_{1n}(\eta;\,\omega),\ldots,G_{qn}(\eta;\,\omega)]',\ \omega\in\mathscr{Z},\ n=1,2,\ldots \tag{3.4.53}$$

is a sequence of $q\times 1$ random vectors, defined on $(\mathscr{Z}, \mathscr{A}_{\mathscr{Z}}, \mathsf{P})$, which is functions of $\eta\in\mathscr{E}$ only. There is a unique vector $(\theta_0,\eta_0)\in\Theta\times\mathscr{E}$ called the "true parameter value".

As in Section 3.3, we assume the second-order expansion of the scores $D_n(\theta,\eta)$ around (θ_0,η_0) and $G_n(\eta)$ around η_0

Assumption 3.4.2 SCORE SECOND-ORDER EXPANSION. *For some non-empty open neighborhood $\mathscr{U}_D$ of (θ_0,η_0) and a pair $(\beta_\theta,\beta_\eta)$ of nonnegative vectors of dimensions p and q respectively,*

$$D_n(\theta,\eta;\,\omega) = D_n(\theta_0,\eta_0;\,\omega) + J(\theta_0,\eta_0)\begin{bmatrix}\mathbf{n}^{-\beta_\theta} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{n}^{-\beta_\eta}\end{bmatrix}\begin{bmatrix}\theta-\theta_0\\ \eta-\eta_0\end{bmatrix} + B_{D,n}(\theta,\theta_0,\eta,\eta_0,\omega) \tag{3.4.54}$$

for $\omega\in\Omega_D$ where Ω_D is an event with probability one where

$$J(\theta,\eta) = [J_\theta(\theta,\eta), J_\eta(\theta,\eta)] \tag{3.4.55}$$

is an $m\times(p+q)$ (nonrandom) function of (θ,η), $\mathbf{n}^{-\beta}$ is a $p\times p$ diagonal matrix with the i-th entry $n^{-\beta_{\theta,i}}$, $i=1,\ldots p$, $\mathbf{n}^{-\beta_\eta}$ is a $q\times q$ diagonal matrix with the j-th entry $n^{-\beta_{\eta,j}}$, $j=1,\ldots q$, and the remainder vector $B_{D,n}(\theta,\theta_0,\eta,\eta_0,\omega)$ is $O_p(\max(\|\theta-\theta_0\|^2,\|\eta-\eta_0\|^2))$.

Assumption 3.4.3 SCORE SECOND-ORDER EXPANSION. *For some non-empty open neighborhood $\mathcal{V}_G$ of η_0 and a q-dimensional nonnegative vector β_η,*

$$G_n(\eta;\,\omega) = G_n(\eta_0;\,\omega) + g(\eta_0)\mathbf{n}^{-\beta_\eta}(\eta-\eta_0) + B_{G,n}(\eta,\eta_0,\omega) \tag{3.4.56}$$

for $\omega \in \Omega_G$ where Ω_G is an event with probability one, $g(\eta_0)$ is an $q \times q$ (nonrandom) function of η, $\mathbf{n}^{-\beta_\eta}$ is a $q \times q$ diagonal matrix with the j-th entry $n^{-\beta_{\eta,j}}$, $j = 1,\ldots q$, and the remainder vector satisfies

$$B_{G,n}(\eta,\eta_0,\omega) = O_p(\|\eta-\eta_0\|^2)\,. \tag{3.4.57}$$

We are particularly interested in the case where

$$\text{rank}[J(\theta,\eta)] = p+q \tag{3.4.58}$$

does not hold and $D_n(\theta_0,\eta_0)$ and $G_n(\eta_0)$ converge to non-degenerate distributions at different rates.

When (3.4.58) holds, one may simply apply the testing framework in Section 3.3 based solely on $D_n(\theta,\eta)$ by considering a $(p+q)$-dimensional vector θ^* defined as

$$\theta^* = (\theta',\eta')' \tag{3.4.59}$$

and reformulating (3.4.51) as

$$H_0 : \psi^*(\theta^*) = 0 \tag{3.4.60}$$

where

$$\psi^*(\theta^*) = \{\psi(\theta) : \theta \text{ is the first } p : \text{elements of } \theta^*\} \tag{3.4.61}$$

(3.4.58), however, requires that $p+q \geq m$ and thus rules out important cases where (θ,η) is not identified solely from $D_n(\theta,\eta)$ and the auxiliary equation $G_n(\eta)$ is required. Alternatively, suppose $D_n(\theta_0,\eta_0)$ and $G_n(\eta_0)$ converge jointly to a Gaussian limit but at

different rates, n^{r_D} and n^{r_G} where $r_G > r_D$. Then, the asymptotic covariance of

$$n^{r_D}\begin{bmatrix} D_n(\theta_0,\eta_0) \\ G_n(\eta_0) \end{bmatrix} \tag{3.4.62}$$

is singular even if the asymptotic covariance of $n^{r_G}G_n(\eta_0)$ is nonsingular since

$$n^{r_D}G_n(\eta_0) \xrightarrow[n\to\infty]{\mathsf{p}} 0\,. \tag{3.4.63}$$

Thus, the testing problem in the presence of the nuisance parameter η requires separate treatment from Section 3.3.

Assumption 3.4.1 specifies the number of equations in $G_n(\eta)$ to be equal to the dimension q of the parameter η. This assumption is only made for ease of exposition and the proposed test procedure can be easily extended to the over-identified case. Such extension is discussed briefly in the latter part of this section.

We now proceed to the rest of assumptions. As mentioned earlier, we allow for the main and auxiliary estimating functions $D_n(\theta,\eta)$ and $G_n(\eta)$ evaluated at (θ_0,η_0) to converge at different rates.

Assumption 3.4.4 RATES OF THE ESTIMATING FUNCTIONS. *$n^{r_D}D_n(\theta_0,\eta_0)$ and $n^{r_G}G_n(\eta_0)$ are stochastically bounded and non-degenerate.*

The following assumptions posit the existence of the restricted estimators $\tilde{\theta}_n^0$ and $\hat{\eta}_n$ which are both consistent under H_0 but may converge at different rates.

Assumption 3.4.5 CONVERGENCE RATE OF THE RESTRICTED PARAMETER. *$\tilde{\theta}_n^0$, $n \geq 1$ is a random sequence on Θ such that*

$$\|\tilde{\theta}_n^0 - \theta_0\| = O_p\left(n^{-r_\theta}\right) \tag{3.4.64}$$

under H_0.

Assumption 3.4.6 CONVERGENCE RATE OF THE NUISANCE PARAMETER. *$\hat{\eta}_n$, $n \geq 1$ is a random sequence on $\mathscr{E}$ such that*

$$\|\hat{\eta}_n - \eta_0\| = O_p\left(n^{-r_\eta}\right) \tag{3.4.65}$$

under H_0.

The following assumptions are counterparts of Assumption 3.3.6-3.3.17 in the presence of the additional parameter η.

Assumption 3.4.7 LIPSCHITZ CONDITION ON $J(\theta,\eta)$. *There exists a nonempty open neighborhood* $\mathscr{U}_{J,1}(\theta_0,\eta_0)$ *such that, for some constant* $C_{J,\theta_0}>0$,

$$\|J(\theta,\eta)-J(\theta_0,\eta)\|\leq C_{J,\theta_0}\|\theta-\theta_0\|, \quad \textit{for any } (\theta,\eta)\in\mathscr{U}_{J,1}(\theta_0,\eta_0), \tag{3.4.66}$$

and there exists a nonempty open neighborhood $U_J(\eta_0)$ *such that, for some constant* $C_{J,\eta_0}>0$,

$$\|J(\theta_0,\eta)-J(\theta_0,\eta_0)\|\leq C_{J,\eta_0}\|\eta-\eta_0\|, \quad \textit{for any } \eta\in U_J(\eta_0) \tag{3.4.67}$$

Assumption 3.4.8 CONVERGENCE RATE OF THE ESTIMATOR OF $J(\theta,\eta)$. $\tilde{J}_n(\theta,\eta)=\left[\tilde{J}_{\theta,n}(\theta,\eta),\tilde{J}_{\eta,n}(\theta,\eta)\right], n\geq 1$ *is a sequence of* $m\times(p+q)$ *random matrices such that there exists a nonempty open neighborhood* $\mathscr{U}_{J,2}(\theta_0,\eta_0)$ *such that, under* H_0,

$$\sup_{(\theta,\eta)\in\mathscr{U}_{J,2}}\|\tilde{J}_{\theta,n}(\theta,\eta)-J_\theta(\theta,\eta)\|=O_P(n^{-r_M}) \tag{3.4.68}$$

and

$$\sup_{(\theta,\eta)\in\mathscr{U}_{J,2}}\|\tilde{J}_{\eta,n}(\theta,\eta)-J_\eta(\theta,\eta)\|=O_P(n^{-r_M}) \tag{3.4.69}$$

Given Assumption 3.4.5- 3.4.7, 3.4.8, we consider the estimator

$$[\tilde{J}_{\theta,n},\tilde{J}_{\eta,n}]:=[\tilde{J}_n(\tilde{\theta}_n^0,\hat{\eta}_n),\tilde{J}_{\eta,n}(\tilde{\theta}_n^0,\hat{\eta}_n)] \tag{3.4.70}$$

of $(J_\theta(\theta_0,\eta_0),J_\eta(\theta_0,\eta_0))$. Similarly, the following regularity conditions are imposed on $g(\eta)$ and its estimator $\tilde{g}(\eta)$.

Assumption 3.4.9 LIPSCHITZ CONDITION ON $g(\theta)$. *There exists a nonempty open neighborhood* $\mathscr{U}_{g,1}(\eta_0)$ *such that for any* $\eta\in\mathscr{U}_{g,1}(\eta_0)$,

$$\|g(\eta)-g(\eta_0)\|\leq C_{g,\eta_0}\|\eta-\eta_0\|, \tag{3.4.71}$$

for some constant $C_{g,\eta_0} > 0$.

Assumption 3.4.10 CONVERGENCE RATE OF THE ESTIMATOR OF $g(\eta)$. $\tilde{g}_n(\eta), n \geq 1$ *is a sequence of* $q \times q$ *random matrices such that there exists a nonempty open neighborhood* $\mathscr{U}_{g,2}(\eta_0)$ *such that, under* H_0,

$$\sup_{\eta \in \mathscr{U}_{g,2}(\eta_0)} \|\tilde{g}_n(\eta) - g(\eta)\| = O_P\left(n^{-r_M}\right) \tag{3.4.72}$$

where $r_M > 0$.

Under Assumption 3.4.6, 3.4.9, 3.4.10, we define the estimator

$$\tilde{g}_n := \tilde{g}_n(\hat{\eta}_n) \tag{3.4.73}$$

of $g(\eta_0)$.
We assume that the matrices $J_\theta(\theta_0, \eta_0)$, $J_\eta(\theta_0, \eta_0)$, and $g(\eta_0)$ have full-rank.

Assumption 3.4.11 NON-DEGENERACY OF THE JACOBIAN MATRICES. *Assumption* 3.4.2 *and* 3.4.3 *are satisfied with some* $(\beta_\theta, \beta_\eta) \in \mathbb{R}_+^{p \times q}$ *such that*

$$\operatorname{rank}[J_\theta(\theta_0, \eta_0)] = p, \tag{3.4.74}$$

$$\operatorname{rank}[J_\eta(\theta_0, \eta_0)] = \min(q, m), \tag{3.4.75}$$

and

$$\operatorname{rank}[g(\eta_0)] = q. \tag{3.4.76}$$

Note that Assumption 3.4.11 requires that $m \geq p$.

Assumption 3.4.12 MATRIX RANK: $(\tilde{J}_{\theta,n}(\tilde{\theta}_n^0, \hat{\eta}_n), \tilde{J}_{\eta,n}(\tilde{\theta}_n^0, \hat{\eta}_n), \tilde{g}_n(\hat{\eta}_n))$. *The matrices* $\tilde{J}_{\theta,n}(\tilde{\theta}_n^0, \hat{\eta}_n)$, $\tilde{J}_{\eta,n}(\tilde{\theta}_n^0, \hat{\eta}_n)$, *and* $\tilde{g}_n$ *have full-rank with probability approaching to one.*

The next assumption is to avoid repetition of regularity conditions that appeared in Section 3.3.

Assumption 3.4.13 SET OF REGULARITY CONDITIONS. *Assumptions* 3.3.10, 3.3.6 *and* 3.3.16 *hold.*

Finally, restrictions on the constants $(r_D, r_G, r_\theta, r_\eta, r_M)$ are imposed.

Assumption 3.4.14 RESTRICTIONS ON THE CONVERGENCE RATES.

$$\min(r_\theta, r_\eta) > \min(r_D, r_G)/2 \tag{3.4.77}$$

and

$$\min(r_\theta, r_\eta) + r_M > \min(r_D, r_G)/2$$

Let

$$\tilde{Q}_n := \tilde{Q}[W_n] = \tilde{P}_n [\tilde{J}'_{\theta,n} W_n \tilde{J}_{\theta,n}]^{-1} \tilde{J}'_{\theta,n} W_n \tag{3.4.78}$$

where

$$\tilde{P}_n = P(\tilde{\theta}_n^0), \quad \tilde{J}_{\theta,n} = \tilde{J}_{\theta,n}(\tilde{\theta}_n^0, \hat{\eta}_n) \tag{3.4.79}$$

and

$$\tilde{J}_{\eta,n} = \tilde{J}_{\eta,n}(\tilde{\theta}_n^0, \hat{\eta}_n),\ \tilde{g}_n = \tilde{g}_n(\hat{\eta}_n)\,. \tag{3.4.80}$$

We consider the following score function:

$$s_n^*(\theta, \eta) = \tilde{Q}_n \{D_n(\theta, \eta) - \tilde{J}_{\eta,n} \tilde{g}_n^{-1} G_n(\eta)\}\,. \tag{3.4.81}$$

which is the basis for our $C(\alpha)$-type statistic.

Note that (3.4.81) can be written as a linear transformation of the estimating function $\begin{bmatrix} D_n(\theta,\eta) \\ G_n(\eta) \end{bmatrix}$:

$$s_n^*(\theta, \eta) = \tilde{T}_n \begin{bmatrix} D_n(\theta, \eta) \\ G_n(\eta) \end{bmatrix}, \quad \tilde{T}_n := \tilde{Q}_n \{\mathbb{I}_{m\times m}, -\tilde{J}_{\eta,n} \tilde{g}_n^{-1}\}\,. \tag{3.4.82}$$

The following lemma characterizes asymptotic properties of $s_n^*(\theta, \eta)$.

Lemma 3.4.1 *Suppose Assumptions* 3.4.12 - 3.4.14 *hold. Under* H_0 *in* (3.4.51)*, the following hold:*

1. *If* $r_D = r_G$,

$$n^{r_D} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) = n^{r_D} Q(\theta_0, \eta_0)[D_n(\theta_0, \eta_0) - J_\eta(\theta_0, \eta_0) g(\eta_0)^{-1} G_n(\eta_0)] + o_P(1) \tag{3.4.83}$$

2. *If* $r_G > r_D$,

$$n^{r_D} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) = n^{r_D} Q(\theta_0, \eta_0) D_n(\theta_0, \eta_0) + o_P(1)\,. \tag{3.4.84}$$

3. *If* $r_G < r_D$,

$$n^{r_G} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) = -n^{r_G} J_\eta(\theta_0, \eta_0) g(\eta_0)^{-1} G_n(\eta_0) + o_P(1)\,. \tag{3.4.85}$$

This result implies that one of $(D_n(\theta_0, \eta_0), G_n(\eta_0))$ is negligible when $r_D \neq r_G$ and the convergence rate of $s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)$ is the smaller one of (r_D, r_G). Asymptotic normality of $n^{\max(r_\theta, r_\eta)} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)$ can be achieved by the following assumption.

Assumption 3.4.15 JOINT ASYMPTOTIC NORMALITY OF $(D_n(\theta_0, \eta_0), G_n(\eta_0))$.

$$\begin{bmatrix} n^{r_D} D_n(\theta_0, \eta_0) \\ n^{r_G} G_n(\eta_0) \end{bmatrix} \xrightarrow[n\to\infty]{L} \mathrm{N}[0,\, \Phi(\theta_0, \eta_0)] \tag{3.4.86}$$

where $\Phi(\theta, \eta)$ *is a* $(m+q) \times (m+q)$ *nonsingular matrix with submatrix blocks* $I(\theta, \eta), \Sigma(\eta), \Xi(\theta, \eta)$ *of size* $m \times m, m \times q$, *and* $q \times q$, *respectively, such that*

$$\Phi(\theta, \eta) = \begin{bmatrix} I(\theta, \eta), & \Xi(\theta, \eta) \\ \Xi(\theta, \eta)' & \Sigma(\eta) \end{bmatrix}. \tag{3.4.87}$$

Furthermore, $(\tilde{I}_n, \tilde{\Xi}_n, \tilde{\Sigma}_n)$ *is a trio of full-rank* $m \times m$, $m \times q$, *and* $q \times q$ *random matrices such that*

$$\tilde{I}_n \xrightarrow[n\to\infty]{\mathrm{p}} I(\theta_0, \eta_0) : \tilde{\Xi}_n \xrightarrow{p} \Xi(\theta_0, \eta_0), \tilde{\Sigma}_n \xrightarrow{p} \Sigma(\eta_0) \tag{3.4.88}$$

under H_0.

Establishing the joint asymptotic distribution of $(D_n(\theta_0, \eta_0), G_n(\eta_0))$ can be challenging in practice. However, when $D_n(\theta_0, \eta_0)$ and $G_n(\eta_0)$ are independent, asymptotic normality of each of $D_n(\theta_0, \eta_0)$ and $G_n(\eta_0)$ implies Assumption 3.4.15 with $\Xi(\theta_0, \eta_0) = 0$.

Independence of $D_n(\theta_0,\eta_0)$ and $G_n(\eta_0)$ holds when $D_n(\theta_0,\eta_0)$ and $G_n(\eta_0)$ are constructed from independent data sets. When the observations are i.i.d., this suggest the use of the sample split method [Angrist and Krueger (1995), Staiger and Stock (1994), Dufour and Jasiak (2001)].

Proposition 3.4.2 *Maintain the assumptions in Lemma* 3.4.1. *In addition, suppose Assumption* 3.4.15 *holds. Then, under* H_0 *in* (3.4.51)*, we have:*

$$n^{2\max(r_\theta,r_\eta)} s_n^*(\tilde{\theta}_n^0,\hat{\eta}_n) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0,\Sigma^*(\theta_0,\eta_0)\right] \tag{3.4.89}$$

where

$$\Lambda(\theta_0,\eta_0) = \begin{cases} T(\theta_0,\eta_0)\Phi(\theta_0,\eta_0)T(\theta_0,\eta_0)' & \text{if } r_D = r_G \\ Q(\theta_0,\eta_0)I(\theta_0,\eta_0)Q(\theta_0,\eta_0)' & \text{if } r_D < r_G \\ \Pi(\theta_0,\eta_0)\Sigma(\eta_0)\Pi(\theta_0,\eta_0)' & \text{if } r_D > r_G, \end{cases} \tag{3.4.90}$$

$$Q(\theta_0,\eta_0) = P(\theta_0)\left[J_\theta(\theta_0,\eta_0)'W_0 J_\theta(\theta_0,\eta_0)\right]^{-1} J_\theta(\theta_0,\eta_0)'W_0, \tag{3.4.91}$$

$$T(\theta_0,\eta_0) = Q(\theta_0,\eta_0)[\mathbb{I}_{m\times m}, -J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}], \tag{3.4.92}$$

$$\Pi(\theta_0,\eta_0) = Q(\theta_0,\eta_0)J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}. \tag{3.4.93}$$

Further,

$$\begin{aligned} \operatorname{rank}[\Lambda(\theta_0,\eta_0)] &= p_1 && \text{if } r_D \le r_G \\ &= \operatorname{rank}[\Pi(\theta_0,\eta_0)] \le \min(p_1,q) && \text{if } r_D > r_G. \end{aligned} \tag{3.4.94}$$

Note that when $r_G > r_D$ Assumption 3.4.15 can be replaced by the following weaker condition as we only require asymptotic normality of $D_n(\theta_0,\eta_0)$ but not of $G_n(\eta_0)$ since the term involving the latter is asymptotically negligible.

Assumption 3.4.16 ASYMPTOTIC NORMALITY OF $D_n(\theta_0,\eta_0)$ AND THE CONVERGENCE RATE OF $G_n(\eta_0)$.

$$n^{r_D} D_n(\theta_0,\eta_0) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0, I(\theta_0,\eta_0)\right] \tag{3.4.95}$$

where $I(\theta_0, \eta_0)$ is a $m \times m$ nonsingular matrix and

$$n^{r_G} G_n(\eta_0) = O_p(n^{-r_G}). \tag{3.4.96}$$

Similarly, when $r_G < r_D$, we only require asymptotic normality of $G_n(\theta_0, \eta_0)$.

Assumption 3.4.17 .

$$n^{r_G} G_n(\eta_0) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, \Sigma(\eta_0)] \tag{3.4.97}$$

where $\Sigma(\theta_0, \eta_0)$ is a $q \times q$ nonsingular matrix and

$$n^{r_D} D_n(\theta_0, \eta_0) = O_p(n^{-r_D}). \tag{3.4.98}$$

Corollary 3.4.3 *In Proposition* 3.4.4*, Assumption* 3.4.15 *can be replaced by Assumption 3.4.16 when $r_D < r_G$ and Assumption 3.4.17 when $r_D > r_G$.*

Assumption 3.4.18 ESTIMATOR OF $\Phi(\theta_0, \eta_0)$. *The sequence of $(m+q) \times (m+q)$ nonsingular matrixes $\tilde{\Phi}_n$ with submatrix blocks $\tilde{I}_n, \tilde{\Sigma}_n, \tilde{\Xi}_n$ of size $m \times m, m \times q$, and $q \times q$, respectively*

$$\tilde{\Phi}_n = \begin{bmatrix} \tilde{I}_n, & \tilde{\Xi}_n \\ \tilde{\Xi}_n' & \tilde{\Sigma}_n \end{bmatrix}. \tag{3.4.99}$$

converges in probability to $\Phi(\theta_0, \eta_0)$ under H_0.

Under Assumption 3.4.18, we denote by $\tilde{\Phi}_n$ the consistent estimator of $\Phi(\theta_0, \eta_0)$ under H_0 defined as

$$\tilde{\Phi}_n = \begin{bmatrix} \tilde{I}_n, & \tilde{\Xi}_n \\ \tilde{\Xi}_n' & \tilde{\Sigma}_n \end{bmatrix}. \tag{3.4.100}$$

Let

$$\tilde{T}_n = \tilde{Q}_n[\mathbb{I}_{m\times m}, -\tilde{J}_{\eta,n}\tilde{g}_n^{-1}] \tag{3.4.101}$$

$$\tilde{\Pi}_n = -\tilde{Q}_n \tilde{J}_{\eta,n} \tilde{g}_n^{-1}. \tag{3.4.102}$$

We consider the following extended generalized test statistic:

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \psi) = n^{2\min(r_\theta, r_\eta)} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)'(\hat{\Lambda}_n)^- s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) \tag{3.4.103}$$

where A^- is the Moore–Penrose generalized inverse (Moore (1920), Penrose (1955)) of A and $\hat{\Lambda}_n \in \left\{\tilde{\Lambda}_n, \tilde{\Lambda}_n^*\right\}$ is a consistent estimator of $\Sigma^*(\theta_0, \eta_0)$:

$$\tilde{\Lambda}_n = \begin{cases} \tilde{T}_n \tilde{\Phi}_n \tilde{T}_n' & \text{if } r_D = r_G \\ \tilde{Q}_n \tilde{I}_n \tilde{Q}_n' & \text{if } r_D < r_G \\ \tilde{\Pi}_n \tilde{\Sigma}_n \tilde{\Pi}_n' & \text{if } r_D > r_G, \end{cases} \tag{3.4.104}$$

and

$$\tilde{\Lambda}_n^* = \begin{cases} \tilde{T}_n \tilde{\Phi}_n \tilde{T}_n' & \text{if } r_D = r_G \\ \tilde{T}_n \tilde{\Phi}_{D,n} \tilde{T}_n' & \text{if } r_D < r_G \\ \tilde{T}_n \tilde{\Phi}_{G,n} \tilde{T}_n' & \text{if } r_D > r_G, \end{cases} \tag{3.4.105}$$

where

$$\tilde{\Phi}_{D,n} = \begin{bmatrix} \tilde{I}_n, & (n^{r_D - r_G}) \tilde{\Xi}_n \\ n^{(r_D - r_G)} \tilde{\Xi}_n' & n^{2(r_D - r_G)} \tilde{\Sigma}_n \end{bmatrix}, \tilde{\Phi}_{G,n} = \begin{bmatrix} n^{2(r_G - r_D)} \tilde{I}_n, & n^{(r_G - r_D)} \tilde{\Xi}_n \\ n^{(r_G - r_D)} \tilde{\Xi}_n' & \tilde{\Sigma}_n \end{bmatrix}. \tag{3.4.106}$$

Assumption 3.4.19 MATRIX RANK: $\tilde{\Pi}_n$.

$$\text{rank}[\tilde{\Pi}_n] = \text{rank}[\Pi(\theta_0, \eta_0)]. \tag{3.4.107}$$

with probability approaching to one.

Then, we have the following result.

Proposition 3.4.4 ASYMPTOTIC DISTRIBUTION OF THE MODIFIED GENERALIZED TEST STATISTIC. *Suppose the assumptions in Proposition* 3.4.4 *and Assumption* 3.4.18 *and* 3.4.18 *hold.*

1. *Suppose either $r_D \leq r_G$, or $r_D > r_G$ and $q \geq m$. Let $\hat{\Lambda}_n \in \{\tilde{\Lambda}_n, \tilde{\Lambda}_n^*\}$. Then,*

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \psi) \xrightarrow[n \to \infty]{L} \mathscr{X}^2(p_1) \tag{3.4.108}$$

2. *Suppose $r_D > r_G$ and $q < m$. Assumption 3.4.19 holds. Then,*

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \tilde{\Lambda}_n \psi) \xrightarrow[n\to\infty]{L} \mathscr{X}^2(p^*) \tag{3.4.109}$$

where

$$p_1 + q - m \leq p^* := \text{rank}[\Pi(\theta_0, \eta_0)] \leq \min(p_1, q)\,. \tag{3.4.110}$$

Proposition 3.4.4 establishes the asymptotic distribution of the statistic $EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \psi)$. In 2., where $r_D > r_G$ and $q < m$, the asymptotic covariance of $n^{\min(r_\theta, r_\eta)} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)$ may be singular and the asymptotic distribution depends on the rank of $\Pi(\theta_0, \eta_0)$. $(\tilde{\Lambda}_n^*)^-$ fails to converge to $(\Lambda(\theta_0, \eta_0))^-$ since $\tilde{\Lambda}_n^*$ has full rank with probability approaching to one. Appropriate regularization of $\tilde{\Lambda}_n^*$ could make its generalized inverse operation continuous, however, such extension is beyond the scope of this paper.

3.5. Local estimating equations and moment conditions

In this section, we consider the application of the generalized $C(\alpha)$ test procedure to problems in the local estimating equation and moment equation setup. We describe the problem and discuss the advantages of the generalized $C(\alpha)$ test over the Wald-type test considered in Calonico, Cattaneo and Titiunik (2014) and the empirical likelihood-based Lagrange multiplier test by Xu (2020). The rest of the subsections consider applications. Section 3.5.2 studies testing on the derivatives of a nonparametric regression function. In Section 3.5.3, we apply the test for the problem of construction of a confidence set for the average treatment effect in the regression discontinuity design considered in Calonico et al. (2014). Section 3.5.4 considers a specification test of the semiparametric stochastic discount factor model.

3.5.1. Hypothesis testing under local estimating equation and moment conditions

The finite-dimensional parameter of interest θ is often defined by local moment conditions of the form

$$\mathbb{E}[m(\theta_0; Z) \,|\, X = x_0] = 0 \tag{3.5.111}$$

where m is a finite-dimensional vector of moment equations for the true value θ_0 of θ locally at a fixed value x_0 of the conditioning variable but not necessarily uniformly. Such instance arises, for example, when θ is the value of a functional evaluated at x_0. More generally, local estimating equations are characterized by equations defined locally at a point. The framework of local estimating equations is first introduced by Carroll et al. (1998) and then extend to allow for non-smooth criterion functions and the presence of nuisance parameters by Xu (2020). Lewbel (2007) considers inference based on local moment conditions in the generalized method of moments framework. Gagliardini et al. (2011) propose the extended method of moments, which accommodates both global and local moment restrictions. When X has no mass at x_0, inference on θ based on local estimating equations is typically carried out by approximating (3.5.111) by kernel smoothing. For example, given a set of observations $\{z_i\}_{i=1}^n := \{(y_i, x_i)\}_{i=1}^n$, (3.5.111) implies an estimating equation

$$D_n(\theta) = \frac{1}{nh_n} \sum_{i=1}^{n} h(\theta_0; z_i) K\left(\frac{x_i - x_0}{h_n}\right) \tag{3.5.112}$$

where K is a kernel function and $h_n \to 0$ as $n \to \infty$. The convergence rate of the estimating equation (3.5.112) depends on the bandwidth parameter h_n and is slower than $n^{1/2}$. It is known that the convergence rate of an estimator of θ_0 based on the estimating equation (3.5.112) is slower than $n^{1/2}$ and each element of such estimator may converge to a non-degenerate distribution at a different rate as demonstrated in the applications in Section 3.5.2-3.5.4 (See also Fan and Gijbels (1996)). Furthermore, the coefficient $J(\theta_0)$ of the linear expansion of $D_n(\theta)$ around $\theta = \theta_0$, in (3.3.29) may not have full rank without appropriately being scaled by the diagonal matrix $\mathbf{n}^{-\beta}$ as shown in succeeding examples. For hypothesis testing in the local estimating function framework, Calonico et al. (2014) consider a Wald-type test and the empirical likelihood Lagrange multiplier test is proposed by Xu (2020). Calonico et al. (2014) considers construction of a confidence set in the regression discontinuity framework.

The generalized $C(\alpha)$ test has the following advantages over these methods: (1) More general test restrictions are allowed. The Wald-type test only allows for restrictions on pa-

rameters that are estimated at the same rate. The Lagrange multiplier test only considers a class of null hypotheses where the parameter of interest takes some fixed hypothetical value. (2) Elements of an restricted estimator can be based on different bandwidths. The alternative method requires the same bandwidth be used for estimation of the parameter vector. In addition, the test statistic may employ a bandwidth of smaller order than those for the restricted estimator and such a choice improves the convergence rate of the test statistic. (3) Only an restricted estimator needs to be estimated while they require an unrestricted estimator. An restricted estimator is often easier to estimate as in the applications below. Furthermore, the asymptotic distribution of the estimator does not need to exist or be known. To illustrate these points, for $\theta \in (\theta_1, \theta_2) \in \mathbb{R}^2$, consider a problem of testing the hypothesis:

$$H_0 : \theta_1 = \theta_2. \tag{3.5.113}$$

Then, a Wald-type statistic takes the form:

$$W_n = \frac{\tilde{\theta}_{1,n} - \tilde{\theta}_{2,n}}{\sqrt{V_n^*}} \tag{3.5.114}$$

where $(\tilde{\theta}_{1,n}, \tilde{\theta}_{2,n})$ are unrestricted estimators of (θ_1, θ_2) and V_n^* is a consistent estimator of the "asymptotic variance" of $\tilde{\theta}_{1,n} - \tilde{\theta}_{2,n}$. Such a test in this context is considered in Calonico et al. (2014) while their statistic includes an additional term for bias correction. Whether such a term is included or not is not essential for the succeeding discussions. Assuming that both $(\tilde{\theta}_{1,n}, \tilde{\theta}_{2,n})$ are asymptotically normal, it is necessary that the two estimators converge at the same rate for the test to have either a correct size or any power. To see this, suppose for some $r_i > 0$

$$\mathrm{Var}(n^{r_i}(\tilde{\theta}_{i,n} - \theta_i)) = \sigma_i > 0, \quad i = 1, 2 \tag{3.5.115}$$

and $(\tilde{\theta}_{1,n}, \tilde{\theta}_{2,n})$ are asymptotically independent and let $V_n^* = \dfrac{\sigma_1^2}{n^{2r_1}} + \dfrac{\sigma_2^2}{n^{2r_2}}$. Then, if $r_1 < r_2$

$$W_n = \frac{n^{r_1}\tilde{\theta}_{1,n} - n^{(r_1 - r_2)} n^{r_2} \tilde{\theta}_{2,n}}{\sqrt{\sigma_1^2 + \sigma_2^2 n^{2(r_1 - r_2)}}} \tag{3.5.116}$$

$$= \frac{n^{r_1}(\tilde{\theta}_{1,n} - \theta_1) + n^{r_1}\theta_1}{\sigma_1} + o_P(1) \tag{3.5.117}$$

so that it diverges unless $\theta_1 = 0$. It is easy to see that under any choice of V_n^*, W_n either converges to zero or diverges regardless of whether H_0 in (3.5.113) holds. The same conclusion holds when $(\tilde{\theta}_{1,n}, \tilde{\theta}_{2,n})$ are not independent.

The fact that we need $r_1 = r_2$ has several implications. First, for estimation of (θ_1, θ_2), the same bandwidth must be used. This can lead to significant estimation bias than the case where different bandwidths are used for estimators of θ_1 and θ_2. In the regression discontinuity design setup in Section 3.5.3, the parameters of interest correspond to the values of two different functions at some fixed point. For each parameter, the optimal choice of the bandwidth (which minimizes the mean squared error of the estimator) depends on some higher order derivative of the underlying function. Thus, the optimal values of the bandwidth can be quite different. In addition, each needs to be estimated from a different set of data points based on partitions of the sample according to the value of some covariate variable. The sizes of the two sub-samples can be disproportionate in practice, which makes the use of the same bandwidth even less desirable. In addition, parameter estimators may not converge at the same rate even when the same bandwidth is used. In particular, derivatives of different orders cannot be estimated at the same rate as we observe in Section 3.5.2. Thus, test restrictions considered in Section 3.5.2 and 3.5.4 may not be considered for the Wald-type test. In order to apply the Lagrange multiplier test, the hypothetical value of $\theta_1 = \theta_2$ needs to be fixed, i.e. it only allows for simple hypotheses. Since the generalized $C(\alpha)$ test requires only a restricted estimator, only one of (θ_1, θ_2) needs to be estimated. Thus, the concern posed for the test based of the form (3.5.113) regarding the choice of the bandwidth is not present. Second, a different bandwidth may be used for each parameter. In fact, the test statistic can also employ a bandwidth such that it converges faster than those used for the restricted estimator. This improves estimation stability and accuracy. Lastly, since it allows for the convergence rates of the estimators of θ_1 and θ_2 to differ, a larger class of test restrictions can be accommodated. As we show in 3.3, restrictions defined by any twice-differentiable function of the parameter can be tested.

In what follows, we consider application of the generalized $C(\alpha)$ test procedure to the problem of testing the derivatives of the conditional expectation function in Section 3.5.2, that of constructing a confidence set for the average treatment effect in the regression dis-

continuity design in Section 3.5.3, and a specification test for the semiparametric stochastic discount factor model. We assume that K is a symmetric probability density function with bounded support on $\mathbb{R}$ for brevity. Define

$$\mu_l = \int z^l K(z)dz, \; \nu_l = \int z^l K^2(z)dz, \qquad l = 0, 1, \ldots. \tag{3.5.118}$$

Note that $\mu_0 = 1$ and $\mu_l = \nu_l = 0$, l even. We note that in all applications, we set the rate of the bandwidths for the restricted estimator to be faster than the MSE optimal rate in order to obtain asymptotic unbiasedness (undersmoothing). Alternatively, an explicit bias-correction approach, such as in Calonico et al. (2014), may be employed.

3.5.2. Testing the derivatives of the conditional expectation function

The problem considered here involves joint hypothesis testing on derivatives of different orders of a function of interest. As shown in Fan and Gijbels (1996), local polynomial estimators of derivatives of different orders converge to non-degenerate limits at different rates. The problem of testing multiple restrictions on parameters under multiple convergence rates of the parameter estimator has not been considered in the literature. We show the generalized $C(\alpha)$ test is applicable to such problem.

For a pair of integrable random variables (Y,X), consider a nonparametric regression of Y of X:

$$Y = g(X) + \varepsilon. \qquad \mathbb{E}[\varepsilon \,|\, X] = 0 \tag{3.5.119}$$

where the conditional expectation function $g(x) = \mathbb{E}[Y : | \, X = x]$ is infinitely continuously differentiable. Denote by $g^{(m)}(x_0)$ the m-th derivative of $g(x)$ evaluated at $x = x_0$. Assume $\sigma^2(x) = \text{Var}(\varepsilon \mid X = x)$ and the density $f(x)$ of X are continuous and bounded away from zero in the neighborhood of x_0. We are interested in testing whether

$$g^{(m)}(x_0) = 0, \qquad \forall m \geq m_0 \tag{3.5.120}$$

for for some positive integer m_0, i.e. all the derivatives of higher order than m_0 are all zero. In particular, when $m_0 = 1$, the marginal effect of x on $m(x)$ at $x = x_0$ is zero under the null hypothesis. The case where $m_0 = 2$ corresponds to a testing on local linearity of m at $x = x_0$. For a positive integer M such that $M \geq m_0$, a Taylor expansion of $g(x)$ of order M

in the neighborhood of x_0 gives

$$g(x) \approx \sum_{s=0}^{M} \frac{1}{s!} \theta_0^{(s)} (x - x_0)^s. \tag{3.5.121}$$

where

$$\begin{aligned} \theta_0 &= \left(\theta_0^{(0)}, \ldots, \theta_0^{(s)}, \ldots, \theta_0^{(M)}\right)' \\ &= \left(g(x_0), \ldots, \frac{1}{s!} g^{(s)}(x_0), \ldots, \frac{1}{M!} g^{(M)}(x_0)\right). \end{aligned} \tag{3.5.122}$$

Then, we have the local moment conditions (3.5.111) where

$$m(\theta_0; Z) = Y - \sum_{s=0}^{M} \theta_0^{(s)} (X - x_0)^s \tag{3.5.123}$$

where $Z = (Y, X)'$. We consider the hypothesis:

$$H_0(m_0) : \theta^{(m)} = 0, \qquad \forall m \geq m_0. \tag{3.5.124}$$

Define a function $\psi : \mathbb{R}^M \to \mathbb{R}^{M - m_0 + 1}$:

$$\psi(\theta) = \left[\theta^{(m)}, \theta^{(m+1)}, \ldots, \theta^{(M)}\right]' \tag{3.5.125}$$

so that the restriction in (3.5.124) is equivalent to $\psi(\theta) = 0$. Given a set of i.i.d. observations $\{x_i, y_i\}_{i=1}^n$ and the local moment equations in (3.5.123), we consider $(M+1) \times 1$ estimating equations $D_n(\theta)$:

$$D_n(\theta) = \begin{bmatrix} \frac{1}{nh_n} \sum_{i=1}^n (y_i - \sum_{s=0}^M \theta^{(s)} (x_i - x_0)^s) K\left(\frac{x_i - x_0}{h_n}\right) \\ \frac{1}{nh_n^2} \sum_{i=1}^n (y_i - \sum_{s=0}^M \theta^{(s)} (x_i - x_0)^s)(x_i - x_0) K\left(\frac{x_i - x_0}{h_n}\right) \\ \vdots \\ \frac{1}{nh_n^{M+1}} \sum_{i=1}^n (y_i - \sum_{s=0}^M \theta^{(s)} (x_i - x_0)^s)(x_i - x_0)^M K\left(\frac{x_i - x_0}{h_n}\right) \end{bmatrix}. \tag{3.5.126}$$

Note that $D_n(\theta)$ can be also interpreted as the vector of first order conditions of the objec-

tive function of the local polynomial regression of order M, up to scaling, which solves

$$S_n^{(M)}(\theta) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=0}^{M} \theta^{(s)}(x_i - x_0)^s \right\}^2 K\left(\frac{x_i - x_0}{h_n}\right) \tag{3.5.127}$$

over $\theta = \left(\theta^{(0)}, \ldots, \theta^{(M)}\right)'$.

Observe that while an unrestricted estimator of θ_0 is obtained by minimizing (3.5.127), an (unconstrained) estimator with a faster rate of convergence can be attained by considering local polynomial regression of order $m_0 - 1$ where

$$\hat{\delta}_n = \left(\hat{\delta}_n^{(0)}, \ldots, \delta_n^{(m_0-1)}\right)' \tag{3.5.128}$$

of a $m_0 \times 1$ parameter δ is obtained by minimizing

$$S_n^0(\beta) = \sum_{i=1}^{n} (y_i - \sum_{s=0}^{m_0-1} \delta^{(s)}(x_i - x_0)^s)^2 K\left(\frac{x_i - x_0}{h_n^*}\right) \tag{3.5.129}$$

and then defining a constrained estimator $\tilde{\theta}_n = \left(\tilde{\theta}_n^{(0)}, \ldots, \tilde{\theta}_n^{(M)}\right)'$ by

$$\tilde{\theta}_n^{(s)} = \begin{cases} \hat{\delta}_n^{(s)} & , s = 0, \ldots, m_0 - 1 \\ 0 & , s = m_0, \ldots, M \end{cases} \tag{3.5.130}$$

Note that the bandwidth parameter h_n^* can be set differently from h_n. By Theorem 3.1 of Fan and Gijbels (1996), if $h_n^* = o(n^{-1/(2m_0+3)})$,

$$\sqrt{n(h_n^*)^{2s+1}}(\tilde{\theta}_n^{(s)} - \theta_0^{(s)}) = o_p(1), \qquad s = 0, \ldots, m_0 - 1 \tag{3.5.131}$$

so that

$$\sqrt{n(h_n^*)^{2m_0-1}}(\tilde{\theta}_n - \theta_0) = O_p(n^{-2/(2m_0+3)}) \tag{3.5.132}$$

or equivalently

$$\|\tilde{\theta}_n - \theta_0\| = O_p(n^{-1/(2m_0+3)}). \tag{3.5.133}$$

Note that the unrestricted estimator based on (3.5.127) converges at the slower rate

$n^{-2/(2M_0+5)}$. This is one advantage of using an restricted estimator in this context. Furthermore, as we see below, the bandwidth h_n for (3.5.126) may converge faster than h_n^* as long as the condition specified as (3.5.152) holds and such choice leads to faster convergence of the test statistic.

By the central limit theorem,

$$\sqrt{nh_n}D_n(\theta_0) \xrightarrow[n\to\infty]{L} \mathrm{N}[0, I(\theta_0)] \tag{3.5.134}$$

where $I(\theta_0)$ is a positive-definite matrix of size $M+1$ defined as

$$I(\theta_0) = \sigma^2(x_0)f(x_0)[v_{i+j-2}]_{1\le i,j\le M+1}. \tag{3.5.135}$$

Now, $D_n(\theta)$ can be expressed as

$$D_n(\theta) = D_n(\theta_0) + J_{\theta,n}(\theta_0)\mathbf{n}^{-\beta}(\theta-\theta_0) \tag{3.5.136}$$

where $J_{\theta,n}(\theta)$ is a symmetric matrix of size $M+1$ defined as

$$J_{\theta,n}(\theta) = \left[\frac{1}{nh_n^i}\sum_{i=1}^{n} K\left(\frac{x_i-x_0}{h_n}\right)(x_i-x_0)^{i+j-2}\right]_{1\le i,j\le M+1}. \tag{3.5.137}$$

and β is a $(M+1)$ dimensional vector satisfying

$$\mathbf{n}^{\beta} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & h_n & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & h_n^M \end{bmatrix}. \tag{3.5.138}$$

Then,

$$\|J_{\theta,n}(\theta_0) - J_{\theta}(\theta_0)\| = o_p((nh)^{-1/2}) \tag{3.5.139}$$

where

$$J_\theta(\theta_0) = f(x_0) \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_M \\ \mu_1 & \mu_2 & \cdots & \mu_{M+1} \\ \vdots & & \ddots & \\ \mu_M & \cdots & \mu_{2M-1} & \mu_{2M} \end{bmatrix}. \tag{3.5.140}$$

Hence, we can write

$$D_n(\theta) = D_n(\theta_0) + J_\theta(\theta_0)\mathbf{n}^{-\beta}(\theta - \theta_0) + o_p\left((nh_n)^{-1/2}\right). \tag{3.5.141}$$

Consider estimators of $(\tilde{J}_{\theta,n}, \tilde{I}_{\theta,n})$ of $(J_\theta(\theta_0), I(\theta_0))$ given by

$$\tilde{J}_{\theta,n} = \hat{f}(x_0) \begin{bmatrix} 1 & \mu_1 & \cdots & \mu_M \\ \mu_1 & \mu_2 & \cdots & \mu_{M+1} \\ \vdots & & \ddots & \\ \mu_M & \cdots & \mu_{2M-1} & \mu_{2M} \end{bmatrix}, \tag{3.5.142}$$

and

$$\tilde{I}_n = \hat{\sigma}^2(x_0)\hat{f}(x_0)[v_{i+j-2}]_{1\le i,j\le M+1}. \tag{3.5.143}$$

where $(\hat{f}(x_0), \hat{\sigma}^2(x_0))$ are consistent estimators of $(f(x_0), \sigma^2(x_0))$. For example, $\hat{f}(x_0)$ may be a usual kernel density estimator

$$\hat{f}(x_0) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n^{(f)}}\right) \tag{3.5.144}$$

Then, by choosing $h_n^{(f)}$ so that $h_n^{(f)} = o(n^{-1/5})$, we have

$$\hat{f}(x_0) - f(x_0) = O_p(n^{-2/5}). \tag{3.5.145}$$

Similarly, define $\hat{\sigma}^2(x_0)$ by

$$\hat{\sigma}^2(x_0) = \sum_{i=1}^{n} \frac{K\left(\frac{x_i - x_0}{h_n^{(\sigma)}}\right) e_i^2}{K\left(\frac{x_i - x_0}{h_n^{(\sigma)}}\right)} \tag{3.5.146}$$

where $e_i = y_i - \tilde{\theta}_n^{(0)}, i = 1, \ldots, n$. Then, if $h_n^{(\sigma)} = o(n^{-1/5})$, we have

$$\hat{\sigma}^2(x_0) - \sigma^2(x_0) = O_p(n^{-2/5}). \tag{3.5.147}$$

Thus,

$$\|\tilde{J}_{\theta,n} - J_\theta(\theta_0)\| = O_p(n^{-2/5}),\ \|\tilde{I}_n - I(\theta_0)\| = O_p(n^{-2/5}). \tag{3.5.148}$$

Finally, the derivative $\tilde{P}_n$ of $\psi(\theta)$ evaluated at $\tilde{\theta}_n$ is given by

$$\tilde{P}_n = [\mathbf{0}_{(M-m_0+1)\times m_0}; \mathbf{1}_{(M-m_0+1)\times(M-m_0+1)}] \tag{3.5.149}$$

where $\mathbf{0}_{(M-m_0+1)\times m_0}$is a $(M - m_0 + 1) \times m_0$ zero matrix and $\mathbf{1}_{(M-m_0+1)\times(M-m_0+1)}$ is an identity matrix of size $(M - m_0 + 1)$. Then, setting $W_n = \tilde{I}_n^{-1}$, the generalized $C(\alpha)$ statistic given in (3.3.46) can be expressed as

$$PC(\tilde{\theta}_n; \psi) = nh \frac{1}{\hat{f}(x_0)\hat{\sigma}^2(x_0)} D_n(\tilde{\theta}_n)' \mathscr{U}^{-1} \tilde{P}_n' (\tilde{P}_n \mathscr{U}^{-1} \mathscr{V} \mathscr{U}^{-1} \tilde{P}_n') \tilde{P}_n \mathscr{U}^{-1} D_n(\tilde{\theta}_n). \tag{3.5.150}$$

where $\mathscr{U} = [\mu_{i+j-2}]_{1\le i,j\le M+1}$ and $\mathscr{V} = [\nu_{i+j-2}]_{1\le i,j\le M+1}$. In particular, if $M = m_0 = 1$, we have

$$PC(\tilde{\theta}_n; \psi) = (nh_n) \frac{\nu_2}{\hat{f}(x_0)\hat{\sigma}^2(x_0)} \left(\frac{1}{nh_n^2} \sum_{i=1}^{n} (y_i - \sum_{s=0}^{1} \tilde{\theta}_n^{(s)} (x_i - x_0)^s)(x_i - x_0) K\left(\frac{x_i - x_0}{h_n}\right) \right)^2 \tag{3.5.151}$$

See Appendix: 3.A for derivations. In light of Assumption 3.3.11, we can choose the order

of h_n by finding a smallest constant b satisfying

$$\frac{1}{2m_0+3} > \frac{1}{2}\left(\frac{1}{2}-b\right), \qquad \frac{1}{2m_0+3}+\frac{2}{5} > \left(\frac{1}{2}-b\right) \tag{3.5.152}$$

and choose h_n so that $h_n = O_p(n^{-b})$. Then, by Theorem 3.3.1,

$$PC(\tilde{\theta}_T; \psi) \xrightarrow[n\to\infty]{L} \chi^2(M-m_0+1). \tag{3.5.153}$$

Note that this test procedure only requires estimation of the derivatives of the order up to m_0-1. On the other hand, other methods, such as the Wald-type test, requires an unrestricted estimator so that all M derivatives need to be estimated. Such an estimator suffers from a slower convergence than our restricted estimator in (3.5.130). In addition, we do not require that the bandwidth h_n^* for the restricted estimator and h_n for the test statistic to be the same. Then, the order of h_n can be smaller than that of h_n^* as long as (3.5.152) holds to improve the convergence rate of the test statistic.

3.5.3. Regression discontinuity design

We consider the problem of constructing a confidence set for the average treatment effect in the regression discontinuity design. Such a problem is considered by Calonico et al. (2014), who employs a Wald-type test. As discussed below, our test procedure avoids a restrictive choice of the smoothing bandwidth required by their Wald-type test. It is particularly advantageous when an unrestricted estimator is estimated from two sets of observations with significantly different sample sizes. Furthermore, the test statistic converges at a faster rate.

Consider the standard sharp regression discontinuity design setup under a randomized experiment (Lee (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010)) where Y is the outcome variable, X is a univariate exogenous variable, and D is a binary variable for treatment which is equal to one if and only if $X \geq c$ for some fixed constant c. Adapting the potential outcomes framework, let Y_1 and Y_0 be the outcomes under/without treatment so that

$$Y = DY_1 + (1-D)Y_2. \tag{3.5.154}$$

Given the i.i.d. data $\{y_i, x_i, D_i\}_{i=1}^n$, we are interested in testing whether the average treat-

ment effect of the kind:

$$\mathbb{E}[Y_1 - Y_0 \,|\, D = 1] \tag{3.5.155}$$

take some fixed value $d_0 \in \mathbb{R}$ Note that under continuity of the regression functions $\mathbb{E}[Y_1 \,|\, X = x]$ and $\mathbb{E}[Y_0 \,|\, X = x]$, the hypothesis is equivalent to

$$H_0(d_0) \,:\, \tau_0^{(+)} - \tau_0^{(-)} = d_0. \tag{3.5.156}$$

where

$$\tau_0^{(+)} = \mathbb{E}[Y_1 \,|\, X = c],\ \tau_0^{(-)} = \mathbb{E}[Y_0 \,|\, X = c] \tag{3.5.157}$$

Note that one can obtain a confidence set for (3.5.155) by inverting a family of hypotheses tests of $\{H_0(d_0)\}_{d_0 \in \mathbb{R}}$. Denote

$$\beta_0^{(+)} = \partial_x \mathbb{E}[Y_1 \,|\, X = x]|_{x=c}. \tag{3.5.158}$$

Then, we have the local moment conditions (3.5.111) where

$$m(\theta_0; Z) = \begin{bmatrix} \mathbf{1}\{X \geq c\}\,(Y - \tau_0^{(+)} - \beta_0^{(+)}(X - c)) \\ \mathbf{1}\{X < c\}\,(Y - \tau_0^{(-)}) \end{bmatrix} \tag{3.5.159}$$

where $Z = (Y, X)$ and $\theta = \left(\tau^{(+)}, \beta^{(+)}, \tau^{(-)}\right)'$.

Given an i.i.d. set $\{(y_i, x_i)\}_{i=1}^n$, we consider the estimating function $D_n^{(Y)}(\theta)$ based on local linear regression:

$$D_n(\theta) := \begin{bmatrix} D_{1,n}(\theta) \\ D_{2,n}(\theta) \\ D_{3,n}(\theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{nh_n} \sum_{i=1}^n \mathbf{1}\{x_i \geq c\}(y_i - \tau^{(+)} - \beta^{(+)}(x_i - c))K\left(\frac{x_i - c}{h_n}\right) \\ \frac{1}{nh_n^2} \sum_i \mathbf{1}\{x_i \geq c\}(y_i - \tau^{(+)} - \beta^{(+)}(x_i - c))(x_i - c)K\left(\frac{x_i - c}{h_n}\right) \\ \frac{1}{nh_n} \sum_{i=1}^n \mathbf{1}\{x_i < c\}(y_i - \tau^{(-)})K\left(\frac{x_i - c}{h_n}\right) \end{bmatrix} \tag{3.5.160}$$

More generally, $D_n^{(Y)}(\theta)$ can be constructed based on a local polynomial regression as in Section 3.5.2.

We consider a restricted estimator $\tilde{\theta}_n$ as follows: first, obtain an unrestricted estimator $(\hat{\tau}_n^{(+)}, \hat{\beta}_n^{(+)})$ of $(\tau^{(+)}, \beta^{(+)})$ by minimizing, over $(\tau^{(+)}, \beta^{(+)})$,

$$S_n^+(\tau^{(+)}, \beta^{(+)}) = \sum_{i=1}^{n} \mathbf{1}\{x_i \geq c\}(y_i - \tau^{(+)} - \beta^{(+)}(x_i - c))^2 K\left(\frac{x_i - c}{h_n^*}\right) \tag{3.5.161}$$

and then set

$$\tilde{\theta}_n := \left(\tilde{\tau}_n^{(-)}, \tilde{\beta}_n^{(-)}, \tilde{\tau}_n^{(-)}\right) = \left(\hat{\tau}_n^{(+)}, \hat{\beta}_n^{(+)}, \tilde{\tau}_n^{(+)} + d_0\right)'. \tag{3.5.162}$$

Notice that a test procedure based on an unrestricted estimator requires estimation of both $\tau^{(+)}$ and $\tau^{(-)}$. Consider an unrestricted estimator $(\hat{\tau}_n^{(-)}, \hat{\beta}_n^{(-)})$ of $(\tau^{(-)}, \beta^{(-)})$ which minimizes

$$S_n^-(\tau^{(-)}, \beta^{(-)}) = \sum_{i=1}^{n} \mathbf{1}\{x_i < c\}(y_i - \tau^{(-)} - \beta^{(-)}(x_i - c))^2 K\left(\frac{x_i - c}{h_n^{**}}\right). \tag{3.5.163}$$

Then, as discussed in Section 3.5.1, a Wald-type test requires that the bandwidths in 132 and (3.5.163) be equal: $h_n^* = h_n^{**}$. However, the data points effectively used to estimate $(\hat{\tau}_n^{(-)}, \hat{\beta}_n^{(-)})$ and $(\tau^{(-)}, \beta^{(-)})$ are not equal: $\sum_{i=1}^n \mathbf{1}\{x_i \geq c\} \approx pn$ and $\sum_{i=1}^n \mathbf{1}\{x_i < c\} \approx (1-p)n$ where $p := \mathbb{P}(X \geq c)$, respectively. In particular, if p is close to 0 or 1, the relative sample size is quite disproportionate and thus using the same bandwidth for the two subsamples leads to significant estimation inefficiency. On the other hand, we only need to estimate (3.5.161) and thus no such concern is posed. In addition, as in Section 3.5.2, h_n for (3.5.160) can be set differently from h_n^*, which leads to a faster convergence of the test statistic. Further, note that in order to apply the Lagrange multiplier test, the hypothetical value of $\theta_1 = \theta_2$ should be fixed under the null hypothesis and thus (3.5.156) may not be directly tested.

By Theorem 3.1 of Fan and Gijbels (1996), one can choose h_n^* so that

$$\|\tilde{\theta}_n - \theta_0\| = O_p(n^{-2/7}) \tag{3.5.164}$$

Define a function $\psi : \mathbb{R}^3 \to \mathbb{R}$:

$$\psi(\theta) = \tau^{(+)} - \tau^{(-)} - d_0 \tag{3.5.165}$$

so that the restriction in (3.5.156) is equivalent to $\psi(\theta) = 0$ and

$$\tilde{P}_n = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \tag{3.5.166}$$

Given the matrix $\mathbf{n}^\beta$ specified as

$$\mathbf{n}^\beta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & h_n & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.5.167}$$

we have

$$J_\theta(\theta_0) = f(x_0) \begin{bmatrix} p & 0 & 0 \\ 0 & p\mu_2 & 0 \\ 0 & 0 & (1-p) \end{bmatrix}. \tag{3.5.168}$$

and

$$I_\theta(\theta_0) = f(x_0) \begin{bmatrix} p^2\sigma_+^2\nu_0 & 0 & 0 \\ 0 & p^2\sigma_+^2\nu_2 & 0 \\ 0 & 0 & (1-p)^2\sigma_-^2\nu_0 \end{bmatrix} \tag{3.5.169}$$

where $\sigma_+^2 = \lim_{x\downarrow c} \mathrm{Var}\,(Y - \mathbb{E}[Y \mid X] \mid X = x)$, $\sigma_-^2 = \lim_{x\uparrow c} \mathrm{Var}\,(Y - \mathbb{E}[Y \mid X] \mid X = x)$. We consider estimators of $(p, \sigma_+^2, \sigma_-^2)$ given by

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \geq c\} \tag{3.5.170}$$

$$\hat{\sigma}_+^2 = \sum_{i=1}^{n} \frac{\mathbf{1}\{x_i \geq c\} K\left(\dfrac{x_i - x_0}{h_n^{(\sigma_+)}}\right) e_i^2}{\mathbf{1}\{x_i \geq c\} K\left(\dfrac{x_i - x_0}{h_n^{(\sigma_+)}}\right)}, \; \hat{\sigma}_-^2 = \sum_{i=1}^{n} \frac{\mathbf{1}\{x_i < c\} K\left(\dfrac{x_i - x_0}{h_n^{(\sigma_-)}}\right) e_i^2}{\mathbf{1}\{x_i < c\} K\left(\dfrac{x_i - x_0}{h_n^{(\sigma_-)}}\right)} \tag{3.5.171}$$

where

$$e_i = \begin{cases} y_i - \tilde{\tau}_n^{(+)} & \text{if } x_i \geq c \\ y_i - (\tilde{\tau}_n^{(+)} + d_0) & \text{if } x_i < c. \end{cases} \tag{3.5.172}$$

and define $\hat{f}(x_0)$ as in (3.5.144). Given $(\hat{p}, \hat{\sigma}_+^2, \hat{\sigma}_-^2, \hat{f}(x_0))$, consider estimator of $(J_\theta(\theta_0), I_\theta(\theta_0))$:

$$\tilde{J}_{\theta,n} = \hat{f}(x_0) \begin{bmatrix} \hat{p} & 0 & 0 \\ 0 & \hat{p}\mu_2 & 0 \\ 0 & 0 & (1-\hat{p}) \end{bmatrix}. \tag{3.5.173}$$

and

$$\tilde{I}_n = \hat{f}(x_0) \begin{bmatrix} \hat{p}^2\hat{\sigma}_+^2\nu_0 & 0 & 0 \\ 0 & \hat{p}^2\hat{\sigma}_+^2\nu_2 & 0 \\ 0 & 0 & (1-\hat{p})^2\hat{\sigma}_-^2\nu_0 \end{bmatrix} \tag{3.5.174}$$

Then, if $(h_n^{(\sigma_+)}, h_n^{(\sigma_-)}, h_n^{(f)})$ are all of order $o(n^{-1/5})$, we have

$$\|\tilde{J}_{\theta,n} - J_\theta(\theta_0)\| = O_p(n^{-2/5}),\ \|\tilde{I}_{\theta,n} - I(\theta_0)\| = O_p(n^{-2/5}). \tag{3.5.175}$$

Then, setting $W_n = \tilde{I}_n^{-1}$, we obtain the generalized $C(\alpha)$ statistic given by

$$PC(\tilde{\theta}_n\,;\psi) = \frac{nh_n}{\hat{f}(x_0)(\hat{\sigma}_+^2+\hat{\sigma}_-^2)\nu_0}\left[\hat{p}^{-1}D_{1,n}(\theta) - (1-\hat{p})^{-1}D_{3,n}(\theta)\right]^2. \tag{3.5.176}$$

The derivation is given in Appendix 3.A. It $h_n = O_p(n^{-b})$ for some b such that $b > \dfrac{1}{14}$ then Assumption 3.3.11 holds and

$$PC(\tilde{\theta}_T;\,\psi) \xrightarrow[n\to\infty]{L} \chi^2(2) \tag{3.5.177}$$

under H_0 in (3.5.182).

3.5.4. Semiparametric stochastic discount factor

Cai, Ren and Sun (2015) consider the nonlinear pricing kernel of the form

$$m_{t+1} = 1 - m(X_t)r_{p,t+1} \tag{3.5.178}$$

where $m(\cdot)$ is specified nonparametrically. Then, it satisfies

$$\mathbb{E}[\{1 - m(X_t)r_{p,t+1}\}\, r_{i,t+1} \mid \Omega_t] = 0 \tag{3.5.179}$$

where $r_{t,i+1}$ is the i-th excess return of the risky assets ($i = 1,\ldots,N$), $r_{p,t+1}$ is the return on the market portfolio in excess of the risk-free asset at time t, Ω_t represents the information set at period t and X_t is an L-dimensional conditioning variables from Ω_t. See Wang (2003) for more details,. Then, we have

$$\mathbb{E}\left[\left\{1-\left(m(x_0)+(\partial m(x_0))'(X_t-x_0)r_{i,t+1}\right)r_{p,t+1}\right\}\middle|X_t=x_0\right]=0, \qquad \forall i=1,\ldots,N \tag{3.5.180}$$

where $\alpha_0 = m(x_0), \beta_0 = \partial m(x_0)$. Note that (3.5.180) must hold for any $i = 1,\ldots,N$. We consider a specification test for the model (3.5.178) in the following manner: For any two assets, which we denote by $i = 1,2$, suppose $\theta_0 = (\alpha_0, \beta_0, a_0, b_0)$ satisfies the local moment conditions (3.5.111) where

$$m(\theta_0; Z_t) = \begin{bmatrix} \left\{1-\left(\alpha_0+\beta_0'(X_t-x_0)\right)r_{p,t+1}\right\}r_{1,t+1} \\ \left\{1-\left(a_0+b_0'(X_t-x_0)\right)r_{p,t+1}\right\}r_{2,t+1} \end{bmatrix} \tag{3.5.181}$$

where $Z_t = (X_t, r_{p_{t+1}}, r_{1,t+1}, r_{2,t+1})$. Then, we consider

$$H_0 : \alpha_0 = a_0, \beta_0 = b_0. \tag{3.5.182}$$

Given a strictly stationary and α-mixing process $\{(X_t, r_{1,t+1}, r_{2,t+1}, r_{p,t+1})\}_{t=1}^T$, we employ orthogonality conditions considered in Cai et al. (2015) to form the estimating function $D_n(\theta)$ defined as

$$D_n(\theta) =:= \begin{bmatrix} D_{1,n}(\theta) \\ D_{2,n}(\theta) \\ D_{3,n}(\theta) \\ D_{4,n}(\theta) \end{bmatrix} \begin{bmatrix} \frac{1}{Th_T}\sum_{t=1}^T\left[\left\{1-\left(\alpha+\beta'(X_t-x_0)\right)r_{p,t+1}\right\}r_{1,t+1}\right]K\left(\frac{X_t-x_0}{h_T}\right) \\ \frac{1}{Th_T^2}\sum_{t=1}^T\left[\left\{1-\left(\alpha+\beta'(X_t-x_0)\right)r_{p,t+1}\right\}r_{1,t+1}(X_t-x_0)\right]K\left(\frac{X_t-x_0}{h_T}\right) \\ \frac{1}{Th_T}\sum_{t=1}^T\left[\left\{1-(a+b'(X_t-x_0))\,r_{p,t+1}\right\}r_{2,t+1}\right]K\left(\frac{X_t-x_0}{h_T}\right) \\ \frac{1}{Th_T^2}\sum_{t=1}^T\left[\left\{1-(a+b'(X_t-x_0))\,r_{p,t+1}\right\}r_{2,t+1}(X_t-x_0)\right]K\left(\frac{X_t-x_0}{h_T}\right) \end{bmatrix} \tag{3.5.183}$$

where $\theta = (\alpha, \beta, a, b)'$.

In what follows, we assume that $L = 1$, i.e. X_t is one-dimensional to simplify exposi-

tion. As in Section 3.5.156, only either (α_0, β_0) or (a_0, b_0) needs to be estimated to build a restricted estimator. Furthermore, observe that the estimators of α and β converge at different rates. Thus, the Wald-type test is not applicable to the test restriction (3.5.182), which involves jointly testing α and β. The Lagrange-multiplier test only allows for simple hypotheses and hence in order to test (3.5.182), one needs to perform a family of tests with all hypothetical values of $\alpha_0(=a_0)$ and $\beta_0(=b_0)$. Such a procedure is significantly more computationally intensive than the generalized $C(\alpha)$ test procedure.

We obtain a restricted estimator $\tilde{\theta}_n$ by first constructing an unrestricted estimator $\left(\tilde{\alpha}_n, \tilde{\beta}_n\right)$ of (α_0, β_0) which solves

$$\frac{1}{Th_T}\sum_{t=1}^{T}\left[\left\{1-\left(\alpha+\beta'(X_t-x_0)\right)r_{p,t+1}\right\}r_{1,t+1}\begin{pmatrix}1\\ X_t-x_0\end{pmatrix}\right]K\left(\frac{X_t-x_0}{h_T^*}\right)=0 \quad (3.5.184)$$

and then define

$$\tilde{\theta}_n=\left(\tilde{\alpha}_T,\tilde{\beta}_T,\tilde{\alpha}_T,\tilde{\beta}_T\right)'. \quad (3.5.185)$$

Then, as in Section 3.5.3 , one can choose h_T^* so that

$$\|\tilde{\theta}_T-\theta_0\|=O_p(T^{-2/7}) \quad (3.5.186)$$

Given the matrix $\mathbf{T}^{\beta}$ specified as

$$\mathbf{T}^{\beta}=\begin{bmatrix}1&0&0&0\\0&h_T&0&0\\0&0&1&0\\0&0&0&h_T\end{bmatrix}, \quad (3.5.187)$$

we have

$$J_\theta(\theta_0)=f(x_0)\begin{bmatrix}\lambda_1&0&0&0\\0&\lambda_1\mu_2&0&0\\0&0&\lambda_2&0\\0&0&0&\lambda_2\mu_2\end{bmatrix}, I_\theta(\theta_0)=f(x_0)\begin{bmatrix}\sigma_1^2\nu_0&0&0&0\\0&\sigma_1^2\nu_2&0&0\\0&0&\sigma_2^2\nu_0&0\\0&0&0&\sigma_2^2\nu_2\end{bmatrix} \quad (3.5.188)$$

where $\lambda_k = \mathbb{E}\left[r_{p,t+1}r_{i,t+1}\,|\,X_t = x_0\right]$, , $\sigma_k^2 = \mathrm{Var}\left[\left\{1 - m(X_t)r_{p,t+1}\right\}r_{k,t+1}\big|\,X_t = x_0\right]$. $(k = 1,2)$. We consider their estimators:

(3.5.189)

$$\tilde{J}_{\theta,T} = \hat{f}(x_0)\begin{bmatrix}\hat{\lambda}_1 & 0 & 0 & 0\\ 0 & \hat{\lambda}_1\mu_2 & 0 & 0\\ 0 & 0 & \hat{\lambda}_2 & 0\\ 0 & 0 & 0 & \hat{\lambda}_2\mu_2\end{bmatrix}, \tilde{I}_T = \hat{f}(x_0)\begin{bmatrix}\hat{\sigma}_1^2\nu_0 & 0 & 0 & 0\\ 0 & \hat{\sigma}_1^2\nu_2 & 0 & 0\\ 0 & 0 & \hat{\sigma}_2^2\nu_0 & 0\\ 0 & 0 & 0 & \hat{\sigma}_2^2\nu_2\end{bmatrix} \tag{3.5.190}$$

where $\hat{f}(x_0)$ is defined in (3.5.144) and $(\hat{\lambda}_k, \sigma_k^2)$ are consistent estimators of $(\lambda_k, \sigma_k^2), k = 1,2$, obtained by solving

$$(\hat{\lambda}_k, \hat{\zeta}_k) = \arg\min_{(\lambda,\zeta)}\left[\sum_{t=1}^{T}(r_{p,t+1}r_{k,t+1} - \lambda - \zeta(X_t - x_0))^2 K\left(\frac{X_t - x_0}{h_T^{(\lambda_k)}}\right)\right] \tag{3.5.191}$$

and

$$(\hat{\sigma}_k^2, \hat{\kappa}_k) = \arg\min_{(\sigma^2,\kappa)}\left[\sum_{t=1}^{T}((e_t^{(k)})^2 - \sigma^2 - \kappa(X_t - x_0))^2 K\left(\frac{X_t - x_0}{h_T^{(\sigma_k^2)}}\right)\right] \tag{3.5.192}$$

where

$$e_t^{(k)} = \left\{1 - \hat{m}(X_t)r_{p,t+1}\right\}r_{k,t+1}. \tag{3.5.193}$$

Then, if $(h_T^{(\lambda_1)}, h_T^{(\lambda_2)}, h_T^{(\sigma_1^2)}, h_T^{(\sigma_2^2)}, h_T^{(f)})$ are all of order $o(T^{-1/5})$, we have

$$\|\tilde{J}_{\theta,T} - J_\theta(\theta_0)\| = O_p(T^{-2/5}),\ \|\tilde{I}_{\theta,T} - I(\theta_0)\| = O_p(T^{-2/5}). \tag{3.5.194}$$

Finally, define a function $\psi : \mathbb{R}^4 \to \mathbb{R}^2$:

$$\psi(\theta) = \begin{bmatrix}\alpha - a\\ \beta - b\end{bmatrix} \tag{3.5.195}$$

so that

$$\tilde{P}_n = \begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}. \tag{3.5.196}$$

Then, setting $W_n = \tilde{I}_n^{-1}$, one can show (See Appendix 3.A) that the generalized $C(\alpha)$ statistic in (3.3.46) can be expressed as

$$PC(\tilde{\theta}_T; \psi) = (Th_T)\frac{1}{\hat{f}(x_0)}\frac{1}{(\hat{\lambda}_1^{-2}\hat{\sigma}_1^2 + \hat{\lambda}_2^{-2}\hat{\sigma}_2^2)}. \tag{3.5.197}$$

$$\left[\frac{1}{\nu_0}\left(\hat{\lambda}_1^{-1}D_{1,T}(\tilde{\theta}_T) - \hat{\lambda}_2^{-1}D_{3,T}(\tilde{\theta}_T)\right)^2 + \frac{1}{\nu_2}\left(\hat{\lambda}_1^{-1}D_{2,T}(\tilde{\theta}_T) - \hat{\lambda}_2^{-1}D_{4,T}(\tilde{\theta}_T)\right)^2\right]. \tag{3.5.198}$$

If $h_T = O(T^{-b})$ where b satisfies

$$b > \frac{1}{14} \tag{3.5.199}$$

then, Assumption 3.3.11 holds and

$$PC(\tilde{\theta}_T; \psi) \xrightarrow[n\to\infty]{L} \chi^2(2) \tag{3.5.200}$$

under H_0 in (3.5.182).

3.6. Two-sample problems under unbalanced sample sizes

When the study concerns itself with multiple populations or the population of interest is categorized into clusters, tests for homogeneity of certain features across different groups are often conducted for various purposes. In panel data models, homogeneity of slope coefficients justifies pooling of data, which improves estimation efficiency. Existing tests include Hausman and Taylor (1981), Pesaran, Smith and Im (1996), Wang, Phillips and Su (2018), and Lian, Qiao and Zhang (2021) in the context of group heterogeneity. Tests on parameter homogeneity are also relevant in implementation of the analysis of covariance (ANCOVA) and in the study of aptitude-treatment interactions. In practice, the number of observations from each group may be vastly different and thus the use of asymptotics based on the premise that the sample sizes of all groups grow at the same rate may not be warranted. In this section, we consider the problem of testing homogeneity of slope coefficients across two different groups under such a setup and apply the extended general-

ized $C(\alpha)$-test ($EC(\alpha)$ test) developed in Section 3.4. As we describe, the problem can be formulated in terms of primary and auxiliary estimating equations, the latter of which depend only on nuisance parameters. While inference and testing problems under estimating equations which converge at different rates are considered in the literature (Lee (2005), Lee (2010), Antoine and Renault (2012)), there is no procedure available for a testing problem where nuisance parameters are only identified from auxiliary estimating equations which converge at a different rate from the primary ones which involve both the parameter of interest and the nuisance parameters. The $EC(\alpha)$ test can be applied to such problem regardless of which set of estimating equations converge at a faster rate and without knowledge of the asymptotic distribution of an unrestricted estimator, which is often difficult to compute even if it exists. Let (Y,X) be a pair of random variables from the population distribution $F_{Y,X}$. Suppose from two subpopulations $F_{Y,X}^{(1)}$ and $F_{Y,X}^{(2)}$ of $F_{Y,X}$, we observe independent sets of i.i.d. observations $\{y_{i,n}^{(1)},x_{i,n}^{(1)}\}_{i=1}^{n}$ and $\{y_{j,m_n}^{(2)},x_{j,m_n}^{(2)}\}_{i=1}^{m_n}$ from $F_{Y,X}^{(1)}$ and $F_{Y,X}^{(2)}$, respectively. Assume the fourth moments of $(y_{i,n}^{(1)},x_{i,n}^{(1)})$ and $(y_{j,m_n}^{(2)},x_{j,m_n}^{(2)})$ exist for all i and j. The number m_n of observations from $F_{Y,X}^{(2)}$ grow at a rate n^{α} where $\alpha > 1$ as n increases and satisfies

$$\frac{m_n}{n^{\alpha}} \underset{n\to\infty}{\longrightarrow} c \tag{3.6.201}$$

for some positive constant $c > 0$. Note that when $\alpha = 1$, the two sample sizes m_n and n grows at the same rate and $\alpha > 1$ ($\alpha < 1$) corresponds to the case where the sampling frequency from $F_{Y,X}^{(2)}$ is higher (lower) than from $F_{Y,X}^{(1)}$. We have a regression model of the form:

$$\begin{aligned} y_{i,n}^{(1)} &= (\theta_0+\gamma_0)x_{i,n}^{(1)}+\varepsilon_{i,n} \\ y_{j,m_n}^{(2)} &= \gamma_0 x_{j,m_n}^{(2)}+u_{i,j} \end{aligned} \tag{3.6.202}$$

where $\varepsilon_{i.n}$ and u_{j,m_n} are expectation errors for $i = 1,\dots n,: j = 1,\dots m_n$. We are interested in testing whether the conditional expectation functions from $F_{Y,X}^{(1)}$ and $F_{Y,X}^{(2)}$ are identical by considering the null hypothesis:

$$H_0 \,:\, \theta_0 = 0. \tag{3.6.203}$$

Following the framework in Section 3.4, consider the primary and auxiliary estimating functions

$$D_n(\theta,\eta)=\frac{1}{n}\sum_{i=1}^{n}(y_{i,n}^{(1)}-(\theta+\eta)x_{i,n}^{(1)})x_{i,n}^{(1)} \tag{3.6.204}$$

and

$$G_n(\eta)=\frac{1}{m}\sum_{j=1}^{m}(y_{i,n}^{(2)}-\eta_0 x_{j,m}^{(2)})x_{j,m}^{(2)} \tag{3.6.205}$$

where η is treated as a nuisance parameter in testing H_0 in 3.6.203. We have asymptotic normality of the estimating functions:

$$\begin{bmatrix}\sqrt{n}D_n(\theta_0,\eta_0)\\ \sqrt{m_n}G_n(\eta_0)\end{bmatrix}\xrightarrow[n\to\infty]{L}\mathrm{N}\left[0,I(\theta_0,\eta_0)\right] \tag{3.6.206}$$

where

$$I(\theta_0,\eta_0)=\begin{bmatrix}I_1(\theta_0,\eta_0) & 0\\ 0 & I_2(\eta_0)\end{bmatrix} \tag{3.6.207}$$

and

$$I_1(\theta_0,\eta_0)=\mathbb{E}[(x_{i,n}^{(1)})^2\varepsilon_{i,n}^2],\qquad I_2(\eta_0)=\mathbb{E}[(x_{u,m_n}^{(2)})^2u_{j,m_n}^2] \tag{3.6.208}$$

Furthermore, the nuisance parameter η_0 is estimated by the least squares estimator

$$\hat{\eta}_{m_n}=\left(\sum_{j=1}^{m_n}(x_{j,m}^{(2)})^2\right)^{-1}\left(\sum_{j=1}^{m_n}x_{j,m}^{(2)}y_{j,m}^{(2)}\right)$$

such that

$$\sqrt{m_n}(\hat{\eta}_{m_n}-\eta_0)=O_p(1).$$

Under H_0, we consider estimator $\tilde{I}_{1,n}$ and $\tilde{I}_{2,m_n}$ of $I_1(\theta_0,\eta_0)$ and $I_2(\eta_0)$

$$\tilde{I}_{1,n}=\frac{1}{n}\sum_{i=1}^{n}(x_{i,n}^{(1)})^2(y_{i,n}^{(1)}-(\theta_0+\hat{\eta}_{m_n})x_{i,n}^{(1)})^2 \tag{3.6.209}$$

and

$$\tilde{I}_{2,n}=\frac{1}{m_n}\sum_{j=1}^{m_n}(x_{j,m_n}^{(2)})^2(y_{j,m_n}^{(2)}-\hat{\eta}_{m_n}x_{j,m_n}^{(2)})^2, \tag{3.6.210}$$

respectively. Then,

$$\|\tilde{I}_{1,n} - I_1(\theta_0, \eta_0)\| = O_p(1/\sqrt{n}), : \|\tilde{I}_{2,m_n} - I_2(\eta_0)\| = O_p(1/\sqrt{m_n}). \tag{3.6.211}$$

Furthermore, let

$$\tilde{J}_{\theta,n} = \tilde{J}_{\eta,n} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,n}^{(1)})^2, : \tilde{g}_n = \frac{1}{m_n}\sum_{j=1}^{m_n}(x_{j,m_n}^{(2)})^2. \tag{3.6.212}$$

Then,

$$\|\tilde{J}_{\theta,n} - J_\theta\| = \|\tilde{J}_{\eta,n} - J_\theta\| = O_p(1/\sqrt{n}) \tag{3.6.213}$$

and

$$\|\tilde{g}_n - g_0\| = O_p(1/\sqrt{m_n}). \tag{3.6.214}$$

The extended generalized $C(\alpha)$ test statistic is based on the estimating function $s_n^*(\theta, \eta)$ defined as

$$s_n^*(\theta, \eta) = \tilde{J}_{\theta,n}^{-1}\left\{D_n(\theta, \eta) - \frac{1}{c}\tilde{J}_{\eta,n}\tilde{g}_n^{-1}G_n(\eta)\right\}. \tag{3.6.215}$$

$$\tilde{Q}_n = \left(\frac{1}{n}\sum_{i=1}^{n}(x_{i,n}^{(1)})^2\right)^{-1}$$

First, we consider the case where $\alpha > 1$. Then,

$$\sqrt{n}s_n^*(\theta, \eta) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0, \Lambda_1(\theta_0, \eta_0)\right] \tag{3.6.216}$$

where

$$\Lambda_1(\theta_0, \eta_0) = \left(\mathbb{E}[(x_{i,n}^{(1)})^2]\right)^{-2}\mathbb{E}[(x_{i,n}^{(1)})^2\varepsilon_{i,n}^2]. \tag{3.6.217}$$

If $\alpha < 1$, we have

$$\sqrt{n^\alpha}s_n^*(\theta, \eta) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0, \Lambda_2(\eta_0)\right]. \tag{3.6.218}$$

where

$$\Lambda_2(\eta_0) = \frac{1}{c^2}\left(\mathbb{E}[(x_{j,m_n}^{(2)})^2]\right)^{-2}\mathbb{E}[(x_{j,m_n}^{(2)})^2 u_{j,m_n}^2]$$

Lastly, if $\alpha = 1$,

$$\sqrt{n}s_n^*(\theta, \eta) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0, \Lambda_3(\theta_0, \eta_0)\right] \tag{3.6.219}$$

where

$$\Lambda_3(\theta_0,\eta_0) = \Lambda_1(\theta_0,\eta_0) + \Lambda_2(\eta_0). \tag{3.6.220}$$

Estimators of $(\Lambda_1(\theta_0,\eta_0), \Lambda_2(\eta_0), \Lambda_3(\theta_0,\eta_0))$ are given, respectively, by

$$\tilde{\Lambda}_{1,n} = \left(\tilde{J}_{\theta,n}\right)^{-2} \tilde{I}_{1,n}, \qquad \tilde{\Lambda}_{2,n} = (\tilde{g}_n)^{-2} \tilde{I}_{2,n}$$

and

$$\tilde{\Lambda}_{3,n} = \tilde{\Lambda}_{1,n} + \tilde{\Lambda}_{2,n}. \tag{3.6.221}$$

The extended generalized $C(\alpha)$ statistic is then

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \psi) = n^{2\min(1,\alpha)} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)'(\hat{\Lambda}_n)^- s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) \tag{3.6.222}$$

where

$$\hat{\Lambda}_n = \begin{cases} \tilde{\Lambda}_{1,n} & \text{, if } \alpha > 1 \\ \tilde{\Lambda}_{2,n} & \text{, if } \alpha < 1 \\ \tilde{\Lambda}_{3,n} & \text{, if } \alpha = 1. \end{cases} \tag{3.6.223}$$

Then, by applying Proposition 3.4.4, we have

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \psi) \xrightarrow[n\to\infty]{L} \chi^2(1). \tag{3.6.224}$$

3.7. Conclusion

In this paper, we have studied generalized $C(\alpha)$ tests on restrictions on a finite-dimensional parameter when estimating equations and nuisance parameter estimators converge at non-standard rates and have shown under general conditions that the null asymptotic distributions of the proposed test statistics are chi-square. In particular, the restricted estimator may converge at a slower rate than the estimating equations. As discussed in Dufour et al. (2016), the generalized $C(\alpha)$ statistic nests the existing $C(\alpha)$-type statistics as special cases and thus our results broaden applicability of these statistics to problems involving nonstandard rates. The advantages of this $C(\alpha)$-type procedure are highlighted in the local estimating function setup in Section 3.5. Existing testing methods require that parameters be

estimated at the same rate. Our testing procedure is not restricted to such cases and thus accommodates more general test restrictions and flexible choices of smoothing bandwidths.

Then, we have proposed the extended generalized $C(\alpha)$ statistic [$EC(\alpha)$] to accommodate a class of models in which the parameters of interest are defined by primary estimating functions which also depend on nuisance parameters such that they are only estimated from auxiliary estimating equations that converge at a different rate from the primary ones. It is shown that while the asymptotic distribution of the test statistic is still chi-square, its degree of freedom depends on which of the convergence rate of the primary and auxiliary equations is slower. When a problem involves two populations that are sampled at disproportionate frequencies, the conventional assumption that each sample size grows at the same rate may be unrealistic. The $EC(\alpha)$ test procedure is applied for a problem of testing homogeneity of the regression functions in such setup.

Lastly, it is of interest to extend the framework in this paper to testing problems involving infinite-dimensional nuisance parameters. Such parameters are typically estimated at a nonparametric rate and the asymptotic distributions of the estimators are often intractable or yet to be known, especially, when machine learning algorithms are employed. In addition, an estimator of the finite-dimensional parameter of interest may not attain $n^{1/2}$-consistent (e.g. Firpo et al. (2009)) in such semiparametric framework. The test statistic in this case can be constructed from an estimating function which satisfies analogous conditions to Neyman orthogonality with respect to a functional parameter introduced by Chernozhukov et al. (2018). Such extension is left for future research.

3.A. Derivations of the test statistic $PC(\tilde{\theta}_n\,;\,\psi)$ in each problem in Section 3.5

3.5.2: Testing the derivatives

Recall that

$$\tilde{J}_{\theta,n} = \hat{f}(x_0)\mathscr{U}\,,\ \tilde{I}_{\theta,n} = \hat{\sigma}^2(x_0)\hat{f}(x_0)\mathscr{V}\,, \tag{3.A.225}$$

$$\tilde{P}_n = [\mathbf{0}_{(M-m_0+1)\times m_0}; \mathbf{1}_{(M-m_0+1)\times(M-m_0+1)}],\ \tilde{W}_n = \tilde{I}_n^{-1}. \tag{3.A.226}$$

where $\mathscr{U} = [\mu_{i+j-2}]_{1\le i,j\le M+1}$ and $\mathscr{V} = [\nu_{i+j-2}]_{1\le i,j\le M+1}$

Then,

$$\begin{aligned}\tilde{Q}_n\tilde{I}_n\tilde{Q}_n' &= \tilde{P}_n\left(\tilde{J}_{\theta,n}'\tilde{I}_{\theta,n}^{-1}\tilde{J}_{\theta,n}\right)^{-1}\tilde{P}_n' \\ &= \frac{\hat{\sigma}^2(x_0)}{\hat{f}(x_0)}\tilde{P}_n\mathscr{U}^{-1}\mathscr{V}\mathscr{U}^{-1}\tilde{P}_n' \end{aligned} \tag{3.A.227}$$

and thus

$$\left[\tilde{Q}_n\tilde{I}_n\tilde{Q}_n'\right]^{-1} = \frac{\hat{f}(x_0)}{\hat{\sigma}^2(x_0)}\left[\tilde{P}_n\mathscr{U}^{-1}\mathscr{V}\mathscr{U}^{-1}\tilde{P}_n'\right]^{-1}. \tag{3.A.228}$$

Note

$$\begin{aligned}\tilde{Q}_nD_n(\tilde{\theta}_n) &= \tilde{P}_n\left(\tilde{J}_{\theta,n}'\tilde{I}_n^{-1}\tilde{J}_{\theta,n}\right)^{-1}\tilde{J}_{\theta,n}'\tilde{I}_n^{-1}D_n(\tilde{\theta}_n) \\ &= \frac{1}{\hat{f}(x_0)}\tilde{P}_n\mathscr{U}^{-1}D_n(\tilde{\theta}_n). \end{aligned} \tag{3.A.229}$$

Then, since $D_n(\theta_0) = O_p((nh_n)^{-1/2})$, we have

$$PC(\tilde{\theta}_n;\psi) = (nh_n)\frac{1}{\hat{f}(x_0)\hat{\sigma}^2(x_0)}D_n(\tilde{\theta}_n)\mathscr{U}^{-1}\tilde{P}_n\left[\tilde{P}_n\mathscr{U}^{-1}\mathscr{V}\mathscr{U}^{-1}\tilde{P}_n'\right]^{-1}\tilde{P}_n\mathscr{U}^{-1}D_n(\tilde{\theta}_n). \tag{3.A.230}$$

Suppose now that $M = m_0 = 1$. Then,

$$\mathscr{U} = \begin{bmatrix}1 & 0\\ 0 & \mu_2\end{bmatrix},\ \mathscr{V} = \begin{bmatrix}\nu_0 & 0\\ 0 & \nu_2\end{bmatrix},\ \tilde{P}_n = \begin{bmatrix}0 & 1\end{bmatrix} \tag{3.A.231}$$

and

$$D_n(\theta) := \begin{bmatrix}D_{1,n}(\theta)\\ D_{2,n}(\theta)\end{bmatrix} = \begin{bmatrix}\frac{1}{nh_n}\sum_{i=1}^n(y_i-\sum_{s=0}^1\theta^{(s)}(x_i-x_0)^s)K\left(\frac{x_i-x_0}{h_n}\right)\\ \frac{1}{nh_n^2}\sum_{i=1}^n(y_i-\sum_{s=0}^1\theta^{(s)}(x_i-x_0)^s)(x_i-x_0)K\left(\frac{x_i-x_0}{h_n}\right)\end{bmatrix}. \tag{3.A.232}$$

Observe

$$\tilde{P}_n\mathscr{U}^{-1}D_n(\tilde{\theta}_n) = \frac{\mu_2^{-1}}{nh_n^2}\sum_{i=1}^n(y_i-\sum_{s=0}^1\tilde{\theta}_n^{(s)}(x_i-x_0)^s)(x_i-x_0)K\left(\frac{x_i-x_0}{h_n}\right) \tag{3.A.233}$$

and

$$\tilde{P}_n\mathscr{U}^{-1}\mathscr{V}\mathscr{U}^{-1}\tilde{P}_n = \begin{bmatrix}0 & \mu_2^{-1}\end{bmatrix}\begin{bmatrix}\nu_0 & 0\\ 0 & \nu_2\end{bmatrix}\begin{bmatrix}0\\ \mu_2^{-1}\end{bmatrix}$$

$$= \quad \mu_2^{-2}\nu_2. \tag{3.A.234}$$

so that by substituting them into 3.A.230 yields

$$PC(\tilde{\theta}_n;\,\psi) = (nh_n)\frac{\nu_2}{\hat{f}(x_0)\hat{\sigma}^2(x_0)}\left(\frac{1}{nh_n^2}\sum_{i=1}^{n}(y_i - \sum_{s=0}^{1}\tilde{\theta}_n^{(s)}(x_i-x_0)^s)(x_i-x_0)K\left(\frac{x_i-x_0}{h_n}\right)\right)^2 \tag{3.A.235}$$

3.5.3: Regression discontinuity design

Recall that

$$\tilde{J}_{\theta,n} = \hat{f}(x_0)\begin{bmatrix}\hat{p} & 0 & 0\\ 0 & \hat{p}\mu_2 & 0\\ 0 & 0 & (1-\hat{p})\end{bmatrix}, \tilde{I}_n = \hat{f}(x_0)\begin{bmatrix}\hat{p}^2\hat{\sigma}_+^2\nu_0 & 0 & 0\\ 0 & \hat{p}^2\hat{\sigma}_+^2\nu_2 & 0\\ 0 & 0 & (1-\hat{p})^2\hat{\sigma}_-^2\nu_0\end{bmatrix} \tag{3.A.236}$$

$$\tilde{P}_n = \begin{bmatrix}1 & 0 & -1\end{bmatrix}, \tilde{W}_n = \tilde{I}_n^{-1} \tag{3.A.237}$$

and $D_n(\theta_0) = O_p((nh_n)^{-1/2})$. Then,

$$\left(\tilde{J}'_{\theta,n}\tilde{I}_{\theta,n}^{-1}\tilde{J}_{\theta,n}\right)^{-1} = (\hat{f}(x_0))^{-1}\begin{bmatrix}(\hat{p})^{-1} & 0 & 0\\ 0 & (\hat{p}\mu_2)^{-1} & 0\\ 0 & 0 & (1-\hat{p})^{-1}\end{bmatrix}\begin{bmatrix}\hat{p}^2\hat{\sigma}_+^2\nu_0 & 0 & 0\\ 0 & \hat{p}^2\hat{\sigma}_+^2\nu_2 & 0\\ 0 & 0 & (1-\hat{p})^2\hat{\sigma}_-^2\nu_0\end{bmatrix}$$
$$\begin{bmatrix}(\hat{p})^{-1} & 0 & 0\\ 0 & (\hat{p}\mu_2)^{-1} & 0\\ 0 & 0 & (1-\hat{p})^{-1}\end{bmatrix} \tag{3.A.238}$$

$$= (\hat{f}(x_0))^{-1}\begin{bmatrix}\hat{\sigma}_+^2\nu_0 & 0 & 0\\ 0 & \hat{\sigma}_+^2\mu_2^{-1}\nu_2 & 0\\ 0 & 0 & \hat{\sigma}_-^2\nu_0\end{bmatrix} \tag{3.A.239}$$

$$\tilde{Q}\tilde{I}_n\tilde{Q}'_n = (\hat{f}(x_0))^{-1}\begin{bmatrix}1 & 0 & -1\end{bmatrix}\begin{bmatrix}\hat{\sigma}_+^2\nu_0 & 0 & 0\\ 0 & \hat{\sigma}_+^2\mu_2^{-1}\nu_2 & 0\\ 0 & 0 & \hat{\sigma}_-^2\nu_0\end{bmatrix}\begin{bmatrix}1\\0\\-1\end{bmatrix} \tag{3.A.240}$$

$$= (\hat{f}(x_0))^{-1}(\hat{\sigma}_+^2 + \hat{\sigma}_-^2)\nu_0. \tag{3.A.241}$$

In addition,

$$\begin{aligned}\tilde{Q}_n D_n(\tilde{\theta}_n) &= (\hat{f}(x_0))^{-1}\begin{bmatrix}1 & 0 & -1\end{bmatrix}\begin{bmatrix}(\hat{p})^{-1} & 0 & 0\\ 0 & (\hat{p}\mu_2)^{-1} & 0\\ 0 & 0 & (1-\hat{p})^{-1}\end{bmatrix}\begin{bmatrix}D_{1,n}(\tilde{\theta}_n)\\ D_{2,n}(\tilde{\theta}_n)\\ D_{3,n}(\tilde{\theta}_n)\end{bmatrix}\\ &= (\hat{f}(x_0))^{-1}\left(\hat{p})^{-1}D_{1,n}(\tilde{\theta}_n) - (1-\hat{p})^{-1}D_{3,n}(\tilde{\theta}_n)\right).\end{aligned}$$

Thus,

$$PC(\tilde{\theta}_n;\psi) = (nh)\frac{2}{\hat{f}(x_0)(\hat{\sigma}_+^2+\hat{\sigma}_-^2)\nu_0}\left(\hat{p})^{-1}D_{1,n}(\tilde{\theta}_n) - (1-\hat{p})^{-1}D_{3,n}(\tilde{\theta}_n)\right)^2. \tag{3.A.242}$$

3.5.4: Stochastic discount factor

Recall that

$$\tilde{J}_{\theta,T} = \hat{f}(x_0)\begin{bmatrix}\hat{\lambda}_1 & 0 & 0 & 0\\ 0 & \hat{\lambda}_1\mu_2 & 0 & 0\\ 0 & 0 & \hat{\lambda}_2 & 0\\ 0 & 0 & 0 & \hat{\lambda}_2\mu_2\end{bmatrix}, \tilde{I}_T = \hat{f}(x_0)\begin{bmatrix}\hat{\sigma}_1^2\nu_0 & 0 & 0 & 0\\ 0 & \hat{\sigma}_1^2\nu_2 & 0 & 0\\ 0 & 0 & \hat{\sigma}_2^2\nu_0 & 0\\ 0 & 0 & 0 & \hat{\sigma}_2^2\nu_2\end{bmatrix} \tag{3.A.243}$$

$$\tilde{P}_n = \begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}, \tilde{W}_n = \tilde{I}_n^{-1} \tag{3.A.244}$$

and $D_n(\theta_0) = O_p((nh_n)^{-1/2})$. Observe

$$\begin{aligned}\left(\tilde{J}'_{\theta,n}\tilde{I}_{\theta,n}^{-1}\tilde{J}_{\theta,n}\right)^{-1} &= (\hat{f}(x_0))^{-1}\begin{bmatrix}(\hat{\lambda}_1)^{-2} & 0 & 0 & 0\\ 0 & (\hat{\lambda}_1\mu_2)^{-2} & 0 & 0\\ 0 & 0 & (\hat{\lambda}_2)^{-2} & 0\\ 0 & 0 & 0 & (\hat{\lambda}_2\mu_2)^{-2}\end{bmatrix}\begin{bmatrix}\hat{\sigma}_1^2\nu_0 & 0 & 0 & 0\\ 0 & \hat{\sigma}_1^2\nu_2 & 0 & 0\\ 0 & 0 & \hat{\sigma}_2^2\nu_0 & 0\\ 0 & 0 & 0 & \hat{\sigma}_2^2\nu_2\end{bmatrix}\\ &= (\hat{f}(x_0))^{-1}\begin{bmatrix}(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2\nu_0 & 0 & 0 & 0\\ 0 & (\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2\mu_2^{-2}\nu_2 & 0 & 0\\ 0 & 0 & (\hat{\lambda}_2)^{-2}\hat{\sigma}_2^2\nu_0 & 0\\ 0 & 0 & 0 & (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2\mu_2^{-2}\nu_2\end{bmatrix}\end{aligned} \tag{3.A.245}$$

$$\tilde{Q}\tilde{I}_n\tilde{Q}'_n = (\hat{f}(x_0))^{-1}\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\begin{bmatrix}(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2\nu_0 & 0 & 0 & 0\\ 0 & (\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2\mu_2^{-2}\nu_2 & 0 & 0\\ 0 & 0 & (\hat{\lambda}_2)^{-2}\hat{\sigma}_2^2\nu_0 & 0\\ 0 & 0 & 0 & (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2\mu_2^{-2}\nu_2\end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \tag{3.A.246}$$

$$= (\hat{f}(x_0))^{-1} \begin{bmatrix} \{(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2 + (\hat{\lambda}_2)^{-2}\hat{\sigma}_2^2\}\nu_0 & 0 \\ 0 & \{(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2 + (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2\}\mu_2^{-2}\nu_2 \end{bmatrix} \tag{3.A.247}$$

and thus

$$\left[\tilde{Q}\tilde{I}_n\tilde{Q}_n'\right]^{-1} = \hat{f}(x_0)\left\{(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2 + (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2\right\}^{-1} \begin{bmatrix} \nu_0^{-1} & 0 \\ 0 & \mu_2^2\nu_2^{-1} \end{bmatrix}. \tag{3.A.248}$$

On the other hand, we have

$$\begin{aligned} \tilde{Q}_T D_T(\tilde{\theta}_T) &= (\hat{f}(x_0))^{-1} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} (\hat{\lambda}_1)^{-1} & 0 & 0 & 0 \\ 0 & (\hat{\lambda}_1\mu_2)^{-1} & 0 & 0 \\ 0 & 0 & (\hat{\lambda}_2)^{-1} & 0 \\ 0 & 0 & 0 & (\hat{\lambda}_2\mu_2)^{-1} \end{bmatrix} \begin{bmatrix} D_{1,T}(\tilde{\theta}_T) \\ D_{2,T}(\tilde{\theta}_T) \\ D_{3,T}(\tilde{\theta}_T) \\ D_{4,T}(\tilde{\theta}_T) \end{bmatrix} \\ &= (\hat{f}(x_0))^{-1} \begin{bmatrix} (\hat{\lambda}_1)^{-1} & 0 & -(\hat{\lambda}_2)^{-1} & 0 \\ 0 & (\hat{\lambda}_1\mu_2)^{-1} & 0 & (\hat{\lambda}_2\mu_2)^{-1} \end{bmatrix} \begin{bmatrix} D_{1,T}(\tilde{\theta}_T) \\ D_{2,T}(\tilde{\theta}_T) \\ D_{3,T}(\tilde{\theta}_T) \\ D_{4,T}(\tilde{\theta}_T) \end{bmatrix} \\ &= (\hat{f}(x_0))^{-1} \begin{bmatrix} (\hat{\lambda}_1)^{-1}D_{1,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{3,T}(\tilde{\theta}_T) \\ \mu_2^{-1}\left\{(\hat{\lambda}_1)^{-1}D_{2,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{4,T}(\tilde{\theta}_T)\right\} \end{bmatrix}. \end{aligned} \tag{3.A.249}$$

Then,

$$\begin{aligned} PC(\tilde{\theta}_T;\psi) &= Th_T \frac{1}{\hat{f}(x_0)} \frac{1}{(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2 + (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2} \begin{bmatrix} (\hat{\lambda}_1)^{-1}D_{1,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{3,T}(\tilde{\theta}_T) \\ \mu_2\left\{(\hat{\lambda}_1)^{-1}D_{2,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{4,T}(\tilde{\theta}_T)\right\} \end{bmatrix}' \\ &\qquad \begin{bmatrix} \nu_0^{-1} & 0 \\ 0 & \mu_2^2\nu_2^{-1} \end{bmatrix} \begin{bmatrix} (\hat{\lambda}_1)^{-1}D_{1,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{3,T}(\tilde{\theta}_T) \\ \mu_2\left\{(\hat{\lambda}_1)^{-1}D_{2,T}(\tilde{\theta}_T) - (\hat{\lambda}_2)^{-1}D_{4,T}(\tilde{\theta}_T)\right\} \end{bmatrix} \\ &= Th_T \frac{1}{(\hat{\lambda}_1)^{-2}\hat{\sigma}_1^2 + (\hat{\lambda}_2)^{-2}\hat{\sigma}_-^2} \left[\frac{1}{\nu_0}\left(\hat{\lambda}_1^{-1}D_{1,T}(\tilde{\theta}_T) - \hat{\lambda}_2^{-1}D_{3,T}(\tilde{\theta}_T)\right)^2\right. \\ &\quad \left. + \frac{1}{\nu_2}\left(\hat{\lambda}_1^{-1}D_{2,T}(\tilde{\theta}_T) - \hat{\lambda}_2^{-1}D_{4,T}(\tilde{\theta}_T)\right)^2\right] \end{aligned} \tag{3.A.250}$$

3.B. Proofs

PROOF OF PROPOSITION 3.3.1 Assumption 3.3.3 implies that for any positive constant $\delta \in (0,1)$, there exists some positive integer $n_0 := n_0(\delta)$ which depends on δ such that

$$\tilde{\theta}_n^0 \in \mathscr{V} := \mathscr{V}_D \cap \left(\cap_{i=1}^2 \mathscr{V}_{J,i}\right) \cap \mathscr{N}_P(\theta_0) \tag{3.B.1}$$

with probability at least $1-\delta$. Fix δ and define

$$A_\delta := \{\omega \in \Omega_J : \tilde{\theta}_n(\omega) \in \mathscr{V}, , \forall n \geq n_0\}. \tag{3.B.2}$$

Throughout the rest of the proof, we assume $\omega \in A_\delta$ and $n \geq n_0$. We have

$$\|\tilde{J}_{\theta,n}(\tilde{\theta}_n^0) - J_\theta(\tilde{\theta}_n^0)\| \leq \sup_{\theta \in \mathscr{V}_{J,2}} \|\tilde{J}_{\theta,n}(\theta) - J_\theta(\theta)\| \leq C_{J,0} n^{-r_M}, \tag{3.B.3}$$

and

$$\|J_\theta(\tilde{\theta}_n^0) - J_\theta(\theta_0)\| \leq C_{J,\theta_0} \|\tilde{\theta}_n^0 - \theta_0\|, \tag{3.B.4}$$

by the assumption 3.3.8 and 3.3.9, so that

$$\|\tilde{J}_{\theta,n}(\tilde{\theta}_n^0) - J_\theta(\theta_0)\| \leq C_{J,0} n^{-r_M} + C_{J,\theta_0} \|\tilde{\theta}_n^0 - \theta_0\|. \tag{3.B.5}$$

by the triangle inequality. It follows from Assumption 3.3.3 that

$$\|\tilde{J}_{\theta,n}(\tilde{\theta}_n^0) - J_\theta(\theta_0)\| = O_p(n^{-\min(r_M, r_\theta)}). \tag{3.B.6}$$

In order to show

$$\|P(\tilde{\theta}_n^0) - P(\theta_0)\| = O_p(n^{-r_\theta}), \tag{3.B.7}$$

observe by Assumption 3.3.6, 3.3.7 that

$$P_l(\tilde{\theta}_n^0) = P(\theta_0) + \mathbb{H}_l(\theta_n^*)(\tilde{\theta}_n - \theta_0), \qquad l = 1, \ldots, p_1 \tag{3.B.8}$$

where θ_n^* is some point between $\tilde{\theta}_n$ and θ_0 and thus in $\mathscr{N}_P(\theta_0)$ and thus

$$\begin{aligned} \|P(\tilde{\theta}_n^0) - P(\theta_0)\| &\leq p_1 \sup_{1\leq l\leq p_l} \sup_{\theta\in\mathscr{N}_P(\theta_0)} \|\mathbb{H}_l(\theta)\| \left\|\tilde{\theta}_n - \theta_0\right\| \\ &\leq p_1 C_{P,\theta_0} \left\|\tilde{\theta}_n - \theta_0\right\|, \qquad \forall n \geq n_P^*. \end{aligned} \tag{3.B.9}$$

Hence,

$$\|P(\tilde{\theta}_n^0) - P(\theta_0)\| O_p(n^{-r_\theta}) \tag{3.B.10}$$

again by appealing to Assumption 3.3.3. By combining (3.B.6) and (3.B.7) along with Assumption 3.3.10,

$$\|\tilde{Q}_n - Q(\theta_0)\| = O_p(n^{-\min(r_\theta, r_M)}). \tag{3.B.11}$$

By Assumption 3.3.10, 3.3.15, $\tilde{Q}_n$ and $\tilde{I}_n$ have rank p_1 and p so that

$$\operatorname{rank}[\tilde{Q}_n \tilde{I}_n \tilde{Q}_n'] = p_1 \tag{3.B.12}$$

and thus $\tilde{Q}_n \tilde{I}_n \tilde{Q}_n'$ is invertible. Then, by Assumption 3.3.12 and 119,

$$\|(\tilde{Q}_n \tilde{I}_n^{-1} \tilde{Q}_n')^{-1} - (Q_0 I_0^{-1} Q_0')^{-1}\| = o_p(1). \tag{3.B.13}$$

Next, we are going to show Next, we are going to show

$$n^{r_\theta} P(\theta_0) \mathbf{n}^{-\beta} (\tilde{\theta}_n^0 - \theta_0) \xrightarrow[n\to\infty]{\mathsf{p}} 0. \tag{3.B.14}$$

Recall

$$\mathbf{n}^{-\beta}(\tilde{\theta}_n^0 - \theta_0) = \check{\theta}_n - \theta_0$$

by definition of $\check{\theta}_n$. By Assumption 3.3.7, we have the following expansion of $\psi(\check{\theta}_n)$:

$$\psi(\check{\theta}_n) = \psi(\theta_0) + P(\theta_0)(\check{\theta}_n - \theta_0) + \mathbf{H}(\check{\theta}_n - \theta_0; \theta_n^*),$$

$$\mathbf{H}(\check{\theta}_n - \theta_0; \theta_n^*) = (\mathbf{H}_1(\check{\theta}_n - \theta_0; \theta^*), \ldots \mathbf{H}_l(\check{\theta}_n - \theta_0; \theta^*), \ldots \mathbf{H}_{p_1}(\check{\theta}_n - \theta_0; \theta^*)) \tag{3.B.15}$$

is a $p_1 \times 1$ vector with the l-th element ($1 \leq l \leq p_1$):

$$\mathbf{H}_l(\check{\theta}_n - \theta_0; \theta^*) = (\check{\theta}_n - \theta_0)' \mathbb{H}_l(\theta_n^*)(\check{\theta}_n - \theta_0), \quad l = 1, \ldots, p_1 \tag{3.B.16}$$

and $\theta_n^* \in \Theta$ is some value between $\check{\theta}_n$ and θ_0. By Assumption 3.3.17, $\psi(\theta_0) = 0$ so that

$$n^{r_D} P(\theta_0)(\check{\theta}_n - \theta_0) = \psi(\check{\theta}_n) - n^{r_D} \mathbf{H}(\check{\theta}_n - \theta_0; \theta^*)$$

and Assumption 3.3.7 combined with Assumption 3.3.3 implies that with probability approaching to one,

$$\|\mathbf{H}(\check{\theta}_n - \theta_0; \theta^*)\| \leq p_1 C_{P,\theta^*} \|\tilde{\theta}_n^0 - \theta_0\|^2 \tag{3.B.17}$$

by similar argument to (3.B.9). Since $\psi(\check{\theta}_n) = O_p(n^{-2r_\theta})$ by Assumption 3.3.3, it follows from Assumption 3.3.17, 3.3.11 that

$$\begin{aligned} \|n^{r_D} P(\theta_0)(\check{\theta}_n - \theta_0)\| &\leq n^{r_D} \psi(\check{\theta}_n) + p_1 C_{P,\theta^*} n^{r_D} \|\tilde{\theta}_n^0 - \theta_0\|^2 . \\ &= O_p(n^{r_D - 2r_\theta}) = o_P(1). \end{aligned} \tag{3.B.18}$$

Now, by Assumption 3.3.4, 3.3.5, we have

$$\begin{aligned} n^{r_D} \tilde{Q}_n D_n(\tilde{\theta}_n^0) &= n^{r_D} \tilde{Q}_n D_n(\theta_0) \\ &\quad + n^{r_D} \tilde{Q}_n J_\theta(\theta_0) \mathbf{n}^{-\beta} (\tilde{\theta}_n^0 - \theta_0) \\ &\quad + n^{r_D} \tilde{Q}_n B_n(\tilde{\theta}_n^0, \theta_0) . \end{aligned} \tag{3.B.19}$$

We are going to show

$$n^{r_D} (\tilde{Q}_n D_n(\tilde{\theta}_n^0) - Q(\theta_0) D_n(\theta_0)) \xrightarrow[n\to\infty]{\mathsf{p}} 0 . \tag{3.B.20}$$

so that by Slutsky's theorem

$$n^{r_D} \tilde{Q}_n D_n(\tilde{\theta}_n^0) \xrightarrow[n\to\infty]{L} \mathrm{N}\left[0, Q(\theta_0) I(\theta_0) Q(\theta_0)'\right] . \tag{3.B.21}$$

Observe

$$\tilde{Q}_n J_\theta(\theta_0) = P(\theta_0) + O_p(n^{-\min(r_M, r_\theta)}) \tag{3.B.22}$$

by (3.B.11) so that

$$\begin{aligned}\|n^{r_D}\tilde{Q}_n J_\theta(\theta_0)\mathbf{n}^{-\beta}(\tilde{\theta}_n^0-\theta_0)\| &\le \|n^{r_D}P(\theta_0)\mathbf{n}^{-\beta}(\tilde{\theta}_n^0-\theta_0)\| + \|O_p(n^{r_D-\min(r_M,r_\theta)-r_\theta})\| \\ &= o_p(1)\,. \end{aligned} \tag{3.B.23}$$

by (3.B.14) and Assumption 3.3.11. Similarly,

$$\|n^{r_D}\tilde{Q}_n B_n(\tilde{\theta}_n^0,\theta_0)\| \le n^{r_D}\|\tilde{Q}_n\|\|B_n(\tilde{\theta}_n^0,\theta_0)\| = O_p(n^{r_D-2r_\theta}) = o_p(1)\,. \tag{3.B.24}$$

We have shown

$$n^{r_D}\tilde{Q}_n D_n(\tilde{\theta}_n^0) = n^{r_D}Q(\theta_0)D_n(\theta_0) + o_P(1) \tag{3.B.25}$$

and thus combining with (3.B.13), we have

$$PC(\tilde{\theta}_n^0;\psi) = n^{2r_\theta}D_n(\theta_0)'Q(\theta_0)'\left[Q(\theta_0)I(\theta_0)Q(\theta_0)'\right]^{-1}Q(\theta_0)D_n(\theta_0) + o_p(1)\,. \tag{3.B.26}$$

where

$$\text{rank}[Q(\theta_0)I(\theta_0)Q(\theta_0)'] = p_1\,. \tag{3.B.27}$$

To prove the final claim, note that under Assumption 3.3.4,

$$\|n^{r_D}\tilde{Q}_n B_n(\tilde{\theta}_n^0,\theta_0)\| \le n^{r_D}\|\tilde{Q}_n\|\|B_n(\tilde{\theta}_n^0,\theta_0)\| = O_p(n^{r_D-r_\theta}). \tag{3.B.28}$$

In addition, under Assumption 3.3.6,

$$\psi(\tilde{\theta}_n^0) = \psi(\theta_0) + P(\theta_0)(\tilde{\theta}_n^0-\theta_0) + B_p(\tilde{\theta}_n^0,\theta_0) \tag{3.B.29}$$

where

$$B_p(\tilde{\theta}_n^0,\theta_0) = o_p(\|\tilde{\theta}_n-\theta_0\|))$$

so that

$$\|n^{r_D}P(\theta_0)(\tilde{\theta}_n^0-\theta_0)\| = o_p(n^{r_D-r_\theta}).$$

□

PROOF OF LEMMA 3.4.1 By Assumption 3.4.5, 3.4.6, for any positive constant

$\delta \in (0,1)$, there exists some positive integer $n_0 := n_0(\delta)$ which depends on δ such that $(\tilde{\theta}_n^0, \hat{\eta}_n) \in \mathscr{U} := \mathscr{U}_D \cap (\cap_{i=1}^2 \mathscr{U}_{J,i})$, $\tilde{\theta}_n^0 \in \mathscr{N}_{P,\theta_0}$, $\hat{\eta}_n \in \mathscr{V} := \mathscr{V}_G \cap (\cap_{i=1}^2 \mathscr{U}_{g,i}(\eta_0))$ with probability at least $1-\delta$. Fix δ and define

$$B_\delta := \{\omega \in \Omega_D \cap \Omega_G : (\tilde{\theta}_n^0, \hat{\eta}_n)(\omega) \in \mathscr{U}, \tilde{\theta}_n^0 \in \mathscr{N}_{P,\theta_0}, \hat{\eta}_n \in \mathscr{V}, \forall n \geq n_0\}. \tag{3.B.1}$$

Throughout this proof, we assume $\omega \in B_\delta$ and $n \geq n_0$.
In the proof of Proposition 3.3.1, we have already shown that

$$P(\tilde{\theta}_n^0) - P(\theta_0) = O_p(n^{-r_\theta}) \tag{3.B.2}$$

and

$$n^{r_D} P(\theta_0)(\tilde{\theta}_n^0 - \theta_0) \xrightarrow[n\to\infty]{\mathsf{p}} 0. \tag{3.B.3}$$

By similar arguments to the proof of the same proposition, we have

$$\|\tilde{Q}_n - Q(\theta_0, \eta_0)\| = O_p\left(n^{-\min(r_\theta, r_\eta, r_M)}\right). \tag{3.B.4}$$

Furthermore,

$$\|\tilde{g}_n - g(\eta_0)\| = O_p\left(n^{-\min(r_\eta, r_M)}\right) \tag{3.B.5}$$

and thus by Assumption 3.4.12,

$$\|\tilde{g}_n{}^{-1} - g(\eta_0)^{-1}\| = O_p\left(n^{-\min(r_\eta, r_M)}\right). \tag{3.B.6}$$

Then, it follows from Assumption 3.4.2 that

$$\begin{aligned}
\tilde{Q}_n D_n(\tilde{\theta}_n^0, \hat{\eta}_n) &= Q(\theta_0, \eta_0) D_n(\theta_0, \eta_0) \\
&\quad + P(\theta_0)\mathbf{n}^{-\beta_\theta}(\tilde{\theta}_n^0 - \theta_0) \\
&\quad + Q(\theta_0, \eta_0) J_\eta(\theta_0, \eta_0)\mathbf{n}^{-\beta_\eta}(\hat{\eta}_n - \eta_0) \\
&\quad + O_p(n^{-\min(r_\theta, r_\eta, r_M) - \min(r_D, r_\theta, r_\eta)}) \\
&= Q(\theta_0, \eta_0) D_n(\theta_0, \eta_0) \\
&\quad + Q(\theta_0, \eta_0) J_\eta(\theta_0, \eta_0)\mathbf{n}^{-\beta_\eta}(\hat{\eta}_n - \eta_0) + o_p(n^{-r_D}).
\end{aligned} \tag{3.B.7}$$

where the first equality follows from (3.B.4) and Assumption 3.4.5, 3.4.6 and the second equality is due to (3.B.3) and Assumption 3.4.14. Similarly,

$$\begin{aligned}\tilde{Q}_n\tilde{J}_{\eta,n}\tilde{g}_n^{-1}G_n(\hat{\eta}_n) &= Q(\theta_0,\eta_0)J_\eta(\theta_0\eta_0)g(\eta_0)^{-1}G_n(\eta_0)\\ &\quad +Q(\theta_0,\eta_0)J_\eta(\theta_0\eta_0)\mathbf{n}^{-\beta_\eta}(\hat{\eta}_n-\eta_0)\\ &\quad +O_p(n^{-\min(r_\theta,r_\eta,r_M)-\min(r_G,r_\theta,r_\eta)})\end{aligned} \tag{3.B.8}$$

Then,

$$n^{r_D}\tilde{Q}_nD_n(\tilde{\theta}_n^0,\hat{\eta}_n)-n^{r_D}\tilde{Q}_n\tilde{J}_{\eta,n}\tilde{g}_n^{-1}G_n(\hat{\eta}_n) \tag{3.B.9}$$

$$= \begin{cases} n^{r_D}Q(\theta_0,\eta_0)D_n(\theta_0,\eta_0) & \text{if } r_D=r_G\\ -n^{r_D}Q(\theta_0,\eta_0)J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}G_n(\eta_0)+o_p(1) & \\ n^{r_D}Q(\theta_0,\eta_0)J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}G_n(\eta_0)+o_p(1) & \text{if } r_D<r_G\\ -n^{r_D}Q(\theta_0,\eta_0)J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}G_n(\eta_0) & \text{if } r_D>r_G \end{cases} \tag{3.B.10}$$

by Assumption 3.4.14. □

PROOF OF PROPOSITION 3.4.2 Asymptotic normality follows immediately from Proposition 3.4.1 and Assumption 3.4.15. To see

$$\text{rank}[\Sigma^*(\theta_0,\eta_0)]=p_1,\quad r_D=r_G \tag{3.B.1}$$

note that $Q(\theta_0,\eta_0)$ is a $p_1\times m$ matrix with rank $p_1$and $\text{rank}[[\mathbb{I}_{m\times m},-J_\eta(\theta_0,\eta_0)g(\eta_0)^{-1}]]\geq\text{rank}[\mathbb{I}_{m\times m}]=m$ so that by Sylvester's rank inequality,

$$\text{rank}[T(\theta_0,\eta_0)]\geq p_1+m-m=p_1\,. \tag{3.B.2}$$

□

PROOF OF PROPOSITION 3.4.4 Suppose either $r_D\leq r_G$, or $r_D>r_G$ and $q\geq m$. Then,

$$n^{\min(r_\theta,r_\eta)}s_n^*(\tilde{\theta}_n^0,\hat{\eta}_n)\xrightarrow[n\to\infty]{L}\text{N}[0,\Lambda(\theta_0,\eta_0)] \tag{3.B.1}$$

where

$$\text{rank}[\Lambda(\theta_0, \eta_0)] = p_1 \,. \tag{3.B.2}$$

Furthermore,

$$\hat{\Lambda}_n \xrightarrow[n\to\infty]{\mathsf{p}} \Lambda(\theta_0, \eta_0) \tag{3.B.3}$$

and $\text{rank}[\hat{\Lambda}_n] = p_1$, hence

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \, \psi) = n^{2\min(r_\theta, r_\eta)} \, s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n)' \Lambda^*(\theta_0, \eta_0)^{-1} s_n^*(\tilde{\theta}_n^0, \hat{\eta}_n) \xrightarrow[n\to\infty]{L} \mathscr{X}^2(p_1) \,. \tag{3.B.4}$$

Suppose $r_D > r_G$ and $q < m$. Assumption 119 combined with 154 implies that

$$(\hat{\Lambda}_n)^- \xrightarrow[n\to\infty]{\mathsf{p}} (\Lambda(\theta_0, \eta_0))^- \tag{3.B.5}$$

by Stewart (1969). Then,

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \, \psi) = n^{2r_G} \, s_{G,n}(\theta_0, \eta_0)' (\Lambda(\theta_0, \eta_0))^- s_{G,n}(\theta_0, \eta_0) + o_P(1) \tag{3.B.6}$$

where

$$s_{G,n}(\theta_0, \eta_0) = Q(\theta_0, \eta_0) J_\eta(\theta_0, \eta_0) g(\eta_0)^{-1} G_n(\eta_0) \xrightarrow[n\to\infty]{L} \text{N}\,[0, \Lambda(\theta_0, \eta_0)] \tag{3.B.7}$$

so that it follows from Theorem 1 of Moore and Spruill (1975),

$$EC(\tilde{\theta}_n^0, \hat{\eta}_n; \hat{\Lambda}_n \, \psi) \xrightarrow[n\to\infty]{L} \mathscr{X}^2(p^*) \tag{3.B.8}$$

where

$$p^* = \text{rank}[\Lambda(\theta_0, \eta_0)] = \text{rank}[\Pi(\theta_0, \eta_0)] \,. \tag{3.B.9}$$

□

Chapter 4

Dynamics of distributions: earnings, income and wealth

4.1. Introduction

Economic agents differ in their preferences, initial endowments and skills and such individual heterogeneity results in distributions of socioeconomic choices and outcomes which evolve over time. The distributions of relevant microeconomic statistics, such as labor earnings, investment, and expenditures, can be well-approximated using data collected from large scale surveys conducted by public and academic institutions, and by corporations. These surveys are often periodically updated and thus present an avenue to gain insight from transition dynamics of the distributions of microeconomic variables.

Comparisons of earnings (as well as income and wealth) distributions are made along several lines of inquiry. One is the question of stochastic dominance (Anderson (1996), Davidson and Duclos (2000, 2013), Barrett and Donald (2003)). First-order stochastic dominance implies that one distribution provides larger earnings for the same population quantile, either as a result of larger aggregate earning or of a more equitable distribution. Comparison of distributions over time can be made in terms of achieving stochastic dominance. Another important characteristic is inequality in the distribution (usually considered as deviation from the uniform) since inequality has a significant impact on social cohesion and stability. The changes in inequality over time impact well-being in a society; recently

documented rise in inequality is considered to be a contributing factor to social unrest and political polarization (Alesina and Perotti (1996)). Another important aspect of the changes in the income distribution is the dynamic of the lower quantiles of the distribution (e.g. quintiles) that are often related to poverty and how they cross the poverty line. (Sen (1976), Foster, Greer and Thorbecke (1984), Atkinson (1987), Davidson and Duclos (2000), Diouf and Dufour (2005)). Often applied transform into aggregate statistics or measures captures only some limited aspect of the data distribution. For example, dynamics of the average household income do not provide any information on income dispersion. Evolution of values of inequality measures, such as the Gini coefficient, provides evidence on changes in rise in inequality but not on what types of distributional shifts contribute to the observed trend of inequality. On the other hand, investigating directly the dependence structure of processes of probability distributions, as random functional elements, provides a fuller picture.

In the literature, dynamics of the income distribution have been studied extensively. Thomas Piketty and his coauthors study evolution of income inequality through certain characteristics of income distributions, such as the top 10% income share (Piketty and Saez (2003, 2014), Atkinson, Piketty and Saez (2011), Alvaredo, Chancel, Piketty, Saez and Zucman (2017)). But, they are primarily interested in the upper tails of the income distribution and thus only focus on particular features of the distribution. Some authors use panel data on individual income and study key features of income dynamics, such as income persistency, through structural parameters of their individual income process model (Kremer and Chen (2002), Bourguignon, Ferreira and Lustig (2004), Nirei and Souma (2007), Altonji, Smith Jr and Vidangos (2013), Guvenen, Karahan, Ozkan and Song (2021)). The analysis relies on availability of individual income history and thus is more suitable for studying relatively short-run dynamics. The limitation of this analysis is the parametric specification of individual processes which may deviate from the true data generating process.

To derive general distributional dynamics, the rich framework and tools of functional data analysis are applied. In functional time series, stochastic processes are typically treated as elements in vector spaces of functions, such as Banach spaces, in particular, Hilbert spaces (Bosq (2000)). The properties of separable Hilbert spaces and operators in Hilbert spaces are well suited to extend the conceptual base of studies of time series from Euclidean

spaces to infinite-dimensional spaces (Bosq (2000), Beare, Seo and Seo (2017) Beare and Seo (2020)). However, random functions of interest may not easily fit into the format of linear processes on Hilbert spaces, in which case application of powerful functional analysis tools, such as the Beveridge-Nelson decomposition, functional principal component analysis, and functional regression, is not warranted for classes of functions, such as distribution or density functions, as those classes are not closed under addition and scalar multiplication in a L^2 space. The existing work on stochastic processes of probability distributions deals with this challenge involving densities in various ways. Petersen and Müller (2016) consider log hazard and log quantile transformation of densities to map them into appropriate Hilbert spaces. Egozcue, Díaz-Barrero and Pawlowsky-Glahn (2006) construct a Hilbert space of bounded probability densities with a compact support from the Aitchison inner product. For more detailed reviews, see Petersen and Müller (2019).

A different approach was recently applied by Chang, Kim and Park (2016), who considered a fully nonparametric model for densities of earnings and examined the corresponding stochastic process of demeaned densities in the Hilbert space with the aim of discovering whether the process exhibits persistence. If the earnings distribution process were mean reverting, and characterized by short-term dynamic features only, then extracting these features (even possibly approximating via a parametric model) would have been sufficient to fully characterize earnings dynamics. If, however, the process exhibits persistence which was assumed to be characterized by a unit root subspace spanned by some eigenfunctions, this would signal that the evolution of the process is driven by the stochastic persistent features. Those features are reflected in the dimension of the eigenspace and the eigenfunctions that span this space.

The fundamental limitation of much of the existing work is that only absolutely continuous distributions are considered. In practice, mass points are commonly observed in data distributions and have significant economic interpretations. Thus, focusing only on the continuous part of the distribution by trimming the mass points out may result in significant loss of information. In the context of income distribution, Saez (2010), among others, confirms the existence of mass points in the US tax return data, possibly related to the Earned Income Tax Credit.Furthermore, to embed the classes of transformed functions into a separable Hilbert space bounded support of the class of distributions was assumed.

The contribution of this paper is three-fold. Firstly, we extend the existing analysis

of earnings dynamics to consider the full set of earnings data, including mass points at zero and at the top-coded values; as well we allow for other possible mass points and possible singularity in the distribution and take into account the dynamics of the support of the distribution without assuming that the supports are uniformly bounded. We examine several stylized examples to illustrate the impact of the support of the distribution and of the possible mass points. We demonstrate that trimming the data could result in misinterpreting the stochastic properties of the process of distributions. Our second contribution is a new transformation into the $L^2[0,1]$ space that accounts for mass points and the varying support of the distribution. We demonstrate the implication that the results will have on the density function (if it exists) to be able to compare with Chang et al. (2016). Third, we provide a direct comparison with the empirical results by Chang et al. (2016) by using the same data set (extended in time). Our test results indicate (similarly to Chang et al. (2016)) that the dimension of the unit root subspace is 2, although with the full data set we get much stronger statistical support for this conclusion. Where the support of the data set was constrained by the top-coded values, a constant that was shifted up once in 1998, the first eigenfunction appeared to be mostly related to this institutional feature. When we adjusted the range of the distribution to be more flexible (reflecting the dynamics of the top average 10%), the behavior of the eigenfunctions changed quite dramatically. This leads us to conclude that the dynamics of the upper quantiles is what drives the persistence of the whole distribution of earnings.

The new transformation approach we propose embeds the probability distributions into $L^2[0,1]$ as linear elements. To this end, we consider transformation of stochastic probability measures into scaled measures, which we call *TZ transformation.* We derive conditions under which TZ transformation is invertible. Then, various methods of functional data analysis, such as functional principal component analysis and functional regression, may be applied to the transformation of the original stochastic sequence to investigate their dependence structure.

Our approach does not require panel data, i.e. measurements of the same individuals over time but functional time series data on the income distribution in a given population. Such data are typically available from tax revenue agencies. Admittedly, this only provides a framework to investigate the dynamics of the aggregate distributions and does not attempt

to characterize how measurements of each unit in a population evolve over time[1]. However, understanding the dependence structures of the aggregate distributions still may lead to policy relevant implications on, for example, inequality and poverty measurements.

The rest of the paper is organized as follows: Section 4.2 presents the framework and main results concerning our transformation approach. In Section 4.3, we discuss construction of stochastic processes of TZ transformed functions and persistency of such processes in terms of the Beveridge-Nelson decomposition. Section 4.4 discusses inference for TZ transformed measures in the time series context and asymptotic properties of the persistency test by Chang et al. (2016) in our setup. Monte Carlo simulations are presented in Section 4.5. In our empirical application, persistency of earning dynamics in the U.S. is investigated in Section 4.6. Section 4.7 concludes.

4.2. TZ transformation

4.2.1. Notation

Denote by $(\Omega, \mathscr{F}, \mathbb{P})$ the underlying probability space, where Ω is the sample space, $\mathscr{F}$ is a σ-algebra of subsets of Ω, and $\mathbb{P}$ is a probability measure on the measurable space $(\Omega, \mathscr{F})$. The scalar product and norm on a L^2 space are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$, $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}, \mathbb{R}_{++} := \{x \in : x > 0\}$ correspond to the extended, nonnegative, and positive real lines, respectively. With some abuse of notation we shall represent functional elements of L^2 as functions $f(x)$ even though they may not be pointwise defined (are equivalence classes of functions); we do this to better track the possible variable transformations.

[1]Note that when one considers a stochastic process of probability distributions, different concepts of dependence arise. Suppose that for each period t, we observe a measurement $x_{i,t}$ of each unit i in the population $[0,1]$. Let F_t be the distribution of $x_{i,t}$ at period t. Then, there are (1) dependence between the distributions at different times t and t' ($t \neq t'$), i.e. F_t and $F_{t'}$, (2) (cross-sectional) dependence between measurements of the same unit i at different times, i.e. $x_{i,t}$ and $x_{i,t'}$ (3) (intertemporal) dependence between measurements of different units i and i' ($t \neq t'$) at the same time t, i.e. $x_{i,t}$ and $x_{i',t}$. In particular, note that the first type of dependence may not provide information regarding (2) cross-sectional dependence of each unit.

4.2.2. Overview

Probability distributions are bounded below by 0 and above by 1 and thus do not live in a linear but convex space. This feature of distribution functions deters application of functional analysis techniques based on representation of linear elements as weighted sums of orthogonal functions. Note that density functions, when they exist, also face an analogous challenge (Petersen and Müller (2016), Beare (2017)).

We propose a representation of probability distributions transformed into elements in $L^2[0,1]$; this is done via a transformation that we call a TZ transformation. The aim of this transformation is to account for a number of distributional features. We mainly focus on the features of distributions of earnings, income and wealth, but similar characteristics can be displayed by other distributions of interest. One feature that we emphasize is the support of the distribution. The changing support may be neglected when the distribution is represented over population shares (as e.g. income) or may be misrepresented through truncation or censoring. Yet the support and its dynamics provides important information regarding the dynamics of top incomes, it is important for establishing stochastic dominance and obtaining better information on inequality. Theoretically, the support of the distribution and changes in the support affects the changes in the Wasserstein distance between the distributions at different time points. In the distributions that we consider the support is typically bounded (stochastically) from above and its lower bound is typically zero; then only one parameter characterizes the support. Another feature that we consider is mass points in the distribution. Typically there is a mass point at zero in the distributions considered, but other mass points are possible. There is also some discussion about scalability of e.g. income and wealth distributions that implies a fractal or other singular structure entailing deviation from absolute continuity. The possible mass points and singularities mean that we cannot employ transformations that involve differentiation such as demeaned density Chang et al. (2016) or log quantile density and log hazard transformations Petersen and Müller (2016). Those approaches are only applicable to absolutely continuous distributions while ours accommodates a larger class of distribution involving mass points and singularity.

The transformation makes it possible to study properties of a time series of probability distributions through their representations in $L^2[0,1]$. In particular, we apply functional principal component analysis to stationary sequences of TZ transformed measures to study the stochastic functional processes. Nonstationarity of stochastic processes of probability

densities is investigated by the unit-roots test based on the rank of the unit-root subspace in Chang et al. (2016). We show that the restrictions that produce absolutely continuous distributions such as trimming away of mass points or censoring may misrepresent the dynamics of the process and, in particular, miss possible non-stationarity. Here, we consider a TZ-transformed process of distributions through the dynamics of processes in $L^2[0,1]$. The TZ transformation is invertible under general conditions and thus any operation applied to the transformed measure can be evaluated in terms of the original probability distribution.

Subsection 4.2.3 introduces a few stylized examples and provides an intuition on TZ transformation.

4.2.3. Intuition for the transformation

We start with a discussion of how to build up a transformation of a distribution function into $L^2[0,1]$ that will take account of the features that characterize distributions of earnings, income and wealth, such as support, and possible mass points and singularity. The class of functions of interest include univariate distribution functions and possibly linear combinations of such functions. Since distribution functions are nondecreasing, the functions of interest are functions of bounded variation. Denote the class of univariate functions of bounded variation by Θ. For a nondecreasing bounded function on $\mathbb{R}$ define its support:

Definition 4.2.1 *Let M be a nondecreasing function on $\mathbb{R}$. The support $supp(M)$ of M is defined as*

$$supp\,(M) = [L_M, U_M] \tag{4.2.1}$$

where

$$L_M = \sup\{x : M(x) \leq \inf M\}, : U_F = \inf\{x : M(x) \geq \sup M\}. \tag{4.2.2}$$

Thus, for example, the typical income distribution F at some point in time will have support on the interval $[0,u]$ where u will be the top income for that time. It could also be the distribution of top-coded data on incomes, where u would represent the censoring value.

Since a function of bounded variation can always be represented as a difference of two non-decreasing functions, its support is the union of the non-intersecting parts of supports of the non-decreasing functions that define it.

We shall illustrate the properties and the intuition behind the proposed transformation by the stylized example of the typical baseline distribution, the uniform.

Example 4.2.1 The distribution F is uniform on $[0,u]$, denoted by $U[0,u]$:

$$F(x) = \begin{cases} \frac{x}{u} & \text{if } x \in [0,u] \\ 0 & \text{if } x < 0; \\ 1 & \text{if } x \geq u. \end{cases} \tag{4.2.3}$$

The uniform distribution can be considered as a baseline distribution for income, wealth and earnings distributions. Distributions with differing supports differ, in that a distribution with u_1 stochastically dominates a uniform with $u_2 < u_1$. As we introduce the features of the proposed TZ transform, we discuss how Example 4.2.1 is impacted.

The idea of the TZ transformation is to construct a map that will transform the distribution function defined on its support into an element of $L^2[0,1]$. Most of the transformations to represent distributions (densities) via elements in a separable Hilbert space, such as L^2 on a bounded support, assume that the support of the distribution is within given bounds, an assumption that we relax by transforming the support into $[0,1]$. For $F(x)$ with support $[0,u]$, the simple linear transformation:

$$\Lambda(x) = x/u \equiv y. \tag{4.2.4}$$

defines the Z transform as

$$Z(F) = \tilde{F}(y) \in L^2[0,1] \tag{4.2.5}$$

by

$$\tilde{F}(y) = F(uy). \tag{4.2.6}$$

The image under such a transformation will not distinguish between different uniform distributions, $F(x) = U[0,u]$ as any such distribution will be mapped into $U[0,1]$. To restore the distinction which may be entirely due to the different support, we further define for some $\gamma > 0$ (with usually γ depending on F; $\gamma = \gamma(F)$) the transformation $\tilde{T} : L^2[0,1] \to L^2[0,1]$ by multiplying with the value of γ:

$$\tilde{T}(\tilde{F}) = \tilde{W}(y) = \gamma \tilde{F}(y). \tag{4.2.7}$$

We shall later consider a γ distinct from u but to simplify the example here, let $\gamma = u$ for the function F. For the Example 4.2.1, the transformation provides distinct images. For this case, there is an immediate interpretation in terms of Wasserstein distance (Vaserstein (1969), Givens and Shortt (1984)). For two uniform distributions, $F_1 = U[0,u_1]$ and $F_2 = U[0,u_2]$ with $u_1 < u_2$, the Wasserstein distance between the two is $\frac{u_2-u_1}{2}$. The transformation gives $\tilde{W}_i(y) = u_i I[0,1]$, $i = 1,2$, and the norm $\left\|\tilde{W}_2(y) - \tilde{W}_1(y)\right\|$ is proportional to the Wasserstein distance.

The $L^2[0,1]$ function obtained by transforming a distribution function as above is non-negative and monotone. To define a more general transformation T, apply a functional shift with some (typically monotonically nondecreasing) function $\lambda(y) \in L^2[0,1]$.

Thus, define a transformation

$$T(\tilde{F}) = W \in L^2[0,1] \tag{4.2.8}$$

by

$$W(y) = \tilde{W}(y) - \lambda(y). \tag{4.2.9}$$

Then, the mapping $T \circ Z$ transforms the distribution F into an element of $L^2[0,1]$, that could take negative as well as positive values and need not be monotone, all depending on how the function $\lambda(y)$ is specified. Thus, the Z transform serves to represent the distribution on $[0,u]$ by mapping it to one with support in $[0,1]$. The T transform scales the distribution by the value of some functional and possibly applies functional shifts, providing non-positive and non-monotonic transforms.

4.2.4. Transformations for the process of distribution functions.

Consider a stochastic sequence of distribution functions $\{F_t(x)\}_{t\in N}$ with (random) support: $\text{supp}(F_t) = [0,u_t]$; and a functional $\gamma_t = \Gamma(F_t)$. For any sequence of distribution functions, almost every point (except possibly a countable set) is a point of continuity for all the functions in the sequence. Then, for a random sequence and a point of continuity x_c, the sequence $F_t(x_c)$ is a random variable (between 0 and 1); expectation $\mathbb{E}F_t(x_c)$ always exists; define $\mathbb{E}F_t(x)$ as a càdlàg (right-continuous and bounded on the left) function that takes values $\mathbb{E}F_t(x_c)$ at every x_c. For $T \geq 1$, denote the average $\frac{1}{T}\Sigma F_t(x)$ by $\hat{\mathbb{E}}_T[F_t(x)]$.

Transformation 0. . The basic transform of the distribution function into a non-negative non-decreasing function $W^0(y)$ in $L^2[0,1]$ is given by

$$W^0(y) = \gamma F(uy) \tag{4.2.10}$$

Remark 4.2.1 We consider here the Z transform that is governed by the upper bound of the support of the distribution F_t, however, more generally u_t could be a variable that differs from this upper bound.

Transformation 1. .

$$\tilde{W}_t^1(y) = F_t(u_t y) - \mathbb{E}F_t(u_t y); W_t^1(y) = \gamma_t \tilde{W}_t^1(y). \tag{4.2.11}$$

This demeans the Z transformed distribution, then applies the T transform.

In finite sample we replace the expectation function by the empirical expectation and $\hat{W}^1(y) = \gamma_t \left[F_t(yu_t) - \hat{E}_T(F_t(yu_t))\right]$. For the rest of this section we do not distinguish between $\mathbb{E}[\cdot]$ and $\hat{\mathbb{E}}_T[\cdot]$ and ignore the possible "hat" in notation. This transformation is related to the demeaned density transformation applied in Chang et al. (2016).

Expectation of $W_t^1(y)$ may differ from zero.

Transformation 2. .

$$W_t^2(y) = \gamma_t F_t(u_t y) - \mathbb{E}\gamma_t F_t(u_t y). \tag{4.2.12}$$

Here, the ***Transform 0.*** is applied, followed by demeaning.

The next lemmas provide some characterizations and relations between the different transformations.

Lemma 4.2.1 The transformations are related by

$$W_t^2(y) = W_t^0(y) - \mathbb{E}W_t^0(y); \tag{4.2.13}$$

$$W_t^1(y) = W_t^2(y) + \left(\hat{\mathbb{E}}_T[\gamma_t F_t(u_t y)] - \gamma_t \hat{\mathbb{E}}_T[F_t(u_t y)]\right). \tag{4.2.14}$$

When γ_t is a constant, $W_t^1(y) = W_t^2(y)$.

Next, we note the expectations.

Lemma 4.2.2 *Assume that $\mathbb{E}\gamma_t$ exists. Then,*

(a) For $W_t^1(y)$ expectation $\mathbb{E}W_t^1(y) = \mathbb{E}\gamma_t F_t(u_t y)) - \mathbb{E}\gamma_t \mathbb{E}F_t(u_t y) = \mathrm{Cov}(\gamma_t, F_t(u_t y))$; the expectation is zero if $\gamma_t, F_t(u_t y)$ are uncorrelated.

(b) The expectation of $W_t^2(y)$ is zero: $\mathbb{E}W_t^2(y) = 0$.

To establish the boundary values for transforms of absolutely continuous distributions recall that in the absolutely continuous case $F_t(0) = 0$; while always $F(u_t) = 1$, when u_t is the upper bound of support of F_t

Lemma 4.2.3 [Boundary values for transforms of absolutely continuous distributions]

Suppose that $F_t(x)$ is absolutely continuous for every t and u_t is the upper bound of support of F_t Then each of the transforms is differentiable in y and satisfies:

(a) For the basic transform (**Transform 0.**), $W^0(0) = 0$; $W^0(1) = \gamma$.

(b) For **Transform 1.**, $W_t^1(0) = W_t^1(0) = 0$ and $W_t^1(1) = W_t^1(1) = 0$.

(c) For **Transform 2.**, $W_t^2(0) = 0$, $W_t^2(1) = \gamma_t - \mathbb{E}\gamma_t$.

We see that the functionals of F that define the properties of the transform are

$$\alpha = F(0), \text{ mass at zero;} \tag{4.2.15}$$

$$\gamma = G(F), \text{ the scaling functional;} \tag{4.2.16}$$

and u, that is usually defined as the upper bound of the support, however, it could also be some othe quantile, or point in $\mathbb{R}$, e.g. such that $F(u) = 1$, or $F(u) = \beta$. In many examples $\gamma = u$. Given a sequence $\{W_t^+\}$ of non-negative non-decreasing cadlag functions on $[0,1]$ define $\gamma_t = W_t^+(1)$; $\alpha_t = \gamma_t^{-1} W_t^+(0)$, then obtain

$$F_t(u_t y) = W_t^+(1)^{-1} W_t^+(y). \tag{4.2.17}$$

Then $W_t^+(y)$ coincides with ***Transform 0.*** of $F_t : W_t^0(y) = W_t^+(y)$. If u_t is known, then $F_t(x)$ is uniquely defined as

$$F_t(x) = W_t^+(1)^{-1} W_t^+\left(u_t^{-1}x\right). \tag{4.2.18}$$

This would hold when $u_t = \gamma_t$ and is given by $W_t^+(1)$. Thus, under some definitions of the underlying functionals, there is one-to one correspondence between the process of distributions and the process $\left\{W_t^+(y)\right\}$ in $L^2[0,1]$ that represents ***Transform 0.***, the demeaned process $\left\{W_t^+(y)\right\}$ provides the process of ***Transform 2.*** for $\{F_t\}$.

This invertibility of the TZ transform implies that we can study the dynamics of the process of distributions $\{F_t(x)\}$ through the dynamics of processes of functions with bounded variation in $L^2[0,1]$. If the distributions are in the class of absolutely continuous distributions, then mass at zero is excluded.

4.2.5. Process of transformed distribution functions and of densities in $L^2[0,1]$.

When the distribution functions are absolutely continuous, the transforms in $L^2[0,1]$ are differentiable functions. Here we are interested in the cases when the derivatives can be defined as elements in $L^2[0,1]$ as this will provide a direct link to the investigation in Chang et al. (2016). Denote a dual of a linear topological space of linear continuous functionals on V, by V^*. The rigged Hilbert space introduced in Gel'fand and Vilenkin (1964) in the special case of interest here for differentiable L^2 functions [Hunter and Nachtergaele (2001)]

$$H^1([0,1]) \subset L^2[0,1] \subset H^1([0,1])^* \tag{4.2.19}$$

gives $H^1([0,1])$ as the subspace of $L^2[0,1]$ of differentiable functions with square-integrable derivatives. The dual space $H^1([0,1])^*$ extends differentiation to all of $L^2[0,1]$, by providing weak derivatives for elements of $L^2[0,1]$. As shown in Hunter and Nachtergaele (2001) for the differentiable elements of $L^2[0,1]$ represented by differentiable functions in $H^1([0,1])$, a week derivative coincides with the ordinary pointwise derivative[2] .

[2]Suppose that $G \in L^2[0,1]$ and consider any $\psi \in H^1([0,1])$, then $\psi' \in L^2[0,1]$ by continuation of $H^1([0,1])$. The functional $\langle G, \psi' \rangle$ is well-defined. Then, there exists $\delta \in H^1([0,1])^*$ that is fully defined

Proposition 4.2.4 *Suppose that the function $W(y) \in H^1([0,1]) \subset L^2[0,1]$ is differentiable with derivative $w(y) \in L^2[0,1]$. Suppose that for some projection operator $\Pi : L^2 \to L^2$ the image of the function $w(y)$ is $w_\Pi(y) \in L^2[0,1]$ and the image of $W(y)$ is $\Pi W(y)$. Then as elements in $L^2[0,1]$ the derivative of the projection, $(\Pi W(y))' = \Pi(w(y))$, the projection of the derivative.*

This result implies that generally for a process of derivatives (e.g. demeaned densities) in $L^2[0,1]$ that can be viewed as derivatives of functions in $H^1([0,1])$, a decomposition by projections onto subspaces is the same for the process of the differentiable functions and their derivatives, e.g. demeaned distribution functions and corresponding demeaned density functions. Thus, consider, for example the (demeaned) square integrable density function $w(y)$ (as in Chang et al. (2016)) on a bounded support. Since the support for all such functions considered there is uniformly bounded, it can be assumed without loss of generality that it is $[0,1]$; the upper bound of supports is 1 for all $F_t(x))$ thus $w \in L^2[0,1]$ and the corresponding function to which w integrates is $W(y) \in H^1([0,1]) \subset L^2[0,1]$. $W(y)$ then is the demeaned probability distribution function on the support $[0,1]$. The implication of the Proposition 4.2.4 is that if a projection Π on a subspace is established for the demeaned density process in $L^2[0,1]$, then the corresponding projection of the demeaned distribution functions in $H^1([0,1]) \subset L^2[0,1]$ provides the same projection for the demeaned density. Then, the dimension of the projected space for the demeaned absolutely continuous distribution function and for its demeaned density are the same and the spaces are spanned by the same functions. Thus our transformations provide a natural generalization of the approach in Chang et al. (2016).

However, when the distribution is not absolutely continuous, the usual approaches to examine its dynamics through that of densities can provide misleading results. If the distribution has mass points at the boundary of support, e.g. at zero or at the top end, to force absolute continuity on this distribution by trimming down to some low quantile and up to some quantile or top code point is often applied. In fact, trimming can be applied to absolutely continuous distributions as well as to ones with singularities and mass points.

---

by

$$\langle \delta, \psi \rangle = -\langle G, \psi' \rangle$$

for any $\psi \in H^1([0,1])$. This δ is called a weak derivative of G. If G were a differentiable function; $G \in H^1([0,1])$, then $\delta = G'$.

The following example demonstrates that trimming can result in a misrepresentation of the dynamics of the process of distributions.

Example 4.2.2 The distribution F is a mixture of $U[0,\gamma]$ and a mass point at $\bar{\kappa}\gamma$ with $\bar{\kappa} \geq 1$.

$$F(x) = \begin{cases} (1-\alpha)\frac{x}{\gamma} & \text{if } x \in [0,\gamma] \\ 0 & \text{if } x < 0; \\ (1-\alpha) & \text{if } \gamma \leq x < \bar{\kappa}\gamma; \\ 1 & \text{if } \bar{\kappa}\gamma \leq x. \end{cases} \tag{4.2.20}$$

Define the stochastic process $\{F_t(x)\}$ by assuming that α and γ are constant over t, but $\{\bar{k}_t\}$ represents a random process on $[1,\infty]$. By trimming a top α percent and rescaling, the uniform distribution $U[0,\gamma]$ is obtained. By ignoring the top part of the distribution one would claim stationarity when in fact it may not hold.

In this example the distribution is singular (mass point), but the impact of trimming will be the same for an absolutely continuous distribution that is a mixture of two distinct distributions, one up to γ, and a different one with support over $x \geq \gamma$. Applying the TZ transform to the trimmed distribution would also provide a series of constant functions. Only by taking into account the whole distribution one can evaluate the true dynamics.

4.3. Features of the stochastic process of TZ transformed distributions in $L^2[0,1]$

Section 4.2 introduced TZ transformations,which yield a representation of a probability distribution as an element of $L^2[0,1]$. In this section, we study more formally the features of stochastic processes of transformed distribution functions. First, we show that TZ transformation preserves stationarity when it is applied to a stochastic sequence and thus (non)stationarity of the original sequence of probability measures can be examined through the corresponding sequence of TZ-transformed measures in $L^2[0,1]$. Then following Chang et al. (2016), we consider the Beveridge-Nelson decomposition and characterize the persistency of a functional process of distributions in terms of the dimension of the unit-root subspace for the transformed process.

4.3.1. Stochastic sequences in $L^2[0,1]$ and stationarity of the transform

A $L^2[0,1]$-valued random element is defined to be a measurable mapping from $(\Omega, \mathscr{F})$ to $(L^2[0,1], \mathscr{B}_{L^2[0,1]})$ where $\mathscr{B}_{L^2[0,1]}$ is the Borel σ-algebra of $L^2[0,1]$ (recall that Borel $\sigma-$algebra is spanned by all open sets of the separable normed space). As in Bosq (2000),, define a discrete time stochastic process as a sequence of such measurable mappings indexed by t, where the index set could be finite: $\{t\} = \{1,..,T\}$, or infinite $\{t\} = N$, or Z with N - the set of natural numbers and Z of all integers.

Let $\{F_t\}$ be a stochastic sequence of distributions, or more generally, uniformly bounded functions in the metric space of bounded functions on $\mathbb{R}$ and γ_t, u_t be some continuous functionals of F_t, $\{\lambda_t\}$ be a stochastic sequence of non-decreasing functions in $L^2[0,1]$. Then, since any continuous mapping is measurable, the mapping $F_t(x) \to W_t(y)$ defined by

$$W_t(y) = \gamma_t F_t(u_t y) - \lambda_t(y) \tag{4.3.21}$$

is measurable and $\{W_t\}$ is a stochastic sequence in $L^2[0,1]$.

A stochastic sequence $\{F_t\}$ is said to be **strictly stationary** if for any $m \in \mathbb{N}_+, h \in \mathbb{Z}$,

$$(F_{t_1}, F_{t_2}, \ldots, F_{t_m}) \sim (F_{t_1+h}, F_{t_2+h}, \ldots, F_{t_m+h}). \tag{4.3.22}$$

Then, the following lemma provides a sufficient condition for the the transformed process $\{W_t\}$ to be stationary.

Lemma 4.3.1 *(a) For the sequence $\{W_t\}$ of transformed distribution functions in (*4.3.21*) to be strictly stationary it is sufficient that the sequence $\{F_t, \lambda_t\}$ be strictly stationary. (b) If the sequence $\{W_t\}$ of transformed distribution functions is given by* **Transformation 1.** *or* **2.***, then $\{F_t, \lambda_t\}$ forms a strictly stationary sequence.*

Lemma 4.3.1 indicates that stationarity of $\{F_t\}$ can be examined through $\{W_t\}$. Stationarity of $\{F_t\}$ implies stationarity of the $\{\lambda_t\}$ sequence with λ_t as it appears in ***Transformation 1.*** and ***2.***. Thus, for those transforms, strict stationarity of the process of distributions provides strict stationarity of the transforms.

Weak stationarity of a process is defined via moment functionals and operators.

The **mean functional** $\mathbb{E}X$ of a $L^2[0,1]$-valued random variable X is an element of $L^2[0,1]$ such that

$$\langle \mathbb{E}X, f\rangle = \mathbb{E}\langle X, f\rangle, \qquad \forall f \in L^2[0,1]. \tag{4.3.23}$$

Note that if $\mathbb{E}\langle X, f\rangle$ exists for any $f \in L^2[0,1]$, $\mathbb{E}X$ exists by the Riesz representation theorem and is represented by an element in $L^2[0,1]$ (as it coincides with its dual).

A **mean stationary process** $\{W_t\}$ is such that $\phi_t = \mathbb{E}W_t \in L^2[0,1]$ exists and is constant for any t : $c_t = \text{const}$.

Consider the Hilbert space $L^2_{\mathbb{P}}[0,1] := L^2_{[0,1]}(\Omega, \mathscr{F}, \mathbb{P})$ of $L^2[0,1]$-valued random variables such that their second moment exists and define the inner product $(\cdot,\cdot)$ by

$$(X,Y) = \mathbb{E}[\langle X, Y\rangle], \qquad \forall X, Y \in L^2_{\mathbb{P}}[0,1]. \tag{4.3.24}$$

The following result derives a condition such that the ***TZ transform 1. and 2.*** $\{W_t^m\}$, $m = 1,2$ belongs to $L^2_{\mathbb{P}}[0,1]$.

Lemma 4.3.2 *Let* $\{W_t^m\}$, $m = 1,2$ *be a TZ transform of probability measures* F_t

If $\mathbb{E}\gamma_t^2$ *exists then* W_t^m *is a sequence in* $L^2_{\mathbb{P}}[0,1]$.

The results follows from showing that for W_t the second moment $\mathbb{E}\|W_t\|^2 = \mathbb{E} < W_t, W_t >$ exists. Note that $\mathbb{E}\|\mathbb{F}_t\|, \mathbb{E}\|\mathbb{F}_t\|^2$ exists always. If $\mathbb{E}[\gamma_t^2]$ exists, then $\mathbb{E}[\gamma_t F_t]$ exists.

A **second-order stationary process** $\{W_t\} \subset L^2_{\mathbb{P}}[0,1]$ is such that in addition to mean stationarity $\mathbb{E} < W_t, W_s >= \rho(|t-s|)\}$. Note that second-order stationarity of $\{F_t, \lambda_t\}$ does not generally imply second-order stationarity of $\{W_t\}$ as opposed to what is shown in Lemma 4.3.1 for strict stationarity.

4.3.2. Nonstationary process of the TZ transform and the Beveridge-Nelson decomposition

Here, we follow the general methodology that was developed in Chang et al. (2016) for the functional principal component analysis. In that paper, it was applied to a stochastic process of demeaned densities while we pursue the application to the TZ transformed distribution functions, generically denoted here W_t, in $L^2([0,1])$.

The space $L^2([0,1])$ is decomposed into two subspaces H_N and H_S:

$$L^2([0,1]) = H_N \oplus H_S. \tag{4.3.25}$$

The spaces H_S and H_N correspond to the stationarity and nonstationary subspaces, respectively, defined as follows: for any nonzero $v \in H$, the coordinate process,

$$\langle v, W_t \rangle$$

has a unit root for all $v \in H_N$, while it is stationary for all $v \in H_S$. Assume H_N is M-dimensional where $0 \leq M < \infty$. Denote by Π_N and Π_S the projections on H_N and H_S, respectively.

Assumption 4.3.1 *The process $\{W_t\}$ allows for the Beveridge-Nelson decomposition:*

$$W_t = W_{t-1} + u_t \tag{4.3.26}$$

where

$$u_t = \Phi(L)\varepsilon_t = \sum_{s=0}^{\infty} \phi_s \varepsilon_{t-s}. \tag{4.3.27}$$

satisfying the following: (i) $\sum_{s=1}^{\infty} s\|\phi_s\| < \infty$, (ii) $\Pi_N \phi(1)$ is of rank M and $\Pi_S \phi(1) = 0$, (iii) $\{\varepsilon_t\}$ is an i.i.d. sequence with mean zero and positive-definite variance Σ and satisfies $\mathbb{E}\|\varepsilon_t\|^p < \infty$ for some $p \geq 4$.

Then, by Chang et al. (2016), the decomposition holds for the process $\{W_t\}$:

$$W_t = \mathrm{W}_t^N + \mathrm{W}_t^S \tag{4.3.28}$$

where

$$W_t^N := \Pi_N W_t = \Pi_N \Phi(1) \sum_{i=1}^{t} \varepsilon_t - \Pi_N \bar{u}_t, \tag{4.3.29}$$

$$W_t^S := \Pi_S W_t = -\Pi_S \bar{u}_t, \tag{4.3.30}$$

and

$$\bar{u}_t = \sum_{i=0}^{\infty} \bar{\Phi}_i \varepsilon_{t-i} \qquad \text{and } \bar{\Phi}_i = \sum_{j=i+1}^{\infty} \Phi_j. \tag{4.3.31}$$

This representation provides the basis for functional principal component analysis in Chang et al. (2016), where the eigenvalues of the sample variance operator

$$Q^T = \sum_{t=1}^{T} W_t \otimes W_t \tag{4.3.32}$$

were shown in the stationary subspace to converge to the corresponding population eigenvalues (as in Bosq (2000)), while in the non-stationary subspace the convergence is in distribution to some functionals of Brownian motion (Theorem 3.3 of Bosq (2000)). The eigenvectors provide the convergence to the spaces spanned by them. A definitive characteristic of the process is the dimension M of the non-stationary subspace.

As in Chang et al. (2016) we consider the dimension M of the unit-root subspace in a test of the null hypothesis

$$\mathscr{H}_0(M) : \dim(H_N) = M \tag{4.3.33}$$

against

$$\mathscr{H}_1(M) : \dim(H_N) \leq M - 1 \tag{4.3.34}$$

assuming the knowledge of some upper bound on the dimension given by an integer $\bar{M}$. The test statistic is an estimated M-th generalized eigenvalue of the first M coordinate processes of W_t given by the long run variance, scaled by the sample size. It converges to a nuisance parameter free distribution under the null hypothesis $\mathscr{H}_0(M)$.

We next establish validity of the test of the dimension of the unit root subspace based on the finite-sample estimated TZ transformations.

4.4. Inference

4.4.1. Estimation of TZ transformed measures

Typically, a stochastic functional sequence is not directly observable and needs to be estimated. In particular, for a time series $\{F_t\}$ of probability distributions, the data is rep-

resented by random real-valued observations drawn from a realization $\{F_t\}_{t=1}^T$ where T is the length of the time series. In the context of income distribution, household survey or tax return data provide observations $\{x_{i,t}\}_{i=1}^{n_t}$ of labor earnings drawn from the population distribution F_t at a given time t.

More specifically, the data is generated in two stages of random sampling where (1) $\{F_t\}_{t=1}^T$ is drawn from some meta-distribution and then (2) F_t, $t = 1,\ldots,T$ generates a sample $\{x_{i,t}\}_{i=1}^{n_t}$ of size n_t. It is assumed that the second stage is independent of the first stage conditional on $\{F_t\}_{t=1}^T$. Thus, we may consider the probability space $(\Omega,\mathscr{G},\mathbb{P})$ to be a product space $(\Omega_1\times\Omega_2,\mathscr{G}_1\times\mathscr{G}_2,\mathbb{P}_1\otimes\mathbb{P}_2)$ where $\{F_t\}_{t=1}^T$ and $\{\{x_{i,t}\}_{i=1}^{n_t}\}_{t=1}^T$ are defined on $(\Omega_1,\mathscr{G}_1,\mathbb{P}_1)$ and $(\Omega_2,\mathscr{G}_2,\mathbb{P}_2)$, respectively. This setup is known as the two-stage sampled scheme and has been employed in the literature of learning of distribution (Póczos, Singh, Rinaldo and Wasserman (2013), Oliva, Póczos and Schneider (2013), Szabó, Sriperumbudur, Póczos and Gretton (2016)) as well as in Petersen and Müller (2016).

We assume that each random F_t has a bounded support $[0,u_t]$, $t = 1,\ldots,T$ and consider initially the TZ transform of the form

$$W_t^0(y) = \gamma_t F_t(u_t y), \quad y\in[0,1] \tag{4.4.35}$$

where γ_t, u_t are some functionals, e.g. quantiles of the distribution $F_t(x)$:

$$\gamma_t = \inf\{x : F_t(x)\le\alpha_\gamma\},\ u_t = \inf\{x : F_t(x)\le\alpha_u\}. \tag{4.4.36}$$

We have previously discussed the examples where $\alpha_\gamma = \alpha_u = 1$. The estimator $\hat{W}_t^0$ is defined as

$$\hat{W}_t^0(y) = \hat{\gamma}_t\hat{F}_t(\hat{u}_t y), \quad y\in[0,1] \tag{4.4.37}$$

where $\hat{F}_t$ is the empirical distribution given the sample $\{x_{i,t}\}_{i=1}^{n_t}$ generated from the distribution of F_t, and $\hat{u}_t,\hat{\gamma}_t$ are the corresponding sample quantiles:

$$\hat{\gamma}_t = \inf\{x : \hat{F}_t(x)\le\alpha_\gamma\},\ \hat{u}_t = \inf\{x : \hat{F}_t(x)\le\alpha_u\}. \tag{4.4.38}$$

when γ_t, u_t are defined as (4.4.36) . Thus, when $\alpha_\gamma = \alpha_u = 1$, we consider the sample maximum: $\hat{\gamma} = \hat{u}_t = x_{(n),t}$, where $x_{(k)}$ stands for a k-th order statistic.

Since, in contrast to the transformations in Chang et al. (2016) or Petersen and Müller

(2016), we allow the possibly nonstationary dynamics for the upper quantiles of the distribution, we do not require a uniform bound on the quantiles and uniformly bounded support for the distribution.

We next address the question of consistency of $\hat{W}_t^0(y)$ as an estimator of the transformation $W_t^0(y)$. In Chang et al. (2016), their Assumption 4.1 asserted uniform (over all $t = 1, ..., T$) convergence in L^2 norm for the density estimators. Here, we provide conditions on the relative growth of γ_t with T and that of T and n_t that guarantees such convergence for the TZ transforms. We do not require absolute continuity, but restrict the class of distributions we consider to enable us to provide a sufficient condition for the uniform convergence.

Definition 4.4.1 *Let $0 < \alpha \leq 1$, $0 < C < \infty$. $\mathscr{F}^{(\alpha,C,N_d)}$ is a class of univariate probability distributions such that the following holds:*

1. *any distribution $F \in \mathscr{F}^{(\alpha,C,N_d)}$ has support $[0, u_F]$ with $0 < u_F < \infty$;*
2. *any distribution $F \in \mathscr{F}^{(\alpha,C,N_d)}$ has a finite number of points of discontinuity not exceeding some $N_d < \infty$;*
3. *for $F \in \mathscr{F}^{(\alpha,C,N_d)}$ at any point of continuity x_0 there exists some $\varepsilon > 0$ and $C < \infty$ such that $x_0 - \varepsilon < x < x_0 + \varepsilon$ implies*

$$|F(x) - F(x_0)| \leq C\,|x - x_0|^{\alpha}. \tag{4.4.39}$$

Assumption 4.4.1 *The stochastic sequence of distribution functions $\{F_t\}_{t=1}^T$ is such that $F_t \in \mathscr{F}^{(\alpha,C,N_d)}$ for any t.*

The boundedness of support for each of the distributions could be relaxed at the expense of a more complex interplay between tail convergence and the transformation to a function supported on $[0,1]$; in light of the envisaged applications to distributions of income and wealth it seems reasonable to assume an upper bound on the maximal possible value at each point in time and thus we do not pursue the generalizations here. Any probability distribution has at most a countable number of points of discontinuity and the condition 2 further limits the set of discontinuity points to be finite. This condition could also be

relaxed. The restriction of Hölder-type conditions at points of continuity could accommodate absolutely continuous distributions with possibly locally unbounded densities as well as singular distributions such as fractals.

Example 4.4.1 The following are examples of probability distributions to illustrate the parts 1-3 of the definition of the class $\mathscr{F}^{(\alpha,C,N_d)}$.

1. Let $\{F_{D,t}\}$ be a sequence of discrete distribution with at most N_d mass points. Then, $\{F_{D,t}\} \subset \mathscr{F}^{(\alpha,C,N_d)}$ with arbitrary $\alpha \in [0,1]$ and $C > 0$.

2. Consider a sequence $\{U_t\}$ of uniform distributions where the support of U_t is $[0,u_t]$, with $u_t > \underline{u}$ lower bounded for some $\underline{u} > 0, t \in \mathbb{Z}$. Then, $\{U_t\} \subset \mathscr{F}^{(\alpha,C,N_d)}$ with $N_d = 0, \alpha = 1, C = \dfrac{1}{\underline{u}}$.

3. If $\{F_{P,t}\}$ were a sequence of Pareto distributions with the scale parameter 1 and the shape parameter $\beta_t, t \in \mathbb{Z}$ where $\beta_t \geq \underline{\beta}$ for some $\underline{\beta} > 0$. Then, $N_d = 0, \alpha = \min(\underline{\beta}, 1), C = 1$ but clearly the bounded support assumption does not hold. On the other hand, suppose a sequence of censored Pareto distributions $\{\bar{F}_{P,t}\}$ where $\bar{F}_{P,t}$ is a censored distribution of $F_{P,t}$ to the interval $[1,u_t]$ for some possibly random scalar $u_t \geq 1$, i.e.
$$\bar{F}_{P,t}(x) = \begin{cases} F_{P,t}(x) & x < u_t, \\ 1 & x \geq u_t. \end{cases} \tag{4.4.40}$$
Then, $\{\bar{F}_{P,t}\} \subset \mathscr{F}^{(\alpha,C,N_d)}$ with $N_d = 1, \alpha = \min(\underline{\beta}, 1), C = 1$.

We impose restrictions on the growth rate of $\sup_{1\leq t\leq T} \gamma_t$ as well as the convergence rates for estimated $\{\hat{\gamma}_t\}$ and $\{\hat{u}_t\}$.

Assumption 4.4.2 *Set α to be a positive constant in Assumption 4.4.1, $n(T) = \min_{1\leq t\leq T}\{n_t\}$ and*

$$\sup_{1\leq t\leq T} \gamma_t = O_p(\lambda_T), \quad \sup_{1\leq t\leq T} |\gamma_t - \hat{\gamma}_t| = O_p(\xi_T), \quad \sup_{1\leq t\leq T} (u_t - \hat{u}_t)^\alpha = O_p(\kappa_T) \tag{4.4.41}$$

with $\lambda_T, \xi_T, \kappa_T, T \in \mathbb{N}$ deterministic sequences of nonnegative values where as as $T \to \infty$, $n(T) \to \infty$ and

$$\xi_T = o_p(1); \frac{T}{\sqrt{n(T)}}\lambda_T = o_p(1); \lambda_T \kappa_T = o_p(1) \tag{4.4.42}$$

Assumption 4.4.2 requires that the cross-sectional sample sizes $n_t, t = 1, \ldots, T$ grow at a much faster rate relative to T, in particular, it is necessary that $n(T)$ grows at a faster rate than T^2. The smoothness α of distributions also plays a role: the smaller value of α entails that $n(T)$ needs to grow at a faster rate as deficient smoothness of distributions at points of continuity leads to difficulties in estimation.

Additionally, we assume cross-sectional independence of the sets $\{x_{i,t}\}_{i=1}^{n_t}$ of samples for each $t = 1, \ldots .T$.

Assumption 4.4.3 *For each $t = 1, \ldots, T$, $\{x_{i,t}\}_{i=1}^{n_t}$ is independent and identically distributed according to F_t.*

Note that any intertemporal dependence between any pair $(x_{i,t}, x_{i',t'})$, $i = 1, \ldots n_t$, $i' = 1, \ldots, n_{t'}$ for $t \neq t'$ is permitted and thus panel data is accommodated.

The Theorem below establishes uniform (in $L^2[0,1]$) consistency of $\hat{W}_t^0$ under the general specification of W_t^0.

Theorem 4.4.1 (Uniform L^2 convergence of $\{\hat{W}_t^0\}_{t=1}^T$ under general specification of γ_t) *Consider the TZ transform of the form (4.4.35) where $\{\gamma_t, u_t\}$ and their estimators $\{\hat{\gamma}_t, \hat{u}_t\}$ satisfy Assumption 4.4.2. Suppose further that Assumption 4.4.1-4.4.3 hold. Then,*

$$\sup_{1 \leq t \leq T} \|\hat{W}_t^0 - W_t^0\| = o_p(1), \qquad \frac{1}{T}\sum_{t=1}^{T} \|\hat{W}_t^0 - W_t^0\| = o_p(1) \tag{4.4.43}$$

holds.

Consistency of $\hat{W}_t^0$ is crucial as shown in Chang et al. (2016) to establish the validity of the statistic for the test for the dimension of the nonstationary subspace.

4.4.2. Asymptotic properties of the eigenvalue-based persistency test

We can now address the main questions regarding the dynamic process of the (TZ transformed) distributions, namely, how to establish possible persistence, the dimension of the

nonstationary subspace, and what are the drivers of this persistence. We apply the methodology developed in Chang et al. (2016) for an arbitrary stochastic process of functions in L^2 over some bounded support to the process of estimated ***TZ transforms 2.*** (we omit the superscript 2 from now on):

$$\tilde{W}_t = \hat{W}_t^0 - \frac{1}{T}\sum_{t=1}^{T} \hat{W}_t^0, \tag{4.4.44}$$

The test of the dimension for the H_N subspace starts with assuming a maximal possible dimension $\bar{M}$ and then proceeds with hypotheses

$$H_0(M) \,:\, \dim(H_N) = M \tag{4.4.45}$$

versus

$$H_1(M) \,:\, \dim(H_N) \leq M - 1 \tag{4.4.46}$$

for $M = \bar{M}, \bar{M} - 1, .., 1$ in descending order. If the null hypothesis is rejected for some M in this set, then the corresponding dimension of the space H_N is $M - 1$. When $M - 1 = 0$, there is no persistence in the process.

The test statistic utilizes the spectral decomposition of the sample variance operator

$$Q^T = \sum_{t=1}^{T} \tilde{W}_t \otimes \tilde{W}_t. \tag{4.4.47}$$

Define $(\lambda_i(Q_T), v_i(Q_T)),: i = 1, \ldots T$ to be the pairs of eigenvalues and eigenvectors of Q_T where $\lambda_1(Q_T) \geq \cdots \geq \lambda_T(Q_T)$. As discussed in Chang et al. (2016), the limit distribution of the test statistic $T^{-2}\lambda_m(Q_T)$ is not nuisance parameter free; in that paper an asymptotically nuisance parameter free test statistic is proposed; then the critical values can be obtained via simulation. We next describe their construction for the test statistic applied here to $\hat{W}_T$. First consider an arbitrary set of vectors $\{v_i,\ i = 1, ...M\}$ that span H_N; the specific choice of the vectors does not matter. For the vector

$$\tilde{z}_t = \left(\left(v_1, \hat{W}_t\right), \ldots, \left(v_M, \hat{W}_t\right)\right)' \tag{4.4.48}$$

and $Z_T = (z_1, ..., z_M)'$, set Q_M^T as $Q_M^T = Z_T' Z_T$.

Then define the long-run variance estimator

$$\tilde{\Omega}_T^M = \sum_{|i|\leq l} \bar{\omega}_l(i)\tilde{\Gamma}_T(i) \tag{4.4.49}$$

where $\bar{\omega}_l$ is a weight function and $\tilde{\Gamma}_T(i) = T^{-1}\sum_{t=1}^T \Delta\tilde{z}_t \Delta\tilde{z}'_{t-i}$.

The generalized eigenvalues for $\tilde{\Omega}_T^M$ are computed from a consistent estimator $\tilde{Q}_M^T$ of Q_M^T reweighted with respect to the consistent estimate $\tilde{\Omega}_T^M$ of Ω_M. Denote the estimated eigenvalues $\lambda_i(\tilde{Q}_T^M, \tilde{\Omega}_T^M)$.

The test statistic is defined by

$$\tau_T^M = T^{-2}\lambda_M(\tilde{Q}_T^M, \tilde{\Omega}_T^M) \tag{4.4.50}$$

where $\lambda_M(\tilde{Q}_T^M, \tilde{\Omega}_T^M)$ is the generalized eigenvalue of $\tilde{Q}_T^M$ with respect to $\tilde{\Omega}_T^M$. The null distribution of the limiting eigenvealues and eigenvectors is generated by eigenvalues and eigenvectors of

$$Q_M^* = \int_0^1 B_M^*(r) B_M^*(r)' dr, \tag{4.4.51}$$

with $B_M^*(r) = \Omega_M^{-1/2} B_M(r)$ representing the standardized Brownian motion distribution (with original covariance Ω_M). The consistency result for the test statistic is the same as in Chang et al. (2016), with the difference that rather than making the high level assumption on consistency of the functional estimators, we provided here in Theorem 4.4.1 a primitive sufficient condition to restrict the distribution class and an explicit condition on the dynamics of some of the functionals associated with the distributions that ensured consistency of our TZ transforms.

Theorem 4.4.2 *[Asymptotic distribution and consistency of τ_T^M]. If $\{W_t\} := \{W_t^0 - \mathbb{E}W_t^0\}$ permits the Beveridge-Nelson decomposition in (*4.3.28*) with $H_0(M)$ holding, and the conditions (*4.4.43*) in Theorem 4.4.1 hold, there is convergence*

$$\tau_T^M \xrightarrow{d} \lambda_M^* \tag{4.4.52}$$

where λ_M^* *is the smallest eigenvalue of*

$$Q_M^* = \int_0^1 B_M^*(r) B_M^*(r)' dr \tag{4.4.53}$$

and $B_M^*(r)$ *is the M-dimensional standard Brownian motion. Furthermore,*

$$\tau_T^M \xrightarrow{p} 0 \tag{4.4.54}$$

under $H_1(M)$.

In the following section, we evaluate the finite sample performance of the test in application to TZ transformed distributions with mass points.

4.5. Monte Carlo simulation

In this section we explore the potential of the test by Chang et al. (2016) applied here to TZ transformed distributions to uncover persistence and the dimension of the unit root space. We examine performance of the test through Monte Carlo experiments. The distributions are generated as mixtures of an absolutely continuous distribution and a mass. Two types of mass are examined here, one is a fixed mass at a varying point, the other is a mass at zero of varying weight. Each example has some relevance to earnings distributions. Indeed, the first is a stylized example of a discontinuity in the distribution that could be due to a labor supply response to an anticipated income means tested transfer (see the explanation in Zinde-Walsh (2008): case b of Example 1). The second type is a mass at zero that is present in distributions of earnings and wealth. As noted before, restricting attention to only absolutely continuous distributions would ignore the impact of such mass. The purpose of the simulation experiment is to verify that indeed the method proposed here will provide evidence on the persistence that may sometimes be entirely due to these types of mass.

In the simulations the processes for empirical distributions of different sizes are considered and for purpose of comparison, the test is also performed assuming the stochastic process for the true distributions.

4.5.1. Data generating processes

The two classes of data generating processes, referred to as **DGP-1** and **DGP-2** are described below. For each class, we consider stationary and nonstationary specifications and two different choices of a class of distributions for the continuous part of the process (uniform and truncated normal distributions). Denote by $U[a,b]$ the uniform distribution with support $[a,b]$ where $a,b \in \mathbb{R}$ ($a \leq b$). Denote by $\mathscr{N}_{[a,b]}(\mu,\sigma^2)$ the truncation of the normal distribution with mean μ and variance σ^2 ($\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{++}$) to $[a,b]$ obtained by renormalizing the density on this interval. Let $H_{x_0}(x): \mathbb{R} \to \{0,1\}$ be the Heaveside step function at $x = x_0$ defined as

$$H_{x_0}(y) = \begin{cases} 0 & \text{if } x < x_0 \\ 1 & \text{if } x \geq x_0. \end{cases} \tag{4.5.55}$$

DGP-1: Mixitures of stationary continuous distributions and time-varying masses

The first type of process **DGP-1**, represents a generalization of the process considered in Example 4.2.1.

Let $\{G_t\}_{t=1}^{\infty}$ be a stationary process of absolutely continuous distributions G_t on bounded supports. We assume $\{G_t\}_{t=1}^{\infty}$ is either (i) a process of uniform distributions $\{U[0,q_t]\}$ where the univariate process $\{q_t\}$ follows

$$q_t = .5 + .5q_{t-1} + v_t, \qquad v_t \overset{i.i.d.}{\sim} \mathscr{N}(0,:(.2)^2) \tag{4.5.56}$$

or (ii) a process of truncated normal distributions $\{\mathscr{N}_{[0,q_t]}(\mu_t,\sigma_t^2)\}$ where the multivariate process $\{\mu_t,\sigma_t,u_t\}$ follows

$$\begin{aligned} \mu_t &= \bar{\mu} + \rho_\mu \mu_{t-1} + v_{\mu,t} \\ \log\sigma_t &= \bar{\sigma} + \rho_\sigma \log\sigma_{t-1} + v_{\sigma,t} \\ q_t &= q_{.99}(\mu_t,\sigma_t) \end{aligned} \tag{4.5.57}$$

where

$$(v_{\mu,t}, v_{\sigma,t}) \overset{i.i.d.}{\sim} \mathscr{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_\mu^2 & 0\\ 0 & \tau_\sigma^2\end{bmatrix}\right). \tag{4.5.58}$$

and $q_{.99}(\mu,\sigma)$ is the 99 percentile of the normal distribution with mean μ and variance σ^2. Here, we specify the set of parameters as

$$(\bar{\mu},\rho_\mu,\bar{\sigma},\rho_\sigma,\tau_\mu,\tau_\sigma) = (.5,.5,0,.5,.2,.1). \tag{4.5.59}$$

Define a unit root univariate process $\{u_t\}_{t=1}^{\infty}$ specified as follows:

$$u_t = \bar{u} + u_{t-1} + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,(.2)^2). \tag{4.5.60}$$

The sequence $\{F_t\}$ is constructed by defining each element F_t as a mixture of G_t and a mass at u_t with weight $\alpha \in [0,1]$:

$$F_t = (1-\alpha)G_t + \alpha H_{u_t}. \tag{4.5.61}$$

It is clear that $\{F_t\}_{t=1}^{T}$ is nonstationary if and only if $\alpha > 0$. We are interested in whether a slight deviation from stationarity (small but positive α) can be detected by the persistency test. Thus, we consider such nonstationary specification where α is set to .01 as well as stationary specification $\alpha = 0$.

DGP-2: Mixitures of continuous distributions and masses at the fixed left boundary $x = 0$ with time-varying weights

For the second class of process **DGP-2**, we consider a sequence of mixtures of continuous distributions as in **DGP-1** and mass points. Here the location of the mass is fixed at 0, the mixture weight α_t at period t is time-variant. The process of G_t is specified as a process of absolutely continuous distributions which may or may not be stationary, accordingly referred to as **DGP-2-S** for stationary and **DGP-2-N** for non-stationary. Each element F_t of the process $\{F_t\}_{t=1}^{\infty}$ of interest is specified as

$$F_t = (1-\alpha_t)G_t + \alpha_t H_0. \tag{4.5.62}$$

where G_t is an absolutely continuous distribution on support $[0,u_t]$. We assume that α_t depends on u_t deterministically:

$$\alpha_t = \Phi_{\chi^2(q)}(u_t) \tag{4.5.63}$$

where $\Phi_{\chi^2(q)}$ is the probability distribution of the chi-square distribution with degree of freedom q. Then, $\{\alpha\}_{t=1}^{\infty}$ is a sequence of $[0,1]$-valued variables.

As in DGP-1, we let $\{G_t\}_{t=1}^{\infty}$ be either a sequence of uniform or truncated normal distributions. When $\{G_t\}_{t=1}^{\infty}$ is set to be the former: $\{U[0,u_t]\}$, $\{u_t\}_{t=1}^{\infty}$ follows an autoregressive mode of order 1:

$$u_t = \bar{u} + \rho_u u_{t-1} + \varepsilon_t; \tag{4.5.64}$$

$$\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,(.01)^2) \tag{4.5.65}$$

where the model parameters $(\bar{u},\rho_u,q)$ take the following values:

$$(\bar{u},\rho_u,q) = \begin{cases} (1 \quad\quad ,.5,6) & \textbf{(DGP-2-S)} \\ (2\cdot 10^{-4},\mathbf{1},6). & \textbf{(DGP-2-N)}. \end{cases}$$

When $\{G_t\}_{t=1}^{\infty}$ is specified a sequence of truncated normal distributions $\{\mathcal{N}_{[0,u_t]}(\mu_t,\sigma_t^2)\}$ following (4.5.57)-(4.5.58), the model parameters are specified as follows:

$$(\bar{\mu},\rho_\mu,\bar{\sigma},\rho_\sigma,\tau_\mu,\tau_\sigma,q) = \begin{cases} (.5\ ,.5,0,.2,,2,.01,6) & \textbf{(DGP-2-S)} \\ (10^{-4},\mathbf{1},-1,.2,.01,.01,6). & \textbf{(DGP-2-N)} \end{cases}$$

4.5.2. Implementation

We consider both the case where the true distribution functions in the process $\{F_t\}_{t=1}^{T}$ are assumed observable, and the case when the distributions are estimated from the observed data generated in finite sample. For the latter, we generate a set of i.i.d. observations $\{x_{i,t}\}_{i=1}^{n}$ of size $n \in \{100,1000\}$ from F_t, for $t = 1,\ldots,T$. The case where the process is directly observed is referred to as $n = \infty$. The length T of the process $\{F_t\}_{t=1}^{T}$ ranges from 100, 250, to 1000. The number of Monte Carlo replication is set to be 500.

To form a sequence $\{W_t\}_{t=1}^{T}$ of TZ transformed measures from $\{F_t\}_{t=1}^{T}$, we apply

Transformation 2. to each F_t, $t = 1, \ldots, T$

$$W_t(y) \quad = \quad \gamma_t F_t(\gamma_t u_t) - \frac{1}{T}\sum_{t=1}^{T} \gamma_t F_t(\gamma_t u_t) \tag{4.5.66}$$

where we specify $\gamma_t = u_t$. An estimator $\hat{W}_t$ of W_t given $\{x_{i,t}\}_{i=1}^n$ is constructed as in Section 4.4.

We also consider the persistency test based on the (demeaned) density process by Chang et al. (2016). Note, however, that the density of F_t, $t = 1, \ldots, T$ does not exist except for the stationary specification of **DGP-1**. Thus, we construct the process $\{f\}_{t=1}^T$ by truncating the mass point of each F_t, $t = 1, \ldots, T$ and defining f_t as the truncated absolutely continuous distribution. Then, we define the (demeaned) density process $\{f_t^*\}_{t=1}^T$ by

$$f_t^* = f_t - \frac{1}{T}\sum_{t=1}^{T} f_t \qquad t = 1, \ldots, T. \tag{4.5.67}$$

For each f_t, $t = 1, \ldots, T$, we construct a kernel density estimator with the Gaussian kernel. The bandwidth for the kernel estimator is selected by least squares cross validation.

To represent the TZ transformed process $\{W_t\}_{t=1}^T$ in $L^2[0,1]$, we employ a basis of $1,024$ B-spline functions of order 4 on $[0,1]$. For the density process $\{f_t^*\}_{t=1}^T$, we also employ a basis of $1,024$ B-spline functions of order 4 but on $[0, C_f]$ where C_f is a constant such that the support of f_t^* is included in $[0, C_f]$ for any $t = 1, \ldots, T$. For both processes, the roughness penalty of smoothing by the basis is chosen by minimizing the generalized cross validation score. A B-spline basis of order p is a set of functions that are polynomial of order $p-1$ between knots and $(p-2)$ times differentiable. See De Boor (1978) for more details and numerical properties of this basis .

4.5.3. Results

Several tables report the results of applying the test of dimension of the nonstationary subspace. The empirical rejection probability for $H_0(M)$ ($M = 1, 2$) for **DGP-1** is presented in Table 4.1 ($M = 1$) and Table 4.2 ($M = 2$). We see from the upper part of the tables that correspond to the stationary cases that the test has strong power in rejecting the null hy-

pothesis when the process is stationary ($\alpha = 0$) for both the TZ transformed and demeaned density processes.

The lower subtables of 4.1-4.2 present the case for $\alpha = .01$, i.e. when there is a slight deviation from stationarity. The test based on the TZ transformed process does not reject $H_0(1)$ as long as the cross sectional sample size is not too small ($n = 1000, \infty$). For a small size, e.g. $n = 100$ in each period, the cross-sectional sample may not include any observation at the mass point, the probability of this happening is 36.6%. In such cases the resulting estimated TZ transforms only reflect the stationary part of the process. However, once the cross-sectional sample size gets to be moderately large ($n = 1000$), the test behaves quite similarly to the case with no estimation uncertainty ($n = \infty$). On the other hand, $H_0(2)$ is always rejected. This suggest that the TZ transform based test correctly finds the dimension of the unit-root space to be 1. As expected, the density based test always rejects $H_0(1)$ (and $H_0(2)$) as only the stationarry part of the process is incorporated.

For the stationarity cases for **DGP-2-S** in the upper parts of Tables 4.3-4.4, similarly to the case of **DGP-1,** the null is strongly rejected. For the nonstationarity case (**DGP-2-N**), the TZ transform based test is more likely to not reject $H_0(1)$ than the density based test. The results for $H_0(2)$ are more subtle. When the process is estimated ($n = 100, 1000$), both tests reject $H_0(2)$ with probability close to 1. This is due to the fact that the variation of both processes explained by the second eigenfunction of the unit-root subspace is so small that it is not distinguishable when the processes are observed with noise. This suggests that the cross-sectional sample size n must grow much faster than T, as was indicated in Section 4.4. On the other hand, when there is no uncertainty ($n = \infty$), the TZ transform based test appears to offer stronger support of the existence of the second dimension of the unit-root subspace. To investigate further, it is important to note that the dimension of the the unit root subspace can be easily interpretable for the TZ transformed process while it is not always the case for the density process. We show in the appendix that the first eigenfunction of the TZ transformed process is associated with the nonstationary dynamics of the support $\{\gamma_t\}$ and the second eigenfunction captures the interaction between the mass points $\{\alpha_t\}$ and $\{\gamma_t\}$. The density process does not incorporate the dynamics of $\{\alpha_t\}$ and thus that of the size of the mass at zero. This likely explains the better detection of the second dimension of the test for TZ transformed distribution. Furthermore, while the first eigenfunction of the density process has a similar interpretation as for the TZ transformed

process, it is not clear what the second eigenfunction captures in the absence of $\{\alpha_t\}$. In particular, when the TZ transformed process is constructed from the uniform distribution, it can be shown that the dimension of the the unit root subspace is one (Appendix). For that case, the dynamics of both density and TZ transformed processes are each uniquely determined by the univariate process $\{\gamma_t\}$. The dimension of the unit root subspace for the TZ transformed process coincides with that of the underlying process $\{\gamma_t\}$ that drives nonstationary and thus is more interpretable.

These results show that the TZ transformed based test performs well in detecting nonstationarity characterized by the varying support or the dynamics of the mass points. Thus the test based on the TZ transform offers advantages over density-based approaches or any method which ignores mass points or requires truncation. The dimension of the unit-root subspace could be more interpretable for the the TZ transform, when driven by the dynamics of support and mass points.

Table 4.1. **DGP-1:** Empirical rejection probability of $H_0 : \dim(H_N) = 1$ vs $H_0 : \dim(H_N) = 0$

(a) Stationary specification

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000
$T = 250$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000
$T = 1000$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000

(b) Nonstationary specification

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	**0.000**	**0.006**	1.000	1.000
	$n = \infty$	**0.000**	**0.000**	1.000	1.000
$T = 250$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	**0.000**	**0.012**	1.000	1.000
	$n = \infty$	**0.000**	**0.000**	1.000	1.000
$T = 1000$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	**0.000**	**0.000**	1.000	1.000
	$n = \infty$	**0.000**	**0.000**	1.000	1.000

Table 4.2. **DGP-1:** Empirical rejection probability of
$H_0 : \dim(H_N) = 2$ vs $H_0 : \dim(H_N) = 1$

(a) Stationary specification

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000
$T = 250$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000
$T = 1000$	$n = 100$	1.000	1.000	1.000	1.000
	$n = 1000$	1.000	1.000	1.000	1.000
	$n = \infty$	1.000	1.000	1.000	1.000

(b) Nonstationary specification

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	0.9780	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	0.9360	0.9980	1.0000
$T = 250$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	0.9960	1.0000	1.0000
$T = 1000$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	1.0000	1.0000	1.0000

Table 4.3. **DGP-2:** Empirical rejection probability of
$H_0 : \dim(H_N) = 1$ vs $H_0 : \dim(H_N) = 0$

(a) Stationary specification (DGP-2-S)

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	1.0000	1.0000	1.0000
$T = 250$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	1.0000	1.0000	1.0000
$T = 1000$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	1.0000	1.0000	1.0000

(b) Nonstationary specification (DGP-2-N)

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	0.8980	1.0000	0.8760	0.9500
	$n = 1000$	0.4580	0.5560	0.4320	0.4020
	$n = \infty$	0.0440	0.1120	0.1000	0.3900
$T = 250$	$n = 100$	0.8800	1.0000	0.8920	0.9340
	$n = 1000$	0.4460	0.4260	0.4040	0.3660
	$n = \infty$	0.0340	0.1020	0.0880	0.3880
$T = 1000$	$n = 100$	0.8680	1.0000	0.9120	0.8660
	$n = 1000$	0.3380	0.3680	0.4340	0.2900
	$n = \infty$	0.0420	0.1080	0.1120	0.4160

Table 4.4. **DGP-2:** Empirical rejection probability of
$H_0 : \dim(H_N) = 2$ vs $H_0 : \dim(H_N) = 1$

(a) Stationary specification (DGP-2-S)

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	1.0000	1.0000	0.9980	1.0000
	$n = 1000$	0.9840	1.0000	0.9980	1.0000
	$n = \infty$	0.9640	1.0000	0.9960	1.0000
$T = 250$	$n = 100$	1.0000	1.0000	0.9980	1.0000
	$n = 1000$	1.0000	1.0000	0.9980	1.0000
	$n = \infty$	1.0000	1.0000	0.9900	1.0000
$T = 1000$	$n = 100$	1.0000	1.0000	1.0000	1.0000
	$n = 1000$	1.0000	1.0000	1.0000	1.0000
	$n = \infty$	1.0000	1.0000	1.0000	1.0000

(b) Nonstationary specification (DGP-2-N)

		TZ (Uniform)	(Truncated normal)	Density (Uniform)	(Truncated normal)
$T = 100$	$n = 100$	0.9940	1.0000	0.9840	0.9980
	$n = 1000$	1.0000	1.0000	0.8540	1.0000
	$n = \infty$	**0.1680**	**0.4460**	**0.5600**	0.7500
$T = 250$	$n = 100$	0.9980	1.0000	0.9880	1.0000
	$n = 1000$	1.0000	1.0000	0.8200	0.9900
	$n = \infty$	**0.0760**	**0.3580**	**0.5780**	0.7980
$T = 1000$	$n = 100$	0.9980	1.0000	0.9980	0.9980
	$n = 1000$	1.0000	1.0000	0.7940	0.8840
	$n = \infty$	**0.2200**	**0.4800**	0.6040	0.8160

4.6. Empirical application: Intertemporal dynamics of the cross-sectional distributions of earnings

4.6.1. Premise

This section applies the transformation approach to examine persistency an intertemporal dynamics of the distributions of individual weekly earnings in the U.S. based on the Current Population Study dataset. As in the Monte Carlo simulations in Section 4.5, we employ a basis of $1,024$ B-spline functions of order 4 on $[0,1]$.

4.6.2. Data description

We consider the population to consist of the labor force, the employed and those in involuntary unemployment, as the population in each period. Then, the population distribution F_t^* at period t is defined as a mixture of a mass at $x=0$ and the distribution $F_t^{(\mathrm{EMP})}$ of earnings by the employed with a mixture weight being the unemployment rate UEMP_t:

$$F_t^* = \mathrm{UEMP}_t + (1-\mathrm{UEMP}_t)\, F_t^{(\mathrm{EMP})}. \tag{4.6.68}$$

However, the data is censored by a top-coded value u_t for each t and thus the data observation F_t is given by

$$F_t = \mathrm{UEMP}_t + (1-\mathrm{UEMP}_t)\, F_t^{(\mathrm{EMP,tc})}. \tag{4.6.69}$$

where

$$F_t^{(\mathrm{EMP,tc})}(x) = \begin{cases} F_t^{(\mathrm{EMP})}(x) & 0 \le x < u_t \\ 1 & x \ge u_t. \end{cases} \tag{4.6.70}$$

The dataset contains cross sectional observations of employment status and weekly earnings at a monthly frequency from January 1994 to July 2022, which corresponds to $T=342$ periods. The values of weekly earnings are not present for the unemployed and those who were employed but chose not to disclose their earnings. The sample size n_t at period $t=1,\ldots,T$ ranges from $45,447$ to $74,425$. For each $t=1,\ldots,T$, we construct an estimator $\widehat{\mathrm{UEMP}}_t$ of the unemployment ratioUEMP_t by dividing the number of the unemployed by n_t. We apply the assumption that the data are missing at random, and thus we drop obser-

vations with missing values and compute the empirical distribution $\hat{F}_t^{(\mathrm{EMP,tc})}$ of $F_t^{(\mathrm{EMP,tc})}$. The number n_t^* of observations used to compute $\hat{F}_t^{(\mathrm{EMP,tc})}$ ranges from $9,693$ to $15,826$ for $t = 1, \ldots, T$.

Then, the estimate $\hat{F}_t$ of the data distribution F_t is given by

$$\hat{F}_t = \widehat{\mathrm{UEMP}}_t + \left(1 - \widehat{\mathrm{UEMP}}_t\right) \hat{F}_t^{(\mathrm{EMP,tc})}, \qquad t = 1, \ldots, T. \tag{4.6.71}$$

Let $\hat{u}_t$ be an estimator of u_t defined as

$$\hat{u}_t = \inf\left\{x : \hat{F}_t^{(\mathrm{EMP,tc})}(x) = 1\right\}. \tag{4.6.72}$$

For each t, we consider the TZ transform W_t (***Transformation 2***)**:**

$$W_t = \gamma_t F_t\left(u_t y\right) - \frac{1}{T}\sum_{s=1}^{T} \gamma_s F_s\left(u_s y\right), \quad y \in [0,1] \tag{4.6.73}$$

where the choice of γ_t and its estimator $\hat{\gamma}_t$ is discussed below in Section 4.6.3. Then, given $\{\hat{F}_t, \hat{u}_t, \hat{\gamma}_t\}_{t=1}^T$, the estimator $\hat{W}_t$ of W_t is given by

$$\hat{W}_t = \hat{\gamma}_t \hat{F}_t\left(\hat{u}_t y\right) - \frac{1}{T}\sum_{s=1}^{T} \hat{\gamma}_s \hat{F}_s\left(\hat{u}_s y\right), \quad y \in [0,1], \quad t = 1, \ldots, 342. \tag{4.6.74}$$

We consider three specifications of γ_t and the persistency test in Section 4.4.2 is performed for each specification.

4.6.3. Choices of γ_t under top-coding

Weekly earnings are top-coded at \$1,923 until 1997 and at \$2,885 from 1998 onwards. We discuss possible choices of the scale parameter γ_t in the presence of such top coding. The first possible choice $\gamma_t^{(\mathrm{tc})}$of γ_t is specified as the upper bound u_t of the data distribution F_t, i.e.

$$\gamma_t^{(\mathrm{tc})} = u_t \tag{4.6.75}$$

Then, letting t^* be the period which corresponds to December 1997, we have

$$\gamma_t^{(\mathrm{tc})} = \begin{cases} 1,923 & , t \leq t^* \\ 2,885 & , t > t^*. \end{cases} \tag{4.6.76}$$

A natural estimator of $\gamma_t^{(\mathrm{tc})}$ is given by the empirical counterpart $\hat{u}_t$ as defined in (4.6.72).

However, the sequence $\{\gamma_t^{(\mathrm{tc})}\}_{t=1}^T$ provides little information on how the support of the underlying distribution F_t^* evolves over time. As an alternative measure of the scale, we may use information on high quantiles of the data distribution $F_t^{(\mathrm{EMP,tc})}$. The 100th percentile is not available in the top coded data set Figure 4.1 shows the trajectories of percentiles of the empirical distributions. Except for the 100th percentile, any of the shown higher percentiles of the data distribution are not subject to top-coding and thus they are also percentiles of unobserved $F_t^{(\mathrm{EMP})}$. Thus, the dynamics of such percentiles may be more informative for learning the dynamics of $\{F_t^*\}_{t=1}^T$ In particular, we employ the 95th percentile to define

$$\gamma_t^{(95)} = \min\left\{x : F_t^{(\mathrm{EMP,tc})}(x) \geq .95\right\}. \tag{4.6.77}$$

Then, as long as $\gamma_t^{(95)} < u_t$, we have

$$\gamma_t^{(95)} = \min\left\{x : F_t^{(\mathrm{EMP})}(x) \geq .95\right\}. \tag{4.6.78}$$

In the sample, $\gamma_t^{(95)}$ may be estimated from the empirical distribution $\hat{F}_t^{(\mathrm{EMP,tc})}$.

Lastly, we employ information about the income distributions obtained from an external source. Piketty and Saez (2003) provides the table of top percentiles of annual labor income and average income of high income groups in the U.S. based on tax return data from 1913 to 1998. More recent data (up to 2018) are available in one of the authors' website (`https://eml.berkeley.edu/saez/`). Such data are not subject to censoring and provide information on the upper tail of the income distribution that is not present in the household survey used here as the main dataset. We convert the average annual earnings of the top groups (ranging from top 10% to .5%) from 1994 to 2018 into weekly values by dividing by 52 weeks. The time series plotted in Figure 4.2 appear to have similar upward trends as in Figure 4.1. We employ the weekly average earnings of the top 10% as a scale. Since such value is observed annually: define the scale $\gamma_t^{(\mathrm{PS})}$ as follows: if t corresponds to

l-th month of year m, then $\gamma_t^{(\mathrm{PS})}$ corresponds to the converted weekly average value of the the top 10% earnings in year m for any $l = 1, \ldots, 12$. Since the data is available up to 2018, we consider the time series up to December, 2018 (the number of periods being 288) when γ_t is specified as $\gamma_t^{(\mathrm{PS})}$.

4.6.4. Results

The persistency test described in Section 4.4.2 is applied to the estimated sequences of transformed distributions with three different specifications of scale parameter $\gamma_{\mathbf{t}} \in \{\gamma_{\mathbf{t}}^{(\mathrm{tc})}, \gamma_t^{(95)}, \gamma_t^{(\mathrm{PS})}\}$. In Table 4.5, the p-values for $\mathscr{H}_0(M) : H_N = M$, $M = 1, \ldots, 4$ are reported. For all specifications, the null hypothesis is not rejected and the estimated dimension of the unit root subspace is 2 at a 10% significance level. While the p-value for $M = 3$ under $\gamma_t = \gamma_t^{(\mathrm{PS})}$ is not significant at a 5% level, the third eigenfunction associated with the unit root subspace explains only .5% of the total variation of the process (the first two eigenfunctions contribute to almost all the rest). Thus, even if the true dimension is indeed 3, the contribution of the third eigenfunction is negligible.

Chang et al. (2016) also report the unit root subspace for the demeaned density process to be 2 by using the same dataset (which is here extended in time).

The first two eigenfunctions and principal components associated with the unit root subspace are presented in Figure 4.4(a). For each specification of γ_t, the first two eigenfunctions (φ_1, φ_2), obtained from functional principal component analysis described in Appendix 4.A, span the unit root subspace H_N. Then, the principal components $\{\xi_{1,t}\}_{t=1}^T$and $\{\xi_{2,t}\}_{t=1}^T$ associated with H_N are defined as the coordinate processes of $\{\hat{W}_t\}$ with respect to φ_j, $j = 1, 2$:

$$\xi_j = \left\langle \hat{W}_t, \varphi_j \right\rangle, \qquad j = 1, 2 \tag{4.6.79}$$

for $t = 1, \ldots, T$,

According to Figure 4.4(a), the first eigenfunction dominates over the top 65% (approximately) of the distribution while the second eigenfunction dominates the bottom 35% for all specifications of γ_t. The large majority of variation of the process is explained by the first eigenfunction (91.3% for $\gamma_t^{(\mathrm{tc})}$, 97.5% for $\gamma_t^{(95)}$, 93.0% for $\gamma_t^{(\mathrm{PS})}$) while the rest is mostly captured by the second eigenfunction. To interpret this in terms of the original sequence $\{F_t\}$, note that for each point $y \in [0, 1]$ in the support of the TZ transform $W_t(y)$,

there exists some quantile $q_t(y) \in [0,1]$ such that $q_t(y) = F_t(u_t y)$. Then, for $y = .35$, $q_t(y)$ ranges from .567 to .899. This indicates that the variation of the transformed process is driven by the dynamics of the upper tail of the earning distribution.

The trajectories of the principal components associated with the unit root subspace in Figure 4.4(b) show the main driver of persistency of the process each specification of γ_t captures. For $\gamma_t = \gamma_t^{(\mathrm{tc})}$, the first principal component sees a large jump at the period when the top-coded value changes and otherwise is relatively flat. This shows persistency of the process is largely due to the change in the top-coded value, which is an artifact of the data rather than a fundamental characteristic of the underlying process. The first principal component for the density process in Chang et al. (2016) (, referred to as the first nonstationary coordinate process in their Fig 1) shows a similar trajectory driven by the top-coded value. While Chang et al. (2016) do not incorporate a mass point at the top-coded value, an increase in its value affects the structure of the data by expanding the support of reported earnings. This shifts the upper tail of the distribution to the right, which leads to a large increase in the principal component associated with the upper half of the distribution.

On the other hand, for $\gamma_t = \gamma_t^{(95)}$, the first component shows a clear upward trend. Thus, persistency is driven by the upward shift of the top of the distribution captured by $\gamma_t^{(95)}$, $t = 1, \ldots, T$. For $\gamma_t = \gamma_t^{(\mathrm{PS})}$, the corresponding process also shows an overall upward trend. We note that the second principal component has a slightly downward slope for all cases and thus the bottom part of the distribution also shows some level of persistency while its contribution to the variation of the process is limited.

We conclude that by incorporating the changing support through specification of γ_t, the principal component of the TZ transformed process manifests the driver of persistency of the underlying income process. Otherwise, one may attribute observed persistency to institutional factors in the data representation.

Table 4.5. p-values for the persistency test

	$\gamma_t^{(\mathrm{tc})}$	$\gamma_t^{(95)}$	$\gamma_t^{(\mathrm{PS})}$
$M=1$	.343	.988	.607
$M=2$	.812	.915	.536
$M=3$	.007	.033	.057
$M=4$	.000	.005	.006

Figure 4.1. Time series of percentiles of the top-coded empirical distributions sets

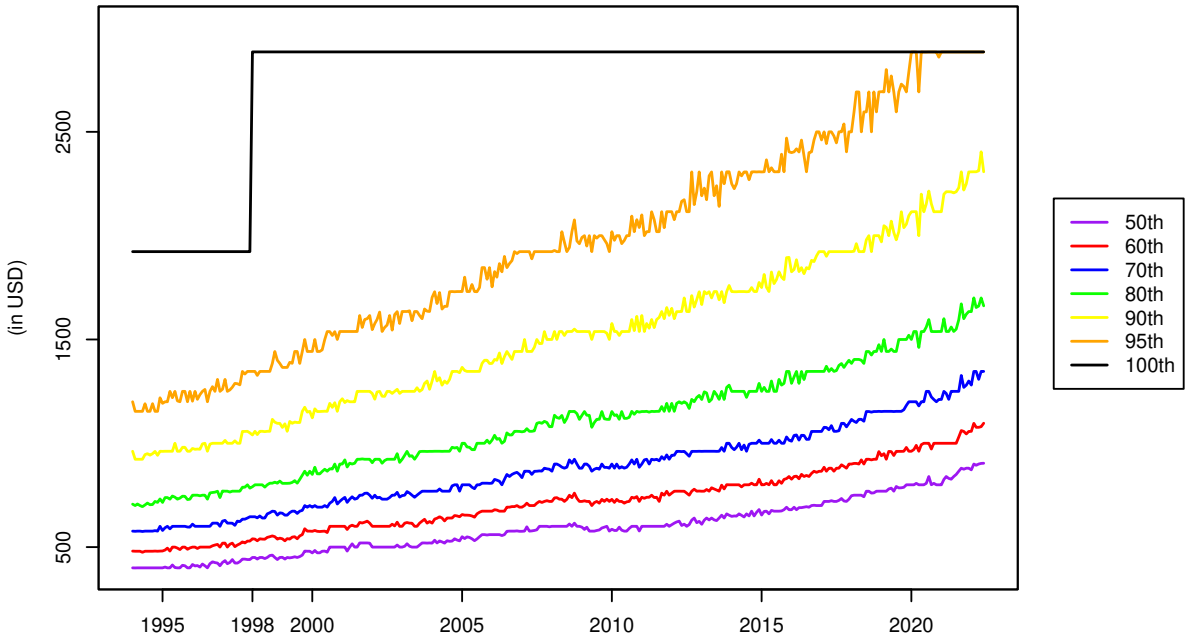
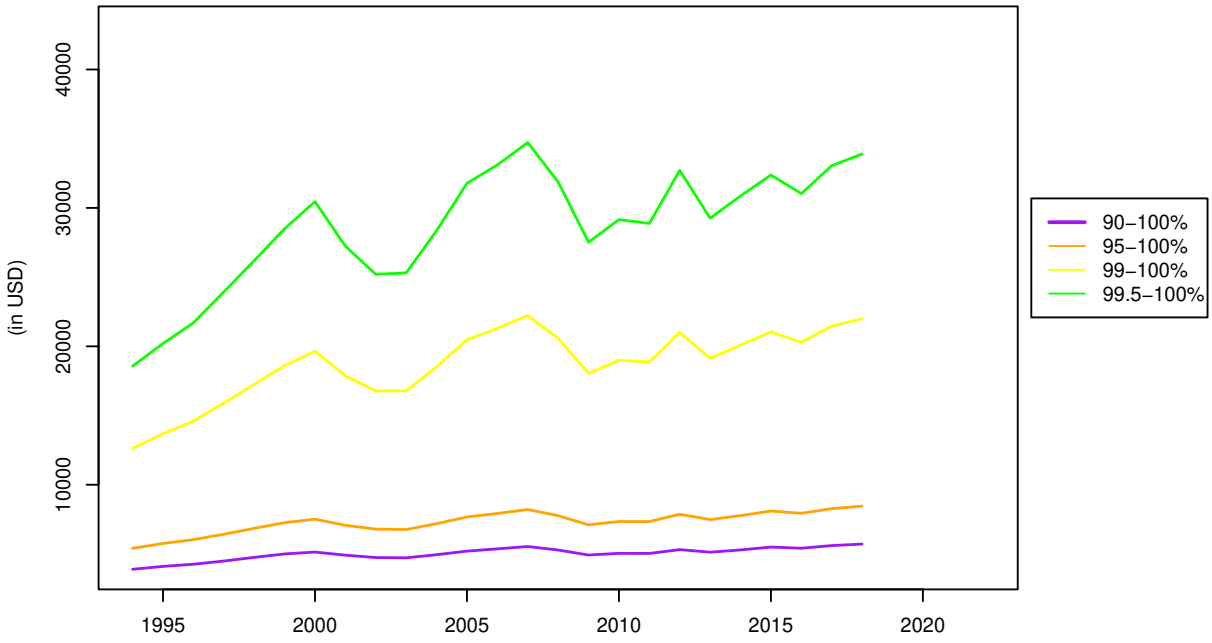
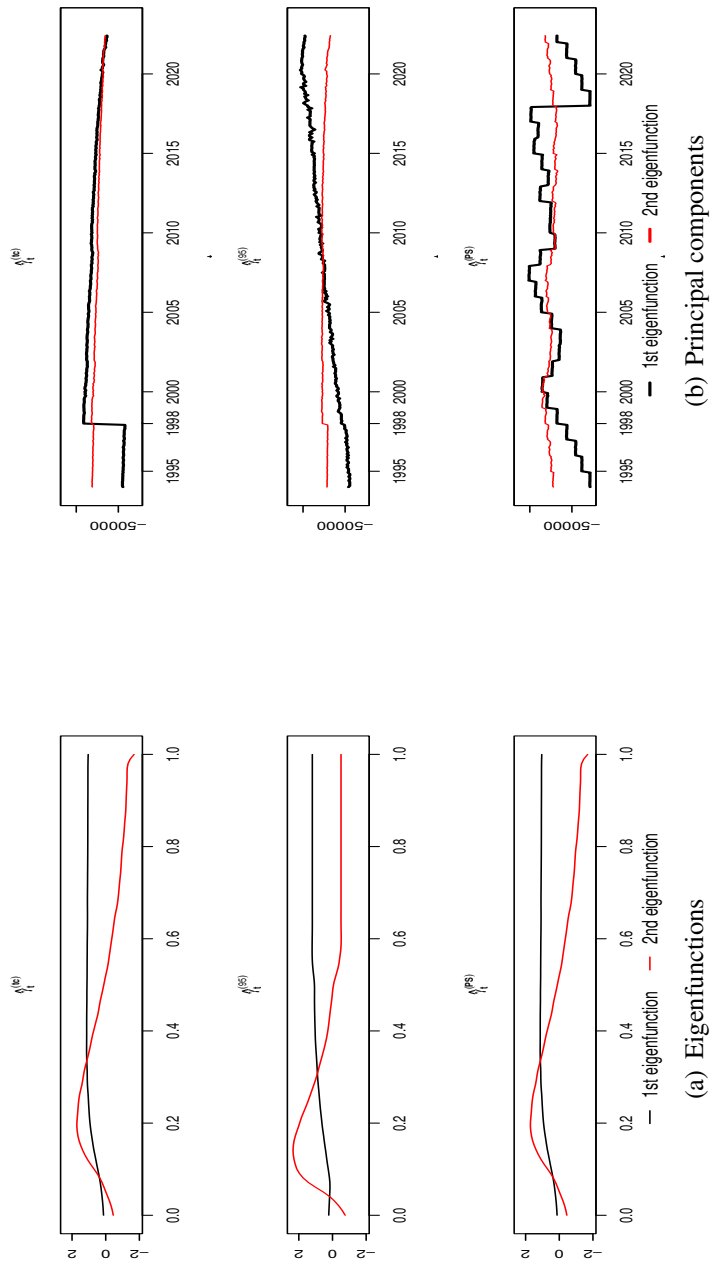


Figure 4.2. Average weekly income of top percentile groups in the U.S. according to tax return data





(a) Eigenfunctions

(b) Principal components

Figure 4.3. Eigenfunctions and principal components associated with the unit root subspace

4.7. Conclusion

We have introduced a transformation approach for stochastic processes of probability distributions which allows to incorporate important features of dynamics of economics variables, such as the presence of mass points and varying support. Statistical properties, such as unit roots, of a process of distributions are investigated via its representation in $L^2[0,1]$. We have also discussed connections between derivatives of transformed measures and demeaned densities considered in Chang et al. (2016). This establishes our approachs as a generalization of that methodology applied to demeaned density functions. In our empirical application, we confirmed persistency of earning dynamics in the U.S., reported in Chang et al. (2016) and found stronger statistical evidence when varying (expanding) support of the process is taken into account. We show that the dynamics of the support, on top percentiles of that distribution is the main persistent feature.

4.A. Functional principal component analysis

Let $\{Z_t\}$ be a square-integrable weakly stationary sequence of TZ-transformed measures on $L^2([0,1])$.

Define the covariance operator

$$C_Z(f) \quad = \quad \mathbb{E}[\langle Z_t, f\rangle Z_t], \quad f \in L^2([0,1]) \tag{4.A.80}$$

which is a bounded linear operator from $L^2([0,1])$ to $L^2([0,1])$. Note one can also express

$$C_Z(f) = \int K_Z(\cdot, y) f(y) dy, \qquad f \in L^2([0,1]) \tag{4.A.81}$$

where

$$K_Z(y_1, y_2) = \mathbb{E}[Z_t(y_1) Z_t(y_2)]. \tag{4.A.82}$$

Since $L^2([0,1])$ is separable, C_W is a Hilbert-Schmidt operator[3] and thus admits the decom-

[3]Let $\{\xi_j\}_{j\in\mathbb{N}}$ be any orthonormal basis of a Hilbert space $\mathscr{H}$ with norm $\|\cdot\|_{\mathscr{H}}$. A linear operator $F : \mathscr{H} \to \mathscr{H}$ is Hilbert-Schmidt if

$$\sum_{i\in\mathbb{N}} \|F\xi_i\|^2_{\mathscr{H}} < \infty.$$

position for any an orthonormal basis $\{\xi_j\}_{j\in\mathbb{N}}$ of $L^2([0,1])$:

$$C_Z(f) = \sum_{j=1}^{\infty} \lambda_j \left\langle f, \xi_j \right\rangle \xi_j \tag{4.A.83}$$

where $\{\lambda_j\}_{j\in\mathbb{N}}$ is a sequence of positive values such that

$$\sum_{j=1}^{\infty} \lambda_j = \mathbb{E}\|Z_t\|^2 < \infty. \tag{4.A.84}$$

Similarly,

$$K_Z(y_1, y_2) = \sum_{j=1}^{\infty} \lambda_j \xi_j(y_1)\xi_j(y_2). \tag{4.A.85}$$

By the Karhunen–Loeve theorem (Bosq (2000) Theorem 1.5, p.25), W_t admits a representation as a linear combination of $\{\xi_j\}_{j\in\mathbb{N}}$:

$$Z_t = \sum_{j=1}^{\infty} \beta_{t,j}\xi_j(y). \tag{4.A.86}$$

where $\beta_{t,j} = \left\langle Z_t, \xi_j \right\rangle$ and

$$\mathbb{E}[\beta_{t,j}] = 0, \quad \forall j \in \mathbb{N}, \tag{4.A.87}$$

$$\mathbb{E}[\beta_{t,j}\beta_{t,l}] = \begin{cases} \lambda_j & , j = l \\ 0 & , j \neq l \end{cases}. \tag{4.A.88}$$

Functional principal component analysis obtains eigenfunctions ξ_j, $j \in \mathbb{N}$ of C_Z, i.e. orthonormal basis functions such that their associated eigenvalues are nonzero, in an iterative manner:

$$\xi_k = \arg\max_{\mathscr{V}_k} \mathrm{Var}\left(\langle Z_t, \xi \rangle\right) \tag{4.A.89}$$

where

$$\mathscr{V}_k = \begin{cases} \{\xi \in L^2([0,1]) : \|\xi\| = 1\} & ,k = 1 \\ \left\{\xi \in L^2([0,1]) : \|\xi\| = 1, \left\langle \xi, \xi_j \right\rangle = 0, \forall j = 1,\ldots,k-1\right\} & ,k \geq 2. \end{cases} \tag{4.A.90}$$

For each $j \in \mathbb{N}$, the eigenvalue λ_j solves

$$\int K_Z(\cdot,z)\xi_j(z)dy = \lambda_j \xi_j. \tag{4.A.91}$$

4.B. Dimension of the unit root subspace in DGP2-N

Suppose that each element F_t of the sequence $\{F_t\}$ is a mixture of a mass at $x = 0$ and and $U[0,u_t]$ with mixture weight α_t:

$$F_t(x) = \begin{cases} 0 & \text{if } x < 0 \\ \alpha_t + (1-\alpha_t)\dfrac{x}{u_t} & \text{if } 0 \leq x \leq u_t \\ 1 & \text{if } x > u_t. \end{cases} \tag{4.B.92}$$

where $\{\alpha_t\}$ and $\{u_t\}$ are sequences of random variables on $[0,1]$ and $\mathbb{R}_{++}$, respectively. Then, W_t^2 *(the **Transform 2**)* is given by

$$W_t^2 = \sum_{j=1}^{2} \beta_{t,j}\varphi_j(y) \tag{4.B.93}$$

where

$$\begin{gathered} \varphi_1(y) = y, \quad \varphi_2(y) = 1 \\ \beta_{t,1} = (\gamma_t - \mathbb{E}[\gamma_t]) - \beta_{t,2},\ \beta_{t,2} = \gamma_t\alpha_t - \mathbb{E}[\gamma_t\alpha_t] \end{gathered} \tag{4.B.94}$$

Then, by applying the Gram–Schmidt process, we have

$$W_{M_0,t}^2 = \sum_{j=1}^{2} \bar{\beta}_{t,j}\bar{\varphi}_j \tag{4.B.95}$$

where

$$\bar{\varphi}_1(y) = \sqrt{3}y,\ \bar{\varphi}_2(y) = 2(1-\frac{3}{2}y) \tag{4.B.96}$$

and

$$\bar{\beta}_{t,1} = \frac{1}{\sqrt{3}}\left\{(\gamma_t - \mathbb{E}[\gamma_t]) + \frac{1}{2}(\gamma_t\alpha_t - \mathbb{E}[\gamma_t\alpha_t])\right\}, \tag{4.B.97}$$

$$\bar{\beta}_{t,2} = \frac{1}{2}(\gamma_t\alpha_t - \mathbb{E}[\gamma_t\alpha_t]). \tag{4.B.98}$$

Note that $\|\bar{\varphi}_1(y)\| = \|\bar{\varphi}_2(y)\| = 1$ and $\langle\bar{\varphi}_1(y), \bar{\varphi}_2(y)\rangle = 0$ and thus $\{W^2_{M_0,t}\}$ is spanned by with orthonormal base functions $\{\bar{\varphi}_1, \bar{\varphi}_2\}$. Suppose now that $\{\gamma_t - \mathbb{E}[\gamma_t]\}$ is a unit root process. Then, the dimension M of the unit root subspace H_N depends on the dynamics of $\{\alpha_t\}$. Suppose $\alpha_t = p$ for some positive constant $p \in (0,1)$. Then,

$$\bar{\beta}_{t,1} = \frac{1}{\sqrt{3}}\left(1+\frac{5}{2}p\right)(\gamma_t - \mathbb{E}[\gamma_t]), \bar{\beta}_{t,2} = \frac{p}{2}(\gamma_t - \mathbb{E}[\gamma_t]) \tag{4.B.99}$$

so that $M = 2$. Suppose α_t is zero for any t. Then,

$$\bar{\beta}_{t,1} = \frac{1}{\sqrt{3}}(\gamma_t - \mathbb{E}[\gamma_t]), \bar{\beta}_{t,2} = 0 \tag{4.B.100}$$

so that $M = 1$. Similarly, if $\gamma_t > 1$, $\forall t$ and $\alpha_t = 1/\gamma_t$, then we again have (4.B.100) and thus $M = 1$.

4.C. Proofs

PROOF OF LEMMA 4.2.1 The assertion follows directly from the definition of each transformation. In particular, when γ_t is constant,

$$\hat{\mathbb{E}}_T[\gamma_t F_t(u_t y)] - \gamma_t\hat{\mathbb{E}}_T[F_t(u_t y)] = 0 \tag{4.C.101}$$

so that $W^1_t(y) = W^2_t(y)$. □

PROOF OF PROPOSITION 4.2.4 For $\psi \in D^1[0,1]$ (and thus in $L_2[0,1]$) the generalized

derivative, w_t, of W_t provides

$$(\mathrm{w}_t, \Pi_N \psi) = -(\mathrm{W}_t, \partial \Pi_N \psi) = -(\mathrm{W}_t, \Pi_N \partial \psi). \tag{4.C.102}$$

Since such ψ span H_N in L_2 norm and thus also span $D^1[0,1]$ in the weak topology, this fully defines $(\Pi_N \mathrm{w}_t, \psi)$, and thus the generalized function $\mathrm{w}_t^N = \Pi_N \mathrm{w}_t$ is defined on $D^1[0,1]$. Analogously, w_t^S is defined and the decomposition follows. □

PROOF OF LEMMA 4.3.1 (a) If $\{F_t\}$ is a stationary sequence, then so are the values of the functionals. Thus $\gamma_t F_t(u_t y)$ is stationary. If additionally the λ_t forms a stationary sequence, then W_t forms a stationary sequence. (b) Under the conditions $W_t(y)$ is a difference of two non-decreasing functions. When $W_t(y)$ is stationary, each of the non-decreasing functions that is uniquely determined (up to possibly a constant) has to be stationary, thus $W_t^+ = \gamma_t F_t(u_t y)$ is stationary. The sequence $F_t(u_t y) = (W_t^+(1))^{-1} W_t^+(y)$ is then stationary. The sequence $\{\lambda_t\}$ as given by the ***Transforms 1.*** or ***2.*** is thus also stationary. □

PROOF OF THEOREM 4.4.1 Represent

$$\begin{aligned} \hat{W}_t^0 - W_t^0 &= \hat{\gamma}_t \hat{F}_t(\hat{u}_t y) - \gamma_t F_t(u_t y) \\ &= (\hat{\gamma}_t - \gamma_t)\hat{F}_t(\hat{u}_t y) + \gamma_t \left[\hat{F}_t(\hat{u}_t y) - F_t(\hat{u}_t y)\right] + \gamma_t \left[F_t(\hat{u}_t y) - F_t(\hat{u}_t y)\right] \\ &= B_{1,t} + B_{2,t} + B_{3,t}. \end{aligned} \tag{4.C.103}$$

We evaluate uniform over $1 \le t \le T$ convergence in norm for each $B_{i,T}, i = 1,2,3$ in turn.

First,

$$\sup_{1 \le t \le T} \left\| B_{1,t} \right\| \le \sup_{1 \le t \le T} |\hat{\gamma}_t - \gamma_t| \tag{4.C.104}$$

since $\sup_{1 \le t \le T} \|\hat{F}_t(\hat{u}_t y)\| \le 1$.

To bound $B_{2,t}$, note

$$\sup_{1 \le t \le T} \|\hat{F}_t(\hat{u}_t y) - F_t(\hat{u}_t y)\| \le \sup_{1 \le t \le T} \sup_{0 \le x \le \bar{u}_t} \|\hat{F}_t(x) - F_t(x)\| \tag{4.C.105}$$

where $\bar{u}_t$ is the upper bound on support of $F_t(x)$ and thus for the empirical distribution $\hat{F}(x)$ as well. We accommodate the possibility that u_t represents a value that may differ

from the upper bound of the distribution F_t.

By the DKW inequality (Dvoretzky, Kiefer and Wolfowitz (1956), Massart (1990)),

$$\sup_{0\le x\le \bar{u}_t} \left|\hat{F}_t(x) - F_t(x)\right| \le \frac{2}{\sqrt{n}} \tag{4.C.106}$$

with probability approaching to one for $t = 1,\dots,T$. Thus, it follows that

$$\sup_{1\le t\le T} \|\hat{F}_t(\gamma_t y) - F_t(\gamma_t y)\| \le 2T/\sqrt{n}. \tag{4.C.107}$$

Consequently,

$$\sup_{1\le t\le T} \left\|B_{2,t}\right\| \le \frac{2T}{\sqrt{n}} \sup_{1\le t\le T} \gamma_t. \tag{4.C.108}$$

To establish the bound on the norm of $B_{3,t}$ define $D_t^x := \left\{x_{t,j}^d\right\}_{j=1}^{N_{d,t}} \subset [0,\bar{u}_t]$ to be the set of all points of discontinuity for F_t where $x_{t,j}^d$ denotes the $j-th$ in magnitude point of discontinuity and $N_{d,t}$ is the cardinality of $D_t^x, t = 1,\dots,T; N_{d,t} \le N_d$. Then the complement to this set is a union of $N_{d,t}+2$ open intervals of points of continuity, $\cup_{j=1}^{N_{d,t}+1}\left(x_{t,j-1}^d, x_{t,j}^d\right)$, where we set $x_{t,0}^d = 0$ and $x_{t,N_{d,t}+1} = \bar{u}_t$. (If $x_{t,1}^d = 0$, the first interval for $j = 1$ is empty, and if $x_{t,N_{d,t}} = \bar{u}_t$, then the last interval is empty). Denote by δ_t^x the minimum length of an interval: $\delta_t^x = \inf\left(x_{t,j+1}^d - x_{t,j}^d\right)$.

Correspondingly, define $\hat{D}_t^y = \left\{y_{t,j}^d\right\}_{j=1}^{N_{d,t}}$, where $y_{t,j}^d$ is such that either $u_t y_{t,j}^d = x_{t,j}^d$ or $\hat{u}_t y_{t,j}^d = x_{t,j}^d$ and let $\delta_t^y = \bar{u}_t^{-1}\delta_t^x$. Consider an arbitrary $\varepsilon_t < \frac{1}{2}\delta_t^y$.

Represent the interval $[0,1]$ as a finite union of intervals (open and closed):

$$[0,1] = \cup_{y_j\in\hat{D}_t^y}\left[y_j - \varepsilon_t, y_j + \varepsilon_t\right] \cup_{j=1}^{N_{d,t}+1}\left(y_{t,j-1}^d + \varepsilon_t, y_{t,j}^d - \varepsilon_t\right) \tag{4.C.109}$$

Then to bound $\sup_{1\le t\le T}\left\|B_{3,t}\right\|$ we write

$$\sup_{1\le t\le T} \|\gamma_t F_t(\hat{u}_t y) - \gamma_t F_t(u_t y)\| \le \sup_{1\le t\le T}\gamma_t \sup_{1\le t\le T} \|F_t(\hat{u}_t y) - F_t(u_t y)\| \tag{4.C.110}$$

and evaluate the L^2 norm for $\|F_t(\hat{\gamma}_t y) - F_t(\gamma_t y)\|$.

Consider

$$\|F_t(\hat{u}_t y) - F_t(u_t y)\|^2 \leq \Sigma_{j=1}^{N_{d,t}+1} \left[\int_{y_{t,j-1+\varepsilon_t}^d}^{y_{t,j}^d - \varepsilon_t} (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy + \int_{y_{tj-1}-\varepsilon_t}^{y_{tj-1}+\varepsilon_t} (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy \right]. \tag{4.C.111}$$

For the first part

$$\Sigma_{j=1}^{N_{d,t}+1} \int_{y_{t,j-1+\varepsilon_t}^d}^{y_{t,j}^d - \varepsilon_t} (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy \leq \int_0^1 (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy \tag{4.C.112}$$

Since there are no discontinuities over the domaine of integration then by (4.4.39) $F_t(\hat{u}_t y) - F_t(u_t y) \leq C(\hat{u}_t - u_t)^\alpha$ so that

$$\Sigma_{j=1}^{N_{d,t}+1} \int_{y_{t,j-1}^d}^{y_{t,j}^d} (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy \leq 2N_d C^2 (\hat{u}_t - u_t)^{2\alpha}. \tag{4.C.113}$$

For the second part consider

$$\Sigma_{j=1}^{N_{d,t}+1} \int_{y_{tj-1}-\varepsilon_t}^{y_{tj-1}+\varepsilon_t} (F_t(\hat{u}_t y) - F_t(u_t y))^2 dy \leq 2N_d \varepsilon_t^2. \tag{4.C.114}$$

where the integral is negligible as $\varepsilon_t \to 0$. Set $\varepsilon_t^2 < \kappa_T$. Then

$$\sup_{1\leq t\leq T} \left\| B_{3,t} \right\| \leq \sup_{1\leq t\leq T} \gamma_t \left(2N_d\right)^{1/2} (C+1)\, \kappa_T. \tag{4.C.115}$$

Then combining the bounds

$$\sup_{1\leq t\leq T} \left\| \hat{W}_t^0 - W_t^0 \right\| \leq O_p \left(\xi_T + \frac{2T}{\sqrt{n(T)}} \lambda_T + \lambda_T (2N_d)^{1/2} (C+1) \kappa_T \right), \tag{4.C.116}$$

thus under condition that $\xi_T = o_p(1)$; $\dfrac{T}{\sqrt{n(T)}} \lambda_T = o_p(1)$; $\lambda_T \kappa_T = o_p(1)$ as $T \to \infty$ consistency for $\sup_{1\leq t\leq T} \left\| \hat{W}_t^0 - W_t^0 \right\|$ follows:

$$\sup_{1\leq t\leq T} \left\| \hat{W}_t^0 - W_t^0 \right\| = o_p(1). \tag{4.C.117}$$

Since

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{W}_t^0 - W_t^0\| \leq \sup_{1\leq t\leq T}\left\|\hat{W}_t^0 - W_t^0\right\| \tag{4.C.118}$$

consistency for the average of norms follows. □

PROOF OF THEOREM 4.4.2 The assertion follows directly from Theorem 4.3 of Chang et al. (2016) combined with (4.4.43). □

Chapter 5

Semiparametric innovation-based tests of orthogonality and causality between two infinite-order cointegrated series

Forthcoming in *Advances in Econometrics: Essays in Honor of Joon Y. Park, 2022*

5.1. Introduction

Studying the dynamic relationship between two multivariate series is a fundamental objective of time series analysis in statistics and econometrics. For example, in econometrics, this can help one to understand the associated economic mechanisms. In this context, a basic problem consists in testing independence (or the absence of serial cross-correlation) between two vector processes. The seminal paper on this problem is due to Haugh (1976), who proposed a general procedure for testing independence between two covariance-stationary ARMA time series. His method is based on considering cross-correlations between residuals obtained after fitting univariate ARMA models on each series. Since the innovations of an ARMA model follow a white noise by assumption, this considerably simplifies the underlying distributional theory, and the corresponding tests are relatively simple to apply. Further, the corresponding statistics have a direct interpretation in terms of process innovations (or reduced-form shocks), a feature of interest in econometrics since innovations can often be interpreted as "shocks" to economic systems.

Consequently, the possibility of focusing on "shock cross-correlations" should be useful in econometric research.

The work of Haugh (1976) has been extended by several authors; see Hong (1996*a*), El Himdi and Roy (1997), Pham, Roy and Cédras (2003), Hallin and Saidi (2005), Bouhaddioui and Roy (2006, **?**), Hallin and Saidi (2007), Saidi (2007), and Bouhaddioui and Dufour (2008). Most of these studies focus on independence between two multivariate finite-order vector autoregressive (VAR) or vector autoregressive moving-average (VARMA) models. El Himdi and Roy (1997) extended the procedure developed by Haugh (1976) in order to test non-correlation between two time series in the context of multivariate stationary and invertible VARMA models. This result was used by Hallin and Saidi (2005) to develop a test which takes into account a possible pattern in the signs of cross-correlations at different lags. In a nonparametric setup, **?** proposed a test for independence between two autoregressive time series which is based on autoregressive rank scores, while Hong (1998) proposed a test based on empirical distribution functions.

The stationarity condition is often unrealistic and constitutes a strong limitation. Even though stationarity may be achieved in many cases by differencing each series (so that distributional complications are avoided), this type of transformation can distort our ability to identify or accurately measure parameters and relations of interest. It is typically more interesting to be able to work with the original series without prefiltering (like differencing). This is especially important if we wish to study cointegrating relationships.

Engle and Granger (1987) introduced the concept of cointegration, which is used in many studies across several fields. In the case of a finite-order autoregressive cointegrated vector, Ahn and Reinsel (1990) developed an efficient estimation method for Gaussian processes. Yap and Reinsel (1995) proposed full- and reduced-rank Gaussian estimation procedures for cointegrated VARMA processes. For a good discussion of the related models, see Lütkepohl (2001). By exploiting the estimation methods proposed by Yap and Reinsel (1995), Pham et al. (2003) generalized the main result of El Himdi and Roy (1997) to the case of two cointegrated (or partially nonstationary) VARMA series. They proposed test statistics based on residual cross-correlation matrices $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(12)}(j)$, $|j| \leq M$ [where M does not depend on the sample size n] between the two residual series $\hat{\boldsymbol{a}}_t^{(1)}$ and $\hat{\boldsymbol{a}}_t^{(2)}$ resulting from fitting the *true* VARMA models to each of the original series $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$. Under the hypothesis of non-correlation between the two series, they show that an arbitrary vector of

residual cross-correlations asymptotically follows a multivariate normal distribution.

In practice, a finite-order VAR model can be a rough approximation to the true data generating process of a multivariate time series. The "true" model may easily not be reducible to a parsimonious model with a small number of unknown parameters. From this perspective, a more flexible alternative approach assumes that the data are generated by an infinite-order autoregressive process. Such models lead one to consider a truncated (potentially long) autoregression as an approximation of the underlying process. In statistics and econometrics, one typically derives the properties of estimators and test criteria under the assumption of correct specification, even if model assumptions are clearly not fulfilled. For example, in VARMA estimation, it is well known that misspecification of the AR or MA orders can lead to inconsistent estimators. Further, the estimation of VARMA models is highly nonlinear and raises difficult identification complications (in the sense of multiple observationally equivalent representations).

The autoregressive model fitting approach has been successfully applied by several authors: Akaike (1969), Berk (1974) and Parzen (1974) for spectral density estimation, Parzen (1974), Lütkepohl (1985), Lewis and Reinsel (1985) and Bhansali (1996) for prediction, Park (1990) and Saikkonen (1992) for inference in cointegrated systems; see also **?**, Lütkepohl (2005) and Park, Shin and Wang (2010) . In previous work [**?**], we have generalized the work of El Himdi and Roy (1997) to the case of two stationary multivariate infinite-order autoregressive series VAR(∞). This result allows one to develop tests against serial cross-correlation at a particular lag or at a fixed number of lags j such as $|j| \leq M$, where M does not depend on the sample size n.

In the univariate stationary case, Hong (1996*c*) introduced an important extension of Haugh's procedure by proposing a class of spectral test statistics. His approach is semiparametric and valid for two infinite-order autoregressive series AR(∞). It is based on fitting an autoregressive model of order p to a series of n observations from each infinite-order autoregressive process. Following Berk (1974), the order p of the fitted autoregression is a function of the sample size. This approach was also used by Hong (1999), Duchesne and Roy (2003), Duchesne (2005) and Shao (2009) for the case of two univariate long memory processes. In Bouhaddioui and Roy (2006), it is extended to VAR(∞) models, hence protecting against misspecification of the underlying VARMA model. In contrast with Haugh's test, which is based on the residual cross-correlations at lag j such that $|j| \leq M$, the port-

manteau test $\mathscr{Q}_n$ is consistent for a large class of serial cross-correlations alternatives of an arbitrary form between the two series.

In this article, we propose a multivariate version of the weighted portmanteau statistic $\mathscr{Q}_n$, based on the sample cross-correlation matrices $\boldsymbol{R}^{(12)}_{\hat{\boldsymbol{a}}}(j)$, $|j| \leq n-1$, between the residuals $\hat{\boldsymbol{a}}^{(1)}_t$ and $\hat{\boldsymbol{a}}^{(2)}_t$. The latter are obtained by approximating two multivariate IVAR(∞) series with finite-order autoregressions whose order increases with the sample size at an appropriate rate. The test statistics continue to have a $\mathscr{N}(0,1)$ asymptotic distribution under the hypothesis of independence of the two series. The tests are consistent against serial cross-correlation of arbitrary form.

5.2. Framework and preliminary results

Following the notations of Saikkonen (1992), Saikkonen and Lütkepohl (1996) and Bouhaddioui and Dufour (2008), we consider a d-dimensional process $\boldsymbol{X} = \{X_t : t \in \mathbb{Z}\}$ partitioned into two subprocesses $X_i = \{\mathbf{X}_{it} : t \in \mathbb{Z}\}$, $i = 1, 2$, with d_1 and d_2 components respectively ($d_1 + d_2 = d$). The data generating process has the form:

$$\boldsymbol{X}_{1t} = \boldsymbol{C}_1 \boldsymbol{X}_{2t} + \boldsymbol{\varepsilon}_{1t}\,, \tag{5.2.1}$$

$$\Delta \boldsymbol{X}_{2t} = \boldsymbol{\varepsilon}_{2t}\,, \tag{5.2.2}$$

where $\boldsymbol{C}_1$ is a fixed $d_1 \times d_2$ matrix, Δ is the usual difference operator, and $\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}'_{1t}, \boldsymbol{\varepsilon}'_{2t})'$ is a stationary process with zero mean and continuous spectral density matrix positive definite at frequency zero. X_{2t} is an integrated vector process of order one (with no cointegrating relationship), while X_{1t} and X_{2t} are cointegrated. By taking first differences in (5.2.1), we see that

$$\Delta \boldsymbol{X}_t = \begin{bmatrix} -\mathbb{I}_{d_1} & \boldsymbol{C}_1 \\ \mathbf{o} & \mathbf{o} \end{bmatrix} \boldsymbol{X}_{t-1} + \boldsymbol{\nu}_t = \boldsymbol{J}\boldsymbol{\Theta}' \boldsymbol{X}_{t-1} + \boldsymbol{\nu}_t \tag{5.2.3}$$

where $\mathbb{I}_d$ is the identity matrix of order d, $\boldsymbol{J}' = [-\mathbb{I}_{d_1} \vdots \mathbf{o}]$, $\boldsymbol{\Theta}' = [\mathbb{I}_{d_1} \vdots -\boldsymbol{C}_1]$, $\boldsymbol{\nu}_t = (\boldsymbol{\nu}'_{1t}, \boldsymbol{\nu}'_{2t})'$ is a nonsingular transformation of $\boldsymbol{\varepsilon}_t$ defined by

$$\boldsymbol{\nu}_{1t} = \boldsymbol{\varepsilon}_{1t} + \boldsymbol{C}_1 \boldsymbol{\varepsilon}_{2t}\,, \quad \boldsymbol{\nu}_{2t} = \boldsymbol{\varepsilon}_{2t}\,, \tag{5.2.4}$$

$$\boldsymbol{X}_t := \begin{bmatrix} \boldsymbol{X}_{1t} \\ \boldsymbol{X}_{2t} \end{bmatrix}, \quad \nu_t := \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \end{bmatrix}. \tag{5.2.5}$$

The notation $\boldsymbol{A} = [\boldsymbol{A}_1 \vdots \boldsymbol{A}_2]$ means that the matrix $\boldsymbol{A}$ is partitioned into a matrix $\boldsymbol{A}_1$ consisting of the first d_1 columns and a matrix $\boldsymbol{A}_2$ with d_2 columns.

We suppose that ν_t (hence also ε_t) has an infinite-order autoregressive representation

$$\sum_{l=0}^{\infty} \boldsymbol{G}_l \nu_{t-l} = \boldsymbol{a}_t \tag{5.2.6}$$

where $\boldsymbol{G}_0 = \mathbb{I}_d$, $\boldsymbol{a}_t$ is a sequence of independent and identically distributed random vectors such that $\mathbb{E}(\boldsymbol{a}_t) = 0$ and $\mathbb{E}(\boldsymbol{a}_t \boldsymbol{a}_t') = \boldsymbol{\Sigma}_{\boldsymbol{a}}$ is positive definite, and the roots of the equation

$$\det\{\boldsymbol{I}_d - \sum_{l=1}^{\infty} \boldsymbol{G}_l z^l\} = 0 \tag{5.2.7}$$

all lie outside the unit circle $|z| = 1$; $\det\{\boldsymbol{A}\}$ denotes the determinant of the square matrix $\boldsymbol{A}$. We also assume that the following summability condition holds:

$$\sum_{l=1}^{\infty} l^{\bar{\delta}} \|\boldsymbol{G}_l\| < \infty \quad \text{for some } \bar{\delta} \geq 1 \tag{5.2.8}$$

where $\|\cdot\|$ is the Euclidean matrix norm defined by $\|\boldsymbol{A}\|^2 = \mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})$. This is a standard condition for weakly stationary processes, which ensures that the process is well defined. It also implies that the process ν_t and, consequently X_t, can be approximated by an autoregression of finite order $p_n = p(n)$ where n is the sample size and p_n can grow with n. More explicitly, we assume that p_n satisfies the following condition.

Assumption 5.2.1 *There is a sequence of positive integers p_n such that*

$$n^{-1/3} p_n \to 0 \quad \textit{and} \quad \sqrt{p_n} \sum_{l=p_n+1}^{\infty} \|\boldsymbol{G}_l\| \to 0 \quad \textit{as } n \to \infty. \tag{5.2.9}$$

The condition $p_n = o(n^{1/3})$ for the rate of increase of p_n ensures that enough sample information is asymptotically available for estimators to have standard limiting distributions. The condition $\sqrt{p_n} \sum_{j=p_n+1}^{\infty} \|\boldsymbol{G}_j\| \to 0$ imposes a lower bound on the growth rate of p_n,

which ensures that the approximation error of the true underlying model by a finite-order autoregression gets small when the sample size increases. A more detailed discussion of these conditions is available in Burnham and Anderson (2002) and Lütkepohl (2005).

Using the equations (5.2.3) - (5.2.6) and rearranging terms, we obtain the autoregressive *error correction model* (ECM) representation

$$\Delta \boldsymbol{X}_t = \boldsymbol{\Psi}\boldsymbol{\Theta}'\boldsymbol{X}_{t-1} + \sum_{l=1}^{p_n} \boldsymbol{\Pi}_l \boldsymbol{\Delta}\boldsymbol{X}_{t-l} + \boldsymbol{e}_t(n),\ t = p_n+1,\ p_n+2,\ \ldots, \tag{5.2.10}$$

$$\boldsymbol{e}_t(n) = \boldsymbol{a}_t - \sum_{l=p_n+1}^{\infty} \boldsymbol{G}_l \boldsymbol{\nu}_{t-l}, \quad \boldsymbol{\Psi} = -\sum_{l=0}^{p_n} \boldsymbol{G}_l \boldsymbol{J}\,, \tag{5.2.11}$$

where $\boldsymbol{\Psi}$ is a $d \times d_1$ a full-column rank matrix (at least for p_n large enough). Details for this derivation can be found in Saikkonen and Lütkepohl (1994) and Saikkonen and Luukkonen (1997). Note the coefficient matrices $\boldsymbol{\Pi}_l$ $(l = 1, \ldots, p_n)$ are functions of $\boldsymbol{\Theta}$ and $\boldsymbol{G}_l$ $(l = 1, 2, ...)$, and they depend on p_n. Furthermore, the sequence $\boldsymbol{\Pi}_l$ $(l = 1, \ldots, p_n)$ is absolutely summable as $p_n \to \infty$.

The autoregressive ECM in (5.2.10) can be rewritten in a pure vector autoregressive (VAR) form

$$\boldsymbol{X}_t = \sum_{l=1}^{p_n+1} \boldsymbol{\Phi}_l \boldsymbol{X}_{t-l} + \boldsymbol{e}_t(n) \tag{5.2.12}$$

where $\boldsymbol{\Phi}_1 = \mathbb{I}_d + \boldsymbol{\Psi}\boldsymbol{\Theta}' + \boldsymbol{\Pi}_1$, $\boldsymbol{\Phi}_l = \boldsymbol{\Pi}_l - \boldsymbol{\Pi}_{l-1}$, $l = 2, \ldots, p_n$ and $\boldsymbol{\Phi}_{p_n+1} = -\boldsymbol{\Pi}_{p_n}$. Although the $\boldsymbol{\Pi}_l$ depend on p_n, the same is not true for the $\boldsymbol{\Phi}_l$ except for $\boldsymbol{\Phi}_{p_n+1}$.

Saikkonen and Lütkepohl (1996) derived the asymptotic properties of the multivariate least square (LS) estimators of the VAR coefficients under a standard assumption. Let

$$\boldsymbol{\Phi}(p_n) = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{p_n}] \tag{5.2.13}$$

be the matrix of the first p_n autoregressive parameter matrices in the representation (5.2.12), and denote by $\hat{\boldsymbol{\Phi}}(p_n) = [\hat{\boldsymbol{\Phi}}_1, \ldots, \hat{\boldsymbol{\Phi}}_{p_n}]$ the corresponding LS estimator. The following proposition gives a direct result on the asymptotic properties of the estimator $\hat{\boldsymbol{\Phi}}(p_n)$. It can be proved using the techniques very similar to those used by Saikkonen (1992, part (i) of Theorem 3.2); see also Saikkonen and Lütkepohl (1996, Theorem 2).

Proposition 5.2.1 ASYMPTOTIC PROPERTIES OF THE AUTOREGRESSIVE PARAMETER ESTIMATORS. *Let* $\{X_t\}$ *be a process which satisfies* (5.2.3) - (5.2.6) *with*

$$\mathbb{E}|a_{it}a_{jt}a_{kt}a_{lt}| < \gamma_4 < \infty, \quad 1 \le i, j, k, l \le d\,. \tag{5.2.14}$$

where $\boldsymbol{a}_t := (a_{1t}, \ldots, a_{dt})'$. *If Assumption* 5.2.1 *holds, then*

$$\|\hat{\boldsymbol{\Phi}}(p_n) - \boldsymbol{\Phi}(p_n)\| = O_p(p_n^{1/2}/n^{1/2})\,. \tag{5.2.15}$$

This proposition is formulated for the first p_n coefficient matrices, whereas the fitted model is a VAR(p_n+1) where p_n goes to infinity with the sample size n. Dropping the last lag in deriving the consistency of the estimators will not affect the asymptotic distribution of the test statistic; see Lütkepohl (2005). Details on the estimates of the $\boldsymbol{\Phi}_l$ matrices are given in Saikkonen and Lütkepohl (1996). This result can be viewed as a generalization of Theorem 1 in Lewis and Reinsel (1985) to infinite-order stationary vector autoregressive processes.

Let us now consider two processes $\boldsymbol{X}^{(h)} = \{\boldsymbol{X}_t^{(h)} : t \in \mathbb{Z}\}$, $h = 1, 2$, with m_1 and m_2 components respectively, each of which satisfies an IVAR(∞) model of the form (5.2.3) - (5.2.6) with $m_h = d_1^{(h)} + d_2^{(h)}$, $h = 1, 2$, where $d_1^{(h)}$ and $d_2^{(h)}$ replace d_1 and d_2 for $\boldsymbol{X}^{(h)}$. The coefficients of the two processes may differ. We wish to decide whether $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are independent against an alternative where they are correlated at some lag. Following Pham et al. (2003), the independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ can be tested by testing non-correlation between the corresponding innovation processes $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$. This leads one to consider the hypothesis:

$$\mathscr{H}_0 : \boldsymbol{\rho}_{\boldsymbol{a}}^{(12)}(j) = \mathbf{0}\,, \text{ for all } j \in \mathbb{Z}\,, \tag{5.2.16}$$

where

$$\boldsymbol{\rho}_{\boldsymbol{a}}^{(12)}(j) = [\boldsymbol{D}(\boldsymbol{\Sigma_1})]^{-1/2}\boldsymbol{\Gamma}_{\boldsymbol{a}}^{(12)}(j)[\boldsymbol{D}(\boldsymbol{\Sigma_2})]^{-1/2}, \quad \boldsymbol{\Gamma}_{\boldsymbol{a}}^{(hi)}(j) = \boldsymbol{E}\big[\boldsymbol{a}_t^{(h)}\big(\boldsymbol{a}_{t-j}^{(i)}\big)'\big]\,,\ j \in \mathbb{Z}\,, \tag{5.2.17}$$

$$\boldsymbol{\Sigma}_h = \boldsymbol{\Gamma}_{\boldsymbol{a}}^{(hh)}(0)\,,\ \boldsymbol{D}(\boldsymbol{\Sigma_h}) = \operatorname{diag}\{\boldsymbol{\Sigma}_h\}\,,\ \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Sigma}_2 \end{bmatrix},\ h, i = 1, 2\,. \tag{5.2.18}$$

$\rho_{\boldsymbol{a}}^{(12)}(j)$ represents the cross-correlation matrix at lag j between the two innovation processes. On setting

$$\boldsymbol{b}_t := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_t = \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1/2} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Sigma}_2^{-1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_{1t} \\ \boldsymbol{a}_{2t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1}/2\boldsymbol{a}_{1t} \\ \boldsymbol{\Sigma}_2^{-1/2}\boldsymbol{a}_{2t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{1t} \\ \boldsymbol{b}_{2t} \end{bmatrix}, \quad (5.2.19)$$

we see that $\rho_{\boldsymbol{b}}^{(12)}(j) = \boldsymbol{\Gamma}_{\boldsymbol{b}}^{(12)}(j) = \rho_{\boldsymbol{b}}^{(12)}(j)$, for all $j \in \mathbb{Z}$, so that $\mathscr{H}_0$ is equivalent to

$$\rho_{\boldsymbol{b}}^{(12)}(j) = 0\,, \text{ for all } j \in \mathbb{Z}\,. \quad (5.2.20)$$

This equivalence plays a central role for proving the required distributional results stated below.

5.3. Test statistics and asymptotic null distributions

Based on a realization $X_1^{(h)}, \ldots, X_n^{(h)}$ of length n, for $h = 1, 2$, a finite-order autoregressive model VAR$(p_n^{(h)} + 1)$ is fitted to each one of these two series. The order $p_n^{(h)}$ depends on the sample size n. The resulting residuals are given by

$$\hat{\boldsymbol{a}}_t^{(h)} = \begin{cases} \boldsymbol{X}_t^{(h)} - \sum_{l=1}^{p_n^{(h)}+1} \hat{\boldsymbol{\Phi}}_l^{(h)}(n)\boldsymbol{X}_{t-l}^{(h)} & \text{if } t = p_n^{(h)} + 2, \ldots, n\,, \\ \mathbf{0} & \text{if } t \le p_n^{(h)} + 1\,, \end{cases} \quad (5.3.1)$$

where the matrices $\hat{\boldsymbol{\Phi}}_l^{(h)}(n)$ are the OLS estimators of $\boldsymbol{\Phi}_l^{(h)}(n)$, and $h = 1, 2$. We can also use the conditional maximum likelihood estimator of the error correction form of the model as discussed by Ahn and Reinsel (1990) and Reinsel (1993), or some other estimator with the same rate of convergence. We now consider the residual sample (cross-)covariance matrices

$$\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(hi)}(j) = \begin{cases} n^{-1}\sum_{t=j+1}^{n} \hat{\boldsymbol{a}}_t^{(h)}(\hat{\boldsymbol{a}}_{t-j}^{(i)})' & \text{if } 0 \le j \le n-1 \\ n^{-1}\sum_{t=-j+1}^{n} \hat{\boldsymbol{a}}_{t+j}^{(h)}(\hat{\boldsymbol{a}}_t^{(i)})' & \text{if } -n+1 \le j \le 0 \end{cases} \quad (5.3.2)$$

where $h, i = 1, 2$, and the corresponding cross-correlation matrices

$$\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(hi)}(j) = [\boldsymbol{D}\big(\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(hh)}(0)\big)]^{-1/2}\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(hi)}(j)\,[\boldsymbol{D}\big(\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(ii)}(0)\big)]^{-1/2} \quad (5.3.3)$$

where $\boldsymbol{D}\big(C_{\hat{\boldsymbol{a}}}^{(hh)}(0)\big) = \mathrm{diag}\{C_{\hat{\boldsymbol{a}}}^{(hh)}(0)\}$. The orthogonality tests we consider are based on $C_{\hat{\boldsymbol{a}}}^{(12)}(j)$ and $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(12)}(j)$. In the sequel, we suppose that $X^{(h)}$ satisfies (5.2.3) for $h = 1, 2$. We wish to test the null hypothesis $\mathscr{H}_0$ using the cross-correlation matrices $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(hi)}(j)$, $j \in \mathbb{Z}$.

In the univariate case, Hong (1996*c*) proposed a portmanteau-type statistic based on the sum of the weighted squared cross-correlations $r_{\hat{a}}^{(12)}(j)$ at all possible lags between the residual series:

$$\mathscr{Q}_n = \frac{n\sum_{j=1-n}^{n-1} k^2(j/M) r_{\hat{a}}^{(12)}(j)^2 - S_n(k)}{\{2D_n(k)\}^{1/2}} \tag{5.3.4}$$

where $k(\cdot)$ is an arbitrary kernel function [see Table 5.2 for examples] and M is a smoothing parameter, while $S_n(k)$ and $D_n(k)$ are normalization coefficients which depend on the kernel $k(\cdot)$:

$$S_n(k) = \sum_{j=1-n}^{n-1} \big(1 - \frac{|j|}{n}\big) k^2(j/M),\; D_n(k) = \sum_{j=2-n}^{n-2} \big(1 - \frac{|j|}{n}\big)\big(1 - \frac{|j|+1}{n}\big)\, k^4(j/M)\,. \tag{5.3.5}$$

They correspond to the asymptotic mean and variance of the weighted sum. In multivariate time series, the squared cross-correlation $r_{\hat{a}}^{(12)}(j)^2$ in (5.3.4) is replaced by a quadratic form in the vector $\boldsymbol{r}_{\hat{\boldsymbol{a}}}^{(12)}(j) = \mathrm{vec}[\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(12)}(j)]$. For $\mathscr{H}_0$, the test statistic is based on the following sum of weighted quadratic forms at all possible lags:

$$\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) = \sum_{j=1-n}^{n-1} k^2(j/M) Q_{\hat{\boldsymbol{a}}}^{(12)}(j)\,, \tag{5.3.6}$$

$$Q_{\hat{\boldsymbol{a}}}^{(12)}(j) := n\,\boldsymbol{r}_{\hat{\boldsymbol{a}}}^{(12)}(j)'\big[\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(22)}(0)^{-1} \otimes \boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(11)}(0)^{-1}\big]\boldsymbol{r}_{\hat{\boldsymbol{a}}}^{(12)}(j)\,, \quad \hat{\boldsymbol{\Sigma}} := \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_1 & \mathbf{o} \\ \mathbf{o} & \hat{\boldsymbol{\Sigma}}_1 \end{bmatrix}, \tag{5.3.7}$$

where $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(hh)}(0)$ is a consistent estimator of the correlation matrix $\boldsymbol{\rho}_{\boldsymbol{a}}^{(h)}$ of the process $\boldsymbol{a}^{(h)}$, and $k(\cdot)$ is a suitable kernel function. The parameter M is a truncation point when the kernel has compact support, or a smoothing parameter when the kernel support is unbounded. We suppose that M is function of n ($M = M_n$) such that $M_n \to \infty$ and $M_n/n \to 0$ as $n \to \infty$. The most commonly used kernels typically give more weight to lower lags and less weight to higher ones. An exception is the truncated uniform kernel $k_T(z) = \mathbb{I}[|z| \leq 1]$, where $\mathbb{I}(A)$ represents the indicator function of the set A, which gives the same weight to all lags. The

asymptotic distribution of $Q_{\hat{\boldsymbol{a}}}(j)$ is given in Bouhaddioui and Dufour (2008). In the sequel, we suppose that the kernel function k and the order $p_n^{(h)}$respectively satisfy the following assumptions.

Assumption 5.3.1 *The kernel function $k : \mathbb{R} \to [-1,1]$ is a symmetric function, continuous at zero, with at most a finite number of discontinuity points, such that $k(0) = 1$ and $\int_{-\infty}^{+\infty} k^2(z)dz < \infty$.*

Assumption 5.3.2 *The orders $p_n^{(h)}$, $h = 1,2$, satisfy the following conditions:*

$$(i) \quad p_n^{(h)} = o\big(n^{1/2}/M^{1/4}\big), \quad (ii) \quad n \sum_{j=p_n^{(h)}+1}^{\infty} ||\boldsymbol{\Phi}_j^{(h)}||^2 = o\big(n^{1/2}/M^{1/4}\big). \tag{5.3.8}$$

Note that the two conditions (i) and (ii) imply that the order $p_n^{(h)}$ satisfies Assumption 5.2.1. The property $k(0) = 1$ implies that the weights assigned to the lower lags are close to unity. The square integrability of $k(\cdot)$ implies that $k(z) \to 0$ as $|z| \to \infty$, so that less weight is given to $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(12)}(j)$ as j increases. Note that all the kernels used in spectral analysis satisfy Assumption 5.3.1; see Priestley (1981, Section 6.2.3). For hypothesis $\mathscr{H}_0$, the test statistic is a standardized version of $\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}})$:

$$\mathscr{Q}_n = \frac{\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - m_1 m_2 S_n(k)}{\sqrt{2m_1 m_2 D_n(k)}}, \tag{5.3.9}$$

where the smoothing parameter $M_n \to \infty$ and $M_n/n \to 0$ when $n \to \infty$.

This test statistic can be viewed as a normalized version of the $\mathscr{L}_2$-norm of a kernel-based estimator of the cross-coherency function between the two innovation series. $S_n(k)$ and $D_n(k)$ represent the asymptotic mean and variance of $\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}})$ under $\mathscr{H}_0$. If $k(\cdot)$ is the truncated uniform kernel, apart from the standardization factors $S_n(k)$ and $D_n(k)$, $\mathscr{Q}_n$ corresponds to the multivariate version of Haugh's statistic used by Pham et al. (2003) for the finite-order cointegrated case, and by Bouhaddioui and Dufour (2008) for the infinite-order case, namely

$$P_M = \sum_{j=-M}^{M} Q_{\hat{\boldsymbol{a}}}(j). \tag{5.3.10}$$

In this case, M is a fixed integer that does not depend on the sample size n. The properties of P_M in the stationary VAR(∞) context and cointegrated IVAR(∞) are studied respectively in **?** and Bouhaddioui and Dufour (2008). As it will be seen below, many kernels k yield tests that are more powerful than P_M.

In the case of testing independence, under some conditions on the smoothing parameter M and if the kernel k verifies Assumption 5.3.1, one sees easily that

$$M^{-1}S_n(k) \to S(k)\ ,\ M^{-1}D_n(k) \to D(k), \tag{5.3.11}$$

where

$$S(k) = \int_{-\infty}^{+\infty} k^2(z)dz\ ,\ D(k) = \int_{-\infty}^{+\infty} k^4(z)dz. \tag{5.3.12}$$

An alternative statistic is obtained by replacing $S_n(k)$ and $D_n(k)$ by their asymptotic approximations $MS(k)$ an $MD(k)$ respectively and is defined by

$$\mathscr{Q}_n^* = \frac{\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - Mm_1m_2S(k)}{\sqrt{2Mm_1m_2D(k)}}. \tag{5.3.13}$$

Both $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ have the same asymptotic null distribution and power properties.

The statistic $\mathscr{Q}_n$ can also be expressed in term of the autocovariances $\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(hh)}(0)$ and the cross-covariances $\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(12)}(j)$ of the same residual series. Invoking Lemma 4.1 of El Himdi and Roy (1997), the quadratic form $\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}})$ can be written as follows in terms of the residual covariances:

$$\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) = n\sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)'\left[\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(22)}(0)^{-1} \otimes \boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(11)}(0)^{-1}\right]\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) \tag{5.3.14}$$

with $\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) = \text{vec}[\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(12)}(j)]$. Let us now consider the "pseudo-statistic"

$$\mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma}) = n\sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)'\left(\boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j) \tag{5.3.15}$$

where $\boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)$ is defined as $\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)$ with the residuals $\hat{\boldsymbol{a}}_t^{(1)}$ and $\hat{\boldsymbol{a}}_t^{(2)}$ replaced by the un-

observable innovation series $\boldsymbol{a}_t^{(1)}$ and $\boldsymbol{a}_t^{(2)}$, $t = 1, \ldots, n$, and

$$\mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) = n \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' (\boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) \,. \tag{5.3.16}$$

Thus, with $\hat{\boldsymbol{\Sigma}}_h = C_{\hat{\boldsymbol{a}}}^{(hh)}(0)$, $h = 1, 2$, we can write the statistic $\mathscr{Q}_n$ as

$$\begin{aligned} \mathscr{Q}_n &= \frac{\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - m_1 m_2 S_n(k)}{\sqrt{2 m_1 m_2 D_n(k)}} \\ &= \frac{\mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma}) - m_1 m_2 S_n(k)}{\sqrt{2 m_1 m_2 D_n(k)}} + \frac{\mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) - \mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma})}{\sqrt{2 m_1 m_2 D_n(k)}} + \frac{\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma})}{\sqrt{2 m_1 m_2 D_n(k)}} \end{aligned} \tag{5.3.17}$$

Since the quantity $\mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma})$ depends only on the stationary process $\boldsymbol{a}$, the result of Lemma 3.1 in Bouhaddioui and Roy (2006) is still valid. We conclude that

$$\frac{\mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma}) - m_1 m_2 S_n(k)}{\sqrt{2 m_1 m_2 D_n(k)}} \xrightarrow{L} \mathscr{N}(0, 1) \,. \tag{5.3.18}$$

The asymptotic distribution of $\mathscr{Q}_n$ follows from the next two propositions.

Proposition 5.3.1 APPROXIMATION OF THE PSEUDO-STATISTIC. *Suppose* $X^{(1)} = \{X_t^{(1)} : t \in \mathbb{Z}\}$ *and* $X^{(2)} = \{X_t^{(2)} : t \in \mathbb{Z}\}$ *satisfy the IVAR*(∞) *model* (5.2.3) - (5.2.6) *along with Assumption* 5.3.1 *and the bounded moment condition*

$$\mathbb{E}|a_{it}^{(h)} a_{jt}^{(h)} a_{kt}^{(h)} a_{lt}^{(h)}| < \gamma_4 < \infty, 1 \leq i, j, k, l \leq m_h \,. \tag{5.3.19}$$

Let $M = M_n$, *with* $M_n \to \infty$ *and* $M_n/n \to 0$ *as* $n \to \infty$, *and suppose that* $p_n^{(h)}$, $h = 1, 2$, *satisfy Assumption* 5.3.2. *If the processes* $\boldsymbol{a}^{(1)}$ *and* $\boldsymbol{a}^{(2)}$ *are independent, then*

$$\mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) - \mathscr{T}(\boldsymbol{a}, \boldsymbol{\Sigma}) = o_p(M^{1/2}) \,. \tag{5.3.20}$$

Proposition 5.3.2 ASYMPTOTIC EQUIVALENCE OF THE TEST STATISTIC. *Under the*

assumptions of Proposition 5.3.1*, we have*

$$\frac{\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma})}{\sqrt{2m_1m_2D_n(k)}} \xrightarrow{p} 0\,. \tag{5.3.21}$$

Our main result is a simple consequence of Propositions 5.3.1 - 5.3.2, as follows.

Theorem 5.3.3 NULL ASYMPTOTIC DISTRIBUTION. *Under the assumptions of Proposition* 5.3.1*, the statistic* $\mathscr{Q}_n$ *defined by* (5.3.9) *has an asymptotic* $\mathscr{N}(0,1)$ *distribution,* i.e. $\mathscr{Q}_n \xrightarrow{L} \mathscr{N}(0,1)$.

5.4. Consistency of the generalized tests

We now investigate the asymptotic power of the test $\mathscr{Q}_n$ under fixed alternatives. We consider a fixed alternative $\mathscr{H}_1$ of serial cross-correlation between the two innovation processes $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$ with the following assumption.

Assumption 5.4.1 *The two innovation processes*

$$\boldsymbol{a}_t^{(1)} = (a_{1,t}^{(1)}, \ldots, a_{m_1,t}^{(1)})' \text{ and } \boldsymbol{a}_t^{(2)} = (a_{1,t}^{(2)}, \ldots, a_{m_2,t}^{(2)})', \quad t \in \mathbb{Z}, \tag{5.4.1}$$

are jointly fourth-order stationary, and their cross-correlation structure is such that $\boldsymbol{\Gamma}_{\boldsymbol{a}}^{(12)}(j) \neq \mathbf{o}$ *for at least one value of* j*, with*

$$\sum_{j=-\infty}^{+\infty} \|\boldsymbol{\Gamma}_{\boldsymbol{a}}^{(12)}(j)\|^2 < \infty\,, \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} |\kappa_{uvuv}(0,i,j,l)| < \infty, \tag{5.4.2}$$

where $\kappa_{uvuv}(0,i,j,l)$ *is the fourth cumulant of the joint distribution of* $a_{u,t}^{(1)}$*,* $a_{v,t+i}^{(2)}$*,* $a_{u,t+j}^{(1)}$*,* $a_{v,t+l}^{(2)}$.

The following theorem gives conditions for the consistency of $\mathscr{Q}_n$ under a fixed alternative.

Theorem 5.4.1 GLOBAL POWER. *Let* $X^{(1)}$ *and* $X^{(2)}$ *be two multivariate processes which follow the IVAR*(∞) *model* (5.2.3) - (5.2.6)*, and suppose that their innovation processes*

$\boldsymbol{a}^{(1)}$ *and* $\boldsymbol{a}^{(2)}$ *satisfy Assumption* 5.4.1. *If the kernel* $k(\cdot)$ *satisfies Assumption 5.3.1 and if* $p_n^{(h)}$, $h = 1, 2$, *satisfy*

$$p_n^{(h)^2} = o\left(\frac{n}{M}\right) \qquad , \qquad \sum_{j=p_n^{(h)}+1}^{\infty} \|\boldsymbol{\Phi}_j^{(h)}\|^2 = o\left(M^{-1}\right), \tag{5.4.3}$$

then, for any sequence of constants $C(n,M)$ *such that* $C(n,M) = o(n/M^{1/2})$,

$$P[\mathscr{Q}_n > C(n,M)] \to 1\,. \tag{5.4.4}$$

This theorem entails that the test based on $\mathscr{Q}_n$ is consistent against the general class of dependence alternatives described by Assumption 5.4.1. The slower M grows, the faster $\mathscr{Q}_n$ goes to infinity. To investigate the relative efficiency of $\mathscr{Q}_n$, one can use the Bahadur's asymptotic slope criterion defined in Bahadur (1960); see also Hong (1996*a*, 1996*c*) and Bouhaddioui and Roy (2006). As in Bouhaddioui and Roy (2006), we can show that the relative efficiency of the kernel $k_2(\cdot)$ with respect to $k_1(\cdot)$ when $M = n^{\nu}$ is given by

$$ARE_B(k_2, k_1) = \left\{\frac{D(k_1)}{D(k_2)}\right\}^{1/(2-\nu)}. \tag{5.4.5}$$

We can then proceed like Bouhaddioui (2002) and Hong (1996*a*, 1996*c*) to derive the kernel that maximizes the asymptotic slope over appropriate classes of kernel functions. For example, consider the following class of kernels:

$$\kappa(\tau) = \{k(\cdot) : \text{Assumption 5.3.1 is satisfied, } k^{(2)} = \tau^2/2,\ K(\lambda) \geq 0 \text{ for } \lambda \in (-\infty, +\infty)\} \tag{5.4.6}$$

where

$$k^{(2)} = \lim_{z\to 0}[1-k(z)]/z^2 \text{ and } K(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} k(z)e^{-iz\lambda}dz. \tag{5.4.7}$$

This class contains the Daniell, Parzen and quadratic-spectral kernels (among others). Using Theorem 1 of Ghosh and Huang (1991), we can see that the Daniell kernel [see Table 5.2] maximizes the asymptotic slope of $\mathscr{Q}_n$ over $\kappa(\tau)$; for a similar argument, see Bouhaddioui (2002). As mentioned in Bouhaddioui and Roy (2006), a test with a greater asymptotic slope may be expected to have a greater power for a fixed alternative than one with

a smaller asymptotic slope. However, there is no clear analytical relationship between the slope of a test and its power function. For a specific alternative, we cannot conclude that a test with greater asymptotic slope should be automatically preferred to one with a smaller asymptotic slope without further analysis of the finite-sample properties of the two test statistics.

5.5. Local power analysis

In this section ,we study the power of the test proposed above against a class of local alternatives of the form

$$\mathscr{H}_{\boldsymbol{a}}(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}) : \boldsymbol{\Gamma}_{\boldsymbol{b}}^{(12)}(j) = \frac{M^{1/4}}{n^{1/2}} \boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}(j) \text{ , for all } j \in \mathbb{Z},$$

where $\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)} = \{\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}(j)\}_{j\in\mathbb{Z}}$ is a sequence of $m_1 \times m_2$ cross-correlation matrices such that only finite elements of $\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}$ are non-zero elements. Let

$$\lambda_{\boldsymbol{b}}^{(12)}(j) = \text{vec}[\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}(j)], \tag{5.5.1}$$

$$\beta(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}) = \sum_{j=-\infty}^{\infty} \lambda_{\boldsymbol{b}}^{(12)}(j)' \lambda_{\boldsymbol{b}}^{(12)}(j). \tag{5.5.2}$$

The following theorem establishes the asymptotic distribution of $\mathscr{Q}_n$ under the local alternative $\mathscr{H}_{\boldsymbol{a}}(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)})$.

Theorem 5.5.1 LOCAL POWER. *Let $X^{(1)}$ and $X^{(2)}$ be two multivariate processes which follow the IVAR(∞) model* (5.2.3) - (5.2.6)*, and suppose that their innovation processes $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$ satisfy Assumption* 5.4.1*. If the kernel $k(\cdot)$ satisfies Assumption* 5.3.1 *and if $p_n^{(h)}$, $h = 1, 2$, satisfy*

$$p_n^{(h)^2} = o(n/M), \qquad \sum_{j=p_n^{(h)}+1}^{\infty} \|\boldsymbol{\Phi}_j^{(h)}\|^2 = o\left(M^{-1}\right), \tag{5.5.3}$$

then, under $\mathscr{H}_{\boldsymbol{a}}(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)})$,

$$\mathscr{Q}_n \xrightarrow{L} \mathscr{N}[\beta(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)})/\sqrt{2m_1m_2D(k)},\, 1]\,. \tag{5.5.4}$$

where $\beta(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)})$ *is defined in* (5.5.2).

Theorem 5.5.1 shows that the test $\mathscr{Q}_n$ has non-trivial power against a class of local alternatives converging to $\mathscr{H}_0$ at the rate of $\dfrac{M^{1/4}}{n^{1/2}}$. The power depends on the kernel function k through $D(k)$. Similarly, we note that increasing slowly the parameter M, the divergence of the test statistic to infinity is faster and consequently, the test is more powerful.

5.6. Simulation study

In the previous sections, we have studied the asymptotic distribution of the test statistics. Here we investigate the finite-sample properties of the proposed test statistics, in particular their exact level and power. To do this, we performed a small Monte Carlo study. In addition to the test statistics discussed in the preceding sections, we also consider the nonstationary multivariate version of the Haugh statistic [previously studied by Pham et al. (2003)]:

$$P_M^* = \sum_{j=-M}^{M} \frac{n}{n-|j|} Q_{\hat{\boldsymbol{a}}}(j) \tag{5.6.1}$$

where $Q_{\hat{\boldsymbol{a}}}^{(12)}(j)$ is given by (5.3.7). P_M^* is a slightly modified version of P_M defined by (5.3.10).

5.6.1. Description of the experiment

In the simulation experiment, we considered bivariate series $\{X_t^{(1)}\}$ and $\{X_t^{(2)}\}$ generated from (joint) 4-dimensional **VAR**(2), **VARMA**(1,1) and $\mathbf{VAR}_{\delta}(1)$ models (see Table 5.1). In the first two models, the two subprocesses $X^{(1)}$ and $X^{(2)}$ are independent bivariate **VAR**(2) or **VARMA**(1,1) and served for the level study and the corresponding submodels are partially nonstationary and invertible. The third one, in which there is instantaneous correlation between the two innovation series, was used for the power study. The correlation depends on a parameter δ and the values $\delta = 1.0,\ 1.5$ and 2 were chosen. For each model,

two series lengths ($n = 100,\ 200$) were considered. With the statistics $\mathcal{Q}_n$ and $\mathcal{Q}_n^*$ defined by (5.3.9) and (5.3.13), we used the four kernels described in Table 5.2. For each kernel, the following three truncation values M were employed: $M = [ln(n)]$, $[3n^{0.2}]$ and $[3n^{0.3}]$ ($[a]$ denotes the integer part of a). These rates are discussed in Hong (1996*a*, p. 849). They lead respectively to $M = 5, 8, 12$ for the series length $n = 100$, and to $M = 5, 9, 15$ for $n = 200$. The same truncation values were used for P_M^*.

Table 5.1. Time series models used in the simulation study

Models	Equations
VAR(2)	$\begin{bmatrix} \boldsymbol{X}_t^{(1)} \\ \boldsymbol{X}_t^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Phi}_1^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{t-1}^{(1)} \\ \boldsymbol{X}_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi}_2^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Phi}_2^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{t-2}^{(1)} \\ \boldsymbol{X}_{t-2}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_t^{(1)} \\ \boldsymbol{a}_t^{(2)} \end{bmatrix}$
VARMA$(1,1)$	$\begin{bmatrix} \boldsymbol{X}_t^{(1)} \\ \boldsymbol{X}_t^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Phi}_1^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{t-1}^{(1)} \\ \boldsymbol{X}_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Psi}^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Psi}^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_{t-1}^{(1)} \\ \boldsymbol{a}_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_t^{(1)} \\ \boldsymbol{a}_t^{(2)} \end{bmatrix}$
VAR$_\delta(1)$	$\begin{bmatrix} \boldsymbol{X}_t^{(1)} \\ \boldsymbol{X}_t^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Phi}_1^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{t-1}^{(1)} \\ \boldsymbol{X}_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_t^{(1)} \\ \boldsymbol{a}_t^{(2)} \end{bmatrix}$
noise covariance matrices	
	$\boldsymbol{\Sigma}_{\boldsymbol{a}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(1)} & \mathbf{o} \\ \mathbf{o} & \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(2)} \end{bmatrix}$ $\qquad$ $\boldsymbol{\Sigma}_{\boldsymbol{a},\delta} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(1)} & \boldsymbol{\Sigma}_{\boldsymbol{a},\delta}^{(12)} \\ \boldsymbol{\Sigma}_{\boldsymbol{a},\delta}^{(21)} & \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(2)} \end{bmatrix}$
Parameters values	
	$\boldsymbol{\Phi}_1^{(1)} = \begin{bmatrix} 0.4 & 0.0 \\ -1.0 & 1.0 \end{bmatrix}$ $\boldsymbol{\Phi}_1^{(2)} = \begin{bmatrix} 1.0 & 0.0 \\ -0.8 & 0.5 \end{bmatrix}$ $\boldsymbol{\Phi}_2^{(1)} = \begin{bmatrix} 0.6 & -0.5 \\ 0.3 & 0.4 \end{bmatrix}$ $\boldsymbol{\Phi}_2^{(2)} = \begin{bmatrix} -0.5 & -0.8 \\ -0.4 & 0.2 \end{bmatrix}$
	$\boldsymbol{\Psi}^{(1)} = \begin{bmatrix} -0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix}$ $\boldsymbol{\Psi}^{(2)} = \begin{bmatrix} 0.8 & 0.3 \\ 0.1 & 0.6 \end{bmatrix}$ $\Sigma_a^{(1)} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$ $\Sigma_a^{(2)} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$
	$\Sigma_{a,\delta}^{(12)} = \begin{bmatrix} 0.1\delta & 0 \\ 0 & 0.05\delta \end{bmatrix}$

Table 5.2. Kernels used with the test statistics $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$

Truncated Uniform (TR):	$k(z) = \begin{cases} 1, & \lvert z\rvert \le 1, \\ 0, & \text{otherwise.} \end{cases}$
Bartlett (BAR):	$k(z) = \begin{cases} 1-\lvert z\rvert, & \lvert z\rvert \le 1, \\ 0, & \text{otherwise.} \end{cases}$
Daniell (DAN):	$k(z) = \frac{\sin(\pi z)}{\pi z}$, $z \in \mathbb{R}$.
Parzen (PAR):	$k(z) = \begin{cases} 1-6z^2+6\lvert z\rvert^3, & \text{if } \lvert z\rvert \le 0.5, \\ 2\left(1-\lvert z\rvert\right)^3, & \text{if } 0.5 \le \lvert z\rvert \le 1, \\ 0, & \text{otherwise.} \end{cases}$
Bartlett-Priestley (BP):	$k(z) = \frac{3}{(\pi z)^2}\{\frac{\sin(\pi z)}{\pi z} - \cos(\pi z)\}$, $z \in \mathbb{R}$.

In the level study, 5000 independent realizations were generated from both models **VAR**(2) and **VARMA**(1,1) for each series length n. Computations were made in the following way.

(1) First, pseudo-random variables from the $\mathscr{N}(0,1)$ distribution were obtained with the

pseudo-random normal generator of the S-plus package and were transformed into independent $\mathscr{N}[\mathbf{o}, \boldsymbol{\Sigma}_{\mathbf{a}}]$ pseudo-random vectors using the Cholesky decomposition. Second, the $\boldsymbol{X}_t$ values were obtained by directly solving the model difference equation.

(2) For the **VAR**(2) model, the least squares estimates of the coefficients of the true models were obtained using the procedure described in Reinsel (1993). The autoregressive order was obtained by minimizing the AIC criterion for $p \leq P$, where P is set to $n^{1/3}$. We chose AIC criterion which seems behave better than other criteria such HQ or SC specifically in LR cointegration tests, see Lütkepohl and Saikkonen (1999). For the **VARMA**(1,1), each subseries was approximated by a possible high-order **VAR** model. From Pham et al. (2003), the value of the **VAR** order was obtained by minimizing the Hannan-Quinn criterion using conditional least square estimation. The residual cross-correlation matrix $\boldsymbol{R}_{\hat{\boldsymbol{a}}}^{(12)}(j)$'s as defined by (5.3.3) is then computed.

(3) For each realization, the test statistics $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ were compared for each of the four kernels and the three values of M. The same values of M were used for the statistic P_M^*. The values of the statistics $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ were compared with the $\mathscr{N}(0,1)$ critical values and those of P_M^* to the $\chi^2_{4(2M+1)}$ critical values.

(4) Finally, for each model, each series length and nominal level, the empirical frequencies of rejection of the null hypothesis of non-correlation were obtained from the 5000 realizations. The results in percentage are reported in Table 5.3. The standard error of the empirical level is 0.14% for the nominal level 1%, 0.31% for 5% and 0.42% for 10%.

Computations for the power analysis were made in a similar way using the $\mathbf{VAR}_{\delta}(1)$ model with different values of δ.

Table 5.3. Empirical level (in percentage) of the $\mathscr{Q}_n$, $\mathscr{Q}_n^*$ and P_M^* tests,
with different kernels and truncation values.
Gaussian **VAR**(2) and **VARMA**$(1,1)$ models. Number of realizations: 5000

				$\mathscr{Q}_n$					$\mathscr{Q}_n^*$					P_M^*
	n	M	$\alpha\%$	DAN	PAR	BAR	BP	TR	DAN	PAR	BAR	BP	TR	
VAR(2)	100	5	1	0.7	0.6	0.8	0.7	0.6	1.2	0.9	0.7	0.7	1.3	0.7
			5	5.8	3.9	5.7	5.2	4.4	5.9	4.3	5.8	6.1	3.7	4.2
			10	9.6	8.0	9.5	10.6	8.3	10.3	8.8	9.4	10.7	9.0	8.8
		8	1	1.3	0.6	0.9	1.2	0.7	1.4	1.2	1.0	1.5	0.6	0.8
			5	5.6	4.1	5.9	5.6	4.0	5.4	4.0	5.2	4.8	4.0	4.3
			10	10.7	9.2	10.8	10.7	7.4	10.6	9.6	11.0	10.4	8.2	8.4
		12	1	0.8	0.7	0.8	1.2	0.6	1.3	0.8	1.4	1.5	0.7	0.8
			5	5.4	4.8	5.3	5.4	4.2	5.6	4.5	4.9	5.7	4.2	4.5
			10	10.4	8.7	11.2	10.8	7.8	10.8	10.4	11.2	10.5	8.1	8.4
	200	5	1	0.8	1.2	0.8	1.2	0.8	0.7	0.8	1.2	1.3	0.7	0.9
			5	5.7	5.2	5.8	5.5	4.1	5.5	4.2	5.9	5.7	4.4	4.2
			10	9.1	9.2	10.4	10.6	8.3	8.4	10.2	10.6	10.2	8.7	8.9
		9	1	1.2	1.1	0.9	0.8	0.7	1.4	0.9	0.8	1.2	0.7	0.7
			5	6.1	4.3	5.5	5.7	4.4	6.3	4.6	5.5	5.9	4.5	4.1
			10	10.9	9.5	10.5	11.0	7.6	11.2	9.3	10.6	10.7	8.6	9.2
		15	1	1.4	0.8	1.2	1.4	1.2	0.9	1.2	1.4	0.8	0.6	0.6
			5	6.0	4.5	6.2	5.4	4.1	5.8	4.7	5.8	5.6	4.3	4.5
			10	10.6	10.3	11.2	10.6	7.9	11.0	10.5	10.8	10.4	8.2	8.9
VARMA(1,1)	100	5	1	1.3	1.1	0.7	0.8	0.7	1.2	0.7	1.4	1.2	0.6	0.8
			5	5.7	4.7	6.2	4.5	4.3	5.8	4.4	5.8	4.6	3.9	4.3
			10	9.6	8.6	9.3	10.4	8.3	9.6	9.0	9.5	10.8	8.2	8.4
		8	1	1.4	0.7	0.8	1.2	0.7	1.3	0.8	1.2	0.9	0.8	1.3
			5	5.6	4.4	5.9	5.6	3.9	5.4	4.1	5.5	5.5	4.3	5.6
			10	10.6	8.5	11.3	10.6	7.3	9.4	9.0	11.0	10.7	8.0	9.4
		12	1	0.9	1.2	0.7	0.8	0.6	1.1	0.9	0.9	1.3	0.7	1.4
			5	5.4	5.1	6.0	5.6	4.2	5.6	5.4	5.8	5.6	4.1	4.5
			10	9.4	8.8	10.4	10.2	7.9	9.1	8.2	9.1	10.6	7.5	8.3
	200	5	1	0.8	1.3	0.7	0.9	0.7	1.2	0.8	1.2	1.2	0.7	1.3
			5	5.6	4.7	5.4	5.9	4.0	6.2	4.8	5.7	6.3	4.6	5.9
			10	9.0	9.3	10.6	11.0	8.9	10.5	9.2	10.5	9.6	8.2	8.9
		9	1	1.3	0.7	1.2	1.1	0.8	0.9	0.8	1.3	0.8	0.8	0.9
			5	6.1	5.2	4.2	6.1	4.3	5.7	5.1	5.5	6.3	4.3	5.6
			10	9.4	10.5	11.0	10.7	8.4	10.7	9.5	10.8	10.3	8.7	8.9
		15	1	1.4	1.1	0.8	0.9	0.7	1.3	0.9	0.9	0.8	0.7	0.8
			5	6.2	4.6	5.2	6.0	4.3	5.3	5.1	5.3	6.0	4.6	5.5
			10	10.3	10.5	10.8	10.6	7.9	10.7	10.2	11.2	10.7	8.4	9.1

5.6.2. Level

5.6.2.1. Gaussian innovations

Results from the level study are presented in Table 5.3. We make the following observations. The asymptotic $\mathscr{N}(0,1)$ distribution provides a good approximation of the exact distributions of $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ at all nominal levels considered, kernels and truncation values.

Almost all empirical levels are within three standard errors of the corresponding nominal levels and the majority are within two standard errors. The statistic $\mathscr{Q}_n^*$ is slightly better approximated than $\mathscr{Q}_n$ since most of its empirical levels are within two standard errors of the nominal level.

These results are similar to those obtained for orthogonality tests between stationary series; see Bouhaddioui and Roy (2006). At the 1% and 10% nominal levels, both statistics have a small tendency to under or over-reject. There is no significant difference between the kernels. The best approximations are obtained with the Bartlett and Bartlett-Priestley kernels, while the performance of the Parzen kernel is inferior. With the Bartlett kernel, the empirical size is always within two standard errors of the nominal size. For the truncated uniform kernel, the size of $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ are very close to the size of P_M^*, which is normal since $\mathscr{Q}_n$ and $\mathscr{Q}_n^*$ are linear transformations of P_M and P_M^* is a finite-sample version of P_M. For the models considered, the values of the truncation parameter M has no significant effect on the size of the tests. Finally, when the series length n goes from 100 to 200, the approximation improves very slightly.

5.6.2.2. Non-Gaussian innovations

We now examine simulation results where innovations follow a multivariate contaminated normal distribution. We consider the distribution

$$p\mathscr{N}_m[\mathbf{o}, \boldsymbol{\Gamma}] + (1-p)\mathscr{N}_m[\mathbf{o}, \boldsymbol{\Lambda}]$$

to denote the m-dimensional contaminated normal distribution in which the $\mathscr{N}_m(0, \boldsymbol{\Gamma})$ distribution is contaminated with probability $1-p$, by the $\mathscr{N}_m[0, \boldsymbol{\Lambda}]$ distribution. We can verify that the fourth-order cumulants of this distribution depend on p, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$. Thus, we consider in this part of the simulation two innovations series $\boldsymbol{a}_t^{(1)}$ and $\boldsymbol{a}_t^{(2)}$ generated independently according to the following two distributions:

$$p_1\mathscr{N}_{m_1}[\mathbf{o}, \mathbb{I}_{m_1}] + (1-p_1)\mathscr{N}_{m_1}[\mathbf{o}, \boldsymbol{\Omega}_{\boldsymbol{a}}^{(1)}]\ ,\ p_2\mathscr{N}_{m_2}[\mathbf{o}, \mathbb{I}_{m_2}] + (1-p_2)\mathscr{N}_{m_2}[\mathbf{o}, \boldsymbol{\Omega}_{\boldsymbol{a}}^{(2)}]\,,$$

$$\boldsymbol{\Omega}_{\boldsymbol{a}}^{(1)} = \begin{bmatrix} 25 & 5 \\ 5 & 4 \end{bmatrix} \text{ and } \boldsymbol{\Omega}_{\boldsymbol{a}}^{(2)} = \begin{bmatrix} 25 & 7.5 \\ 7.5 & 4 \end{bmatrix}. \tag{5.6.2}$$

Simulations were made for different values of the pair (p_1, p_2) and for two models of Table 5.1, where $\Sigma_{\boldsymbol{a}}^{(1)}$ and $\Sigma_{\boldsymbol{a}}^{(2)}$ are now the covariance matrices of the two contaminated normal distributions in (5.6.2). The results in Table 5.4 are obtained by using $(p_1, p_2) = (0.7, 0.9)$; the results for the other values of (p_1, p_2) are similar. From Table 5.4, we see that the non-normality of the innovations does not significantly affect the behavior of the test statistic $\mathscr{Q}_n$ with the associate kernel function and truncation parameter for the two sizes $n = 100$ and $n = 200$.

Table 5.4. Empirical level (in percentage) of the test $\mathscr{Q}_n$, $\mathscr{Q}_n^*$ and P_M^* with different kernels and truncation values. **VAR**(2) and **VARMA**$(1,1)$ models with non-Gaussian innovations. Number of realizations: 5000

				$\mathscr{Q}_n$					$\mathscr{Q}_n^*$					P_M^*
	n	M	α%	DAN	PAR	BAR	BP	TR	DAN	PAR	BAR	BP	TR	
VAR(2)	100	5	1	1.3	0.7	1.2	1.3	0.6	0.8	1.3	0.9	0.8	1.4	1.3
			5	5.4	4.6	5.8	5.3	4.1	5.5	4.4	5.9	5.8	4.0	4.2
			10	9.8	8.4	10.5	10.7	8.2	10.5	9.0	9.1	9.3	8.5	8.9
		8	1	0.7	1.2	0.8	1.3	0.7	1.2	0.8	0.8	1.3	0.7	0.8
			5	6.0	5.4	4.6	5.8	3.8	5.7	4.2	5.6	4.4	4.0	4.2
			10	11.0	9.4	10.6	9.5	8.2	10.8	9.4	10.8	10.6	8.4	8.8
		12	1	1.2	0.9	0.7	1.3	0.7	1.4	1.2	0.8	1.3	0.6	0.8
			5	5.8	5.6	5.2	5.6	4.0	4.6	4.8	5.3	5.4	3.8	4.2
			10	11.3	10.9	11.0	10.6	8.4	10.6	9.8	10.8	9.5	8.3	8.8
	200	5	1	1.2	0.9	0.8	1.3	0.7	0.8	1.3	1.1	0.8	0.8	1.2
			5	6.0	5.8	5.4	5.6	3.9	6.1	5.9	5.5	5.3	4.0	4.4
			10	10.6	9.0	10.2	10.4	8.4	9.4	10.8	11.0	10.6	8.4	9.2
		9	1	0.7	0.9	0.7	0.8	0.8	1.3	0.7	0.7	1.1	0.8	0.8
			5	5.8	5.6	5.2	4.7	4.2	6.0	4.8	5.8	5.8	4.2	4.6
			10	11.2	9.3	9.6	10.6	8.8	11.4	9.7	10.3	10.9	8.6	9.4
		15	1	1.3	1.1	0.8	0.7	0.7	1.1	1.3	0.9	0.8	0.6	0.7
			5	5.6	5.8	6.0	5.6	4.2	5.6	4.4	6.0	6.2	4.1	4.6
			10	11.2	10.6	10.2	10.8	8.6	11.0	10.8	10.3	10.2	8.6	9.0
VARMA(1,1)	100	5	1	0.8	1.2	1.3	0.7	0.6	1.1	0.8	1.2	1.2	0.6	0.7
			5	5.9	6.1	5.6	4.4	4.0	5.7	5.9	4.8	4.8	4.0	4.4
			10	10.6	9.2	9.6	11.0	8.5	10.9	10.4	9.2	11.0	8.0	9.0
		8	1	1.4	1.2	1.2	0.8	0.7	1.2	1.4	1.3	0.8	0.7	1.4
			5	6.0	4.2	5.6	5.8	3.8	6.2	4.0	6.1	6.3	4.2	6.0
			10	11.6	9.6	10.4	10.8	8.0	11.2	9.4	11.2	10.6	8.0	9.6
		12	1	0.8	1.3	0.8	0.9	0.7	1.2	1.1	0.9	1.1	0.8	1.3
			5	5.8	5.3	5.8	6.0	4.4	6.0	5.2	5.4	5.8	4.0	5.8
			10	10.8	9.2	11.4	10.6	8.1	11.2	9.4	9.3	11.0	8.4	8.8
	200	5	1	1.1	1.2	0.9	1.3	0.7	1.2	1.3	1.1	0.8	0.8	1.2
			5	6.1	5.4	4.8	6.1	4.2	5.9	4.7	5.4	6.0	4.4	5.8
			10	10.6	10.3	11.3	11.5	8.4	11.3	10.4	11.0	10.8	8.4	9.2
		9	1	1.3	1.2	0.9	1.2	0.8	1.2	1.3	1.1	0.9	0.7	1.3
			5	5.9	5.9	4.6	5.4	4.1	5.7	6.1	5.2	5.8	4.4	5.8
			10	11.4	10.8	10.6	10.6	8.8	11.2	10.8	10.4	9.8	8.6	9.3
		15	1	0.9	1.3	0.8	1.2	0.8	1.3	1.2	1.3	1.1	0.7	1.3
			5	5.4	5.8	6.2	5.6	4.0	5.5	5.6	5.8	5.4	4.2	5.8
			10	11.0	10.8	9.8	10.2	8.2	10.6	10.6	10.2	10.4	8.6	9.3

5.6.3. Power

Results on power are presented in Table 5.5. In $\mathbf{VAR}_\delta(1)$, the cross-correlation at lag 0 between the two innovation series increases with δ and, as expected, the powers of the three tests increase with δ. Since the relative behaviors of the various tests are similar for

the three values of δ considered [$\delta = 1,\ 1.5,\ 2$], we only present the results for $\delta = 2$. Similarly, we only present results for $\mathcal{Q}_n^*$, since $\mathcal{Q}_n$ and $\mathcal{Q}_n^*$ have exhibit similar behaviors with respect to kernels and truncation values.

Table 5.5. Power of the tests $\mathcal{Q}_n$, $\mathcal{Q}_n^*$ and P_M^* based on asymptotic critical values with different kernels and different truncation values. $\mathbf{VAR}_\delta(1)$ model with $\delta = 2$.

			$\mathcal{Q}_n^*$					P_M^*
n	M	$\alpha\%$	DAN	PAR	BAR	BP	TR	
100	5	1	57.3	53.5	54.6	52.6	35.3	24.6
		5	63.2	60.1	56.4	58.6	36.8	26.8
		10	72.6	70.8	62.5	64.3	38.2	27.5
	8	1	49.6	46.1	51.4	48.0	27.5	22.6
		5	58.4	53.2	55.8	51.6	31.2	23.8
		10	63.7	60.8	62.6	61.7	34.6	25.8
	12	1	43.6	38.5	41.8	42.6	23.3	18.9
		5	50.2	44.7	40.3	43.0	26.4	21.2
		10	56.8	50.6	48.8	46.5	28.8	23.7
200	5	1	78.4	74.5	74.8	76.2	54.8	50.6
		5	85.6	82.6	81.6	85.8	56.4	54.1
		10	93.4	89.5	87.5	90.2	60.4	56.8
	9	1	69.5	65.2	63.0	66.8	42.4	40.7
		5	75.6	76.6	72.4	78.2	46.2	44.6
		10	80.8	78.5	77.6	82.8	50.4	46.4
	15	1	56.8	52.4	54.8	56.1	36.8	32.8
		5	60.1	57.4	53.2	60.1	40.2	35.0
		10	64.8	54.4	54.2	62.6	44.8	40.4

From Table 5.5, we draw the following observations. First, power decreases as M increases. Indeed, the model considered here is characterized by the lag 0 serial correlation. In such a situation, we expect that the tests assigning more weight to small lags will be more powerful than those assigning weights to a large number of lags. For the three significance levels and the three truncation values, the Daniell kernel provides the most powerful test, while the Parzen, Bartlett and Bartlett-Priestley kernels yield similar powers for $\mathcal{Q}_n^*$. However, the power of $\mathcal{Q}_n^*$ with the truncated uniform kernel is much smaller and is comparable to the power of P_M^*. For the chosen model, the new tests $\mathcal{Q}_n$ or $\mathcal{Q}_n^*$ with kernels

other than the truncated uniform are preferable to the nonstationary multivariate version of Haugh's test P_M^*. Finally, the powers of all tests increase when the sample size varies from 100 to 200.

5.7. Conclusion

In this paper, we have proposed a semiparametric approach to test the non-correlation (or independence in the Gaussian case) between infinite-order cointegrated series IVAR(∞). The approach is semiparametric in the sense that if the two series are VARMA, we do not need to separately estimate the *true* model for each of the series. Instead, we fit a vector autoregression to each series, and the test statistics are based on residual cross-correlations at all possible lags. The weights assigned to the lags are defined by a kernel function and a smoothing parameter. Under the hypothesis of independence or non-causality of the two series, the asymptotic normality of the tests statistics are established. The finite-sample properties of the test were investigated by a Monte Carlo experiment which shows that the level is reasonably well controlled for both series lengths 100 and 200. Furthermore, with the model considered, the four kernels DAN, PAR, BAR, BP lead to similar powers and are more powerful than the truncated uniform kernel which corresponds to the multivariate version of the portmanteau test proposed by Bouhaddioui and Dufour (2008).

5.A. Proofs

The following notations are adopted. The Euclidean scalar product of $\mathbf{x}_t$ and $\mathbf{x}_s$ is defined by $\langle \mathbf{x}_t, \mathbf{x}_s \rangle = \mathbf{x}_t' \mathbf{x}_s$ and the Euclidean norm of $\mathbf{x}_t$ by $\|\mathbf{x}_t\| = \sqrt{\langle \mathbf{x}_t, \mathbf{x}_t \rangle}$. The scalar Δ denotes a generic positive bounded constant which may differ from place to place.

PROOF OF PROPOSITION 5.3.1 First, let

$$\boldsymbol{\Xi} := [\boldsymbol{\Xi}_1 \vdots \cdots \vdots \boldsymbol{\Xi}_p \vdots \boldsymbol{\Xi}_{p+1,1}] = [\boldsymbol{\Psi} \vdots \boldsymbol{\Pi}_1 \vdots \cdots \vdots \boldsymbol{\Pi}_p] \boldsymbol{D}_p := \boldsymbol{\Pi}\, \boldsymbol{D}_p \tag{5.A.1}$$

where $\boldsymbol{D}_p$ is a suitable nonsingular transformation matrix containing the unknown matrix

C_1. The ECM representation (5.2.10) can be written as

$$\Delta \boldsymbol{X}_t = \boldsymbol{\Psi}_0 \boldsymbol{X}_{2,t-1} + \sum_{l=1}^{p} \boldsymbol{\Xi}_l \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\Xi}_{p+1,1} \boldsymbol{\varepsilon}_{1,t-p-1} + \boldsymbol{e}_t(n). \tag{5.A.2}$$

The matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}_0$ are defined in Saikkonen (1992, equation (A.2)). Set

$$\boldsymbol{\Lambda} := [\boldsymbol{\Xi} \vdots \boldsymbol{\Psi}_0], \quad \boldsymbol{W}_t := \boldsymbol{W}_t(p) := [\boldsymbol{\Upsilon}_t', \boldsymbol{X}_{2,t-1}'], \tag{5.A.3}$$

$$\boldsymbol{\Upsilon}_t' := \boldsymbol{\Upsilon}_t(p)' := [\boldsymbol{\varepsilon}_{t-1}', \ldots, \boldsymbol{\varepsilon}_{t-p}', \boldsymbol{\varepsilon}_{1,t-p-1}']. \tag{5.A.4}$$

Consider the linear transformation

$$\boldsymbol{b}_t := \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_t, \quad \hat{\boldsymbol{b}}_t = \boldsymbol{\Sigma}^{-1/2} \hat{\boldsymbol{a}}_t, \tag{5.A.5}$$

where $\boldsymbol{\Sigma}$ is defined in (5.2.18). Since $C_{\hat{\boldsymbol{b}}}^{(12)}(j) = \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(12)}(j) \boldsymbol{\Sigma}_2^{-1/2}$. Using the property $\mathrm{vec}(ABC) = (C' \otimes A)\mathrm{vec}(B)$, we have:

$$\begin{aligned} \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) &= n \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \left(\boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}\right) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) \\ &= n \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)' \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) = \mathscr{T}_{\hat{\boldsymbol{b}}}^{(12)}. \end{aligned} \tag{5.A.6}$$

Thus, to prove the result, it is sufficient to show that

$$\mathscr{T}_{\boldsymbol{b}}^{(12)} - \mathscr{T}_{\hat{\boldsymbol{b}}}^{(12)} = o_p(M^{1/2}). \tag{5.A.7}$$

The result follows by decomposing the latter difference in two parts,

$$\begin{aligned} \mathscr{T}_{\boldsymbol{b}}^{(12)} - \mathscr{T}_{\hat{\boldsymbol{b}}}^{(12)} &= n \sum_{j=1-n}^{n-1} k^2(j/M) (\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 + 2 \langle \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j), \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) \rangle) \\ &= [T_n^{(1)} + T_{n-}^{(1)}] + T_n^{(2)} \end{aligned} \tag{5.A.8}$$

where

$$T_n^{(1)} := n\sum_{j=0}^{n-1} k^2(j/M)\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)-\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2, \quad T_{n-}^{(1)} := n\sum_{j=1-n}^{-1} k^2(j/M)(\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)-\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2, \tag{5.A.9}$$

$$T_n^{(2)} := 2\left\langle \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j), \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)-\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\right\rangle), \tag{5.A.10}$$

and showing that each part is $o_p(M^{1/2})$. Consider the positive lags $j \geq 0$, since for negative lags, the proof is similar by symmetry.

Define $\hat{\boldsymbol{\delta}}_t = \boldsymbol{b}_t^{(1)} - \hat{\boldsymbol{b}}_t^{(1)}$ and $\hat{\boldsymbol{\eta}}_t = \boldsymbol{b}_t^{(2)} - \hat{\boldsymbol{b}}_t^{(2)}$. From (5.3.2), we have

$$\begin{aligned} T_n^{(1)} &= n\sum_{j=0}^{n-1} k^2(j/M)\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)-\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 \\ &= n\sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}(\boldsymbol{b}_t^{(1)}\boldsymbol{b}_{t-j}^{(2)'} - \hat{\boldsymbol{b}}_t^{(1)}\hat{\boldsymbol{b}}_{t-j}^{(2)'})\|^2, \end{aligned} \tag{5.A.11}$$

and using the Cauchy-Schwarz inequality, we obtain

$$T_n^{(1)} = n\sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}(\boldsymbol{b}_t^{(1)}\hat{\boldsymbol{\eta}}_{t-j}' + \hat{\boldsymbol{\delta}}_t\boldsymbol{b}_{t-j}^{(2)'} - \hat{\boldsymbol{\delta}}_t\hat{\boldsymbol{\eta}}_{t-j}')\|^2 \leq 4n(T_{1n}+T_{2n}+T_{3n}) \tag{5.A.12}$$

with $T_{1n} = \sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}\boldsymbol{b}_t^{(1)}\hat{\boldsymbol{\eta}}_{t-j}'\|^2$, $T_{2n} = \sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}\hat{\boldsymbol{\delta}}_t\boldsymbol{b}_{t-j}^{(2)'}\|^2$ and $T_{3n} = \sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}\hat{\boldsymbol{\delta}}_t\hat{\boldsymbol{\eta}}_{t-j}'\|^2$. It suffices to show that the terms T_{jn}, $j = 1,2,3$, are $o_p(M^{1/2}/n)$. We can then write:

$$\begin{aligned} \hat{\boldsymbol{\delta}}_t &= (\hat{\boldsymbol{b}}_t^{(1)} - \boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{e}_t^{(1)}) + (\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{e}_t^{(1)} - \boldsymbol{b}_t^{(1)}) = \boldsymbol{\Sigma}_1^{-1/2}\{[\hat{\boldsymbol{a}}_t^{(1)} - \boldsymbol{e}_t^{(1)}] + [\boldsymbol{e}_t^{(1)} - \boldsymbol{a}_t^{(1)}]\} \\ &= \boldsymbol{\Sigma}_1^{-1/2}\{(\hat{\boldsymbol{\Lambda}}^{(1)} - \boldsymbol{\Lambda}^{(1)})\boldsymbol{W}_t^{(1)} + \xi_t(p_1)\} \end{aligned} \tag{5.A.13}$$

where $\boldsymbol{e}_t^{(1)} := \boldsymbol{e}_t^{(1)}(n)$, $\boldsymbol{\Lambda}^{(h)}$ and $\boldsymbol{W}_t^{(h)}$, $h = 1,2$ are defined as in (5.A.2) for each process, $\hat{\boldsymbol{\Lambda}}$ is the LS estimator of $\boldsymbol{\Lambda}$ and $\xi_t(p_1) = \sum_{l=p_1+1}^{\infty}\boldsymbol{\Phi}_l\boldsymbol{X}_{t-l}^{(1)}$ represents the bias of the VAR(p_1) approximation of $\{\boldsymbol{X}_t^{(1)}\}$.

The second equality is from Saikkonen and Lütkepohl (1996, page 832). Also, using the

result of Proposition 5.2.1, we deduce that

$$\|\hat{\boldsymbol{\Lambda}}^{(1)} - \boldsymbol{\Lambda}^{(1)}\|^2 = O_p(\frac{p_1}{n}). \tag{5.A.14}$$

By equation 3.15 in **?**, we have $\mathbb{E}\left(\|\xi_t(p_n^{(h)})\|^2\right) = O\left(\sum_{l=p_n^{(h)}+1}^{\infty} \|\boldsymbol{\Phi}_l^{(h)}\|\right)^2$, $h = 1, 2$. Based on the result (3.17) in **?** and equation (5.2.15), we obtain:

$$T_{1n} = \sum_{j=0}^{n-1} k^2(j/M) \|\frac{1}{n} \sum_{t=j+1}^{n} \boldsymbol{b}_t^{(1)} \hat{\eta}_{t-j}'\|^2 = O_p(\frac{p_n^{(2)} M}{n^2}) \{\frac{1}{M} \sum_{j=0}^{n-1} k^2(j/M)\} \tag{5.A.15}$$

Since $p_n^{(2)} = o(n/M^{1/2})$, we have $T_{1n} = o_p(M^{1/2}/n)$. By symmetry, we can prove that $T_{2n} = o_p(\frac{M^{1/2}}{n})$. For the third term T_{3n}, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
T_{3n} &= \sum_{j=0}^{n-1} k^2(j/M) \|n^{-1} \sum_{t=j+1}^{n} \hat{\eta}_t \hat{\delta}_{t-j}'\|^2 \\
&\leq \|\boldsymbol{\Lambda}^{(1)} - \hat{\boldsymbol{\Lambda}}^{(1)}\|^2 \|\boldsymbol{\Lambda}^{(2)} - \hat{\boldsymbol{\Lambda}}^{(2)}\|^2 \sum_{j=0}^{n-1} k^2(j/M) \|n^{-1} \sum_{t=j+1}^{n} \boldsymbol{W}_t^{(1)}(p_1) \boldsymbol{W}_{t-j}^{(2)}(p_2)'\|^2 \\
&+ \|\boldsymbol{\Lambda}^{(1)} - \hat{\boldsymbol{\Lambda}}^{(1)}\|^2 \sum_{j=0}^{n-1} k^2(j/M) \|n^{-1} \sum_{t=j+1}^{n} \boldsymbol{W}_t^{(1)}(p_1) \xi_{t-j}(p_2)'\|^2 \\
&+ \|\boldsymbol{\Lambda}^{(2)} - \hat{\boldsymbol{\Lambda}}\|^2 \sum_{j=0}^{n-1} k^2(j/M) \|n^{-1} \sum_{t=j+1}^{n} \xi_t(p_1) \boldsymbol{W}_{t-j}^{(2)}(p_2)'\|^2 \\
&+ \sum_{j=0}^{n-1} k^2(j/M) \|n^{-1} \sum_{t=j+1}^{n} \xi_t(p_1) \xi_{t-j}(p_2)'\|^2 .
\end{aligned} \tag{5.A.16}$$

Using the equations (3.19) - (3.22) in Bouhaddioui and Roy (2006), the assumptions $p_n^{(h)} = o(n^{1/2}/M^{1/4})$, $n\sum_{l=p_n^{(h)}+1}^{\infty} \|\boldsymbol{\Phi}_l^{(h)}\|^2 = o(n^{1/2}/M^{1/4})$ and the result (5.2.15), we conclude that $T_{3n} = o_p(M^{1/2}/n)$. Therefore, we obtain

$$T_n^{(1)} = n \sum_{j=0}^{n-1} k^2(j/M) \|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 = o_p\left(M^{1/2}\right). \tag{5.A.17}$$

Finally, using Cauchy-Schwarz inequality once more, we have

$$|T_n^{(2)}| \leq n \sum_{j=1-n}^{n-1} k^2(j/M) |\langle \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j), \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) \rangle| \leq n \sum_{l=4}^{6} T_{ln}, \tag{5.A.18}$$

with

$$T_{4n} = \sum_{j=0}^{n-1} k^2(j/M) \|\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\| \| \frac{1}{n} \sum_{t=j+1}^{n} \hat{\delta}_t \, (\boldsymbol{b}_{t-j}^{(2)})' \|, \tag{5.A.19}$$

$$T_{5n} = \sum_{j=0}^{n-1} k^2(j/M) \|\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\| \| \frac{1}{n} \sum_{t=j+1}^{n} \boldsymbol{b}_t^{(1)} \hat{\eta}'_{t-j} \|, \tag{5.A.20}$$

$$T_{6n} = \sum_{j=0}^{n-1} k^2(j/M) \|\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\| \| \frac{1}{n} \sum_{t=j+1}^{n} \hat{\delta}_t \, \hat{\eta}'_{t-j} \| \, . \tag{5.A.21}$$

Thus, it is sufficient to show that the terms T_{jn}, $j = 4, 5, 6$, are $o_p(M^{1/2}/n)$. By conditioning on $(\boldsymbol{b}_s^{(2)})_{s=-\infty}^{n}$ and using Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\big[T_{4n} \,|\, (\boldsymbol{b}_s^{(2)})_{s=-\infty}^{n}\big] &\leq \sum_{j=1-n}^{n-1} k^2(j/M) \\
&\times \; [\mathbb{E}(\{(\frac{1}{n}\sum_{\tau=1}^{n} \|\boldsymbol{b}_\tau^{(1)} \boldsymbol{b}_{\tau-j}^{(2)^T}\|)(\frac{1}{n}\sum_{t=j+1}^{n} \|\hat{\boldsymbol{\delta}}_t \boldsymbol{b}_{t-j}^{(2)^T}\|)\}^2 \,|\, (\boldsymbol{b}_s^{(2)})_{s=-\infty}^{n})]^{1/2} \\
&\leq \frac{M\Delta}{n^2} \{\frac{1}{M}\sum_{1-n}^{n-1} k^2(j/M)\} (\frac{1}{n}\sum_{\tau=1}^{n} \|\boldsymbol{b}_\tau^{(2)}\|^2)^{1/2} (\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\|\hat{\boldsymbol{\delta}}_t\|^2)^{1/2} \\
&= O_p\big(\frac{M (p_n^{(2)})^{1/2}}{n^{5/2}}\big) = o_p\big(\frac{M^{1/2}}{n^{3/2}}\big) \, .
\end{aligned} \tag{5.A.22}$$

The first equality is obtained by using the conditions on $p_n^{(2)}$, $\boldsymbol{\Phi}^{(2)}$, and the assumption of independence of the two innovation series. Then, $T_{4n} = o_p(M^{1/2}/n)$. By symmetry, we have also $T_{5n} = o_p(M^{1/2}/n)$. Finally, from Markov inequality, we have

$$\sum_{j=1}^{n-1} k^2(j/M) \|\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 = O_p(M/n) \tag{5.A.23}$$

hence, using the Cauchy-Schwarz inequality and the result for T_{3n}, we obtain that $T_{6n} = o_p(M/n)$. Thus, $T_n^{(2)} = o_p(M^{1/2})$ and the proof of Proposition 5.3.1 is completed. □

PROOF OF PROPOSITION 5.3.2 Since $D_n(k) = MD(k)\{1+o(1)\}$, it is sufficient to show that

$$\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) = O_p(M/n^{1/2}). \tag{5.A.24}$$

Using the fact that $\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(hh)}(0) - \boldsymbol{\Sigma}_n^h = O_p(n^{-1/2})$ for $h = 1, 2$ [see Lütkepohl and Saikkonen (1997, p.133)], it follows that

$$\left[\boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(22)}(0)^{-1} \otimes \boldsymbol{C}_{\hat{\boldsymbol{a}}}^{(11)}(0)^{-1}\right] - \left[\boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}\right] = \boldsymbol{O}_p(n^{-1/2}). \tag{5.A.25}$$

Thus,

$$\begin{aligned} \mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) &= n \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \boldsymbol{O}_p(n^{-1/2}) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) \\ &= O_p(n^{1/2}) \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j). \end{aligned} \tag{5.A.26}$$

To complete the proof, it remains to prove that

$$\mathscr{B}(n) = \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) = O_p(M/n). \tag{5.A.27}$$

First, let us decompose $\mathscr{B}(n)$ in two parts

$$\begin{aligned} \mathscr{B}(n) &= \sum_{j=1-n}^{n-1} k^2(j/M) \{\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)' \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)\} + \sum_{j=1-n}^{n-1} k^2(j/M) \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)' \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j) \\ &= \mathscr{B}_1 + \mathscr{B}_2. \end{aligned} \tag{5.A.28}$$

By an argument similar to the one used to prove (5.A.7) in Proposition 5.3.1, we have:

$$\mathscr{B}_1(n) = \sum_{j=1-n}^{n-1} k^2(j/M) \{\boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j)' \boldsymbol{c}_{\hat{\boldsymbol{a}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)' \boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)\} = o_p(M^{1/2}/n), \tag{5.A.29}$$

and, by Markov inequality, it follows that

$$\mathscr{B}_2(n) = \sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j)'\boldsymbol{c}_{\boldsymbol{a}}^{(12)}(j) = O_p(M/n)\,. \tag{5.A.30}$$

Combining the results for $\mathscr{B}_1(n)$ and $\mathscr{B}_2(n)$, we obtain that

$$\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}}) - \mathscr{T}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma})) = O_p(n^{1/2})O_p(M/n) = O_p(M/n^{1/2}), \tag{5.A.31}$$

and the proof of Proposition 5.3.2 is completed. □

PROOF OF THEOREM 5.4.1 First, we note that the statistic $\mathscr{Q}_n$ is a normalized version of $\mathscr{T}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Sigma}})$ which can be viewed as the $\mathscr{L}_2$-norm of a kernel-based estimator of the cross-coherency function between the two innovations processes. Thus, the statistic $\mathscr{Q}_n$ can be expressed as

$$\mathscr{Q}_n = \frac{n\|\boldsymbol{s}_{\hat{\boldsymbol{a}}}^{(12)}\|_2^2 - m_1 m_2 S_n(k)}{\sqrt{2m_1 m_2 D_n(k)}} \tag{5.A.32}$$

where $\boldsymbol{s}_{\hat{\boldsymbol{a}}}^{(12)}$ is the estimator of the cross-coherency function between the two innovations processes given by

$$\|\boldsymbol{s}_{\boldsymbol{a}}^{(12)}\|_2^2 = \sum_{j=-\infty}^{\infty} \boldsymbol{\gamma}_{\boldsymbol{a}}^{(12)}(j)'(\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\gamma}_{\boldsymbol{a}}^{(12)}(j) \tag{5.A.33}$$

where $\boldsymbol{\gamma}_{\boldsymbol{a}}^{(12)}(j) := vec[\Gamma_{\boldsymbol{a}}^{(12)}(j)]$For details, see Section 4 in Bouhaddioui and Roy (2006). By definition of $\mathscr{Q}_n$, we can write

$$\begin{aligned}\left(\frac{M^{1/2}}{n}\right)\mathscr{Q}_n &= \frac{M^{1/2}\|\boldsymbol{s}_{\hat{\boldsymbol{a}}}^{(12)}\|_2^2 - \frac{M^{1/2}}{n} m_1 m_2 S_n(k)}{\{2m_1 m_2 D(k)\}^{1/2}} \\ &= \frac{\|\boldsymbol{s}_{\hat{\boldsymbol{a}}}^{(12)}\|_2^2}{\{2m_1 m_2 M^{-1} D_n(k)\}^{1/2}} - \frac{n^{-1}S_n(k)}{\{2M^{-1}D_n(k)\}^{1/2}}\,(m_1 m_2)^{1/2}\end{aligned} \tag{5.A.34}$$

From (5.3.11), the last term of the previous equation goes to zero when $M/n \to 0$ as $n \to \infty$. Using the linear transformation $\boldsymbol{b}_t = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_t$, as in Proposition 5.3.1, we have

$\|\boldsymbol{s}_{\boldsymbol{a}}^{(12)}\| = \|\boldsymbol{s}_{\boldsymbol{b}}^{(12)}\|$. Also, since the processes $\boldsymbol{b}^{(1)}$ and $\boldsymbol{b}^{(2)}$ are stationary and by Lemma A.7 in Bouhaddioui and Roy (2006), we have that

$$\|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|^2 - \|\boldsymbol{s}_{\boldsymbol{b}}^{(12)}\|^2 \xrightarrow{p} 0 \tag{5.A.35}$$

where $\|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|$ is defined as $\|\boldsymbol{s}_{\hat{\boldsymbol{b}}}^{(12)}\|$, the residual series $(\hat{\boldsymbol{b}}_t^{(1)}, \hat{\boldsymbol{b}}_t^{(2)})_{t=1}^n$ being replaced by the innovation series $(\boldsymbol{b}_t^{(1)}, \boldsymbol{b}_t^{(2)})_{t=1}^n$. Thus, to prove the consistency result (5.4.4), it is sufficient to verify that $\|\boldsymbol{s}_{\hat{\boldsymbol{b}}}^{(12)}\|_2^2 - \|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|_2^2 \xrightarrow{p} 0$, which follows from the following lemma. □

Lemma 5.A.1 *Under the assumptions of Theorem 5.4.1, we have*

$$\|\tilde{\boldsymbol{s}}_{\hat{\boldsymbol{b}}}^{(12)}\|_2^2 - \|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|_2^2 \xrightarrow{p} 0 \tag{5.A.36}$$

PROOF OF LEMMA 5.A.1 By definition of $\boldsymbol{s}_{\hat{\boldsymbol{b}}}^{(12)}$ and $\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}$, and by similar calculations to those for the proof in Proposition 5.3.1, we obtain

$$\begin{aligned}
\|\boldsymbol{s}_{\hat{\boldsymbol{b}}}^{(12)}\|_2^2 - \|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|_2^2 &= \sum_{j=1-n}^{n-1} k^2(j/M)(\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j)\|^2 - \|\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2) \\
&= \sum_{j=1-n}^{n-1} k^2(j/M)\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 \\
&+ 2\sum_{j=1-n}^{n-1} k^2(j/M)\langle \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j), \boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\rangle \,. \qquad (5.A.37)
\end{aligned}$$

It is sufficient to prove that the first term goes to zero in probability, because the second term can be bounded by a product of the first term and a finite quantity, using the Cauchy-Schwarz inequality. With the notations of Proposition 5.3.1, we can write

$$\sum_{j=1-n}^{n-1} k^2(j/M)\|\boldsymbol{c}_{\hat{\boldsymbol{b}}}^{(12)}(j) - \boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)\|^2 \leq 4\sum_{l=1}^{3} T_{ln}, \tag{5.A.38}$$

where T_{ln}, $l = 1, 2, 3$, are defined in Proposition 5.3.1. We first prove that $T_{1n} \to 0$ in

probability. By the Cauchy-Schwarz inequality, we obtain

$$T_{1n} \leq M\{\frac{1}{M}\sum_{j=0}^{n-1} k^2(j/M)\}\{\frac{1}{n}\sum_{t=1}^{n}\|\boldsymbol{b}_t^{(1)}\|^2\}\{\frac{1}{n}\sum_{t=1}^{n}\|\hat{\eta}_t\|^2\}. \tag{5.A.39}$$

By definition of $\hat{\eta}_t$, it follows that

$$\frac{1}{n}\sum_{t=j}^{n}\|\hat{\eta}_t\|^2 \leq \frac{1}{n}\sum_{t=1}^{n}\{\|(\boldsymbol{\Lambda}^{(2)}-\hat{\boldsymbol{\Lambda}}^{(2)})\boldsymbol{W}_t^{(2)}\|^2+\|\xi_t(p_n^{(2)})\|^2\}. \tag{5.A.40}$$

Since $\|\boldsymbol{\Gamma}_{\boldsymbol{a}}^{(11)}(l)\|$ is uniformly bounded by a positive constant Δ, and the parameters $\{\boldsymbol{\Phi}_l\}$ are a linear function of the original parameters $\{\boldsymbol{G}_l\}$, then the bias approximation can be bounded by

$$\mathbb{E}\|\xi_t(p_n^{(2)})\|^2 \leq \Delta(\sum_{l=p_2+1}^{\infty}\|\boldsymbol{\Phi}_l^{(2)}\|)^2 = o(n^{-1}). \tag{5.A.41}$$

See also the result (A.12) in Saikkonen (1992). Under the assumptions on the process $\boldsymbol{b}$, on $p_n^{(2)}$ and on the parameters $(\boldsymbol{\Phi}_l^{(2)})$, we have

$$T_{1n} = O_p\big(\frac{M(p_n^{(2)})^2}{n}\big) + O_p(M\sum_{l=p_2+1}^{\infty}\|\boldsymbol{\Phi}_l^{(2)}\|^2) = o_p(1). \tag{5.A.42}$$

By symmetry, we can verify that $T_{2n} = o_p(1)$. For T_{3n}, we can write

$$\begin{aligned} T_{3n} &= \sum_{j=0}^{n-1} k^2(j/M)\|\frac{1}{n}\sum_{t=j+1}^{n}\hat{\delta}_t\hat{\eta}_{t-j}'\|^2 \\ &\leq M\{\frac{1}{M}\sum_{j=0}^{n-1}k^2(j/M)\}\{\frac{1}{n}\sum_{t=1}^{n}\|\hat{\delta}_t\|^2\}\{\frac{1}{n}\sum_{t=1}^{n}\|\hat{\eta}_t\|^2\}. \end{aligned} \tag{5.A.43}$$

By symmetry, we can prove that $\frac{1}{n}\sum_{t=1}^{n}\|\hat{\delta}_t\|^2 = O_p\big((p_n^{(1)})^2/n\big) + O_p(1)\sum_{l=p_1+1}^{\infty}\|\boldsymbol{\Phi}_l^{(1)}\|^2$, and using the same assumptions as those for T_{1n}, we obtain that $T_{3n} = o_p(1)$. Finally, we conclude that

$$\|\tilde{\boldsymbol{s}}_{\hat{\boldsymbol{b}}}^{(12)}\|^2 - \|\tilde{\boldsymbol{s}}_{\boldsymbol{b}}^{(12)}\|^2 = o_p(1). \tag{5.A.44}$$

This completes the proof of Lemma 5.A.1 and then Theorem 5.4.1.

PROOF OF THEOREM 5.5.1 By the proof of Theorem 5.4.1,

$$\mathscr{Q}_n = \frac{n\|\boldsymbol{s}_{\boldsymbol{b}}^{(12)}\|_2^2 - m_1 m_2 S_n(k)}{\sqrt{2m_1 m_2 D_n(k)}} + o_p(1). \tag{5.A.45}$$

where

$$\begin{aligned}
n\|\boldsymbol{s}_{\boldsymbol{b}}^{(12)}\|_2^2 &= \sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) \\
&= n\sum_{j=1-n}^{n-1} k^2(j/M)(\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) - \boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j))'(\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) - \boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j)) \\
&\quad +2n\sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j) \\
&\quad -n\sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j) \\
&= n\sum_{j=1-n}^{n-1} k^2(j/M)(\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) - \boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j))'(\boldsymbol{c}_{\boldsymbol{b}}^{(12)}(j) - \boldsymbol{\gamma}_{\boldsymbol{b}}^{(12)}(j)) \\
&\quad +M^{1/2}\sum_{j=1-n}^{n-1} k^2(j/M)\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j) + o_p(M^{1/4}).
\end{aligned}$$

Since there exists $j^* \in \mathbb{Z}$ such that $\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}(j) = \mathbf{o}$, $\forall j\,(|j| > j^*)$, we have

$$\begin{aligned}
\sum_{j=-\infty}^{\infty} k^2(j/M)\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j) &= \sum_{j=-j^*}^{j^*} k^2(j/M)\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j) \\
&\to \sum_{j=-\infty}^{\infty} \boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j)'\boldsymbol{\lambda}_{\boldsymbol{b}}^{(12)}(j) := \beta(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)}).
\end{aligned}$$

Thus,

$$\frac{n\|\boldsymbol{s}_{\boldsymbol{b}}^{(12)}\|_2^2 - m_1 m_2 S_n(k)}{\sqrt{2m_1 m_2 D_n(k)}} \xrightarrow{L} Z + \frac{\beta(\boldsymbol{\Lambda}_{\boldsymbol{b}}^{(12)})}{\sqrt{2m_1 m_2 D(k)}}$$

where $Z \sim \mathscr{N}(0,1)$. □

Conclusion

In this thesis, we studied statistical methods which are valid without arbitrary assumptions on the underlying data generating process. In Chapter 2, we consider a bound approach for weakly identified nonparametric regression $\mathbb{E}[Y \mid X]$ based on a parsimonious approximate model complemented by approximation bounds. The proposed bounds only depends on the moments of observables and can be easily estimated. Inference takes the form of sets and favorable finite sample properties of the approach were reported in the Monte Carlo simulations. In Chapter 3, we propose generalized $C(\alpha)$-type test procedures under non-standard rates of convergence. We allow for estimating equations and restricted estimators of nuisance parameters to converge at slower rates than the parametric rate $n^{1'2}$ and show that the asymptotic distributions of the proposed statistics are asymptotically chi-squared. We consider applications of the framework to testing under local estimating functions and asymptotically unequal sample sizes. In Chapter 4, we consider a transformation approach for dynamics of probability distributions. The proposed transformation incorporates such key features as mass points and varying support that existing density-based methods, such as Chang et al. (2016), are not equipped to capture. The empirical application to the dynamics of the earning distribution in the U.S. showed that the dynamics of the support or top quantiles is the main driver for persistence. Chapter 5 examines independence of two integrated vector autoregression series under weak assumptions on the form of the underlying process. A simulation studies show favorable finite sample performance of the proposed test in terms of the size and power.

Bibliography

Ahn, S. K. and Reinsel, G. C. (1990), 'Estimation for partially nonstationary multivariate autoregressive models', *Journal of the American statistical association* **85**(411), 813–823.

Akaike, H. (1969), 'Power spectrum estimation through autoregressive model fitting', *Annals of the Institute of Statistical Mathematics* **21**, 407–419.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

Alesina, A. and Perotti, R. (1996), 'Income distribution, political instability, and investment', *European Economic Review* **40**(6), 1203–1228.

Altonji, J. G., Smith Jr, A. A. and Vidangos, I. (2013), 'Modeling earnings dynamics', *Econometrica* **81**(4), 1395–1454.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E. and Zucman, G. (2017), 'Global inequality dynamics: New findings from wid. world', *American Economic Review* **107**(5), 404–09.

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

Anderson, G. (1996), 'Nonparametric tests of stochastic dominance in income distributions', *Econometrica* **64(5)**, 1183–1193.

Anderson, O. D., ed. (1982), *Time Series Analysis: Theory and Practice 1*, North-Holland Publishing.

Andrews, D. W. (1991*a*), 'Asymptotic normality of series estimators for nonparametric and semiparametric regression models', *Econometrica* **59(2)**, 307–345.

Andrews, D. W. (1991*b*), 'cross-validation in regression with heteroskedastic errors', *Journal of Econometrics* **47**, 359–377.

Andrews, D. W. K. (1991*c*), 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* **58**, 817–858.

Andrews, D. W. and Pollard, D. (1994), 'An introduction to functional central limit theorems for dependent stochastic processes', *International Statistical Review/Revue Internationale de Statistique* **62(1)**, 119–132.

Andrienko, Y., Nemtsov, A. et al. (2005), 'Estimation of individual demand for alcohol', *Economics Education and Research Consortium Working Paper Series* **5**(10).

Angrist, J. D. and Krueger, A. B. (1995), 'Split-sample instrumental variables estimates of the return to schooling', *Journal of Business & Economic Statistics* **13**(2), 225–235.

Ansley, C. F. (1980), 'Computation of the theoretical autocovariance function for a vector ARMA process', *Journal of Statistical Computation and Simulation* **12**, 15–24.

Antoine, B. and Renault, E. (2012), 'Efficient minimum distance estimation with multiple rates of convergence', *Journal of Econometrics* **170**(2), 350–367.

Arellano, M., Blundell, R. and Bonhomme, S. (2017), 'Earnings and consumption dynamics: a nonlinear panel data framework', *Econometrica* **85**(3), 693–734.

Armstrong, T. B. and Kolesár, M. (2020), 'Simple and honest confidence intervals in nonparametric regression', *Quantitative Economics* **11**(1), 1–39.

Atkinson, A. B. (1987), 'On the measurement of poverty', *Econometrica* **55(4)**, 749–764.

Atkinson, A. B., Piketty, T. and Saez, E. (2011), 'Top incomes in the long run of history', *Journal of Economic Literature* **49**(1), 3–71.

Bahadur, R. R. (1960), 'Stochastic comparison of tests', *Annals of Mathematical Statistics* **31**, 276–295.

Bahadur, R. R. and Savage, L. J. (1956), 'The nonexistence of certain statistical procedures in nonparametric problems', *Annals of Mathematical Statistics* **27**(4), 1115–1122.

Balasubramanian, V., Ho, S.-S. and Vovk, V. (2014), *Conformal prediction for reliable machine learning: theory, adaptations and applications*, Newnes.

Baltagi, B., ed. (2001), *Companion to Theoretical Econometrics*, Blackwell Companions to Contemporary Economics, Basil Blackwell.

Banerjee, A., Guo, X. and Wang, H. (2005), 'On the optimality of conditional expectation as a bregman predictor', *IEEE Transactions on Information Theory* **51**(7), 2664–2669.

Banks, J., Blundell, R. and Lewbel, A. (1997), 'Quadratic engel curves and consumer demand', *Review of Economics and statistics* **79**(4), 527–539.

Barrett, G. F. and Donald, S. G. (2003), 'Consistent tests for stochastic dominance', *Econometrica* **71**(1), 71–104.

Bartoo, J. B. and Puri, P. S. (1967), 'On optimal asymptotic tests of composite statistical hypotheses', *Annals of Mathematical Statistics* **38**(6), 1845–1852.

Basawa, I. (1985), Neyman-le cam tests based on estimating functions, *in* 'Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer', Vol. 2, Wadsworth, pp. 811–825.

Basawa, I. V., Godambe, V. P. and Taylor, R. L., eds (1997), *Selected Proceedings of the Symposium on Estimating Functions*, Vol. 32 of *IMS Lecture Notes Monograph Series*, Institute of Mathematical Statistics, Hayward, California.

Beare, B. K. (2017), 'The chang-kim-park model of cointegrated density-valued time series cannot accommodate a stochastic trend', *Econ Journal Watch* **14**(2), 133.

Beare, B. K., Seo, J. and Seo, W.-K. (2017), 'Cointegrated linear processes in hilbert space', *Journal of Time Series Analysis* **38**(6), 1010–1027.

Beare, B. K. and Seo, W.-K. (2020), 'Representation of i (1) and i (2) autoregressive hilbertian processes', *Econometric Theory* **36**(5), 773–802.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015), 'Some new asymptotic theory for least squares series: Pointwise and uniform results', *Journal of Econometrics* **186**(2), 345–366.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017), 'Program evaluation and causal inference with high-dimensional data', *Econometrica* **85**(1), 233–298.

Bera, A. and Bilias, Y. (2001), 'Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results', *Journal of Statistical Planning and Inference* **97**, 9–44.

Bera, A. K. and Ullah, A. (1991), 'Rao's score test in econometrics', *CentER Discussion Paper* **1991-43**.

Bera, A. K. and Yoon, M. J. (1993), 'Specification testing with locally misspecified alternatives', *Econometric theory* **9(4)**, 649–658.

Beran, J., Bhansali, R. J. and Ocker, D. (1998), 'On unified model selection for stationary and non-stationary short- and long- memory autoregresive processes', *Biometrika* **85**, 921–934.

Berger, A. and Wallenstein, S. (1989), 'On the theory of $C(\alpha)$ -tests', *Stattistics & probability letters* **7(5)**, 419–424.

Berk, K. N. (1974), 'Consistent autoregressive spectral estimates', *Annals of Statistics* **2**, 489–502.

Bernshtein, A. V. (1976), 'On optimal asymptotic tests for composite hypotheses under non-standard conditions', *Theory of Probability and its Applications* **21**, 34–47.

Bernshtein, A. V. (1978), 'On optimal asymptotic tests of homogeneity', *Theory of Probability and its Applications* **22**, 377–383.

Bernshtein, A. V. (1980*a*), 'On the construction of majorizing tests', *Theory of Probability and its Applications* **25**, 16–26.

Bernshtein, A. V. (1980*b*), 'On verifying composite hypotheses with nuisance parameters in the multivariate case', *Theory of Probability and its Applications* **25**, 287–298.

Bernshtein, A. V. (1981), 'Asymptotically similar criteria', *Journal of Soviet Mathematics* **17**(3), 1825–1857.

Berry, S., Levinsohn, J. and Pakes, A. (1995), 'Automobile prices in market equilibrium', *Econometrica* **63(4)**, 841–890.

Berry, S. T. (1994), 'Estimating discrete-choice models of product differentiation', *The RAND Journal of Economics* **25(2)**, 242–262.

Bertanha, M. and Moreira, M. J. (2020), 'Impossible inference in econometrics: Theory and applications', *Journal of Econometrics* **218(2)**, 247–270.

Bhansali, R. J. (1978), 'Linear prediction by autoregressive model fitting in the time domain', *Annals of Statistics* **6**, 224–231.

Bhansali, R. J. (1996), 'Asymptotically efficient autoregressive model selection for multistep prediction', *Annals of the Institute of Statistical Mathematics,* **48**, 577–602.

Bhat, B. R. and Nagnur, B. N. (1965), 'Locally asymptotically most stringent tests and Lagrangian multiplier tests of linear hypotheses', *Biometrika* **52**(3-4), 459–468.

Billingsley, P. (1995), *Probability and Measure*, third edn, John Wiley & Sons, New York.

Blundell, R., Browning, M. and Crawford, I. (2008), 'Best nonparametric bounds on demand responses', *Econometrica* **76**(6), 1227–1262.

Blundell, R., Duncan, A. and Pendakur, K. (1998), 'Semiparametric estimation and consumer demand', *Journal of Applied Econometrics* **13**(5), 435–461.

Blundell, R. W., Browning, M. and Crawford, I. A. (2003), 'Nonparametric engel curves and revealed preference', *Econometrica* **71**(1), 205–240.

Bontemps, C. (2019), 'Moment-based tests under parameter uncertainty', *Review of Economics and Statistics* **101**(1), 146–159.

Bontemps, C. and Meddahi, N. (2012), 'Testing distributional assumptions: A GMM approach', *Journal of Applied Econometrics* **27**, 978–1012.

Bosq, D. (2000), *Linear processes in function spaces: theory and applications*, Vol. 149, Springer.

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press.

Boudjellaba, H., Dufour, J.-M. and Roy, R. (1992), 'Testing causality between two vectors in multivariate ARMA models', *Journal of the American Statistical Association* **87**(420), 1082–1090.

Boudjellaba, H., Dufour, J.-M. and Roy, R. (1994), 'Simplified conditions for non-causality between two vectors in multivariate arma models', *Journal of Econometrics* **63**(2), 271–287.

Bouhaddioui, C. (2002), Tests d'indépendance de deux séries multivariées autorégressives d'ordre infini, PhD thesis, Département de mathématiques et de statistique, Université de Montréal.

Bouhaddioui, C. and Dufour, J.-M. (2008), 'Tests for non-correlation of two infinite-order cointegrated vector autoregressive series', *Journal of Applied Probability and Statistics* **3**(1), 78–94.

Bouhaddioui, C. and Roy, R. (2006), 'A generalized portmanteau test for independence of two infinite order vector autoregressive series', *Journal of Time Series Analysis* **27**(4), 505–544.

Bourguignon, F., Ferreira, F. H. and Lustig, N. (2004), *The microeconomics of income distribution dynamics in East Asia and Latin America*, World Bank Publications.

Bowden, R. (1973), 'The theory of parametric identification', *Econometrica* **41(6)**, 1069–1074.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Brockwell, P. J. and Davis, R. A. (2013), *Time series: theory and methods*, Springer Science & Business Media.

Brown, B. (1971), 'Martingale central limit theorems', *Annals of Mathematical Statistics* **42**, 59–66.

Bühler, W. J. and Puri, P. S. (1966), 'On optimal asymptotic tests of composite hypotheses with several constraints', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **5**(1), 71–88.

Buja, A., Berk, R. A., Brown, L. D., George, E. I., Pitkin, E., Traskin, M., Zhao, L. and Zhang, K. (2015), 'Models as approximations-a conspiracy of random regressors and model deviations against classical inference in regression', *Statistical Science* .

Bunke, H. and Bunke, O. (1974), 'Identifiability and estimability', *Statistics: A Journal of Theoretical and Applied Statistics* **5**(3), 223–233.

Burnham, K. P. and Anderson, D. R. (2002), *Model selection and multimodel inference : a practical information-theoretic approach*, Springer.

Cacoullos, T. and Papathanasiou, V. (1985), 'On upper bounds for the variance of functions of random variables', *Statistics & probability letters* **3**(4), 175–184.

Cacoullos, T. and Papathanasiou, V. (1992), 'Lower variance bounds and a new proof of the central limit theorem', *Journal of Multivariate Analysis* **43**(2), 173–184.

Cai, T. T., Levine, M. and Wang, L. (2009), 'Variance function estimation in multivariate nonparametric regression with fixed design', *Journal of Multivariate Analysis* **100**(1), 126–136.

Cai, Z., Ren, Y. and Sun, L. (2015), 'Pricing kernel estimation: A local estimating equation approach', *Econometric Theory* **31**(3), 560–580.

Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018), 'On the effect of bias estimation on coverage accuracy in nonparametric inference', *Journal of the American Statistical Association* **113**(522), 767–779.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

Caner, M. (2006), 'M-estimators with non-standard rates of convergence and weakly dependent data', *Journal of Statistical Planning and Inference* **136**(4), 1207–1219.

Card, D., Lee, D. S., Pei, Z. and Weber, A. (2015), 'Inference on causal effects in a generalized regression kink design', *Econometrica* **83**(6), 2453–2483.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998), 'Local estimating equations', *Journal of the American Statistical Association* **93**(441), 214–227.

Cattaneo, M. D., Farrell, M. H. and Feng, Y. (2020), 'Large sample properties of partitioning-based series estimators', *Annals of Statistics* **48**(3), 1718–1741.

Cavaliere, G. (2005), 'Limited time series with a unit root', *Econometric Theory* **21**(5), 907–945.

Cavaliere, G. and Xu, F. (2014), 'Testing for unit roots in bounded time series', *Journal of Econometrics* **178**, 259–272.

Chandra, T. K. and Joshi, S. (1983), 'Comparison of the likelihood ratio, Rao's and Wald's tests and a conjecture of C.R. Rao', *Sankhyā: The Indian Journal of Statistics, Series A* **45(2)**, 226–246.

Chang, Y., Kaufmann, R. K., Kim, C. S., Miller, J. I., Park, J. Y. and Park, S. (2020), 'Evaluating trends in time series of distributions: A spatial fingerprint of human effects on climate', *Journal of Econometrics* **214**(1), 274–294.

Chang, Y., Kim, C. S. and Park, J. Y. (2016), 'Nonstationarity in time series of state densities', *Journal of Econometrics* **192**(1), 152–167.

Chant, D. (1974), 'On asymptotic tests of composite hypotheses in nonstandard conditions', *Biometrika* **61**(2), 291–298.

Charalambos, D. A. and Border, K. C. (2006), *Infinite dimensional analysis: a hitchhiker's guide*, Springer.

Chaudhuri, S. and Zivot, E. (2011), 'A new method of projection-based inference in GMM with weakly identified nuisance parameters', *Journal of Econometrics* **164**(2), 239–251.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of Econometrics* **6**, 5549–5632.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations', *Annals of Mathematical Statistics* **23(4)**, 493–507.

Chernoff, H. (1964), 'Estimation of the mode', *Annals of the Institute of Statistical Mathematics* **16**(1), 31–41.

Chernoff, H. (1981), 'A note on an inequality involving the normal distribution', *Annals of Probability* **9(3)**, 533–535.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *Econometrics Journal* **21**(1), C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2022), 'Locally robust semiparametric estimation', *Econometrica* **90(4)**, 1501–1535.

Chernozhukov, V., Hong, H. and Tamer, E. (2007*a*), 'Estimation and confidence regions for parameter sets in econometric models 1', *Econometrica* **75**(5), 1243–1284.

Chernozhukov, V., Hong, H. and Tamer, E. (2007*b*), 'Parameter set inference in a class of econometric models', *Econometrica* **75**(5), 1243–1284.

Cheung, Y.-W. and Ng, L. K. (1996), 'A causality-in-variance test and its application to financial market prices', *Journal of Econometrics* **72**, 33–48.

Chibisov, D. M. (1973), Asymptotic expansions for Neyman's $C(\alpha)$ tests, *in* 'Proceedings of the Second Japan-USSR Symposium on Probability Theory', Springer, Berlin, pp. 16–45.

Cleveland, W. S. and Devlin, S. J. (1988), 'Locally weighted regression: an approach to regression analysis by local fitting', *Journal of the American statistical association* **83**(403), 596–610.

Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Engle, R.F., White, H. (1999), Oxford University Press, Oxford.

Cox, D. D. (1988), 'Approximation of least squares regression on nested subspaces', *Annals of Statistics* **16(2)**, 713–732.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

Dagenais, M. G. and Dufour, J.-M. (1991), 'Invariance, nonlinear models and asymptotic tests', *Econometrica* **59**, 1601–1615.

Davidson, R. and Duclos, J.-Y. (2000), 'Statistical inference for stochastic dominance and for the measurement of poverty and inequality', *Econometrica* **68**(6), 1435–1464.

Davidson, R. and Duclos, J.-Y. (2013), 'Testing for restricted stochastic dominance', *Econometric Reviews* **32**(1), 84–125.

Davidson, R. and Flachaire, E. (2007), 'Asymptotic and bootstrap inference for inequality and poverty measures', *Journal of Econometrics* **141**(1), 141–166.

Davidson, R. and MacKinnon, J. G. (1991), 'Artificial regressions and $C(\alpha)$ tests', *Economic Letters* **35**, 149–153.

Davidson, R. and MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

Dawson, D. A. (1975), 'Stochastic evolution equations and related measure processes', *Journal of Multivariate Analysis* **5**(1), 1–52.

De Boor, C. (1978), *A practical guide to splines*, Vol. 27, Springer.

Dégerine, S. and Lambert-Lacroix, S. (2003), 'Characterization of the partial autocorrelation function of nonstationary time series', *Journal of Multivariate Analysis* **87**, 46–59.

Deistler, M. and Seifert, H.-G. (1978), 'Identifiability and consistent estimability in econometric models', *Econometrica* **46(4)**, 969–980.

Den Haan, W. J. and Levin, A. T. (1997), 'A practitioner's guide to robust covariance matrix estimation', *Handbook of Statistics* **15**, 299–342.

Deutsch, F. R. (2012), *Best approximation in inner product spaces*, Springer.

Diouf, M. A. and Dufour, J. M. (2005), Improved nonparametric inference for the mean of a bounded random variable with applicaiton to poverty measures, Technical report.

Druedahl, J., Jørgensen, T. and Graber, M. (2021), 'High frequency income dynamics', *SSRN 3816424* .

Dua, D. and Graff, C. (2017), 'UCI machine learning repository'.
*http://archive.ics.uci.edu/ml

Duchesne, P. (2005), 'Testing for serial correlation of unknown form in cointegrated time series models?', *Annals of the Institute of Statistical Mathematics* **57**, 575–595.

Duchesne, P. and Roy, R. (2003), 'Robust tests for independence of two time series', *Statistica Sinica* **13**, 827–852.

Dufour, J.-M. (1997), 'Some impossibility theorems in econometrics with applications to structural and dynamic models', *Econometrica* **65(6)**, 1365–1387.

Dufour, J.-M. (2002), 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics', *Journal of Econometrics* **133(2)**, 443–477.

Dufour, J.-M. (2003), 'Identification, weak instruments, and statistical inference in econometrics', *Canadian Journal of Economics/Revue canadienne d'économique* **36**(4), 767–808.

Dufour, J.-M. and Dagenais, M. G. (1992), 'Nonlinear models, rescaling and test invariance', *Journal of Statistical Planning and Inference* **32(1)**, 111–135.

Dufour, J.-M. and Hallin, M. (1992), 'Simple exact bounds for distributions of linear signed rank statistics', *Journal of Statistical Planning and Inference* **31(3)**, 311–333.

Dufour, J.-M. and Jasiak, J. (2001), 'Finite sample limited information inference methods for structural equations and models with generated regressors', *International Economic Review* **42**(3), 815–844.

Dufour, J.-M., Pelletier, D. and Renault, É. (2006), 'Short run and long run causality in time series: Inference', *Journal of Econometrics* **132**(2), 337–362.

Dufour, J.-M. and Renault, É. (1998), 'Short run and long run causality in time series: Theory', *Econometrica* **66**, 1099–1125.

Dufour, J.-M., Trognon, A. and Tuvaandorj, P. (2016), Generalized $C(\alpha)$ tests for estimating functions with serial dependence, *in* 'Advances in Time Series Methods and Applications', Springer, pp. 151–178.

Dufour, J.-M., Trognon, A. and Tuvaandorj, P. (2017), 'Invariant tests based on m-estimators, estimating functions, and the generalized method of moments', *Econometric Reviews* **36**(1-3), 182–204.

Dufour, J.-M. and Valéry, P. (2009), 'Exact and asymptotic tests for possibly non-regular hypotheses on stochastic volatility models', *Journal of Econometrics* **150**, 193–206.

Durbin, J. (1960), 'Estimation of parameters in time series regression models', *Journal of Royal Statistical Society: Series B (Methodological)* **22(1)**, 139–153.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956), 'Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator', *Annals of Mathematical Statistics* **27(3)**, 642–669.

Efron, B. (2014), 'Estimation and accuracy after model selection', *Journal of the American Statistical Association* **109**(507), 991–1007.

Egozcue, J. J., Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006), 'Hilbert space of probability density functions based on aitchison geometry', *Acta Mathematica Sinica* **22**(4), 1175–1182.

El Himdi, K. and Roy, R. (1997), 'Tests for noncorrelation of two multivariate ARMA time series', *Canadian Journal of Statistics* **25**, 233–256.

Engle, R. and Granger, C. (1987), 'Co-integration and error correction: Representation, estimation and testing', *Econometrica* **55**, 251–276.

Eriksson, K., Estep, D. and Johnson, C. (2013), *Applied mathematics: Body and soul: Volume 1: Derivatives and geometry in IR3*, Springer.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, Routledge.

Farrell, M. H., Liang, T. and Misra, S. (2021), 'Deep neural networks for estimation and inference', *Econometrica* **89(1)**, 181–213.

Firpo, S., Fortin, N. M. and Lemieux, T. (2009), 'Unconditional quantile regressions', *Econometrica* **77**(3), 953–973.

Foster, J., Greer, J. and Thorbecke, E. (1984), 'A class of decomposable poverty measures', *Econometrica* **52(3)**, 761–766.

Foutz, R. V. (1976), 'On the consistency of locally asymptotically most stringent tests', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **4**(2), 211–219.

Gagliardini, P., Gourieroux, C. and Renault, E. (2011), 'Efficient derivative pricing by the extended method of moments', *Econometrica* **79**(4), 1181–1232.

Gasser, T. and Müller, H.-G. (1979), *Kernel estimation of regression functions*, Springer, pp. 23–68.

Gasser, T., Muller, H.-G. and Mammitzsch, V. (1985), 'Kernels for nonparametric curve estimation', *Journal of the Royal Statistical Society: Series B (Methodological)* **47(2)**, 238–252.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986), 'Residual variance and residual pattern in nonlinear regression', *Biometrika* **73**(3), 625–633.

Gel'fand, I. M. (1956), 'Generalized random processes: a theory and the white gaussian process', *Dokl. Akd. Nuuk SSSR* **100**, 855–856. (in Russian).

Gel'fand, I. M. and Vilenkin, N. Y. (1964), *Generalized Functions*, Vol. 4: Applications of harmonic analysis, Academic Press.

Geweke, J. (1981), 'The approximate slopes of econometric tests', *Econometrica* **49**, 1427–1442.

Geweke, J. (1984), 'Inference and causality in economic time series', *Handbook f Econometrics* **2**, 1102–1144.

Ghosh, B. K. and Huang, W. M. (1991), 'The power and optimal kernel of the Bickel-Rosenblatt test for goodness of fit', *Annals of Statistics* **17**, 999–1009.

Giles, J. A. (2002), Time series analysis testing for two-step Granger noncausality in trivariate VAR models, *in* 'Handbook Of Applied Econometrics And Statistical Inference', CRC Press, chapter 18, pp. 371–399.

Givens, C. R. and Shortt, R. M. (1984), 'A class of wasserstein metrics for probability distributions.', *Michigan Mathematical Journal* **31**(2), 231–240.

Glad, I. K. (1998), 'Parametrically guided non-parametric regression', *Scandinavian Journal of Statistics* **25**(4), 649–668.

Godambe, V. P. (1960), 'An optimum property of regular maximum likelihood estimation', *Annals of Mathematical Statistics* **31**, 1208–1212. Ackowledgement 32 (1960), 1343.

Godambe, V. P., ed. (1991), *Estimating Functions*, Clarendon Press, Oxford, U.K.

Gómez, V. (2016), *Multivariate Time Series with Linear State Space Structure*, Springer.

Gordin, M. I. (1969), 'Central limit theorem for stationary processes', *Doklady Akademii Nauk SSSR* **188**(4), 739.

Gourieroux, C. and Monfort, A. (1995), *Statistics and econometric models*, Vol. 1, Cambridge University Press.

Gouriéroux, C. and Monfort, A. (1997), *Time Series and Dynamic Models*, Cambridge University Press.

Gourieroux, C., Monfort, A. and Trognon, A. (1984), 'Pseudo maximum likelihood methods: Theory', *Econometrica* **52(3)**, 681–700.

Granger, C. (1969*a*), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* **37**, 424–459.

Granger, C. W. (1969*b*), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* **37(3)**, 424–438.

Green, K. C. and Armstrong, J. S. (2015), 'Simple versus complex forecasting: The evidence', *Journal of Business Research* **68**(8), 1678–1685.

Grenander, U. (1981), Abstract inference, Technical report.

Griliches, Z. and Intrilligator, M. D., eds (1984), *Handbook of Econometrics, Volume 2*, Amsterdam.

Guvenen, F., Karahan, F., Ozkan, S. and Song, J. (2021), 'What do data on millions of us workers reveal about lifecycle earnings dynamics?', *Econometrica* **89**(5), 2303–2339.

Haan, W. J. D. and Levin, A. T. (1996), 'Inferences from parametric and non-parametric covariance matrix estimation procedures', *NBER Techical Working Paper* .

Hahn, J., Todd, P. and Van der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**(1), 201–209.

Hall, A. R. (2004), *Generalized Method of Moments*, Advanced Texts in Econometrics, OUP.

Hall, P. (1992), 'Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density', *Annals of Statistics* **20(2)**, 675–694.

Hall, P. and Horowitz, J. (2013), 'A simple bootstrap method for constructing nonparametric confidence bands for functions', *Annals of Statistics* **41**(4), 1892–1921.

Hall, W. J. and Mathiason, D. J. (1990), 'On large-sample estimation and testing in parametric models', *International Statistical Review* **58**(1), 77–97.

Hallin, M., Jurečkovà, J., Picek, J. and Zahaf, T. (1999), 'Nonparametric tests of independence of two autoregressive time series based on autoregression rank scores', *Journal of Statistical Planning and Inference* **75**, 319–330.

Hallin, M. and Saidi, A. (2005), 'Testing independance and causality between multivariate ARMA times series', *Journal of Time Series Analysis* **26**, 83–106.

Hallin, M. and Saidi, A. (2007), 'Optimal tests for non-correlation between multivariate time series', *Journal of The American Statistical Association* **102**(479), 938–951.

Hannan, E. (1970), *Multiple Time Series*, John Wiley and Sons.

Hannan, E. J. (1976), 'The asymptotic distribution of serial covariances', *Annals of Statistics* **4**(2), 396–399.

Hansen, B. E. (2014), 'Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation', *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* .

Hansen, L. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**, 1029–1054.

Haugh, L. D. (1976), 'Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach', *Journal of the American Statistical Association* **71**, 378–385.

Hausman, J. A. and Taylor, W. E. (1981), 'Panel data and unobservable individual effects', *Econometrica* **49(6)**, 1377–1398.

Hayashi, F. (2011), *Econometrics*, Princeton University Press.

Heyde, C. C. (1997), *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*, Springer.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hong, Y. (1996*a*), 'Consistent testing for serial correlation of unknown form', *Econometrica* **64**, 837–864.

Hong, Y. (1996*b*), A separate mathematical appendix for testing for independence between two covariance stationary time series, Technical report, Ithaca, New York.

Hong, Y. (1996*c*), 'Testing for independence between two covariance stationary time series', *Biometrika* **83**(3), 615–625.

Hong, Y. (1998), 'Testing for pairwise serial independence via the empirical distribution function', *Journal of Royal Statistical Society: Series B (Statistical Methodology)* **60(2)**, 429–453.

Hong, Y. (1999), Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach, Technical Report 448.

Horowitz, J. L. (1992), 'A smoothed maximum score estimator for the binary response model', *Econometrica* **60(3)**, 505–531.

Horowitz, J. L. (2009), *Semiparametric and nonparametric methods in econometrics*, Vol. 12, Springer.

Hosking, J. R. M. (1980), 'The multivariate portemanteau statistic', *Journal of the American Statistical Association* **75**, 602–608.

Hunter, J. K. and Nachtergaele, B. (2001), *Applied analysis*, World Scientific Publishing Company.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998), 'Smoothing parameter selection in nonparametric regression using an improved akaike information criterion', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2), 271–293.

Imbens, G. W. and Lemieux, T. (2008), 'Regression discontinuity designs: A guide to practice', *Journal of Econometrics* **142**(2), 615–635.

Imbens, G. W. and Manski, C. F. (2004), 'Confidence intervals for partially identified parameters', *Econometrica* **72**(6), 1845–1857.

Itô, K. (1954), 'Stationary random distributions', *Memoirs of the College of Science, University of Kyoto. Series A: Mathematics* **28**(3), 209–223.

Jaggia, S. and Trivedi, P. K. (1994), 'Joint and separate score tests for state dependence and unobserved heterogeneity', *Journal of Econometrics* **60**(1), 273–291.

Jann, B. (2019), Influence functions for linear regression (with an application to regression adjustment), Technical report.

Jennrich, R. I. (1969), 'Asymptotic properties of non-linear least squares estimators', *Annals of Mathematical Statistics* **40**(2), 633–643.

Jonker, M. and Van der Vaart, A. (2014), 'On the correction of the asymptotic distribution of the likelihood ratio statistic if nuisance parameters are estimated based on an external source', *International Journal of Biostatistics* **10**(2), 123–142.

Kakwani, N. (1980), 'On a class of poverty measures', *Econometrica* **48(2)**, 437–446.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering* **82**(1), 35–45.

Kanaya, S. (2017), 'Convergence rates of sums of α-mixing triangular arrays: with an application to nonparametric drift function estimation of continuous-time processes', *Econometric Theory* **33**(5), 1121–1153.

Kim, J. and Pollard, D. (1990), 'Cube root asymptotics', *Annals of Statistics* **18(1)**, 191–219.

King, R. G., Plosser, C. I. and Rebelo, S. T. (1988), 'Production, growth and business cycles', *Journal of Monetary Economics* **21**(2/3), 196–232.

Kneip, A. and Utikal, K. J. (2001), 'Inference for density families using functional principal component analysis', *Journal of the American Statistical Association* **96**(454), 519–542.

Koch, P. and Yang, S.-S. (1986), 'A method for testing the independance of two time series that accounts for a potential pattern in the cross-correlation function', *Journal of the American Statistical Analysis* **81**, 533–544.

Kocherlakota, S. and Kocherlakota, K. (1991), 'Neyman's $C(\alpha)$ test and Rao's efficient score test for composite hypotheses', *Statistical & Probability Letters* **11**, 491–493.

Koenker, R. and Hallock, K. F. (2001), 'Quantile regression', *Journal of Economic Perspectives* **15**(4), 143–156.

Konishi, S. and Kitagawa, G. (2008), *Information criteria and statistical modeling*, Springer Science & Business Media.

Kosorok, M. R. (2008), *Introduction to empirical processes and semiparametric inference.*, Springer.

Kremer, M. and Chen, D. L. (2002), 'Income distribution dynamics with endogenous fertility', *Journal of Economic Growth* **7**(3), 227–258.

Kydland, F. E. and Prescott, E. C. (1982), 'Time to build and aggregate fluctuations', *Econometrica* **50(6)**, 1345–1370.

Lavergne, P. and Patilea, V. (2013), 'Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory', *Journal of Econometrics* **177**(1), 47–59.

Le Cam, L. (1956), On the asymptotic theory of estimation and testing hypotheses, *in* 'Contribution to the Theory of Statistics', University of California Press.

Le Cam, L. and Traxler, R. (1978), 'On the asymptotic behavior of mixtures of Poisson distributions', *Probability Theory and Related Fields* **44**(1), 1–45.

Lee, D. S. (2008), 'Randomized experiments from non-random selection in us house elections', *Journal of Econometrics* **142**(2), 675–697.

Lee, D. S. and Lemieux, T. (2010), 'Regression discontinuity designs in economics', *Journal of Economic Literature* **48**(2), 281–355.

Lee, L.-f. (2005), 'Classical inference with ml and gmm estimates with various rates of convergence', *manuscript, Department of Economics, Ohio State University, Columbus, Ohio* .

Lee, L.-f. (2010), 'Pooling estimates with different rates of convergence: a minimum $\chi 2$ approach with emphasis on a social interactions model', *Econometric Theory* **26**(1), 260–299.

Lewbel, A. (2007), 'A local generalized method of moments estimator', *Economics Letters* **94**(1), 124–128.

Lewbel, A. (2019), 'The identification zoo: Meanings of identification in econometrics', *Journal of Economic Literature* **57**(4), 835–903.

Lewbel, A. and Pendakur, K. (2008), 'Estimation of collective household models with engel curves', *Journal of Econometrics* **147**(2), 350–358.

Lewis, R. and Reinsel, G. C. (1985), 'Prediction of multivariate time series by autoregressive model fitting', *Journal of Multivariate Analysis* **16**, 393–411.

Li, H., Xie, D. and Zou, H.-F. (2000), 'Dynamics of income distribution', *Canadian Journal of Economics/Revue canadienne d'économique* **33**(4), 937–961.

Li, K.-C. (1987), 'Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set', *Annals of Statistics* pp. 958–975.

Li, K.-C. (1989), 'Honest confidence regions for nonparametric regression', *Annals of Statistics* **17**(3), 1001–1008.

Li, Q. and Racine, J. (2004), 'Cross-validated local linear nonparametric regression', *Statistica Sinica* pp. 485–512.

Li, Q. and Racine, J. S. (2007), *Nonparametric econometrics: theory and practice*, Princeton University Press.

Li, W. K. and McLeod, A. I. (1981), 'Distribution of the residual autocorrelation in multivariate ARMA time series models', *Journal of the Royal Statistical Society: Series B* **43**, 231–239.

Lian, H., Qiao, X. and Zhang, W. (2021), 'Homogeneity pursuit in single index models based panel data analysis', *Journal of Business and Economic Statistics* **39**(2), 386–401.

Loève, M. (1977), *Probability Theory, Volumes I and II*, 4th edn, New York.

Loève, Michel, M. (1960), *Probability Theory*, D. Van Nostrand Co. Inc.

Longworth, D., ed. (1992), *Monetary Seminar: A Seminar Sponsored by the Bank of Canada, May 7-9, 1990*, Bank of Canada, Ottawa.

Lütkepohl, H. (1985), 'The joint asymptotic distribution of multistep prediction errors of estimated vector autoregressions', *Economics Letters* **17**, 103–106.

Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Lütkepohl, H. (1993), Testing for causation between two variables in higher dimensional VAR models, *in* (Schneeweiss and Zimmermann 1993).

Lütkepohl, H. (2001), *Vector Autoregressions*, in (Baltagi 2001), chapter 32, pp. 678–699.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Lütkepohl, H. and Burda, M. M. (1997), 'Modified Wald tests under nonregular conditions', *Journal of Econometrics* **78**, 315–332.

Lütkepohl, H. and Saikkonen, P. (1997), 'Impulse response analysis in infinite order cointegrated vector autoregressive processes', *Journal of Econometrics* **81**, 127–157.

Lütkepohl, H. and Saikkonen, P. (1999), Order selection in testing for the cointegrating rank of a VAR process, in *Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Engle, R.F., White, H. Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Engle, R.F., White, H.* (1999), pp. 168–199.

Mallows, C. L. (1973), 'Some comments on c p', *Technometrics* **15**(4), 661–675.

Manski, C. F. (1975), 'Maximum score estimation of the stochastic utility model of choice', *Journal of Econometrics* **3**(3), 205–228.

Manski, C. F. (1985), 'Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator', *Journal of Econometrics* **27**(3), 313–333.

Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *American Economic Review* **80**(2), 319–323.

Manski, C. F. (2003), *Partial identification of probability distributions*, Springer.

Manski, C. F. and Tamer, E. (2002), 'Inference on regressions with interval data on a regressor or outcome', *Econometrica* **70**(2), 519–546.

Martins-Filho, C., Mishra, S. and Ullah, A. (2008), 'A class of improved parametrically guided nonparametric regression estimators', *Econometric Reviews* **27**(4-6), 542–573.

Massart, P. (1990), 'The tight constant in the dvoretzky-kiefer-wolfowitz inequality', *Annals of Probability* **18(3)**, 1269–1283.

Matzkin, R. L. (2007), 'Nonparametric identification', *Handbook of Econometrics* **6**, 5307–5368.

McCulloch, W. S. and Pitts, W. (1943), 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics* **5**(4), 115–133.

Meidan, R. (1979), 'Reproducing-kernel hilbert spaces of distributions and generalized stochastic processes', *SIAM Journal on Mathematical Analysis* **10**(1), 62–70.

Meidan, R. (1980), 'On the connection between ordinary and generalized stochastic processes', *Journal of Mathematical Analysis and Applications* **76**(1), 124–133.

Meyer, R. (1974), 'On correlation and linearity', *The American Statistician* **28**(3), 103–103.

Moore, D. S. and Spruill, M. C. (1975), 'Unified large-sample theory of general chi-squared statistics for tests of fit', *Annals of Statistics* **3(3)**, 599–616.

Moore, E. H. (1920), 'On the reciprocal of the general algebraic matrix', *Bulletin of the American Mathematical Society* **26**, 394–395.

Moran, P. A. P. (1970), 'On asymptotically optimal tests of composite hypotheses', *Biometrika* **57**(1), 47–55.

Moran, P. A. P. (1973), 'Asymptotic properties of homogeneity tests', *Biometrika* **60**(1), 79–85.

Nadaraya, E. A. (1964), 'On estimating regression', *Theory of Probability & Its Applications* **9**(1), 141–142.

Newbold, P. (1982), Causality testing in economics, *in* 'Time Series Analysis: Theory and Practice I', North-Holland Publishing Company.

Newey, W. K. (1991), 'Uniform convergence in probability and stochastic equicontinuity', *Econometrica* **59(4)**, 1161–1167.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147–168.

Newey, W. K. and McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of Econometrics* **4**, 2111–2245.

Newey, W. K. and West, K. D. (1987*a*), 'Hypothesis testing with efficient method of moments estimators', *International Economic Review* **28**, 777–787.

Newey, W. K. and West, K. D. (1987*b*), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

Neyman, J. (1954), 'Sur une famille de tests asymptotiques des hypothéses statistiques composées', *Trabajos de Estadistica* **5**(2), 161–168.

Neyman, J. (1959), *Probability and Statistics*, Wiley, chapter Optimal asymptotic tests of composite hypotheses, pp. 213–234.

Neyman, J. (1979), 'C(α) tests and their use', *Sankhyā: The Indian Journal of Statistics, Series A* **41**, 1–21.

Neyman, J. and Pearson, E. S. (1928), 'On the use and interpretation of certain test criteria for purposes of statistical inference', *Biometrika* **20A(1/2)**, 263–294.

Nielsen, H. S., Sørensen, T. and Taber, C. (2010), 'Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform', *American Economic Journal: Economic Policy* **2**(2), 185–215.

Nirei, M. and Souma, W. (2007), 'A two factor model of income distribution dynamics', *Review of Income and Wealth* **53**(3), 440–459.

Nobel, A. and Dembo, A. (1993), 'A note on uniform laws of averages for dependent processes', *Statistics & Probability Letters* **17**(3), 169–172.

Office for National Statistics (2002), 'Family expenditure survey, 2000-2001', http://doi.org/10.5255/UKDA-SN-4490-1. [data collection]. UK Data Service. SN: 4490.

Oliva, J., Póczos, B. and Schneider, J. (2013), 'Distribution to distribution regression', *Proceedings of the 30th International Conference on Machine Learning, PMLR* **28(3)**, 1049–1057.

Pagan, A. and Ullah, A. (1999), *Nonparametric econometrics*, Cambridge university press.

Pal, C. (2003), 'Higher order $c(\alpha)$ tests with applications to mixture models', *Journal of Statistical Planning and Inference* **113**(1), 179–187.

Paparoditis, E. (1996), 'Bootstrapping autoregressive and moving average parameter estimates of infinite order vector autoregressive processes', *Journal of Multivariate Analysis* **57**, 277–296.

Park, J. Y. (1990), 'Testing for unit roots and cointegration by variable addition', *Advances in Econometrics* **8**, 107–133.

Park, J. Y. and Qian, J. (2012), 'Functional regression of continuous state distributions', *Journal of Econometrics* **167**(2), 397–412.

Park, J. Y., Shin, K. and Wang, Y.-J. (2010), 'A semiparametric cointegrating regression: Investigating the effects of age distributions on consumption and saving', *Journal of Econometrics* **157**, 165–178.

Parzen, E. (1974), 'Some recent advances in time series modeling', *IEEE Transactions on Automatic Control* **19(6)**, 723–730.

Paul, S. R. and Barnwal, R. K. (1990), 'Maximum likelihood estimation and a $C(\alpha)$ test for a common intraclass correlation', *Journal of the Royal Statistical Society. Series D (The Statistician)* **39**(1), 19–24.

Paulino, C. D. M. and de Bragança Pereira, C. A. (1994), 'On identifiability of parametric statistical models', *Journal of the Italian Statistical Society* **3**(1), 125–151.

Penrose, R. (1955), A generalized inverse for matrices, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 51, Cambridge University Press, pp. 406–413.

Pesaran, H., Smith, R. and Im, K. S. (1996), Dynamic linear models for heterogenous panels, *in* 'The econometrics of panel data', Springer, pp. 145–195.

Petersen, A. and Müller, H.-G. (2016), 'Functional data analysis for density functions by transformation to a hilbert space', *Annals of Statistics* **44**(1), 183–218.

Petersen, A. and Müller, H.-G. (2019), 'Wasserstein covariance for multiple random densities', *Biometrika* **106**(2), 339–351.

Petersen, A., Zhang, C. and Kokoszka, P. (2022), 'Modeling probability density functions as data objects', *Econometrics and Statistics* **21**, 159–178.

Pham, D. T., Roy, R. and Cédras, L. (2000), Tests for non-correlation of two cointegrated ARMA time series, Technical Report CRM-2649.

Pham, D. T., Roy, R. and Cédras, L. (2003), 'Tests for non-correlation of two cointegrated ARMA time series', *Journal of Time Series Analysis* **24**, 553–577.

Philips, T. K. and Nelson, R. (1995), 'The moment bound is tighter than chernoff's bound for positive tail probabilities', *The American Statistician* **49**(2), 175–178.

Phillips, P. C. B. (1989), 'Partially identified econometric models', *Econometric Theory* **5**(2), 181–240.

Phillips, P. C. B. (2005), 'Hac estimation by automated regression', *Econometric Theory* **21**(1), 116–142.

Pierce, D. A. and Haugh, L. D. (1977), 'Causality in temporal systems: Characterizations and survey', *Journal of Econometrics* **5**(3), 265–293.

Pierce, D. A. and Haugh, L. D. (1979), 'The characterization of instantaneous causality, a comment', *Journal of Econometrics* **10**(2), 257–259.

Piketty, T. and Saez, E. (2003), 'Income inequality in the united states, 1913–1998', *Quarterly Journal of Economics* **118**(1), 1–41.

Piketty, T. and Saez, E. (2014), 'Inequality in the long run', *Science* **344**(6186), 838–843.

Ploberger, W. and Phillips, P. C. (2001), *Rissanen's theorem and econometric time series*, Cambridge University Press.

Ploberger, W. and Phillips, P. C. (2003), 'Empirical limits for time series econometric models', *Econometrica* **71**(2), 627–673.

Póczos, B., Singh, A., Rinaldo, A. and Wasserman, L. (2013), Distribution-free distribution regression, *in* 'Artificial Intelligence and Statistics', PMLR, pp. 507–515.

Porter, J. (2003), 'Estimation in the regression discontinuity model', *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison* **2003**, 5–19.

Poskit, D. S. (2007), 'Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases', *Annals of the Institute of Statistical Mathematics* **59**, 697–725.

Pötscher, B. M. (1991), 'Effects of model selection on inference', *Econometric Theory* **7**(2), 163–185.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Vol. 1, Academic Press.

Racette, D. and Raynauld, J. (1992), Un modèle BVAR de prévision de la dépense nominale et d'analyse de la politique monétaire canadienne, *in* (Longworth 1992), pp. 317–325.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis*, Springer.

Rao, B. L. S. P. (1996), 'Optimal asymptotic tests of composite hypotheses for continuous time stochastic processes', *Sankhyā: The Indian Journal of Statistics, Series A* **58**, 8–24.

Rao, C. R. (1948), Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *in* 'Mathematical Proceedings of the Cambridge Philosophical Society', Vol. 44, Cambridge University Press, pp. 50–57.

Rao, C. R. (1973), *Linear statistical inference and its applications*, Vol. 2, Wiley New York.

Ray, R. M. (1974), 'Maxmin $C(\alpha)$ tests against two-sided alternatives', *Annals of Statistics* **2**(6), 1175–1188.

Reinsel, G. C. (1993), *Elements of Multivariate Time Series Analysis*, Springer.

Renault, E., Sekkat, K. and Szafarz, A. (1998), 'Testing for spurious causality in exchange rates', *Journal of Empirical Finance* **5**, 47–66.

Rice, J. (1984), 'Bandwidth choice for nonparametric regression', *Annals of Statistics* **12**(4), 1215–1230.

Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**(5), 465–471.

Rissanen, J. (1982), 'Estimation of structure by minimum description length', *Circuits, Systems and Signal Processing* **1**(3-4), 395–406.

Rissanen, J. (1986), 'Stochastic complexity and modeling', *Annals of Statistics* pp. 1080–1100.

Rissanen, J. (1987), 'Stochastic complexity', *Journal of the Royal Statistical Society: Series B (Methodological)* **49**(3), 223–239.

Robinson, P. M. (1988), 'Root-n-consistent semiparametric regression', *Econometrica* **56(4)**, 931–954.

Romano, J. P. (2004), 'On non-parametric testing, the uniform behaviour of the t-test, and related problems', *Scandinavian Journal of Statistics* **31**(4), 567–584.

Romano, J. P. and Shaikh, A. M. (2010), 'Inference for the identified set in partially identified econometric models', *Econometrica* **78**(1), 169–211.

Ronchetti, E. (1987), 'Robust $C(\alpha)$-type tests for linear models', *Sankhyā Series A* **49**, 1–16.

Roy, R. (1989), 'Asymptotic covariance structure of serial correlation in multivariate time series', *Biometrika* **76**, 824–827.

Runge, C. (1901), 'Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten', *Zeitschrift für Mathematik und Physik* **46**, 20.

Saez, E. (2010), 'Do taxpayers bunch at kink points?', *American Economic Journal: Economic Policy* **2**(3), 180–212.

Saidi, A. (2007), 'Consistent testing for non-correlation of two cointegrated ARMA time series', *Canadian Journal of Statistics* **35**(1), 169–188.

Saikkonen, P. (1992), 'Estimation and testing of cointegrated systems by an autoregressive approximation', *Econometric Theory* **8**, 1–27.

Saikkonen, P. and Lütkepohl, H. (1994), Infinite order cointegrated autoregressive processes: Estimation and inference, Technical Report 5.

Saikkonen, P. and Lütkepohl, H. (1996), 'Infinite-order cointegrated vector autoregressive processes: Estimation and inference', *Econometric Theory* **12**, 814–844.

Saikkonen, P. and Luukkonen, R. (1997), 'Testing cointegration in infinite order vector autoregressive proceses', *Journal of Econometrics* **81**, 93–126.

Santos, A. (2012), 'Inference in nonparametric instrumental variables with partial identification', *Econometrica* **80**(1), 213–275.

Schmitt-Grohé, S. and Uribe, M. (2004), 'Solving dynamic general equilibrium models using a second-order approximation to the policy function', *Journal of Economic Dynamics and Control* **28**(4), 755–775.

Schneeweiss, H. and Zimmermann, K., eds (1993), *Studies in Applied Econometrics*.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *Annals of statistics* **6**(2), 461–464.

Sen, A. (1976), 'Poverty: an ordinal approach to measurement', *Econometrica* **44(2)**, 219–231.

Seo, M. H. and Otsu, T. (2018), 'Local m-estimation with discontinuous criterion for dependent and limited observations', *Annals of Statistics* **46**(1), 344–369.

Seo, W.-K. (2017), 'Cointegrated density-valued linear processes', *arXiv preprint arXiv:1710.07792* .

Shao, X. (2009), 'A generalized portmanteau test for independence between two stationary time series', *Econometric Theory* **25**, 195–210.

Shibata, R. (1980), 'Asymptotically efficient selection of the order of the model for estimating parameters of a linear process', *Annals of Statistics* **8**, 147–164.

Shmueli, G. et al. (2010), 'To explain or to predict?', *Statistical Science* **25**(3), 289–310.

Sims, C. (1980), 'Macroeconomic and reality', *Econometrica* **48**, 1–48.

Singh, A. C. and Zhurbenko, I. G. (1975), 'The power of the optimal asymptotic tests of composite statistical hypotheses', *Proceedings of the National Academy of Sciences* **72**(2), 577–580.

Small, C. G. and McLeish, D. L. (1994), *Hilbert Space Methods in Probability and Statistical Inference*, Wiley.

Smith, R. J. (1987*a*), 'Alternative asymptotically optimal tests and their application to dynamic specification', *Review of Economic Studies* **LIV**, 665–680.

Smith, R. J. (1987*b*), 'Testing the normality assumption in multivariate simultaneous limited dependent variable models', *Journal of Econometrics* **34**, 105–123.

Sobolev, S. (1992), *Cubature Formulas and Modern Analysis: An Introduction Gordon and Breach*, Oxonian Press.

Staiger, D. O. and Stock, J. H. (1994), 'Instrumental variables regression with weak instruments', *Working Paper* .

Stewart, G. (1969), 'On the continuity of the generalized inverse', *SIAM Journal on Applied Mathematics* **17**(1), 33–45.

Stone, C. J. (1977), 'Consistent nonparametric regression', *Annals of Statistics* **5(4)**, 595–620.

Su, L. and Chen, Q. (2013), 'Testing homogeneity in panel data models with interactive fixed effects', *Econometric Theory* **29**(6), 1079–1135.

Szabó, Z., Sriperumbudur, B. K., Póczos, B. and Gretton, A. (2016), 'Learning theory for distribution regression', *Journal of Machine Learning Research* **17**(1), 5272–5311.

Takeuchi, K. (1976), 'The distribution of information statistics and the criterion of goodness of fit of models', *Mathematical Science* **153**, 12–18.

Tang, H.-K. and See, C.-T. (2009), 'Variance inequalities using first derivatives', *Statistics & Probability Letters* **79**(9), 1277–1281.

Tarone, R. E. (1979), 'Testing the goodness of fit of the binomial distribution', *Biometrika* **66**(3), 585–590.

Tarone, R. E. (1985), 'On heterogeneity tests based on efficient scores', *Biometrika* **72**(1), 91–95.

Tarone, R. E. and Gart, J. J. (1980), 'On the robustness of combined tests for trends in proportions', *Journal of the American Statistical Association* **75**(369), 110–116.

Theil, H. (1971), *Principles of Econometrics*, Wiley.

Thistlethwaite, D. L. and Campbell, D. T. (1960), 'Regression-discontinuity analysis: An alternative to the ex post facto experiment.', *Journal of Educational psychology* **51**(6), 309.

Tiao, G. and Box, G. (1981), 'Modeling multiple time series with applications', *Journal of the American Statistical Association* **76**, 802–816.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tjøstheim, D. (1981), 'Granger-causality in multiple time series', *Journal of Econometrics* **17(2)**, 157–176.

Tong, T., Ma, Y., Wang, Y. et al. (2013), 'Optimal variance estimation without estimating the mean function', *Bernoulli* **19**(5A), 1839–1854.

Tuvaandorj, P. and Zinde-Walsh, V. (2014), Limit theory and inference about conditional distributions, *in* 'Essays in Honor of Peter CB Phillips', Emerald Group Publishing Limited.

Ullah, A., Wan, A. K. and Chaturvedi, A., eds (2002), *Handbook of Applied Econometrics and Statistical Inference*, CRC Press.

Vaart, A. v. d. and Wellner, J. A. (2000), Preservation theorems for glivenko-cantelli and uniform glivenko-cantelli classes, *in* 'High dimensional probability II', Springer, pp. 115–133.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

Vaserstein, L. N. (1969), 'Markov processes over denumerable products of spaces, describing large systems of automata', *Problemy Peredachi Informatsii* **5**(3), 64–72.

Vergara, R. C. (2021), 'Generalized stochastic processes as linear transformations of white noise', *arXiv preprint arXiv:2101.01839* .

Von Neumann, J. (1941), 'Distribution of the ratio of the mean square successive difference to the variance', *Annals of Mathematical Statistics* **12**(4), 367–395.

Vorob'ev, L. S. and Zhurbenko, I. G. (1979), 'Bounds for $C(\alpha)$-tests and their applications', *Theory of Probability and its Applications* **24**, 253–268.

Vovk, V., Gammerman, A. and Shafer, G. (2005), *Algorithmic learning in a random world*, Springer.

Vrieze, S. I. (2012), 'Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic).', *Psychological methods* **17**(2), 228.

Wager, S. (2014), 'Asymptotic theory for random forests', *arXiv preprint arXiv:1405.0352* .

Wager, S. and Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wager, S., Hastie, T. and Efron, B. (2014), 'Confidence intervals for random forests: The jackknife and the infinitesimal jackknife', *Journal of Machine Learning Research* **15**(1), 1625–1651.

Wald, A. (1943), 'Tests of statistical hypotheses concerning several parameters when the number of observations is large', *Transactions of the American Mathematical society* **54**(3), 426–482.

Wallis, K. F. (1987), 'Time series analysis of bounded economic variables', *Journal of Time Series Analysis* **8**(1), 115–123.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2015), 'Review of functional data analysis', *arXiv preprint arXiv:1507.05135* .

Wang, K. Q. (2003), 'Asset pricing with conditioning information: A new test', *Journal of Finance* **58**(1), 161–196.

Wang, P. C. C. (1981), 'Robust asymptotic tests of statistical hypotheses involving nuisance parameters', *Annals of Statistics* **9**(5), 1096–1106.

Wang, P. C. C. (1982), 'On the computation of a robust version of the optimal $C(\alpha)$ test', *Communications in Statistics-Simulation and Computation* **11**(3), 273–284.

Wang, W., Phillips, P. C. and Su, L. (2018), 'Homogeneity pursuit in panel data models: Theory and application', *Journal of Applied Econometrics* **33**(6), 797–815.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics, Series A* **26(4)**, 359–372.

Wei, L.-J. (1992), 'The accelerated failure time model: a useful alternative to the cox regression model in survival analysis', *Statistics in medicine* **11**(14-15), 1871–1879.

White, H. (1980*a*), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

White, H. (1980*b*), 'Using least squares to approximate unknown regression functions', *International Economic Review* **21(1)**, 149–170.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50(1)**, 1–25.

Wiener, N. (1956), *The Theory of Prediction*, New York.

Wooldridge, J. M. (1990), 'A unified approach to robust, regression-based specification tests', *Econometric Theory* **6(1)**, 17–43.

Xu, K.-L. (2020), 'Inference of local regression in the presence of nuisance parameters', *Journal of Econometrics* **218**(2), 532–560.

Yakovlev, E. (2018), 'Demand for alcohol consumption in russia and its implication for mortality', *American Economic Journal: Applied Economics* **10**(1), 106–49.

Yang, Y. and Barron, A. (1999), 'Information-theoretic determination of minimax rates of convergence', *Annals of Statistics* **27(5)**, 1564–1599.

Yap, S. F. and Reinsel, G. C. (1995), 'Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model', *Journal of the American Statistical Analysis* **90**, 253–267.

Zellner, A. (2001), *Simplicity, Inference and Modelling*, Cambridge University Press, chapter 14, pp. 242–262.

Zhang, C., Kokoszka, P. and Petersen, A. (2022), 'Wasserstein autoregressive models for density time series', *Journal of Time Series Analysis* **43**(1), 30–52.

Zhang, L. (2008), 'Political economy of income distribution dynamics', *Journal of Development Economics* **87**(1), 119–139.

Zinde-Walsh, V. (2008), 'Kernel estimation when density may not exist', *Econometric Theory* **24**(3), 696–725.

Zinde-Walsh, V. (2009), 'Errors-in-variables models: a generalized functions approach', *arXiv preprint arXiv:0909.5390* .

Zinde-Walsh, V. (2011), 'Presidential address: Mathematics in economics and econometrics', *Canadian Journal of Economics/Revue canadienne d'économique* **44**(4), 1052–1068.

Zinde-Walsh, V. (2013), 'Nonparametric functionals as generalized functions', *arXiv preprint arXiv:1303.1435* .

Zinde-Walsh, V. (2014), 'Measurement error and deconvolution in spaces of generalized functions', *Econometric Theory* **30**(6), 1207–1246.

Zinde-Walsh, V. and Phillips, P. C. (2003), 'Fractional brownian motion as a differentiable generalized gaussian process', *Lecture Notes-Monograph Series* pp. 285–292.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.