Separable variance-covariance structure: estimation, testing and environmental application

Marc Ameur Manceur

Doctor of Philosophy

Department of Plant Science
Faculty of Agricultural and Environmental Sciences

McGill University

Montreal, Quebec, Canada

March 2012

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

Separable variance-covariance structure: Estimation, testing, and environmental application

Multi-dimensional, univariate or multivariate datasets arise when one or several random variables are observed on a spatio-temporal domain. A parsimonious model is often used to facilitate the estimation of variance-covariance parameters. This is the case in particular with the matrix and tensor normal distribution models, which imply a simply and doubly separable variance-covariance structure, respectively. A separable variance-covariance matrix is the Kronecker product of two, three, or more component variance-covariance matrices, each representing variability and dependencies in one dimension (e.g. 1-D space and time; multivariate, 1-D space, and time; 3-D space and time). In this thesis, the focus is on parameter estimation by maximum likelihood (ML), the likelihood ratio test (LRT) of separability, and the application to an original dataset. First, the empirical bias of the ML estimator of a simply separable variance-covariance matrix is shown to follow a non-monotonic 'peak-trough' pattern with increasing sample size, a result apparently not conform to theory. This atypical pattern is explained by decomposing the ergodic (empirical) bias into an estimation bias and a fluctuation bias minus a non-orthogonality factor. Then, an unbiased modified LRT for simple separability of a variance-covariance structure, without or with modeling of the mean, is proposed. A penalty factor improves the chi-square distribution of the LRT statistic in finite samples, which represents a simpler and more general procedure to obtain a valid LRT of separability than existing methods. Thereafter, the tensor normal distribution model is presented in detail, with a decomposition of the bias of the ML estimator of a doubly separable variance-covariance matrix and an unbiased modified LRT for double separability. Finally, an original multi-dimensional dataset of wood density in trunk sections of white spruce (*Picea glauca* (Moench) Voss), as measured from computed tomography scanning data, is used to test and accept the hypothesis of double separability on the variance-covariance structure and to assess direction, height and year effects on mean wood density using modified analysis-of-variance *F*-tests based on Box's 'epsilon'.

Structure de variance-covariance séparable: Estimation, test, et application environnementale

Les jeux de données multidimensionnels, univariés ou multivariés, se présentent lorsqu'une ou plusieurs variables aléatoires sont observées sur un domaine spatio-temporel. Un modèle parcimonieux est souvent utilisé pour faciliter l'estimation de la matrice de variance-covariance. C'est le cas en particulier des modèles de la matrice et du tenseur aléatoires normaux, qui impliquent une structure de variance-covariance simplement ou doublement séparable, respectivement. Une matrice de variance-covariance séparable est le produit de Kronecker de deux, trois, ou plus de matrices de variance-covariance, chacune représentant la variabilité et les dépendances dans une dimension (p. ex. espace 1-D et temps; plusieurs variables, espace 1-D, et temps; espace 3-D et temps). Dans cette thèse, le focus est sur l'estimation des paramètres par maximum de vraisemblance (MV), le test du rapport de vraisemblances (TRV), et l'application des modèles à un jeu de données original. Tout d'abord, il est montré que le biais empirique de l'estimateur MV d'une matrice de variance-covariance simplement séparable décroît de manière non-monotone en suivant un patron 'pic-creux' lorsque la taille d'échantillon augmente, un résultat non conforme à la théorie en apparence. Ce patron atypique est expliqué en décomposant le biais ergodique (empirique) en un biais d'estimation et un biais de fluctuation, moins un facteur de non-orthogonalité. Ensuite, un TRV modifié non-biaisé de séparabilité simple pour une structure de variance-covariance, sans ou avec modélisation de la moyenne, est proposé. Un facteur de pénalité améliore la distribution chi-deux de la statistique TRV en échantillons finis, ce qui représente une procédure plus simple et plus générale d'obtenir un TRV valide de séparabilité que les méthodes existantes. Par après, le modèle du tenseur aléatoire normal est présenté en détail, avec une décomposition du biais de l'estimateur MV d'une matrice de variance-covariance doublement séparable et un TRV modifié non-biaisé de séparabilité double. Enfin, un jeu de données multidimensionnel, fait de mesures de densité du bois obtenues à partir de données de tomodensitométrie pour des sections de troncs d'épinette blanche (*Picea glauca* (Moench) Voss), est utilisé pour tester et accepter l'hypothèse de séparabilité double de la matrice de variance-covariance et pour

évaluer les effets de la direction, de la hauteur et de l'année sur la densité moyenne du bois à l'aide de tests *F* d'analyse de variance modifiés sur base du 'epsilon' de Box.

Acknowledgements

I first want to thank my supervisor Prof. Pierre R. L. Dutilleul, a great mentor and scientist who took interest in my intellectual growth in all its aspects. His steadfast attachment to science and truth made working under his supervision a pleasure and a challenge. His kind nature and composure generated an excellent environment to prosper. His mentoring represents for me an example that I will strive to emulate in my future endeavours.

I also express my gratitude to my co-supervisor Dr. Bernard Pelletier. He helped me balance the applied and theoretical aspects of statistics, and assisted me in the initial stages of conceptualization. Dr. Katrine Stewart also deserves praise, as she ensured that I looked at the big picture and integrated the statistical tools with the environmental sciences.

Dr. Guillaume Larocque's assistance went beyond the call of duty. His assistance in programming was invaluable, as well as his ability to repair computers! Mr. Liwen Han's ability to operate the computed tomography scanner and contribution to data analysis were instrumental to complete the project. I also thank the administrative staff, notably Ms. Carolyn Bowes, for ensuring that I could focus on my research.

I acknowledge the financial assistance of the Natural Sciences and Engineering Research Council of Canada through an Alexander Graham Bell Canada Graduate Scholarship, as well as the financial assistance of McGill University through two graduate fellowships.

I also thank my family, notably my sister whose support was deeply appreciated. My friends, of course, played an important role in keeping insanity at bay and recharging my batteries.

Table of contents

List of tables

List of figures

List of important abbreviations

1-D, 2-D, 3-D	One, two, three dimensions
ANOVA	Analysis of variance
B_E, B_S, B_F, B_T	Ergodic, simulation, fluctuation, pseudo-theoretical biases
CT	Computed tomography
CTN	Computed tomography number
d.f.	Degrees of freedom
EGLS	Estimated generalized least square
i.i.d.	Identically and independently distributed
H_0, H	Null, alternative hypothesis
HU	Hounsfield unit
ML, MLE	Maximum likelihood, maximum likelihood estimator
LRT	Likelihood ratio test
OLS, GLS	Ordinary least square, generalized least square

Thesis format

This thesis is presented in manuscript-based format. This format has been approved by the Faculty of Graduate Studies, at McGill University, as described in "Guidelines for Thesis Preparation and Submission". The structure of the thesis is determined by the number of dimensions of the multi-dimensional normal distribution model studied and particular aspects of estimation, testing, and application. The thesis is composed of nine chapters and seven appendices. In Chapters 1 and 2, the objects of study are introduced, while in Chapters 7, 8, and 9 the results are discussed, summarized, and extended for possible future research. Four manuscripts are included as chapters. In Chapter 3, the empirical bias of the maximum likelihood estimator of a simply separable variance-covariance matrix is dissected under the matrix normal distribution model. A manuscript based on Chapter 3, entitled "Bias decomposition for the maximum likelihood estimator of a simply separable variance-covariance matrix", along with Appendix A in the thesis, is in revision for possible publication in *Statistics and Computing* as of March 2012. In Chapter 4, an unbiased modified likelihood ratio test for simple separability of a variance-covariance matrix is presented. A manuscript based on Chapter 4, entitled "An unbiased modified likelihood ratio test for simple separability of a variance-covariance structure" along with Appendices B and E in the thesis was submitted to *Statistics and Probability Letters* in April 2012. In Chapter 5, the tensor normal distribution model of order 3 and more is presented in detail, and the empirical bias and precision of the maximum likelihood estimator of a doubly separable variance-covariance matrix is studied by simulation. A manuscript based on Chapter 5, entitled "Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion", along with Appendices C and F of the thesis was submitted in November 2011 to *Journal of Computational and Applied Mathematics*. In addition, appendices D and E are associated with Chapter 5 in the thesis. An extended version of Appendix D "Bias decomposition for the maximum likelihood estimator of a doubly separable variance-covariance matrix" is in preparation for submission. In Chapter 6, the models and corresponding estimation and testing procedures are used in an innovative environmental application, where wood density is estimated from computed tomography scanning data. A manuscript based on Chapter 6 entitled "A multi-

dimensional statistical model for wood data analysis, with density estimated from CT scanning data as an example", along with Appendix G of the thesis is in press in the *Canadian Journal of Forest Research* (doi: 10.1139/x2012-053) as of May 2012. The manuscripts have been reformatted and slightly modified for thesis consistency.

Contributions of authors

Chapter 3 is co-authored by Mr. Manceur and Prof. Dutilleul, who wrote the MLE algorithm in the 1990s. Mr. Manceur upgraded the algorithm for use in current software, realized the simulation studies, observed the unexpected pattern, and wrote the initial manuscript. Both authors participated in writing the final manuscript, with contributions by Prof. Dutilleul for the theoretical explanation.

Chapter 4 is co-authored by Mr. Manceur and Prof. Dutilleul. Both designed the unbiased modified likelihood ratio test. Mr. Manceur realized the simulation studies, and wrote the initial manuscript. Both authors participated in writing the final manuscript.

Chapter 5 is co-authored by Mr. Manceur and Prof. Dutilleul. Mr. Manceur programmed the algorithm, performed the simulation studies, and wrote the initial manuscript. Both authors participated in writing the final manuscript.

Chapter 6 is co-authored by Mr. Manceur, Dr. Beaulieu, Mr. Han, and Prof. Dutilleul. Dr. Beaulieu furnished the prepared wood samples and contributed to the development of the approach. Mr. Han operated the computed tomography scanner and contributed to data analysis. Mr. Manceur wrote the initial manuscript. All authors participated in writing the final manuscript.

Chapter 1. Introduction

1.1 General introduction

Large, complex, and multi-dimensional datasets are becoming more prevalent in the natural sciences, especially since new measurement technologies generate large datasets stored in ever expanding computer memory (Fey et al. 2008). For example, the measurement of water pH at several depths in a lake sampled several times in a season provides spatio-temporal data (two dimensions, 2-D), while that of wood density for a number of directions, heights and growth rings result in doubly spatial, temporal data (three dimensions, 3-D). As can be seen from these examples, the random variable of interest is indexed from a spatio-temporal domain in such univariate multi-dimensional datasets.

The time axis is particular since it ranges from past to present and future, in one direction. Space does not possess this property and there are spatial vertical and spatial horizontal dimensions, in which variation may be very different (Dutilleul 2011). Once a spatio-temporal point is reached, whether geographically or within an organism, it is cost efficient to collect data for several variables instead of only one, resulting in a multivariate dataset indexed on a spatio-temporal domain (Njue 2001). The variables could be discrete or continuous quantitative, although in this Ph.D. thesis the variables are assumed to be continuous quantitative and, in general, normally distributed.

Datapoints in a univariate multi-dimensional dataset are likely to be inter-dependent. Your weight today is related to your weight yesterday, so the observation of a person's weight over time results in a vector of dependent data, also known as repeated measures (Crowder and Hand 1990). The inter-dependency of datapoints extends to the multivariate context, as when several variables related to milk quality are observed over time or when fiber length and microfibril angle are measured in addition to wood density in an example with trees.

It is recommended that the inter-dependencies of datapoints be taken into account in statistical analyses. Otherwise, statistical tests may not be valid (a test is valid if the rejection of the null hypothesis, while it is true, is equal to the nominal Type I error risk or significance level fixed by the experimenter), and estimators may not be efficient (their variance may be inflated or deflated, Dutilleul 2011). It is possible to mathematically

remove inter-dependencies from the data (i.e. pre-whitening), and then use classical methods of statistical inference that postulate the independence of datapoints. This approach may require some type of stationarity (Dutilleul 2011), which is discussed in Chapter 2 of this thesis. Alternatively, it may be possible to develop a variant of the statistical method that adequately incorporates the inter-dependencies of the data. However, without modeling, a large number of variance-covariance parameters would need to be estimated, since there is one covariance for each pair of datapoints in addition to two variances. Thus, the variance-covariance matrix, with variances on the diagonal and covariances off the diagonal, is often modeled. Instead of estimating each and every element of the matrix, the parameters of the variance-covariance structure or model are estimated.

Consequently, multi-dimensional normal distribution models are presented, studied, and applied in this thesis, because they allow that inter-dependencies in the data be taken into account in a parsimonious way, as they imply that the variance-covariance matrix is separable: there is one variance-covariance matrix per dimension. This is the case with the matrix normal distribution model, when for example the variance-covariance matrix for a univariate 2-D spatio-temporal dataset is the Kronecker product (see Chapter 2) of two component variance-covariance matrices, one purely spatial and the other purely temporal. Multi-dimensional normal distribution models extend the well-known scalar and vector normal distributions, making them easier to apprehend. It is hoped that they will assist scientists in extracting knowledge from increasingly complex datasets, and encourage the development of parsimonious statistical models.

1.2 General hypotheses

The structure of the thesis is based on four hypotheses:

1. The bias of the maximum (ML) estimator of a simply separable variance-covariance matrix, implied by the matrix normal distribution model, decreases monotonically with increasing sample size.
2. The likelihood ratio test (LRT) presently available in the literature for testing the hypothesis of simple separability for a variance-covariance structure is biased, in that the rejection rate is not equal to the nominal Type I error risk when the null hypothesis of separability is true.

3. The "MLE-3D" algorithm will allow the estimation of the variance-covariance parameters of the tensor normal distribution model of order 3, with small samples.
4. The variance-covariance structure for a data tensor (multi-dimensional dataset) made of wood density measures on white spruce trees in two directions, at two heights and in two growth rings is doubly separable.

1.3 General and specific objectives

The general objective of this Ph.D. thesis is to study separable variance-covariance structures (which are characteristic of multi-dimensional normal distribution models), in order to make a number of contributions to estimation and testing aspects, and to apply the models and procedures developed to an environmental dataset.

The specific objectives of the thesis are:

1. To dissect the bias of the ML estimator of the simply separable variance-covariance matrix, implied by the matrix normal distribution model.
2. To develop the unbiased modified LRT for simple separability of a variance-covariance structure.
3. To present the tensor normal distribution model or more that is characterized by a doubly or more separable variance-covariance matrix.
4. To dissect the bias of the ML estimator of a doubly separable variance-covariance matrix.
5. To present the unbiased modified LRT for double separability of a variance-covariance structure.
6. To estimate wood density from computed tomography (CT) scanning data and demonstrate that multi-dimensional normal distribution models can be applied in the framework of analysis of variance with doubly and triply repeated measures, to assess differences in mean wood density estimates in relation to direction, height and year.

Chapter 2. Literature review

2.1 Statistics

2.1.1 An introduction to multi-dimensional data analysis

There exists a link between the 'size' of data collected in a designed survey or experiment and the corresponding mathematical object on which the statistical analyses are based. For example, wood density measured for one ring, at one height and in one direction inside a tree trunk can be statistically represented as a scalar-valued random variable and assumed to be normally distributed with mean m and variance u (Table 2.1). Measuring wood density for n rings in one direction at one height in the same tree trunk provides one vector of n observed values or one observation from a random n-variate vector $\mathbf{x}$. Assuming the random vector is normally distributed with mean vector $\mathbf{m}$ and variance-covariance matrix $\mathbf{\Sigma}$, the data vectors can be analyzed in a multivariate analysis of variance or a repeated measures analysis of variance (Crowder and Hand 1990).

The vector of n wood density measures, with n re-numbered n_1, can be collected in n_2 directions (e.g. North and South), resulting in one $n_1 \times n_2$ matrix of values or one observation from the random $n_1 \times n_2$ matrix $\mathbf{X}$, which may be assumed to be normally distributed (Table 2.1). Furthermore, a data matrix of wood densities (e.g. a matrix of n_1 rows or rings by n_2 columns or directions) can be collected at n_3 heights (e.g. breast and live crown heights), resulting in a data tensor of order 3 and size $n_1 \times n_2 \times n_3$ or one observation from a random tensor of same order and size $\mathbb{X}$, assumed to be normally distributed. Thus, 3-D data are represented with a 3-D mathematical object, the tensor.

A tensor of order 3 has three dimensions, with the row, the column and the edge representing one dimension each (tree ring, direction, height); see Figure 6.1. Dimensions are sometimes referred to as repeated factors (Dutilleul and Pinel-Alloul 1996), ways (Kroonenberg 2008), or modes (Kolda and Bader 2009), and a size or number of levels is associated with each dimension. There are also several tensorial notations, such as those used by McCullagh (1987), Kolda and Balder (2009), and Dutilleul (1990).

A tree trunk represents a spatio-temporal sampling domain, with direction and height corresponding to the horizontal and vertical spaces and the growth ring representing time (year). The vertical spatial dimension exists in a tree trunk because of primary growth from the apical meristem, while the horizontal spatial dimension exists

thanks to secondary growth from the cambium (see Section 2.2). The vertical elongation of the buds is influenced by the climatic conditions during the previous season (Duff and Nolan 1953), and the secondary growth process implies that wood density at one ring is influenced by wood density at the preceding ring (Fritts 1976). Thus, the elements of the random matrices or tensors involved in such surveys or experiments are often inter-dependent or correlated. When analyzing the data and testing for differences in the mean value depending on the dimension or the position in the dimension, it is not justified to assume that the data were generated independently from their spatio-temporal neighbors. More generally, if the variance also changes with the dimension or the position in the dimension, it is required to estimate the variance-covariance matrix (which embodies both the variability and the inter-dependencies) in order to perform valid tests of significance for the mean. The next two sub-sections present a certain number of elements about possible multi-dimensional models and variance-covariance structures of interest.

2.1.2 Multi-dimensional normal distribution models

In what follows, the multi-dimensional (i.e. matrix and tensor) random objects are considered normally distributed. The scalar normal distribution is commonly used, in part because it is entirely characterized by two parameters, namely its mean and variance. This property holds in the cases of the vector, matrix, and tensor normal distributions, except the variance then becomes a variance-covariance matrix, which is square and symmetrical by definition; in addition to a variance (diagonal) for each element in the random vector, matrix, or tensor, there is a covariance (off-diagonal) for each pair of different elements.

By symmetry (i.e. entries below the diagonal are equal to entries above the diagonal), the number of distinct entries in an $n_1 n_2 \times n_1 n_2$ variance-covariance matrix $\mathbf{\Sigma}$ is at most equal to $n_1 n_2 (n_1 n_2 + 1)/2$. For the moment generating function of the vector normal distribution to exist, the variance-covariance matrix only needs to be positive semi-definite (Muirhead 1982), but for the probability density function to exist and hence, for likelihood-based inference (e.g. solving systems of likelihood equations), $\mathbf{\Sigma}$ must be positive definite, that is, $\mathbf{y}' \mathbf{S}\, \mathbf{y} > 0$ for any non-null column vector $\mathbf{y}$ of size $n_1 n_2$.

In an example where wood density is measured on two rings in the North and South directions or the activity of two parts of the brain is recorded at two times, the expected values of the corresponding random variables, or population means, can be presented in a 2 × 2 matrix:

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}.$$

This mean matrix can be re-arranged as a mean vector by using the vec operator, which transforms an $n_1 \times n_2$ matrix into a column vector of size $n_1 n_2$ (Schott 1997):

$$\text{vec}(\mathbf{M}) = \begin{bmatrix} m_{11} \\ m_{21} \\ m_{12} \\ m_{22} \end{bmatrix}.$$

If the variance-covariance structure of the vectorized random matrix $\mathbf{X}$ is simply separable, the $n_1 n_2 \times n_1 n_2$ variance-covariance matrix is constructed as the Kronecker product $\otimes$ (Schott 1997) of two component variance-covariance matrices that each embodies the variability and inter-dependencies in one dimension:

$$\mathbf{\Sigma} = \mathbf{U}_2 \otimes \mathbf{U}_1 = \begin{bmatrix} u_{2_{11}} & u_{2_{12}} \\ u_{2_{21}} & u_{2_{22}} \end{bmatrix} \otimes \begin{bmatrix} u_{1_{11}} & u_{1_{12}} \\ u_{1_{21}} & u_{1_{22}} \end{bmatrix} = \begin{bmatrix} u_{2_{11}}u_{1_{11}} & u_{2_{11}}u_{1_{12}} & u_{2_{12}}u_{1_{11}} & u_{2_{12}}u_{1_{12}} \\ u_{2_{11}}u_{1_{21}} & u_{2_{11}}u_{1_{22}} & u_{2_{12}}u_{1_{21}} & u_{2_{12}}u_{1_{22}} \\ u_{2_{21}}u_{1_{11}} & u_{2_{21}}u_{1_{12}} & u_{2_{22}}u_{1_{11}} & u_{2_{22}}u_{1_{12}} \\ u_{2_{21}}u_{1_{21}} & u_{2_{21}}u_{1_{22}} & u_{2_{22}}u_{1_{21}} & u_{2_{22}}u_{1_{22}} \end{bmatrix},$$

where the $n_2 \times n_2$ matrix $\mathbf{U}_2$ (variance-covariance matrix for the columns) has at most $n_2(n_2+1)/2$ distinct parameters and the $n_1 \times n_1$ matrix $\mathbf{U}_1$ (variance-covariance matrix for the rows) has at most $n_1(n_1+1)/2$ distinct parameters. These two matrices allow the construction of the $n_1 n_2 \times n_1 n_2$ matrix $\mathbf{\Sigma}$ from fewer parameters, but with constraints such as $\sigma_{11} = u_{2_{11}} u_{1_{11}}$. While the matrix normal distribution model is associated with simple separability, the tensor normal distribution model implies a doubly separable variance-covariance structure: $\mathbf{\Sigma}$ is then the Kronecker product of three component matrices.

In general, the variance-covariance matrix of a vector normal distribution model is not separable, and is unstructured if no specific model, separable or another, is

postulated; in other words, the variance-covariance matrix is positive semi-definite or positive definite:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix},$$

where σ_{11} is the variance of wood density in ring 1 and direction 1 and σ_{12} is the covariance between wood density in ring 1, direction 1 and that in ring 2, direction 1 if $\mathbf{U}_1$ stands for the rings and $\mathbf{U}_2$ for the directions in the tree example.

Separability of the variance-covariance structure assists in the estimation, by reducing the number of parameters (for $n_1 = n_2 = 2$, there are 6 parameters whereas in the unstructured case, there are 10; Figure 2.1) while ensuring positive definiteness. Prior to discussing estimation of the model parameters, a number of separable and non-separable models of variance-covariance structure are the objects of the next section.

2.1.3 Models of variance-covariance structure

Below, three assumptions commonly made about the variance-covariance matrix in a multivariate analysis with data spatially or temporally referenced, or both, are discussed. These assumptions are: stationarity, separability, and circularity. Several other assumptions and models exist (Wolfinger 1996), but these three are of high relevance to this Ph.D. thesis project.

As suggested by the name, stationarity implies some form of 'stability' either for some of the population parameters, those of first and second order in particular, or for the complete distribution of the random variable of interest, when this is not normal. Since the multi-dimensional distribution models studied in this thesis will be essentially normal, it is sufficient to discuss the former type of 'stability' here. Weak stationarity implies that the mean and variance of the random variable of interest is stable over the space-time domain, while the covariance is a function of the lag (i.e. an interval length in time, but a vector between two sampling locations in space). The general form of separability for a variance-covariance structure does not imply stationarity, even at second order (for variances and covariances). See Dutilleul (2011) for details on the concept of stationarity and its various forms.

In the general case of simple separability $\mathbf{\Sigma} = \mathbf{U}_2 \otimes \mathbf{U}_1$, the component matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are unstructured, and so are the component matrices $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ in the general case of double separability $\mathbf{\Sigma} = \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$. However, the component variance-covariance matrices can be modeled. For example, the structure of $\mathbf{U}_1$ can be first-order autoregressive [AR(1)] or compound symmetric (CS), while $\mathbf{U}_2$ is left unstructured (Roy 2006, Roy and Khattree 2005). Such liberty in the choice of the model in each dimension provides flexibility in modeling (Huang et al. 2007), but may involve the assumption of stationarity (e.g. AR(1)). It must be noted that the component variance-covariance matrices, whether modeled or not, are defined up to a positive multiplicative constant: $\mathbf{U}_2 \otimes \mathbf{U}_1 = a\mathbf{U}_2 \otimes (1/a)\mathbf{U}_1 = \mathbf{\Sigma}$ and $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 = a\mathbf{U}_3 \otimes b\mathbf{U}_2 \otimes (1/ab)\mathbf{U}_1 = \mathbf{\Sigma}$, with a, $b > 0$; only the Kronecker product is uniquely defined (Dutilleul 1999, Lee et al. 2010), and is the final object of estimation (Table 2.1).

In a separable variance-covariance structure, only limited interactions between dimensions are possible. In particular, in the spatio-temporal framework, temporal correlations are then the same at each point in space, and so are the spatial correlations at each point in time. Thus, to obtain the best linear prediction of $X_{i_1 i_2}$ (an element of data matrix $\mathbf{X}$) from $X_{i_1 i_2'}$ (same time, different spatial location) and $X_{i_1' i_2}$ (same spatial location, different time), it is sufficient to use $X_{i_1' i_2}$ (O'Hagan 2002). In a predictive context like this, information about spatial neighbors at the same time is superfluous since the pattern of correlation at each spatial location is repeated over time. Using only past data collected at the same spatial location is sufficient.

The restrictive conditions imposed on covariances by a separable variance-covariance structure imply that it does not apply to all datasets, and a test of significance will help identify when it applies and when it does not (Lu and Zimmerman 2005, Mitchell et al. 2006; see also Subsection 2.1.5 and Chapters 4 and 6 here). When a variance-covariance matrix is separable, there are several advantages to this:

(1) because of the smaller number of parameters, ML estimation requires a (much) smaller sample size than for an unstructured variance-covariance matrix – see next subsection;
(2) in each of the dimensions, changes in the mean are tested with a specific variance-covariance matrix in an extension of the repeated measures ANOVA (Dutilleul and Pinel-Alloul 1996) – see Chapter 6;
(3) separability offers computational advantages, notably for matrix inversion (Schott 1997);
(4) the parsimony of a separable variance-covariance structure may provide increased power in statistical tests for the mean (Wolfinger 1996);
(5) modeling the covariance can be of crucial scientific interest.

Circularity is the most general necessary and sufficient condition on variances and covariances for valid *F*-tests for fixed effects in the ANOVA method. An $n \times n$ variance-covariance matrix $\mathbf{\Sigma}$ (separable or not) is circular if $\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C} = \lambda\, \mathbf{I}_{n-1}$, with ^{T} the transpose operator, λ a positive scalar, and $\mathbf{C}$ an $n \times (n - 1)$ matrix of orthonormal contrasts (i.e. particular linear transformations of raw data; the coefficients defining contrasts as linear combinations of raw data add up to zero; these linear combinations are orthonormal when they are orthogonal for the scalar product and their norm as a vector is 1). In other words, when $\mathbf{\Sigma}$ is circular, the variance-covariance matrix of orthonormal contrasts is the identity matrix multiplied by a scalar (Crowder and hand 1990). This condition or variance-covariance structure is also known as covariance of type H, the Huynh-Feldt condition, or sphericity. Deviation from circularity is measured with Box's 'epsilon' (Box 1954a, 1954b; see Subsection 2.1.6 here).

The three variance-covariance structures presented and briefly discussed above are not mutually exclusive, but they do not necessarily imply one another either. For example, (1) a separable variance-covariance matrix need not be stationary; (2) a diagonal variance-covariance matrix is stationary and separable if circular, because the elements of the diagonal matrix are then all equal; (3) if a variance-covariance matrix is separable but not diagonal, then nothing can be said about stationarity. The question of estimation of the model parameters is addressed in the next subsection.

2.1.4 Maximum likelihood estimation

In this Ph.D. thesis, the context to which the estimation of multi-dimensional normal distribution model parameters belongs is the one in which a number K of identical and independently distributed (i.i.d.) observations from a random matrix or tensor are available for statistical inference. Although the elements of each of the data matrices or tensors are inter-dependent in general, the K matrices or tensors are considered as replicates because they come from individuals randomly sampled from one population (e.g. trees of about the same height, located not too close to each other within the same region).

The context above excludes fields of application such as environmental monitoring, when data are collected at stations over periods of time or a partial realization of the underlying spatio-temporal stochastic process or "random function" in geostatistics is collected. There is one replicate (i.e. one random matrix) and elements are likely to be correlated. Several replicates are not possible in the absence of "parallel universes" (Dutilleul 2011). In the context where $K = 1$, a form of stationarity at order 2 is usually assumed in the data analysis, sometimes in addition to separability. In this thesis, it is assumed that at least two replicates are always available for statistical inference, through repeated measures.

The parameters of normal distribution models can be estimated by ML. The ML estimator of the mean in the scalar, vector, matrix, or tensor case is the sample mean in the absence of specific modeling; see Chapter 4 for examples of response surface modeling. Under the vector normal distribution model, the ML estimator of $\mathbf{\Sigma}_{UN}$ is:

$$\hat{\mathbf{\Sigma}}_{UN} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_k - \hat{\mathbf{m}})(\mathbf{x}_k - \hat{\mathbf{m}})^T , \qquad (2.1)$$

where $\hat{\mathbf{m}}$ denotes the ML estimator of the mean vector (i.e. $\bar{\mathbf{x}}$) and the index UN specifies that $\mathbf{\Sigma}$ is unstructured (i.e. has no specific structure). For the vectorized matrix and tensor normal distributions, the equation is the same, except $\mathbf{x}_k$ and $\hat{\mathbf{m}}$ are replaced by a vectorized data matrix or tensor and the corresponding vectorized sample mean matrix or tensor. The estimator (2.1) is available only if $K \geq n + 1$, $K \geq n_1 n_2 + 1$ and $K \geq n_1 n_2 n_3 + 1$, respectively; in the univariate 2-D spatio-temporal case, it means that there

are at least as many replicates as space-time sampling points plus one. This estimator is not the unbiased sample variance-covariance matrix, as the divisor is K instead of $K - 1$.

Actually, ML estimators of variance-covariance parameters are generally biased and their bias is inversely proportional to sample size. If θ denotes the variance-covariance parameter of interest and $\hat{\theta}$ is one of its estimators, the bias of $\hat{\theta}$ is defined as $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$, where E denotes the expectation operator. In Chapter 3, the bias of the ML estimator of a simply separable variance-covariance matrix is studied in detail.

Several algorithms exist for the estimation of separable variance-covariance matrices under the matrix normal distribution model. The most commonly used is the iterative MLE-2D algorithm (Dutilleul 1999), later renamed "flip-flop" (Lu and Zimmerman 2005), which is presented in Chapter 3. This algorithm is expected to produce an output provided $K \geq \max(n_1 / n_2, n_2 / n_1) + 1$, and this condition will be extended for the MLE-3D algorithm and the ML estimation of a doubly separable variance-covariance matrix in Chapter 5.

Werner et al. (2008) introduced a modification of Dutilleul's original MLE-2D algorithm, by stopping the algorithm after one iteration. Estimation with a specific structure on either $\mathbf{U}_1$ or $\mathbf{U}_2$ (e.g. AR(1), CS) has been developed (Roy 2006, Roy and Khattree 2005). Assuming an AR(1) structure for one of the two variance-covariance matrices, Naik and Rao (2001) presented an iterative ML type of estimation algorithm that can be used when there are missing data. In the analysis of magneto-encephalographic data, Huizenga et al. (2002) modeled the temporal variance-covariance matrix as Toeplitz and the spatial one as exponential. That is, they assumed both stationarity at second order and separability to reduce the number of variance-covariance parameters to estimate.

An iterative ML type of estimation algorithm using Newton-Raphson optimization was implemented in SAS PROC MIXED, for example. It was studied by Lu and Zimmerman (2004), who concluded that the MLE-2D algorithm was much faster than the ML Newton-Raphson algorithm. The latter was even found to not converge at all when $\mathbf{U}_2$ was modeled as CS or AR(1), whereas the former converged (Roy and Khattree 2005). This is in accordance with the fact that the ML Newton-Raphson

algorithm in SAS PROC MIXED is of general purpose, whereas the MLE-2D algorithm is specifically designed for ML estimation of simply separable variance-covariance matrices. In the analysis of repeated measures in time and space, a Fisher's scoring algorithm was proposed for iterative residual maximum likelihood estimation (Verbyla and Cullis 1992); it can be used when there are missing data. An iterative least-squares approach was also presented in the brain science literature (Bijma et al. 2005, Huizenga et al. 2002). From now on in this thesis, when referring to ML estimation of a separable variance-covariance matrix, we will refer by default to Dutilleul's MLE-2D algorithm used with two unstructured component variance-covariance matrices, their Kronecker product being separable by construction, and no missing data.

The so-called "covariance matching" provides an exact or approximate Kronecker product of two component variance-covariance matrices $\tilde{\mathbf{U}}_1$ ($n_1 \times n_1$) and $\tilde{\mathbf{U}}_2$ ($n_2 \times n_2$) for a given $n_1 n_2 \times n_1 n_2$ variance-covariance matrix $\boldsymbol{\Sigma}$. In simple terms, a rearrangement function (R) is defined, and $R(\boldsymbol{\Sigma})$ is an $n_2^2 \times n_1^2$ matrix, i.e., it is not square in general. The two component variance-covariance matrices $\tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{U}}_2$ follow from a singular value decomposition of $R(\boldsymbol{\Sigma})$. This method allows the calculation of κ (to be used in Chapter 4), which measures the separability approximation error (or lack of separability) of $\boldsymbol{\Sigma}$: $\kappa = \| \boldsymbol{\Sigma} - \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1 \| / \| \boldsymbol{\Sigma} \|$, where $\| . \|$ denotes the Euclidean norm (also called "Frobenius norm") (Genton 2007). When $\boldsymbol{\Sigma}$ is perfectly separable, $\kappa = 0$, and the upper bound for values of κ is $1/\sqrt{1 - rank(R(\boldsymbol{\Sigma}))}$; κ is approximately 1 for very large matrices. Another separability index was developed by Boik (1991). Covariance matching is discussed in Van Loan (2000), Genton (2007), and Werner et al. (2007). An iterative covariance matching procedure is presented in detail by Bijma et al. (2005). The covariance matching method can accommodate additional structures; for example, if $\boldsymbol{\Sigma}$ is block Toeplitz, then $\tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{U}}_2$ are Toeplitz.

Werner et al. (2007) [see Figure 1 therein] compared several estimation procedures, and showed by simulation that the MLE-2D algorithm provided matrix estimates with the smallest normalized root-mean square error with small sample sizes (i.e. as small as from 2 to 10 when $n_1 = n_2 = 4$). For doubly separable variance-covariance

matrices, there are a limited number of estimation algorithms available, and these are reviewed in Chapter 5 before the extension of the MLE-3D algorithm is presented; see also Appendix F for the MLE-4D algorithm. Whether in 2-D or 3-D (or 4-D), since model parameters are estimated, it is very important to assess that the assumptions of multi-dimensional normal distribution models hold for the dataset.

2.1.5 Likelihood ratio tests for simple separability of a variance-covariance matrix

The assumptions of the matrix normal distribution model are multivariate normality and simple separability of the variance-covariance structure. For ML estimation purposes, an i.i.d. random sample of sufficient size is required, implying that the mean matrix and the two component variance-covariance matrices are the same for all the distributions of matrices of the random sample. Testing of the separability assumption can be based on the discrepancy between the Kronecker product of ML estimates of $\mathbf{U}_2$ and $\mathbf{U}_1$ under the null hypothesis and the ML estimate of $\mathbf{\Sigma}_{UN}$ under the alternative. The rationale is that if $\hat{\mathbf{\Sigma}}_{UN}$ "resembles" $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ enough, then separability should not be rejected for that dataset.

The sample size required for testing the assumption of separability (i.e. $K \geq n_1 n_2 + 1$) is superior to that required for ML estimation under the matrix normal distribution model (i.e. $K \geq \max(n_1 / n_2, n_2 / n_1) + 1$), since $\mathbf{\Sigma}_{UN}$ then needs to be estimated by ML. Thus, it is possible that the sample size allows ML estimation assuming separability, but does not permit testing the assumptions. In that case, separability can only be assumed without testing (Dutilleul and Pinel-Alloul 1996).

Several likelihood ratio tests (LRTs) for simple separability of a variance-covariance matrix have been developed (Table 2.2); by definition, any LRT statistic is the ratio of suprema of the likelihood function under the null and alternative hypotheses (Muirhead 1982). So far, all the LRTs proposed for assessing separability are biased (Muirhead 1982), an issue distinct from the bias for an estimator and occurring when the probability of rejecting the null hypothesis is not minimum under the null hypothesis. A solution is proposed here in Chapter 4, with an unbiased modified LRT for simple separability. Roy and Leiva (2008) had developed an LRT for double separability of a variance-covariance structure when at least one of the component matrices is structured.

In Appendix E of this thesis, an unbiased LRT test for double separability when all the component variance-covariance matrices are unstructured is presented.

2.1.6 Modified ANOVA *F*-tests based on Box's 'epsilon'

Traditionally, the number of variance-covariance parameters contained in $\boldsymbol{\Sigma}$ tends to be reduced to one (i.e. the variance) by design of the experiment (e.g. homogeneous material and field, randomization of treatment assignment), so that a small number of replicates are sufficient to obtain a positive-definite variance-covariance matrix estimate. Why is it that the number of variance-covariance parameters cannot be reduced by experimental design for multi-dimensional datasets when these are spatio-temporal, univariate or multivariate? Because repeated measures are not collected in a 'classical' field lay-out: sampling times (years, months, days, hours) are usually the same for all units, and so are sampling locations (depths, heights, directions). In other words, as ANOVA qualitative factors, time and space are usually crossed with other factors whether these are treatment factors or more generally, classification factors. Furthermore, due to limited resources, there is often a lack of replicates to satisfy the condition $K \geq n_1 n_2 + 1$ or $K \geq n_1 n_2 n_3 + 1$, and estimate the unstructured $n_1 n_2 \times n_1 n_2$ or $n_1 n_2 n_3 \times n_1 n_2 n_3$ variance-covariance matrix. Also, it is not recommended to ignore any heterogeneity of variance (heteroscedasticity) and inter-dependencies (correlations) in the data that are susceptible to make invalid the unmodified ANOVA *F*-tests for the so-called "within-subject effects" (which are related to repeated measures factors). Accordingly, the variance-covariance matrix often needs to be modeled. The separable variance-covariance structure with unstructured component variance-covariance matrices, as implied by the multi-dimensional normal distribution model in its general form, was selected for this Ph.D. thesis because it is parsimonious and does not require stationarity. While properties and limitations will be further discussed, contributions regarding estimation and testing aspects will be provided.

Doubly or triply repeated measures ANOVA can be used under the matrix or tensor normal distribution model, to test for differences among elements of the mean matrix or tensor. Each dimension can then be viewed as a repeated measures factor with a certain number of levels (Dutilleul and Pinel-Alloul 1996). Usually, with spatio-temporal repeated measures, means should be compared by taking into account discrepancies from the circularity assumption (see Subsection 2.1.3) due to heteroscedasticity and

autocorrelation. Such discrepancies are evaluated with Box's 'epsilon' (Box 1954a, 1954b), and under separability, it is possible to calculate one Box's 'epsilon' for each component variance-covariance matrix (for each repeated measures factor) as well as for the Kronecker product (Dutilleul and Pinel-Alloul 1996, Dutilleul 2011). Box's 'epsilon' is used as a multiplicative factor to the numbers of degrees of freedom (d.f.) of the numerator and denominator of the *F*-ratio tests for repeated measures effects. Under circularity (Subsection 2.1.3), Box's 'epsilon' is equal to 1 and there is no need to adjust (reduce) the two numbers of d.f. and modify the ANOVA *F*-test.

Actually, in applications, Box's 'epsilon' needs to be estimated. Following Greenhouse and Geisser (1959), Huynh and Feldt (1976) designed an estimator for situations when the number of individuals (on which repeated measures are made) is small and the true but theoretical value of Box's 'epsilon' is close to 1. It was corrected later to work properly with several groups (Lecoutre 1991). The corrected form of Huyhn-Feldt version of Box's 'epsilon' estimator was shown to be more robust than the uncorrected version, while being equally valid and powerful (Quintana and Maxwell 1994).

An interesting relationship between separability and circularity was explored by Boik (1991), who proved and showed by simulation that modified ANOVA *F*-tests based on Box's 'epsilon' will be valid in the doubly repeated measures ANOVA if the covariance has a separable structure. With increasing departure from separability and holding Box's 'epsilon' value fixed, Boik's simulations suggest that the test size (probability of rejecting the null hypothesis while it is true) decreases.

The statistical models, methods and procedures described above or their latest improvement or upgrade as the result of my Ph.D. thesis project will be applied to 3-D repeated measures of wood density estimated from CT scan data in Chapter 6. Therefore, the next section gives more details about tree growth and the tree species of interest.

2.2 Trees

2.2.1 Tree growth

A tree ring can be seen as the physical embodiment of time: "On cross-sections of trees, structures and patterns can be recognized with regularity and the question of time is apparent" (Wimmer 2002). Also, a variation in the properties of wood is expected within

a ring, at various heights in the tree and at different orientations in the trunk (Schweingruber 2007). If wood cores are taken at different heights or orientations within a tree for measurement of one property or characteristic, then spatio-temporal data arise, and the tree appears as a spatio-temporal sampling domain.

Ring width and its links with temperature and precipitation are the classical subjects of investigation of dendrochronology or dendroclimatology. Chemical and anatomical variables, such as calcium concentration, fibre angle, ratio of late to early xylem, density of resin duct, and wood density, may be linked to wind events, pollution, and management intervention (Wimmer 2002), and may be more responsive to environmental changes than ring width (Smith 2008). Such data are multivariate and potentially spatio-temporal.

The autocorrelation of variables and the cross-correlation between variables are well-known in tree-ring series analysis (Wimmer 2002), and the dendrochronological literature highlights the need for properly evaluating error values: "Proper statistical treatment of the data, for example, accounting for non-independence of variables (often a large issue in structural research), will allow more correct interpretation of the actual patterns" (Gartner et al. 2002). In fact, if cross-correlation is strong, most of the information can be extracted by using only those variables that are easy to measure (Wimmer 2002), thus reducing the workload to obtain measures.

Density is a particularly important variable in wood science, since it determines the properties and values of pulp and solid wood products, and is highly heritable genetically (Pliura 2006). The accurate modeling of wood density variation, both between and within trees, should lead to an improved use of raw materials in the forest industry (Lindstrom 2002).

Measurability is a main challenge and "there is a need for new and rapid techniques to assess the response variables within the wood" (Gartner et al. 2002). It is expensive and time consuming to measure several variables such as wood density and microfibril angles within each ring, so that datasets tend to be of small size and limited to a few species. SilviScan (Evans et al. 1995, Keunecke et al. 2009) allows the measurement of a variety of anatomical and mechanical properties of wood, and its ability to measure these properties in more than one dimension is improving (Defo et al.

2009). However, the data can only be collected on thin, small stripes of wood. Downes et al. (2009) present other measurement technologies, while discussing the importance of tree-ring research for commercial forestry. X-ray CT scanning is a technique that allows the measurement of density in 3-D on larger pieces of wood. Prior to giving details about X-ray CT scanning, the tree species of interest is described below.

2.2.2 White spruce

The pieces of wood CT scanned in this Ph.D. thesis project are trunk sections from white spruce trees (*Picea glauca* (Moench) Voss). In 2009, it was estimated that there were 111 686 hectares of spruce plantations in Canada, about a third of the total area under plantation (National Forestry Database 2009). White spruce is an object of genetic studies (Beaulieu et al. 2011), breeding programs (Zhang et al. 2004), and dendrochronological studies (Szeicz and Macdonald 1996).

Wood density is known to vary in the trunk of a white spruce tree. Generally, from pith to bark, density decreases, stabilizes, and then increases (Corriveau et al. 1990). From the stump to the top of the tree, wood density increases (Wang and Micko 1984, Singh 1984). These patterns may vary with the diameter of the tree trunk (Singh 1984) and the method used to measure wood density (Zhang et al. 2004). Thus, variation in wood density estimated from the CT scan data collected for the trunk sections of white spruce trees could be expected.

2.3 Technologies

2.3.1 Non-destructive techniques for graphical and quantitative analyses of wood properties

Non-destructive methods that allow imaging of the internal structure of wood can be classified according to energy, frequency, or length of the emitted wave that interacts with the wood material. Based on wavelength, X-rays have the highest energy transmitted (i.e. the smallest wavelength, from 1.0*E*-12 to 1.0*E*-9 m; Bucur 2003). Theoretically, sub-microscopic, microscopic, and macroscopic structures can all be observed, depending on the resolution of the machine. For example, X-ray computed microtomography allows the observation of vessels in small wood cubes (Steppe et al. 2004), while CT scanners of the medical type can be used with larger pieces (Fromm et al. 2001).

Increasing the wavelength and decreasing the energy results in visible light (1.0*E*-6 m), and then in infrared (1.0*E*-5 to 1.0*E*-6 m), microwave (1.0*E*-3 to 1.0*E*-1 m), ultrasound (1.0*E*-2 to 1.0*E*3 m), and radio waves (1.0*E*3 to 1.0*E*5 m). Radio waves can be used to visualize forests and full trees, while infrared and micro waves are used to observe macroscopic structures. Visible light, as we know it, can be used to observe microscopic and macroscopic features. Ultrasound can be used to visualize entire trees, as well as macroscopic and microscopic features (Bucur 2003). Magnetic resonance imaging is popular in medical applications because it is safer than X-rays for human uses, but it can only be used with material that contains a large amount of water, such as the soft tissue of an animal, and unlike CT scanning, magnetic resonance imaging is affected by soil iron content (Tollner et al. 1994). Gamma rays can also be used (Macedo et al. 2002), but the source of waves is then radioactive.

Basic elements of the theory behind CT scanning are presented below, because wood material has been CT scanned with a scanner of medical type in the environmental application of this Ph.D. thesis project (Chapter 6). Assume a monochromatic source of X-rays (one energy level) emits N photons which traverse an incremental thickness of material ΔY, so that only $N - \Delta N$ photons are counted by the receptor on the other side of the material. If $N(y) = N_0 e^{-\mu y}$, then the number of photons at position y within the material is a function of the initial number of photons entering the material (N_0), and of μ which represents the photon loss rate per distance unit, the absorption coefficient (Kak and Slaney 2001). If the coefficient value is very large (e.g. lead), most photons will have "disappeared" (Kak and Slaney 2001). At the energies of a medical CT scanner, the Compton effect (i.e. deflection of a photon as it interacts with a free electron in the material) is considered the most important mechanism behind the "disappearance" of photons (Mull 1984). The attenuation coefficient for a heterogeneous material can be calculated as $\mu = w_1\mu_1 + ... + w_i\mu_i + ... + w_f\mu_f$, where μ_i is the absorption coefficient for material i and w_i is the percent by volume of material i. The absorption coefficient of a material is a function of the density and of the atomic composition (Lindgren 1991a).

Measurements obtained by CT scanning are subject to noise, which arises from fluctuations in the numbers of photons counted by receptors; this quantum noise is

inherent to the physical process. In general, the larger the number of photons counted by receptors, the smaller the noise (Hilts and Duzenli 2004). One possible approach to decrease the noise, if necessary, is to increase the energy of the X-rays, Energy = $P\ C\ T$, where the energy is the load on the X-ray tube in kilo-Joules, P is the energy potential in kilo-Volts, C is the electrical current in milli-Amperes, and T is the scan time per rotation in seconds (Hilts and Duzenli 2004). Other approaches have to do with the size of the sample and the length of the image reconstruction interval for 2-D CT images.

2.3.2 CT scanning of wood material

CT scanning technology has been used with wood material in previous studies (Bucur 2003, Lindgren 1991b). However, the material is heterogeneous, which influences the measurement procedure. Ignoring water and micro-nutrients, wood is composed of cellulose, hemicellulose, and lignin, for which the absorption coefficients are: $\mu_c = 0.2634$, $\mu_h = 0.2655$, and $\mu_l = 0.2608$, respectively (Lindgren 1991b). These values are equal to the second decimal, so the total absorption coefficient does not vary much for mixtures of the three components (Lindgren 1991b). Furthermore, it is generally acknowledged that the density of the cell wall in dry wood is stable at 1500 $kg.m^{-3}$. Theoretically, the absorption coefficient for a very small parallelepipedic rectangular volume (i.e. a voxel) of wood thus varies nearly perfectly with fluctuations in wood density due to the morphology and distribution of anatomical elements (e.g. wide lumen). This is empirically verified by a nearly perfect ($R^2 = 0.99$) linear relation between the absorption coefficient (converted to a CT number; see Chapter 6), and oven-dried estimated wood density (Lindgren 1991b, Mull 1994).

Moving away from idealized objects of dry and pure wood, pores in wood can be filled with water, and there may be some heavy elements in the cell wall. In the presence of water in the wood, the absorption coefficient of a voxel is larger, as water replaces air in the pores and absorbs photons (Lindgren 1991b). With wet wood, a measure of the percentage of water in the wood is required, which is possible but increases the technical difficulty (Macedo et al. 2002). Concerning heavy elements in the cell wall, it was calculated that for elements with an atomic number greater than 20, a change of 150 $\mu g.g^{-1}$ of concentration changes the absorption coefficient by 1% with X-ray densitometry

(similar to classical radiography), although the change in physical density is minimal (Kouris et al. 1981).

At the end of this literature review, it appears that the statistical models, methods and procedures to be developed or validated here could prove very useful to wood scientists, and there is great potential for a useful and innovative application of X-ray CT scanning.

Table 2.1: Normal distribution models under study and objects of estimation: Multi-dimensional normal distribution models are extensions of the well-known scalar and vector models.

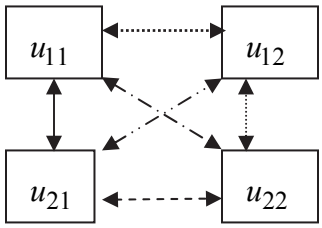
Example	Random object	Model	Variance-covariance parameters, object of estimation
Wood density for one ring in one direction at one height	Random scalar (0-D) variable	$X \sim N(m,u)$	$\mathbf{\Sigma} = u$
Wood density for two rings in one direction at one height	Random vector (1-D) of size n	$\mathbf{x} \sim N_n(\mathbf{m}, \mathbf{U})$	$\mathbf{\Sigma} = \mathbf{U}$
Wood density for two rings in two directions at one height	Random matrix (2-D) of size $n_1 \times n_2$	$\mathbf{X} \sim N_{n_1,n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2)$ $\Leftrightarrow \text{vec}(\mathbf{X}) \sim N_{n_1 n_2}(\text{vec}(\mathbf{M}), \mathbf{U}_2 \otimes \mathbf{U}_1)$	$\mathbf{\Sigma} = \mathbf{U}_2 \otimes \mathbf{U}_1$
Wood density for two rings in two directions at two heights	Random tensor of order three (3-D) and size $n_1 \times n_2 \times n_3$	$\mathbb{X} \sim N_{n_1,n_2,n_3}(\mathbb{M}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ $\Leftrightarrow$ $\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)$	$\mathbf{\Sigma} = \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$

Table 2.2: Likelihood-ratio tests[a] for simple separability of a variance-covariance matrix.

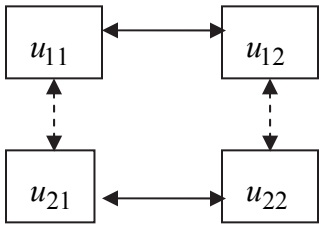
Reference	Modeling of $\mathbf{U}_1$ and $\mathbf{U}_2$ [b]
Naik and Rao (2001)	UN, CS
Svantesson and Wallace (2003)	UN, UN
Lu and Zimmerman (2005)	UN, UN
Mitchell et al. (2005)	UN, UN UN, AR(1)
Roy and Khattree (2005)	UN, CS
Roy (2006)	UN, AR(1)
Mitchell et al. (2006)	UN, UN
Roy and Leiva (2008)	UN, CS UN, AR(1)
Srivastava et al. (2008)	UN, CS UN, UN

[a] All the likelihood-ratio tests listed above are biased.
[b] UN: unstructured; CS: compound symmetric; AR(1): first-order autoregressive.

(a) Unstructured variance-covariance matrix



(b) Separable variance-covariance matrix



(c) Identity variance-covariance matrix

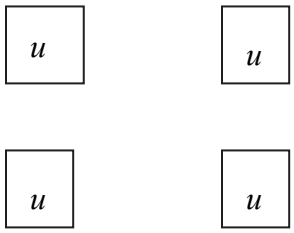


Figure 2.1: Design with two repeated measures factors (two dimensions), with two levels per repeated measures factor ($n_1 = n_2 = 2$). (a) In the unstructured variance-covariance matrix, there are four distinct variances (u_{11}, u_{12}, u_{21}, u_{22}) and six distinct covariances (see the six double arrows of various types). (b) In the separable model, there are four distinct variances and only two distinct covariance parameters (see the two pairs of double arrows). (c) Assuming independence (absence of correlation) and stationarity, there is one variance, and no covariance (no double arrow). The separable model for a variance-covariance structure is parsimonious and allows inter-dependencies among datapoints.

Preface to Chapter 3

In Chapter 2, the matrix normal distribution model, which implies a simply separable variance-covariance structure, was presented. The estimation of variance and covariance parameters in particular by ML was discussed, with emphasis on properties of the ML estimator such as the bias. Thus, the object of study in Chapter 3 is the bias of the Kronecker product of the two estimated component variance-covariance matrices obtained as solutions of the iterative MLE-2D algorithm. It is found that the bias follows an unexpected 'peak-trough' pattern, explained by decomposing the bias – an original conceptual contribution. The simulation results are discussed in relation to statistical theory, links are made with decompositions performed in geostatistics, and matrix algebra details are provided in Appendix A. As of May 2012, a manuscript (co-authored by Mr. Manceur and Prof. Dutilleul) based on this chapter was in revision for possible publication in *Statistics and Computing*. Prof. Dutilleul wrote the MLE algorithm in the 1990s. Mr. Manceur upgraded the algorithm for use in current software, realized the simulation studies, observed the unexpected pattern, and wrote the initial manuscript.

Chapter 3. Bias decomposition for the maximum likelihood estimator of a separable variance-covariance matrix

3.1 Abstract

In the matrix normal distribution model $\mathbf{X} \sim N_{n_1,n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2)$, the variance-covariance structure is separable: the variance-covariance matrix of vec($\mathbf{X}$) is the Kronecker product of the variance-covariance matrices, $\mathbf{U}_2$ for the n_2 columns and $\mathbf{U}_1$ for the n_1 rows. From an i.i.d. random sample $\mathbf{X}_1,...,\mathbf{X}_K$ with $K \geq \max\left(n_1/n_2, n_2/n_1\right) + 1$, it is possible to estimate $\mathbf{U}_2 \otimes \mathbf{U}_1$ by ML, by estimating $\mathbf{U}_1$ and $\mathbf{U}_2$ iteratively with the 'flip-flop' MLE algorithm. In this chapter, we report on the empirical bias of the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$, and show that it may decrease monotonically or not with K depending on the values of n_1 and n_2, before vanishing as K becomes very large. To explain this apparent paradox to the theory of properties of ML estimators, we decompose the empirical bias, renamed as "ergodic bias", into an "estimation bias" and a "fluctuation bias" minus a non-orthogonality factor. Our results also provide valuable information to users of the MLE algorithm for the matrix normal distribution, as to which combinations of n_1, n_2 and K may correspond to higher or lower bias values than expected in theory.

3.2 Introduction

Bias is a measure classically used in the statistical sciences to position an estimator relative to the population parameter of interest. In the one-parameter case, it is theoretically defined as the difference between the expected value of the sample statistic used as an estimator and the value of the parameter that is the object of estimation (Kotz et al. 1982, p. 230). In the case of several parameters contained in a vector or a matrix, the sample statistic is simply or doubly multivariate, and the extended definition of bias, for it to remain a scalar quantity, is based on the norm of the vector or the matrix of expected values of the differences between the estimator's entries and the corresponding parameter values. In this case, the bias values are restricted to be non-negative, as they are norms or distances. When different from zero, bias values may depend on various

factors, such as the method of estimation, the strength of autocorrelation and heteroscedasticity in the sample data, and the sample size.

In this chapter, we focus on the bias of the ML estimator of the variance-covariance matrix under the matrix normal distribution model $\mathbf{X} \sim N_{n_1,n_2}(\mathbf{M},\mathbf{U}_1,\mathbf{U}_2)$ (Dawid 1981, de Waal 1988). The variance-covariance structure is then said to be separable, factorized, or Kronecker structured, as the variance-covariance matrix of the vectorized matrix $\mathbf{X}$, $\text{vec}(\mathbf{X})$, is the Kronecker product ($\otimes$) of the two variance-covariance matrices: $\mathbf{U}_2$, for the n_2 columns (e.g. temporal repeated measures), and $\mathbf{U}_1$, for the n_1 rows (e.g. spatial repeated measures), in this order:

$$\text{var}(\text{vec}(\mathbf{X})) = \mathbf{U}_2 \otimes \mathbf{U}_1. \tag{3.1}$$

Under the vector normal distribution model $\mathbf{x} \sim N_n(\mathbf{m},\mathbf{\Sigma})$, the ML estimator of the variance-covariance matrix $\mathbf{\Sigma}$, obtained from an i.i.d. random sample $\mathbf{x}_1,...,\mathbf{x}_K$ with $K > n$, is biased, and the bias decreases monotonically with K, and vanishes asymptotically (Kendall and Stuart 1967). The system of likelihood equations then has one exact, analytical solution for $\mathbf{\Sigma}$, i.e., the sample variance-covariance matrix multiplied by $(K-1)/K$. This is not the case for $\mathbf{U}_1$ and $\mathbf{U}_2$ under the matrix normal distribution model (Dutilleul 1999), which may have implications for the study of the bias of the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$, built as the Kronecker product of estimators $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_1$ obtained iteratively as final solutions of the system of likelihood equations for $\mathbf{U}_1$ and $\mathbf{U}_2$.

Such study of bias is of importance and interest at this moment, for the following reasons. The iterative MLE algorithm developed by Dutilleul (1999) to solve the system of likelihood equations for $\mathbf{U}_1$ and $\mathbf{U}_2$ is increasingly used, after it was shown that: (i) it converges with datasets for which other estimation methods fail (Roy and Khattree 2005); (ii) it presents the smallest mean square error in small sample sizes (Werner et al. 2008); and (iii) with it, there is no need to parameterize $\mathbf{U}_1$ and $\mathbf{U}_2$ in a particular way prior to proceeding with estimation (Lee et al. 2010). The MLE algorithm for the matrix normal distribution has been found useful in several fields, including biochemistry (Theobald and Wuttke 2008), image analysis (Dryden et al. 2008), and engineering

(Werner et al. 2008), and was mentioned in recent statistical publications on doubly multivariate data analysis (Allen and Tibshirani 2010, Wang and West 2009). If the bias of the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ did not follow the statistical theory as expected, the users of the iterative MLE algorithm should better know it, and there might be potential for discoveries and original explanations about the bias behavior, and an expansion of the bias concept eventually. The simulation results presented in Dutilleul (1999) include standardized empirical bias values for the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ that were relatively constant in small samples sizes and decreased monotonically to approach zero for larger sample sizes. Computational power is now much greater, but the results and findings presented hereafter are not limited to only simulations and numerical aspects.

Therefore, fundamental elements about the MLE algorithm for the matrix normal distribution are presented in Section 3.3. In Section 3.4, the empirical bias (renamed "ergodic bias") of the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ is decomposed into an estimation bias plus a fluctuation bias, minus a non-orthogonality factor. The three biases are studied by simulation in Section 3.5. Results are discussed in Section 3.6, where the atypical part in the behavior of the empirical bias is explained, a link is made with practice concerning uncertainty assessment in geostatistics, and the discussion is completed with the pseudo-theoretical bias. Conclusions are drawn in Section 3.7, while matrix algebra and computational details are given in Appendix A.

3.3 Maximum likelihood estimation for the matrix normal distribution

Under the matrix normal distribution model $\mathbf{X} \sim N_{n_1, n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2)$, the variance-covariance structure is uniquely defined through $\mathbf{U}_2 \otimes \mathbf{U}_1$ (see (3.1)), instead of $\mathbf{U}_1$ and $\mathbf{U}_2$ separately; $\mathbf{U}_2 \otimes \mathbf{U}_1 = 1/a\mathbf{U}_2 \otimes a\mathbf{U}_1$ with $a > 0$. In other words, $\mathbf{U}_1$ and $\mathbf{U}_2$ are defined up to a positive multiplicative constant, and $\mathbf{U}_2 \otimes \mathbf{U}_1$ with $\mathbf{U}_1$ and $\mathbf{U}_2$ both assumed to be positive definite is the object of estimation here, as far as variances and covariances are concerned. In the absence of analytical solutions for $\mathbf{U}_1$ and $\mathbf{U}_2$ in the system of likelihood equations, by ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ we mean the estimator obtained as the Kronecker product of the final solutions $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_1$ of the MLE algorithm used to solve the system of equations in $\mathbf{U}_1$ and $\mathbf{U}_2$ iteratively. The MLE algorithm for the matrix normal distribution (called 'flip-flop' by Lu and Zimmerman

2005) was presented in details by Dutilleul (1999). Its most important characteristics are summarized below.

Using a random sample $\mathbf{X}_1,...,\mathbf{X}_K$ i.i.d. $N_{n_1,n_2}(\mathbf{M},\mathbf{U}_1,\mathbf{U}_2)$ with $\mathbf{U}_1 > 0$ and $\mathbf{U}_2 > 0$, equalling to zero the first derivatives of the log-likelihood function with respect to $\mathbf{U}_1$ and $\mathbf{U}_2$ leads to

$$\begin{cases} \hat{\mathbf{U}}_1 = \sum_{k=1}^{K} (\mathbf{X}_k - \bar{\mathbf{X}})\,(K n_2 \hat{\mathbf{U}}_2)^{-1} (\mathbf{X}_k - \bar{\mathbf{X}})^T \\ \hat{\mathbf{U}}_2 = \sum_{k=1}^{K} (\mathbf{X}_k - \bar{\mathbf{X}})^T (K n_1 \hat{\mathbf{U}}_1)^{-1} (\mathbf{X}_k - \bar{\mathbf{X}}) \end{cases} \tag{3.2}$$

where $\bar{\mathbf{X}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{X}_k$. The existence of estimates $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ at successive iterations of the algorithm requires that the sample size K be sufficiently large, but not necessarily greater than $n_1 n_2$ if $n_1 > 1$ or $n_2 > 1$:

$$K \geq \max\left(\frac{n_1}{n_2}, \frac{n_2}{n_1}\right) + 1. \tag{3.3}$$

The MLE algorithm used to solve (3.2) can be initialized with $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$ or any other justified $n_2 \times n_2$ positive definite matrix. To stop it, the criterion can be on the norm of the difference between the Kronecker products $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ of two successive pairs of estimates or between the estimates taken separately; different positive infinitesimal quantities ϵ_1 and ϵ_2 for the two norms of differences can be used when justified in the latter case.

3.4 Bias decomposition

The bias which is introduced below and decomposed thereafter is empirical in that it cannot be evaluated in finite samples without simulations, due to the random nature of the matrix defining the variance-covariance matrix estimates as two sums of K quadratic forms in matrix-valued observations (see $(K n_2 \hat{\mathbf{U}}_2)^{-1}$ for $\hat{\mathbf{U}}_1$ and $(K n_1 \hat{\mathbf{U}}_1)^{-1}$ for $\hat{\mathbf{U}}_2$ in (3.2)):

$$B_E = \| E(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_2 \otimes \mathbf{U}_1) \|, \tag{3.4}$$

where $\|.\|$ denotes the Euclidean norm. We call B_E the "ergodic bias" because the difference in (3.4) is between the variance-covariance matrix statistic $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ computed from an i.i.d. random sample, and the variance-covariance matrix parameter value $\mathbf{U}_2 \otimes \mathbf{U}_1$, used as a reference point. Although the variance-covariance parameters could be those of a spatio-temporal random function or stochastic process, this type of ergodicity does not correspond to the classical definition that applies in time series analysis, when the statistic value tends to the parameter value as the time series length becomes very large (Blum 1981). Instead, it can be related to the ergodic variance used in geostatistics to assess the uncertainty in the estimation of semivariance parameters (see Section 3.6). Computational details about the evaluation of (3.4), including the replacement of $(Kn_1\hat{\mathbf{U}}_1)^{-1}$ by $(Kn_1(\hat{\mathbf{U}}_1)_V)^{-1}$ and of $(Kn_2\hat{\mathbf{U}}_2)^{-1}$ by $(Kn_2(\hat{\mathbf{U}}_2)_V)^{-1}$ where $(\hat{\mathbf{U}}_1)_V$ and $(\hat{\mathbf{U}}_2)_V$ represent averages over V simulation runs, are given in the Appendix A.

Because $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$, which are involved in the construction of the ML estimator $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$, are the numerical solutions of an iterative algorithm, it is possible that there be 'fluctuations' in the accessibility of the 'true' parameter values from the data in applications, especially when the sample size K is equal to the minimum required (3.3) plus a few units. Accordingly, the average $(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V$ of a large number V (1.0*E*6) of ML estimated variance-covariance matrices would provide a better point of reference and could be used to study a new type of bias. Along those lines of thought, the difference $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_2 \otimes \mathbf{U}_1$ can be decomposed as follows:

$$\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_2 \otimes \mathbf{U}_1 = \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V + (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V - \mathbf{U}_2 \otimes \mathbf{U}_1, \quad (3.5)$$

and we call $B_S = \| E(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V) \|$, the "estimation bias", and $B_F = \| (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V - \mathbf{U}_2 \otimes \mathbf{U}_1 \|$, the "fluctuation bias", with $E((\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V) \approx (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V$.

The bias B_S measures the distance between the expected value of the estimator $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$, as obtained empirically from a theoretical formula (see above and the Appendix A), and a different reference point than the classical 'true' matrix parameter

value, provided by the average of V estimated variance-covariance matrices obtained from the final solutions of the MLE algorithm. The bias B_F for its part measures how the average of V estimated variance-covariance matrices obtained from the final solutions of the MLE algorithm 'fluctuates' around the 'true' matrix parameter value. A large B_F combined with a small B_S corresponds to a combination of K, n_1 and n_2, for which the 'true' matrix parameter value is hardly accessible from sample data while the estimation procedure is successful in providing the average reachable variance-covariance matrix estimate.

Unlike the ANOVA decomposition, the sum of the two biases B_S and B_F is not equal to the ergodic (empirical) bias B_E. The difference $(B_S + B_F) - B_E$, which is positive, provides a "non-orthogonality component" denoted δ, so that the ergodic bias is finally decomposed as follows:

$$B_E = B_S + B_F - \delta . \tag{3.6}$$

To be complete and for comparison purposes, we also studied a "pseudo-theoretical bias", B_T, which differs from B_E by the presence of non-random matrices in the quadratic forms defining $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ in (3.2), that is, by the replacement of $(K n_1 \hat{\mathbf{U}}_1)^{-1}$ by $(K n_1 \mathbf{U}_1)^{-1}$ and of $(K n_2 \hat{\mathbf{U}}_2)^{-1}$ by $(K n_2 \mathbf{U}_2)^{-1}$. No simulation is required to calculate this pseudo-theoretical bias (see the Appendix A).

3.5 Evaluation of the ergodic bias and its components

The three biases B_E, B_S, B_F and the non-orthogonality component δ were evaluated for various combinations of n_1, n_2 and K (see below), using V=1.0E6 estimates $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ provided by as many simulation runs and permissible (non-singular) solutions $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ of the MLE algorithm at convergence ($\epsilon_1 = \epsilon_2$ = 1.0E-6, $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$, 1.0$E$7 iterations maximum). A customized program was written in Matlab 2010a (The MathWorks Inc., Natick, MA). The mean matrix $\mathbf{M}$ was set at zero and $\mathbf{U}_1 = \mathbf{I}_{n_1}$, $\mathbf{U}_2 = \mathbf{I}_{n_2}$, which did not affect our results except by decreasing the time required for the MLE algorithm to converge. The procedure followed to simulate i.i.d. random samples from a matrix normal distribution is the same as in Dutilleul (1999), except that the

Ziggurat Algorithm (Marsaglia and Tsang 2000) was used here to generate pseudo-random numbers from a univariate normal distribution. The smallest sample size considered for the matrix normal distribution, denoted K_{min} below, is equal to $\mathrm{INT}[\max(n_1/n_2, n_2/n_1)]+1$ if $\max(n_1/n_2, n_2/n_1)$ is an integer number and $\mathrm{INT}[\max(n_1/n_2, n_2/n_1)]+2$ otherwise; $\mathrm{INT}[.]$ represents the integer part of the number in brackets. The combinations of number of rows, number of columns and sample size considered for the evaluation of biases are: $n_1 = 2, \ldots, 12$ and $n_2 = 2, 3$ by steps of 1; and $K = K_{min}, \ldots, 15$ by steps of 1, in addition to 20 as well as 1000 in particular cases. Finally, bias values were standardized by the corresponding $\| \mathbf{U}_2 \otimes \mathbf{U}_1 \|$.

Values of K_{min} are plotted as a function of n_1 and n_2 in Figure 3.1. When $n_1 = n_2$, K_{min} is equal to two, and then increases by 'steps' with n_1. The width of each 'step' is equal to the value of n_2.

In Figure 3.2 ($n_2 = 2$) and Figure 3.3 ($n_2 = 3$), the pattern of the ergodic (empirical) bias with increasing sample size K increases from K_{min} for given values of n_1 and n_2 is very surprising [see panel (a)] because B_E does not always decrease monotonically with K! Actually, non-monotonic patterns are observed for the three biases (ergodic, estimation, fluctuation), and present one or two characteristic aspects depending on the bias: only a 'peak' shortly after K_{min} for the ergodic bias; a 'peak' shortly after K_{min}, followed by a 'trough' prior to decreasing slowly and approaching zero around $K = 20$ for the estimation and fluctuation biases [see panels (b) and (c)]. The non-monotonic patterns are sufficiently regular, stable and repeated for not being artefacts and for being related to the values of n_1, n_2 and K because they are observed for particular combinations of the three numbers.

When $n_2 = 2$ (Figure 3.2), a 'peak' over $K_{min} + 1$ and $K_{min} + 2$ is present in the ergodic bias for the even values of n_1 considered here (2, 4, 6, 8, 10, 12), but not for the odd values (3, 5, 7, 9, 11); the ergodic bias decreases monotonically with K for these. The estimation bias presents a 'peak' over $K_{min} + 1$ and $K_{min} + 2$, followed by a 'trough', and this pattern is restricted again to the even values of n_1. For $n_1 = 2, 4, 6$, the

fluctuation bias presents a 'peak' over K_{min} + 1 and K_{min} + 2, followed by a marked 'trough'; for n_1 = 8, 10, 12, there is no 'peak' but a higher value, followed by a marked 'trough'; for all odd values of n_1 (3, 5, 7, 9, 11), there is only a slight 'trough' shortly after K_{min}, the largest value of the fluctuation bias being observed at n_1 = 3. In all cases, the non-orthogonality component takes very large values at the beginning (small *K*), but quickly converges to a value close to zero as K increases.

When n_2 = 3 (Figure 3.3), similar general patterns are observed, with the following differences. The ergodic bias continues to present no 'trough', but 'peaks' are observed at K_{min} + 1 for n_1 = 3, 6, 9, 12 (multiples of three). Accordingly, there are 'peaks' in the estimation bias at K_{min} + 1 for n_1 = 3, 6, 9, 12, and there is a clear 'trough' for smaller values of n_1. The pattern of the fluctuation bias with n_2 = 3 is similar to what it is with n_2 = 2 (Figure 3.2), except that 'peaks' occur at K_{min} + 1 for n_1 = 3, 6 instead of n_1 = 2, 4, 6 and the largest value is observed at n_1 = 5, $n_2 = 3$ and $K = K_{min} = 3$. Again, the non-orthogonality component takes very large values at the beginning (small *K*), but quickly converges to a value close to zero as K increases in all cases.

In view of Figures 3.1(a)-(b), 3.2(a) and 3.3(a), the first 'peak' in the ergodic bias happens for the combination of n_1 and n_2 for which K_{min} is the smallest, that is, n_1 = n_2 = 2 and n_1 = n_2 = 3. The three biases become very small for values of *K* greater than 20 and values of n_1 and n_2 in the range considered here. For example, with n_1 = 5, n_2 = 2 and K = 1000, the values of B_E, B_S, B_F, and δ are 7.90*E*-4, 2.29*E*-4, 1.01*E*-3 and 4.51*E*-4, respectively.

The representation of the three non-theoretical biases in the same graph, as in Figure 3.4(a) for n_1 = n_2 = 2 and in Figure 3.4(b) for n_1 = n_2 = 3, provides insight into how they vary and differ as K increases from K_{min} to 20; for $n_1 = n_2 = 4$ and 5, results are very similar to these for $n_1 = n_2 = 3$, and are not reported. Bias values are larger for *K* = 3 or 4, and the estimation bias provides the largest values, followed by the fluctuation bias and the ergodic bias in this order, which is very interesting and is discussed below.

From $K = 7$ on, the three biases and the non-orthogonality components take much smaller values and tend to coincide in the graphs.

3.6 Discussion

3.6.1 Unusual behavior for empirical bias

The ML estimator of variance-covariance matrix studied here for its bias is the Kronecker product of two estimated variance-covariance matrices, which are dependent on each other via the random central matrices defining them as quadratic forms in normal matrices (3.2). This makes the ML estimator in question particular. Furthermore, the two variance-covariance matrix estimates $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ are the outcome of an iterative algorithm in applications, in the absence of analytical solutions for the system of likelihood equations. Accordingly, simulations were used to evaluate empirically the expected values of centered moments of order 4 involved in the bias analysis. A large number of simulations (1.0*E*6) were necessary to obtain stable values of the ergodic (empirical) bias and its components. A minimum number of 1.0*E*4 simulations are required to observe the 'peak'-and-'trough' pattern (Figures 3.2 and 3.3). Thus, Dutilleul (1999) empirically showed a 'plateau' instead of a 'peak' at small sample sizes, because of limited computing power at the time.

The ergodic (empirical) bias does not always follow a decreasing monotonic pattern as the sample size K increases, a novel finding to the best of our knowledge. Such behavior made of a 'peak' shortly after K_{min} (generally $K_{min} + 1$) is very unusual, but may remind the "anomaly" reported by McCullagh (2008) for the Fisher information of the autocorrelation parameter estimated from k parallel series of n points in space or time, with a decrease of the Fisher information after $k = n/2$. Our decomposition of the ergodic bias into an estimation bias and a fluctuation bias minus a non-orthogonality component was aimed to provide an explanation to the unusual behavior of the ergodic bias in relation to sample size. (Note: A similar decomposition of the ergodic variance in our case would be practically intractable because it would involve eight-order moments for a matrix instead of scalar estimator.) As a matter of fact, a 'peak' may happen in all three biases at the same value of K (see Figure 3.4 for $n_1 = n_2 = 2, 3$ and Figures 3.2 and 3.3 for other values of n_1), meaning that both the expected value of $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ evaluated empirically and the average of ML estimated variance-covariance matrices $(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V$

are then far from the true parameter $\mathbf{U}_2 \otimes \mathbf{U}_1$ (high ergodic and fluctuation biases) and from each other (high estimation bias), even if the convergence criterion in the estimation algorithm is met without exception. 'Troughs' appear with different 'depths' for different combinations of n_1, n_2 and K in the estimation and fluctuation biases (and the non-orthogonality component), and this superimposition of 'troughs' with different 'depths' at different places in B_S, B_F and δ results in the absence of 'troughs' in the ergodic bias $B_E = B_S + B_F - \delta$.

The results discussed above do not preclude the use of the "flip-flop" MLE algorithm in practice. In fact, the results of our extensive simulation study show that this algorithm can successfully estimate a separable variance-covariance matrix from an i.i.d. random sample with a size as small as K_{min}, which is in accordance with former simulation results (Dutilleul 1999, Werner et al. 2008). However, our results demonstrate that depending on the values of n_1 and n_2, sample sizes greater than $K_{min} + 1$ or $K_{min} + 2$ may be required to avoid excessively large bias in the ML estimator and provide a behavior of the ergodic bias in accordance with theory, i.e., decreasing with increasing sample size.

3.6.2 Links with geostatistics

Our decomposition of the bias for the ML estimators of variance and covariance parameters for the matrix normal distribution has links with the analysis of uncertainty in the framework of the linear model of coregionalization, which is based on variances of differences between variogram estimator and semivariance parameter and between the two estimators, called "experimental variogram" and "regional variogram" (Larocque et al. 2007). Two important differences are that in Larocque et al. (2007), the domain of investigation is spatial so that the intensity of sampling (discretely within a region versus continuously within the region or over the entire sampling domain) is of concern rather than aspects of numerical analysis in the estimation, and the decomposition concerns variances of differences instead of biases.

If (i) $\mathbf{\Gamma}(h)$ represents the 'true' parameters in the linear model of coregionalization, that is, the matrix of theoretical direct and cross variograms at distance h for the multivariate random function under study; (ii) $\mathbf{\Gamma}^R(h)$, the regional variograms, are the approximation provided by a partial realization of the multivariate random

function if this was thoroughly sampled over a bounded, continuous domain or region R ; and (iii) $\boldsymbol{\Gamma}^*(h)$, the experimental variograms, are the estimates obtained from a partial realization reduced to a finite number of sampling locations within region R, then the following decomposition holds exactly, without a non-orthogonality component:

$$\text{var}(\boldsymbol{\Gamma}^*(h) - \boldsymbol{\Gamma}(h)) = \text{var}(\boldsymbol{\Gamma}^*(h) - \boldsymbol{\Gamma}^R(h)) + \text{var}(\boldsymbol{\Gamma}^R(h) - \boldsymbol{\Gamma}(h)). \qquad (3.7)$$

In words, the equation above reads as follows: the ergodic variance of variograms, $\text{var}(\boldsymbol{\Gamma}^*(h) - \boldsymbol{\Gamma}(h))$, is equal to their estimation variance, $\text{var}(\boldsymbol{\Gamma}^*(h) - \boldsymbol{\Gamma}^R(h))$, plus their fluctuation variance, $\text{var}(\boldsymbol{\Gamma}^R(h) - \boldsymbol{\Gamma}(h))$. Since $\boldsymbol{\Gamma}(h)$ is constant, the first and third variances reduce to $\text{var}(\boldsymbol{\Gamma}^*(h))$ and $\text{var}(\boldsymbol{\Gamma}^R(h))$, respectively. The only remaining variance of a difference is $\text{var}(\boldsymbol{\Gamma}^*(h) - \boldsymbol{\Gamma}^R(h))$, which measures the dispersion of the discrepancies between experimental and regional variograms resulting from the partial sampling of region R. For a given ergodic variance, a lower estimation variance corresponds to a higher fluctuation variance and vice versa. A higher fluctuation variance means that the variance-covariance structure over the region, represented by $\boldsymbol{\Gamma}^R(h)$, is highly uncertain. This is the case when sampling is conducted within a region too small for the range of spatial autocorrelation, i.e., the range-to-extent ratio is large.

Equation (3.7) has a structure similar to our decomposition of the ergodic (empirical) bias of $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$, equation (3.6). The regional variogram in (3.7) plays the role of $(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V$ in equation (3.6), and represents what is accessible in terms of parameter value, while the experimental variogram plays the role of $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ and corresponds to the inferential result for one dataset or sample of data.

3.6.3 Pseudo-theoretical bias

The pseudo-theoretical bias defined at the end of Section 3.4, which differs from the ergodic (empirical) bias by the use of non-random central matrices in the quadratic forms defining $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$, does not show a behavior with 'peaks' for some combinations of values of n_1 and n_2 (Figure 3.5). In that sense, the former bias might seem 'better' than the latter one, because it is in accordance with theory, i.e., monotonic decrease with increasing sample size. Actually, the pseudo-theoretical bias is not a reliable or representative measure of the differences between the expected value of $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ and

the parameter value of $\mathbf{U}_2 \otimes \mathbf{U}_1$. The replacement of $(Kn_1\hat{\mathbf{U}}_1)^{-1}$ by $(Kn_1\mathbf{U}_1)^{-1}$ and of $(Kn_2\hat{\mathbf{U}}_2)^{-1}$ by $(Kn_2\mathbf{U}_2)^{-1}$ in the central part of quadratic forms in the calculation of the pseudo-theoretical bias may avoid the circularity of the problem ($\mathbf{U}_1$ and $\mathbf{U}_2$ are needed to evaluate the bias of $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$), but such a replacement assumes, at the least, the absence of bias for $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ as estimators of $\mathbf{U}_1$ and $\mathbf{U}_2$. The values of the fluctuation bias for their Kronecker product (Figures 3.2–3.4) indicate that this is not a reasonable assumption.

3.7 Conclusion

To the best of our knowledge, the decomposition of the ergodic (empirical) bias into an estimation bias and a fluctuation bias minus a non-orthogonality factor is new. We proposed it here to explain an unusual pattern in the behavior of the ergodic (empirical) bias of the ML estimator of a separable variance-covariance matrix for the matrix normal distribution. Such a decomposition might be found helpful to explain other 'anomalies', for example, in likelihood-based inference with parallel series of points in space or time (McCullagh 2008). Besides these fundamental aspects, our study also provides insight into practice and applications. In no way is the use of the "flip-flop" MLE algorithm questioned here. Simply but importantly, its users, in increasing number, are now better informed of the properties of its output. A customized MATLAB (The MathWorks Inc.) code is available upon request for those who would like to study the ergodic, estimation and fluctuation biases for combinations of n_1, n_2 and K other than the ones considered here.

3.8 Acknowledgements

The financial assistance of the Natural Sciences and Engineering Research Council of Canada (NSERC), through an Alexander Graham Bell Canada Graduate Scholarship to Mr. Manceur and a Discovery Grant (Mathematics and Statistics) to Prof. Dutilleul, is gratefully acknowledged.

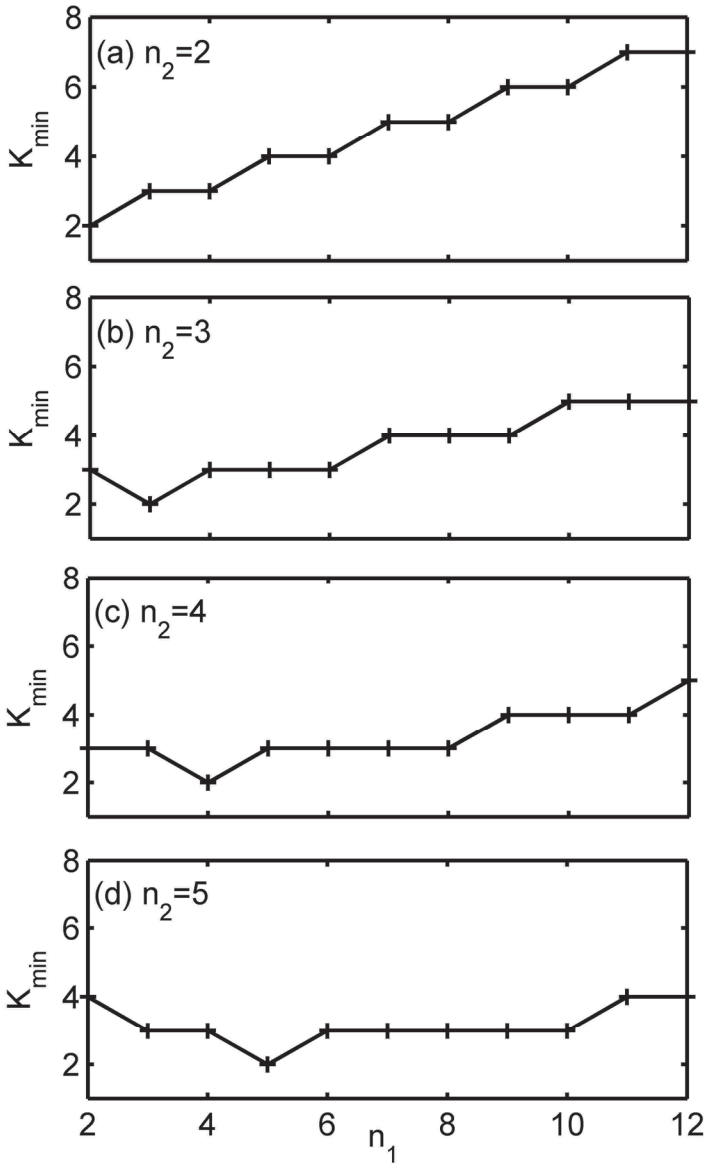


Figure 3.1: Minimum sample size K_{min} (INT[max(n_1/n_2, n_2/n_1)] + 1 if max(n_1/n_2, n_2/n_1) is an integer number and INT[max(n_1/n_2, n_2/n_1)] + 2 otherwise), required for the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ ($\mathbf{U}_1$: $n_1 \times n_1$; $\mathbf{U}_2$: $n_2 \times n_2$) to exist. The smallest K_{min} is at $n_1 = n_2$. For a given value of n_2, K_{min} increases in a 'stairway' with increasing n_1 after $n_1 = n_2$, with a different 'width' depending on the value of n_2.

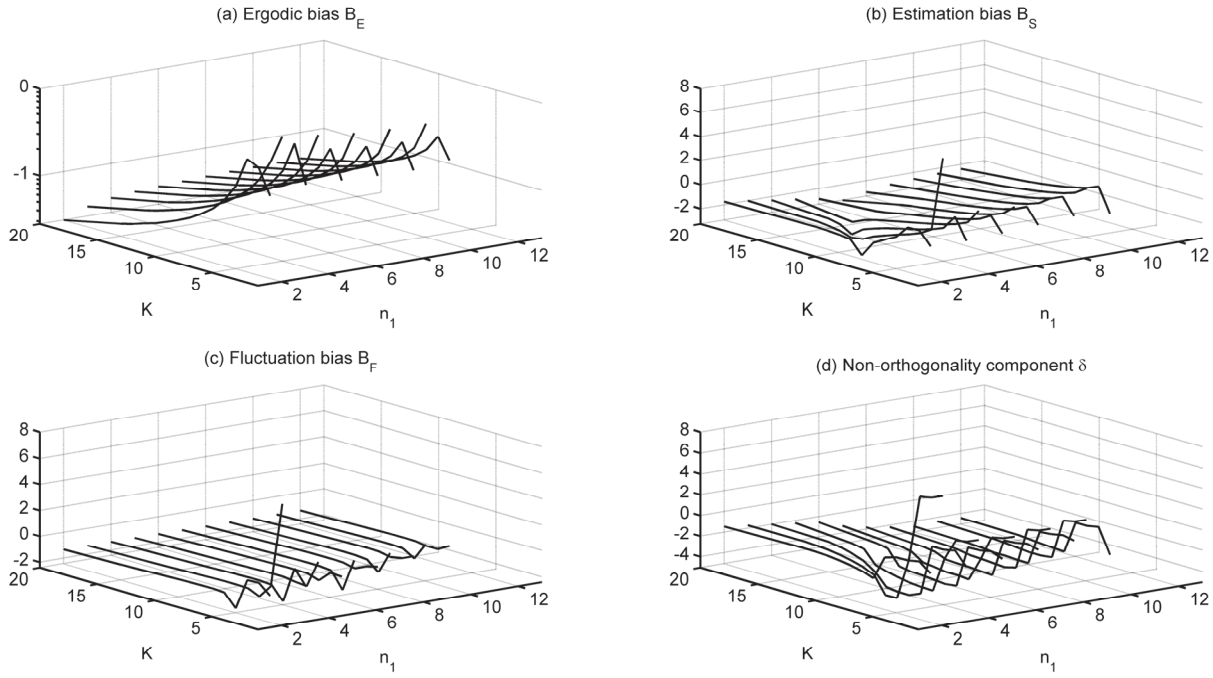


Figure 3.2: Standardized (divided by $||\mathbf{U}_2 \otimes \mathbf{U}_1||$) ergodic, estimation, and fluctuation biases and the non-orthogonality component for the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$; $n_1 = 2, \ldots, 12$; $n_2 = 2$; and sample size $K = K_{min}, \ldots, 15$ by steps of 1, in addition to 20. The vertical axis is in logarithmic scale, basis 10 ('0' means $1.0E0 = 1$; '1' means $1.0E1 = 10$, etc.). There is an atypical 'peak' combined with an atypical 'trough' in some of the biases and the non-orthogonality component for some values of n_1, as the sample size K increases. The pattern is related to the values of K_{min} in Figure 3.1(a).

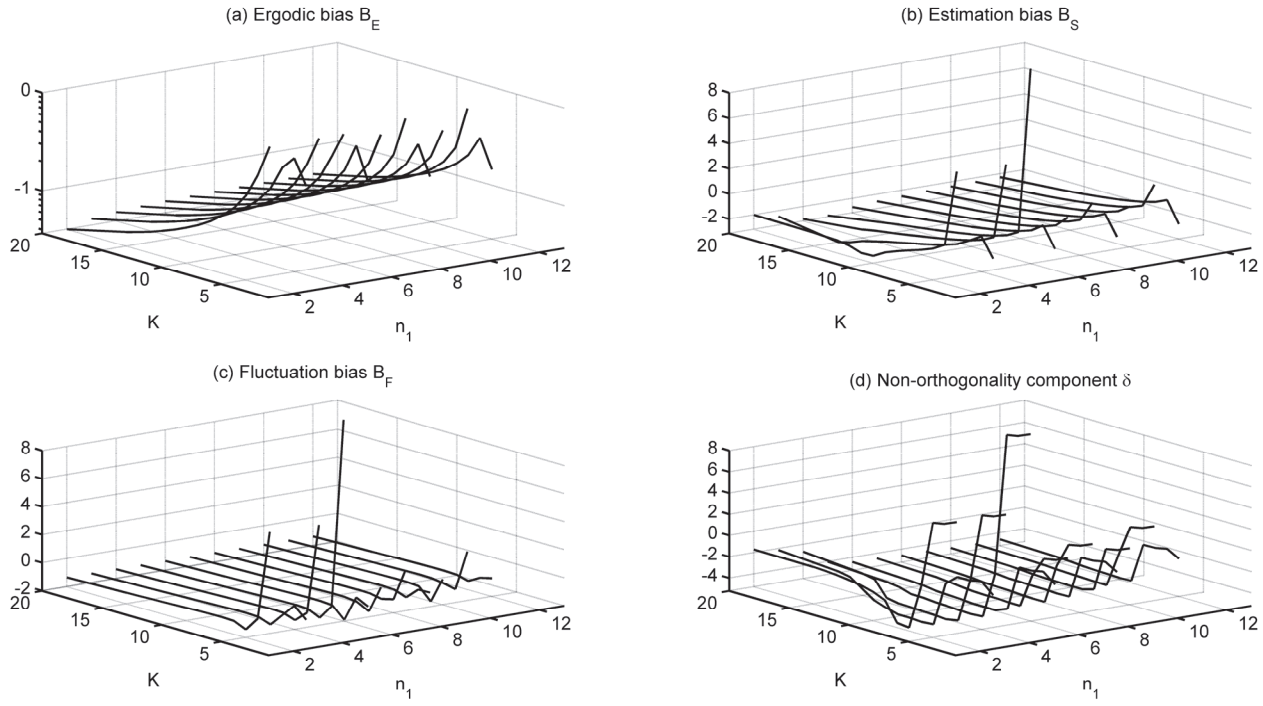


Figure 3.3: Standardized ergodic, estimation, and fluctuation biases and the non-orthogonality component for the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$; $n_1 = 2, \ldots, 12$; $n_2 = 3$; and sample size $K = K_{min}, \ldots, 15$ by steps of 1, in addition to 20. The vertical axis is in logarithmic scale, basis 10. As for $n_2 = 2$ (Figure 3.2), there is an atypical 'peak' combined with an atypical 'trough' in some of the biases and the non-orthogonality component for some values of n_1, as the sample size K increases. The pattern is related to the values of K_{min} in Figure 3.1(b).

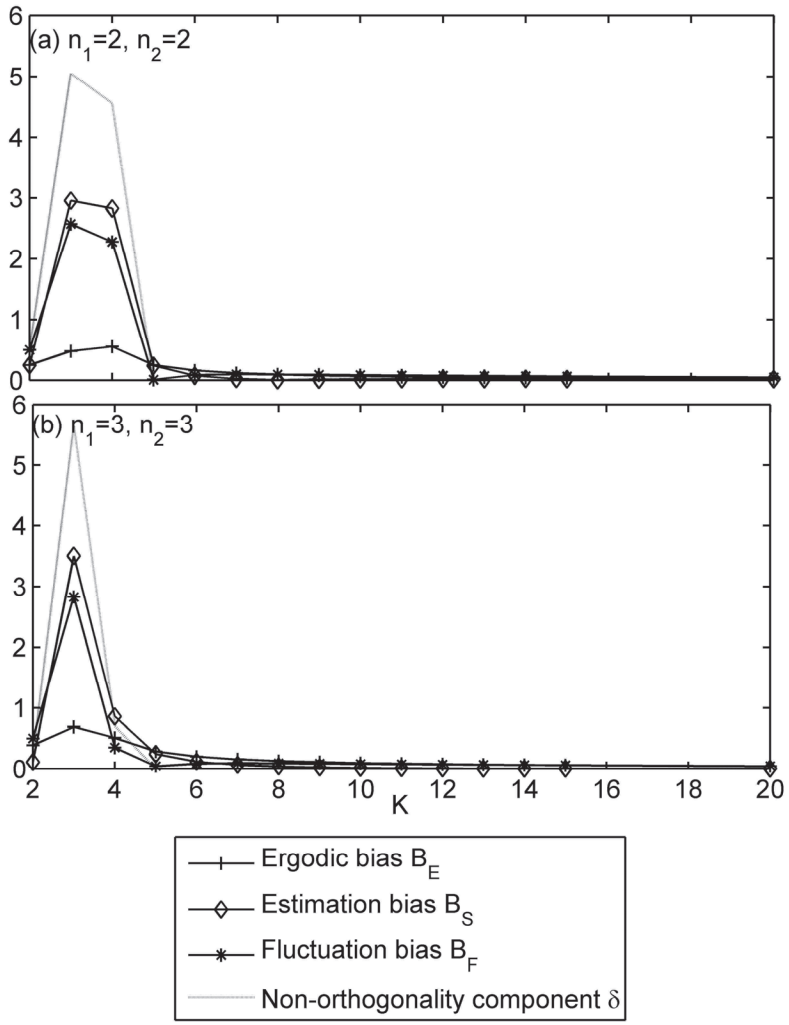


Figure 3.4: Standardized ergodic, estimation and fluctuation biases and the non-orthogonality component for the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ with sample size $K = K_{min}$, …, 15 by steps of 1, in addition to 20. The vertical axis is *not* in logarithmic scale. The first curve from each of the four panels in Figure 3.2 are grouped in panel (a), while the second curve from each panel of Figure 3.3 are grouped in panel (b) here. The three biases tend to present a 'peak' at the same value of the sample size, and the fluctuation bias shows a 'trough' when the estimation bias becomes smaller than the ergodic bias.

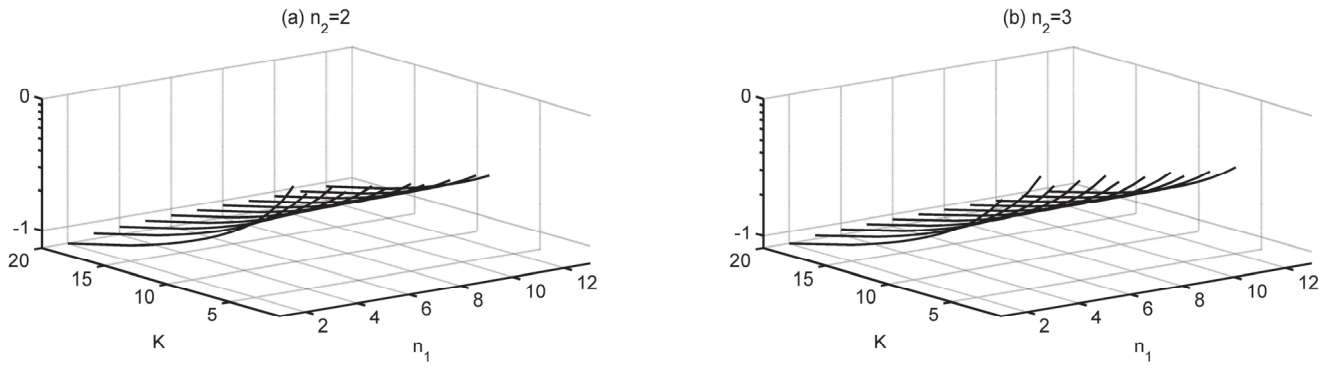


Figure 3.5: Standardized pseudo-theoretical bias of the ML estimator of $\mathbf{U}_2 \otimes \mathbf{U}_1$ with $n_1 = 2, \ldots, 12$ and sample size $K = K_{min}, \ldots, 15$ by steps of 1, in addition to 20. The vertical axis is in logarithmic scale, basis 10. Unlike patterns displayed in Figure 3.2 and 3.3, there is always a monotonic decay in the pseudo-theoretical bias with increasing sample size.

Preface to Chapter 4

The matrix normal distribution model was presented and the bias of the ML estimator of the Kronecker product of the two component variance-covariance matrices was studied in Chapter 3. In Chapter 4, the object of study is the assessment by LRT of the simple separability of a variance-covariance matrix, which is a key assumption of the model. The LRT is modified with a penalty, optimized so that the distribution of the LRT statistic is very well fitted by the theoretical chi-square distribution which, otherwise, holds only asymptotically. As a result, the modified LRT is unbiased. It is also shown that the bias of the unmodified LRT is more severe when the mean is modeled. An estimated generalized least-squares estimator (EGLS) algorithm, allowing modeling of the mean when working with an i.i.d. random sample of the matrix normal distribution, is presented in Appendix B. In April 2012, a manuscript (co-authored by Mr. Manceur and Prof. Dutilleul) based on this chapter was submitted to *Statistics and Probability Letters* for publication. Mr. Manceur and Prof. Dutilleul designed the unbiased modified likelihood ratio test. Mr. Manceur realized the simulation studies, and wrote the initial manuscript. Both authors participated in writing the final manuscript.

Chapter 4. An unbiased modified likelihood ratio test for simple separability of a variance-covariance structure

4.1 Abstract

The assumptions of simple and double separability of the variance-covariance structure, which are respectively made under the matrix and tensor normal distribution models, should be assessed in applications, for example to spatio-temporal repeated measures. When used for this purpose, the traditional likelihood ratio test (LRT) statistic does not follow a chi-square distribution with the expected number of d.f. when the sample size is finite, and is biased because the probability of rejecting the null hypothesis, while the separability of the variance-covariance structure holds, is then greater than the postulated significance level. Therefore, we modified the LRT statistic with an optimal penalty, so that the resulting modified LRT is unbiased by construction. The application of the penalty factor consists in a homothetic transformation of the LRT statistic, and the distribution of the modified LRT statistic is very well approximated by the chi-square distribution with the expected number of d.f. for most finite sample sizes, except very small ones.

4.2 Introduction

In general terms, a statistical test is said to be "biased if the probability of rejecting the hypothesis tested (H_0), when H_0 is valid – the significance level – is greater than the probability of rejection when some other hypothesis (*H*, say) is valid" (Kotz and Johnson, 1982, pp. 230–231). The unbiasedness of statistical tests is important in many if not all contexts, including the analysis of multi-dimensional data (e.g., 2-D spatio-temporal, multivariate temporal) where the inference usually requires certain assumptions. These datasets may be analyzed under the matrix normal distribution model, but this assumes that the variance-covariance structure is (simply) separable. Verifying such an assumption may be the object of inference by itself, possibly with a LRT preferably unbiased.

A separable variance-covariance structure, if it holds for a given dataset, presents some advantages. In particular, it need not be stationary in the general case, and the number of parameters to estimate is reduced compared to the case of an unstructured variance-covariance matrix submitted to no other constraint than positive definiteness.

For given numbers of rows (n_1) and columns (n_2) for a random matrix $\mathbf{X}$, the matrix normal distribution model is defined as:

$$\mathbf{X} \sim N_{n_1,n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2) \text{ if and only if } \text{vec}(\mathbf{X}) \sim N_{n_1 n_2}(\text{vec}(\mathbf{M}), \mathbf{U}_2 \otimes \mathbf{U}_1) \qquad (4.1)$$

where $E(\mathbf{X}) = \mathbf{M}$ $(n_1 \times n_2)$; $\mathbf{U}_1$ $(n_1 \times n_1)$ and $\mathbf{U}_2$ $(n_2 \times n_2)$ are the component variance-covariance matrices among the rows and among the columns of $\mathbf{X}$, respectively; vec(.) is the operator that stacks the columns of an $n_1 \times n_2$ matrix into an $n_1 n_2 \times 1$ vector; $\otimes$ is the Kronecker (direct) product; and var{vec($\mathbf{X}$)} = $\mathbf{U}_2 \otimes \mathbf{U}_1$. Let K denote the size of an i.i.d. random sample available for inference (estimation and testing).

The LRT statistic for simple separability is defined as the ratio between the likelihood function maximized under the null hypothesis (H_0: the variance-covariance matrix is the Kronecker product of two component variance-covariance matrices) and under the alternative hypothesis (H: the variance-covariance matrix is unstructured, or positive definite). The LRT statistic is theoretically distributed as $\chi^2(f)$, where f is equal to the difference between the number of free parameters in the null and alternative models (Muirhead 1982), that is, $f = \frac{n_1 n_2 (n_1 n_2 + 1)}{2} - \frac{n_1 (n_1 + 1)}{2} - \frac{n_2 (n_2 + 1)}{2} + 1$. The unmodified LRT used to assess whether a variance-covariance matrix equals a specified matrix is biased (Muirhead 1982, p. 353).

Several LRTs for simple separability of a variance-covariance structure have been proposed in the literature. Naik and Rao (2001), Njue (2001), Roy and Khattree (2005), Mitchell et al. (2005), Roy (2006), Simpson (2010), and Srivastava et al. (2008) presented LRTs for simple separability of a variance-covariance structure that are biased and in which at least one of the two component matrices is further modeled (e.g., first-order AR(1), CS). Svantesson and Wallace (2003), Lu and Zimmerman (2005), and Mitchell et al. (2006) presented LRTs that are biased and in which both component variance-covariance matrices are unstructured. In an article focusing on estimation, Dutilleul (1999) succinctly presented an unbiased LRT for simple separability of a variance-covariance structure, following Muirhead (1982, p. 357). The statistics of all

these tests, whether biased or not, are not distributed as $\chi^2(f)$ under the null hypothesis when the sample size is finite, a common problem for LRTs. To compensate for this, Mitchell et al. (2006) proposed to modify the quantile of the $\chi^2(f)$ distribution in order to control the significance level of the biased LRT.

Given the increasing collection of multi-dimensional data on which simple as well as double separability could or should be assessed, we focused on the development of modified LRTs for separability that are unbiased and whose test statistics follow approximate $\chi^2(f)$ distributions already for moderate sample sizes. In section 4.3, the estimation method is presented, including aspects of modeling of the mean. In section 4.4, after a description of the biased LRT for simple separability of a variance-covariance structure, the modified LRT, unbiased by construction, is introduced. In section 4.5, the optimal penalty *p*, to be used in the penalty factor that appears in the modified LRT statistic and aims to improve the $\chi^2(f)$ approximation, is studied by simulation. In section 4.6, the power of the new unbiased modified LRT for simple separability of a variance-covariance structure is compared to the rejection rate of the traditional biased LRT, with or without modeling of the mean. In section 4.7, concluding remarks are made. In Appendix B, the EGLS algorithm used to fit a response surface model to the mean matrix with a separable variance-covariance structure is described.

4.3 Estimation

Provided an i.i.d. random sample of size $K \geq \left(n_1 / n_2, n_2 / n_1\right)+1$ is available for inference under (4.1), Dutilleul (1999) presented the MLE (alias 'flip-flop') algorithm for the matrix normal distribution, more specifically for the two component variance-covariance matrices:

$$\begin{cases} \hat{\mathbf{U}}_1 = \sum_{k=1}^{K} (\mathbf{X}_k - \hat{\mathbf{M}})(Kn_2\hat{\mathbf{U}}_2)^{-1}(\mathbf{X}_k - \hat{\mathbf{M}})^T \\ \hat{\mathbf{U}}_2 = \sum_{k=1}^{K} (\mathbf{X}_k - \hat{\mathbf{M}})^T (Kn_1\hat{\mathbf{U}}_1)^{-1}(\mathbf{X}_k - \hat{\mathbf{M}}) \end{cases} \tag{4.2}$$

where $\hat{\mathbf{M}}$ is the estimator of $\mathbf{M}$ and ^{T} is the transpose operator. Under first-order stationarity, $\mathbf{M} = m\mathbf{J}$ and there is one mean parameter to estimate, the scalar *m*, so

$\hat{\mathbf{M}} = \hat{m}\mathbf{J} = \frac{1}{n_1 n_2 K}\left[\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{k=1}^{K}\mathbf{X}_{i_1 i_2, k}\right]\mathbf{J}$, with $\mathbf{J}$ the $n_1 \times n_2$ matrix of ones. Alternatively, $\mathbf{M}$ can be modeled with a linear (3 parameters) or quadratic (6 parameters) response surface model, vec($\mathbf{M}$) = $\mathbf{D\beta}$, and estimated by generalized least squares, vec($\hat{\mathbf{M}}$) = $\mathbf{D}\hat{\boldsymbol{\beta}}_{EGLS}$ (see Appendix B). $\mathbf{M}$ can also be left unmodeled; there are then $n_1 n_2$ mean parameters and the estimator is the classical sample mean, $\hat{\mathbf{M}} = \bar{\mathbf{X}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{X}_k$.

Whether the mean matrix $\mathbf{M}$ is modeled or not, there is no analytical solution to the system of equations above, so the component variance-covariance matrices and their Kronecker product are estimated iteratively (Dutilleul 1999). By comparison, the unstructured variance-covariance matrix $\boldsymbol{\Sigma}_{UN}$ in the vector normal distribution model $\text{vec}(\mathbf{X}) \sim N_{n_1 n_2}(\text{vec}(\mathbf{M}), \boldsymbol{\Sigma}_{UN})$ is estimated by ML using vec($\hat{m}\mathbf{J}$), some variant of $\hat{\boldsymbol{\beta}}_{EGLS}$ or vec($\bar{\mathbf{X}}$) depending on the mean model, if K is greater than or equal to $K_{min} = n_1 n_2 + 1$.

4.4 Likelihood ratio tests for simple separability of a variance-covariance structure

The hypotheses under testing are:

$$H_0 : \boldsymbol{\Sigma}_{UN}^{-1}(\mathbf{U}_2 \otimes \mathbf{U}_1) = \mathbf{I}_{n_1 n_2} \text{ against } H : \boldsymbol{\Sigma}_{UN}^{-1}(\mathbf{U}_2 \otimes \mathbf{U}_1) \neq \mathbf{I}_{n_1 n_2} \tag{4.3}$$

where $\mathbf{I}_{n_1 n_2}$ is the identity matrix of size $n_1 n_2$. Following Muirhead (1982, p. 357), the statistic of the biased unmodified LRT is:

$$\Lambda = K\left[n_2 \log|\hat{\mathbf{U}}_1| + n_1 \log|\hat{\mathbf{U}}_2| - \log|\hat{\boldsymbol{\Sigma}}_{UN}|\right] \tag{4.4}$$

where $|.|$ is the determinant. In a preliminary step, we studied the distribution of Λ by simulation, using the Ziggurat algorithm to generate normal pseudo-random numbers (Marsaglia and Tsang 2000) in Matlab (The MathWorks, 2010). The number of simulation runs was 1.0*E*5; $\mathbf{U}_1 = \mathbf{I}_{n_1}$ (n_1 = 2); $\mathbf{U}_2 = \mathbf{I}_{n_2}$ (n_2 = 2); matrix $\mathbf{I}_{n_2}$ was used to initiate the algorithm; and 1*E*-6, as the tolerance value to stop the iterations. As shown in Figure 4.1 (top line), at values of K close to the minimum required for ML estimation of $\boldsymbol{\Sigma}_{UN}$ (K_{min} = 5), the empirical rejection rate (using critical values based on the $\chi^2(f)$ distribution) is greater than the significance level of 0.05; at K = 50, the rejection rate is

still visibly greater than 0.05. Figure 4.2 presents the theoretical $\chi^2(f)$ distribution as well as the empirical distribution of Λ. The two do not match. In particular, the peak of the former distribution is located at a larger value than in the latter. Consequently, critical values based on the theoretical $\chi^2(f)$ distribution are incorrectly used when Λ is the test statistic.

Sugiura and Nagao (1968) and Muirhead (1982, p. 357) present unbiased LRTs for variance-covariance matrices in general terms (i.e. not specific to the separable structure), using $K-1$ instead of K in the definition of Λ because of the estimation of the mean from the data; $K-1$ is also the number of d.f. of the Wishart distribution of $K\hat{\boldsymbol{\Sigma}}_{UN} = \sum_{k=1}^{K}\{(\mathrm{vec}(\mathbf{X}_k)-\bar{\mathbf{X}})(\mathrm{vec}(\mathbf{X}_k)-\bar{\mathbf{X}})^T\}$ (Rencher 1995). Note that Mitchell et al. (2006) present a test for separability, based on restricted maximum likelihood estimators; this test was not formally claimed to be unbiased by the authors, and is not the unbiased LRT recommended by Muirhead (1982).

To improve the $\chi^2(f)$ fit of the distribution of the LRT statistic, we propose to modify Λ by replacing K with a value that can be different from $K-1$:

$$\Lambda^* = \left(\frac{K-p}{K}\right)\Lambda = (K-p)[n_2 \log|\hat{\mathbf{U}}_1| + n_1 \log|\hat{\mathbf{U}}_2| - \log|\hat{\boldsymbol{\Sigma}}_{UN}|] \qquad (4.5)$$

where p is a penalty applied to adjust the LRT statistic, so that its distribution is fitted by $\chi^2(f)$ as closely as possible and the bias is thus corrected by construction. The use of a penalty factor ($\frac{K-p}{K}$ here), to modify a LRT statistic and fit its distribution, was studied by Barndorff-Nielsen and Cox (1985), for testing the equality of variances from two samples drawn from normal populations; their work was built on early work by Bartlett (1937). The next section explores by simulation the optimal value of p for use in (4.5).

4.5 Optimal value of penalty in unbiased modified likelihood ratio test

The value of penalty p in equation (4.5) is optimized for the distribution of Λ^* to be best approximated by $\chi^2(f)$. This implies that the rejection rate will be close to the significance level when the critical value $\chi^2_{1-\alpha}(f)$ is used in the modified LRT. A second simulation study was performed to investigate the effects of the significance level

($\alpha = 0.01, 0.05, 0.1$), the numbers of rows and columns in the data matrices (n_1 = 2, 4 and n_2 = 2, 3, 4, 5, 6, 10), the sample size ($K = K_{min}, K_{min} + 1, \ldots, 50, 100, 500$), and the estimation of **M** (Mean Model 1, stationary, **M** = m**J** with m = 2; Mean Model 2, linear or quadratic response surface, vec(**M**) = **D**$\boldsymbol{\beta}$ – see Appendix B for coefficient values; Mean Model 3, no modeling, **M** is filled with positive and negative pseudo-random numbers) on the optimal value of penalty p. For all mean models, the optimal penalty p was selected from values of 0 to $n_1 n_2 + 0.9$ by increments of 0.1.

Overall, the mean value of the optimal penalty p does not appear to be dependent on the significance level, as it is 7.1, 7.1, and 7.2 for α = 0.01, 0.05, and 0.10, respectively. Thus, the optimal penalty p values averaged over α are reported in Table 4.1 for different combinations of values of n_1, n_2, and K. For given values of n_1, n_2, and K, the optimal penalty p increases as the number of mean parameters to estimate increases, from 1 ($\hat{m}$**J**) to $n_1 n_2$ ($\bar{\mathbf{X}}$). For all values of K, the increase in optimal penalty p, from a response surface model fitted by EGLS to the use of no model for the mean, remains small, even when the difference in the number of parameters is large. In the case $n_1 = n_2 = 2$, however, the optimal penalty p is the same or nearly the same when a linear response surface model is used and when the mean matrix is not modeled; the difference in the number of mean parameters is then one.

The optimal penalty p is dependent on n_1 and n_2. However, when comparing its values for $n_1 = 2$, $n_2 = 6$ vs. $n_1 = 4$, $n_2 = 3$ (the product of dimension sizes is 12), they are very similar. It is the same thing for the pairs $n_1 = 2$, $n_2 = 10$ and $n_1 = 4$, $n_2 = 5$ (product: 20). The optimal penalty p for the pair $n_1 = 4$, $n_2 = 4$ (product: 16) falls between the optimal penalty p values for $n_1 n_2$ = 12 and $n_1 n_2$ = 20. Thus, the optimal penalty p increases with the product of dimensions $n_1 n_2$. For given values of n_1 and n_2 and given mean model and estimation procedure, the optimal penalty p decreases nonlinearly with the sample size: the decrease is rapid for K comprised between K_{min}

and about K_{min} + 10, then the decrease is less marked and a plateau is almost reached for values of K beyond 100.

When investigating the interaction between the three factors (i.e. sample size K, product of dimensions $n_1 n_2$, mean model and estimation procedure), the largest value of the optimal penalty p is found for K_{min} when $n_1 n_2$ is largest and the mean matrix is unmodeled and estimated by $\bar{\mathbf{X}}$ (Table 4.1). The effect of the optimal penalty p on the distribution of the test statistic is illustrated in Figure 4.2: the distribution of Λ^* is fitted very closely with $\chi^2(f)$, thanks to a homothetic transformation of Λ; see equation (4.5). In Figure 4.1, the modified LRT appears slightly conservative at values of K close to K_{min}, but quickly the empirical rejection rate stabilizes at 0.05 when using critical values from $\chi^2(f)$. In conclusion, the LRT thus modified is unbiased, or valid (i.e. the theoretical and empirical significance levels match).

4.6 Power study

The empirical power for α = 0.05 and K = 2000 was studied by simulation (using a procedure similar to that of section 4.5) with $n_1 = 4$ and $n_2 = 3$ (Table 4.2). Separable and non-separable variance-covariance matrices were defined following Boik (1991). The degree of non-separability was measured with the κ factor, whose range is restricted to non-negative values (Genton 2007). Its value could vary from 0 (perfectly separable) to 0.96 in this study. The variance-covariance matrices with known κ were positive definite, and were used to generate i.i.d. random samples from the vector normal distribution. Estimation procedures for the mean (essentially the EGLS fitting of response surface models) were adapted accordingly.

The empirical probabilities of rejecting H_0 when it is true ($\kappa = 0$, significance level of the test) and when it is false ($\kappa > 0$, power of the test) are reported in Table 4.2 for the three mean models and estimation procedures, with Λ and Λ^* (Mean Model 1: optimal penalty p = 1.30; Mean Model 2: optimal penalty p = 3.90; Mean Model 3: optimal penalty p = 4.30). The significance level of the LRT based on Λ^* is 0.05, and for a given mean model, its power is slightly lower than the empirical rejection rate of the biased LRT based on Λ. While the minimum rejection rate of the LRT based on Λ is achieved at H_0, the bias of this test comes from the fact that its actual significance level

is greater than 0.05 even when the sample size is very large. For any value of κ, the difference between Λ and Λ^* is largest when the number of mean parameters to estimate is largest (Table 4.2).

4.7 Concluding remarks

We have modified the statistic of the LRT for simple separability of a variance-covariance structure in order to improve the fitting of its distribution with the $\chi^2(f)$ distribution with the expected number of d.f. By comparison, Mitchell et al. (2006) used a biased LRT (4.4), and modified the critical value to control the significance level, without modeling the mean matrix. In our approach, a penalty is applied to the LRT statistic (4.5), which has for effect to relocate the distribution of the LRT statistic and ensure it is closest to the $\chi^2(f)$ distribution. The LRT thus modified is unbiased and the use of $\chi^2(f)$-based critical values is justified. This was demonstrated with a stationary mean model, a linear or quadratic mean model, and an unmodeled mean matrix.

4.8 Acknowledgements

The financial assistance of the Natural Sciences and Engineering Research Council of Canada (NSERC), through an Alexander Graham Bell Canada Graduate Scholarship to Mr. Manceur and a Discovery Grant (Mathematics and Statistics) to Prof. Dutilleul, is gratefully acknowledged.

Table 4.1: Penalty (p) found to be optimal for modifying the LRT statistic for simple separability of a variance-covariance structure, for three mean models (1: stationary; 2: response surface – linear when $n_1 = 2$, and quadratic when $n_1 = 4$; 3: unmodeled matrix), six pairs of dimension values (n_1, n_2), and sample sizes (K) starting from the minimum required for testing ($K = n_1 n_2 + 1$) up to 500.

K	Mean model 1	Mean Model 2	Mean Model 3
$n_1 = 2, n_2 = 2$			
5	2.65	3.75	3.75
6	2.65	3.65	3.65
7	2.65	3.65	3.60
8	2.62	3.65	3.57
9	2.58	3.65	3.57
10	2.62	3.57	3.57
25	2.58	3.57	3.57
50	2.60	3.57	3.57
100	2.60	3.57	3.55
500	2.60	3.48	3.55
$n_1 = 4, n_2 = 4$			
17	8.68	10.38	10.55
18	8.25	9.52	9.68
19	8.02	9.08	9.27
20	7.82	8.82	9.00
21	7.68	8.58	8.78
22	7.57	8.43	8.68
25	7.32	8.12	8.37
50	6.63	7.43	7.72
100	6.40	7.13	7.58
500	6.45	6.32	7.37

K	Mean Model 1	Mean Model 2	Mean Model 3
$n_1 = 2, n_2 = 6$			
13	7.05	8.55	8.72
14	6.72	7.80	8.07
15	6.55	7.45	7.73
16	6.40	7.22	7.55
17	6.32	7.08	7.38
20	6.13	6.78	7.15
25	5.95	6.57	6.98
50	5.80	6.23	6.78
100	5.75	6.23	6.62
500	5.07	5.97	6.53
$n_1 = 4, n_2 = 3$			
13	6.68	8.28	8.38
14	6.35	7.67	7.68
15	6.18	7.32	7.35
16	6.05	7.10	7.18
17	5.93	6.95	7.02
20	5.75	6.72	6.78
25	5.58	6.50	6.58
50	5.37	6.18	6.30
100	5.37	6.07	6.25
500	5.37	5.55	6.10

K	Mean Model 1	Mean Model 2	Mean Model 3
$n_1 = 2, n_2 = 10$			
21	11.58	13.23	13.52
22	11.07	12.13	12.58
23	10.78	11.58	12.08
24	10.57	11.27	11.78
25	10.38	11.05	11.57
26	10.28	10.88	11.43
30	9.93	10.42	10.98
50	9.37	9.72	10.37
100	9.05	9.40	10.08
500	8.98	9.00	10.32
$n_1 = 4, n_2 = 5$			
21	10.68	12.52	12.75
22	10.20	11.48	11.72
23	9.92	10.95	11.20
24	9.68	10.65	10.90
25	9.48	10.35	10.65
26	9.35	10.17	10.48
30	8.95	9.68	10.03
50	8.33	8.87	9.33
100	8.03	8.38	9.07
500	8.05	8.30	8.42

Table 4.2: Empirical probability of rejecting H_0 (simple separability of a variance-covariance structure) as a function of discrepancy κ, with a biased unmodified LRT based on Λ vs. an unbiased modified LRT based on Λ^*, for three mean models (1: stationary; 2: quadratic response surface; 3: unmodeled matrix), with $n_1 = 4$, $n_2 = 3$, $K = 2000$, and $\alpha = 0.05$ ($\chi^2_{0.95}(63) = 82.5287$).

	Mean Model 1		Mean Model 2		Mean Model 3	
κ	Λ-based LRT	Λ^*-based LRT	Λ-based LRT	Λ^*-based LRT	Λ-based LRT	Λ^*-based LRT
0.0000	0.0504	0.0500	0.0512	0.0499	0.0514	0.0500
0.0050	0.0548	0.0544	0.0536	0.0523	0.0537	0.0524
0.0100	0.0644	0.0640	0.0649	0.0633	0.0648	0.0633
0.0146	0.0829	0.0823	0.0837	0.0819	0.0837	0.0819
0.0210	0.1269	0.1261	0.1284	0.1256	0.1281	0.1255
0.0296	0.2372	0.2362	0.2358	0.2319	0.2367	0.2329
0.0404	0.4790	0.4776	0.4827	0.4783	0.4832	0.4787
0.0694	0.9887	0.9886	0.9888	0.9884	0.9886	0.9882

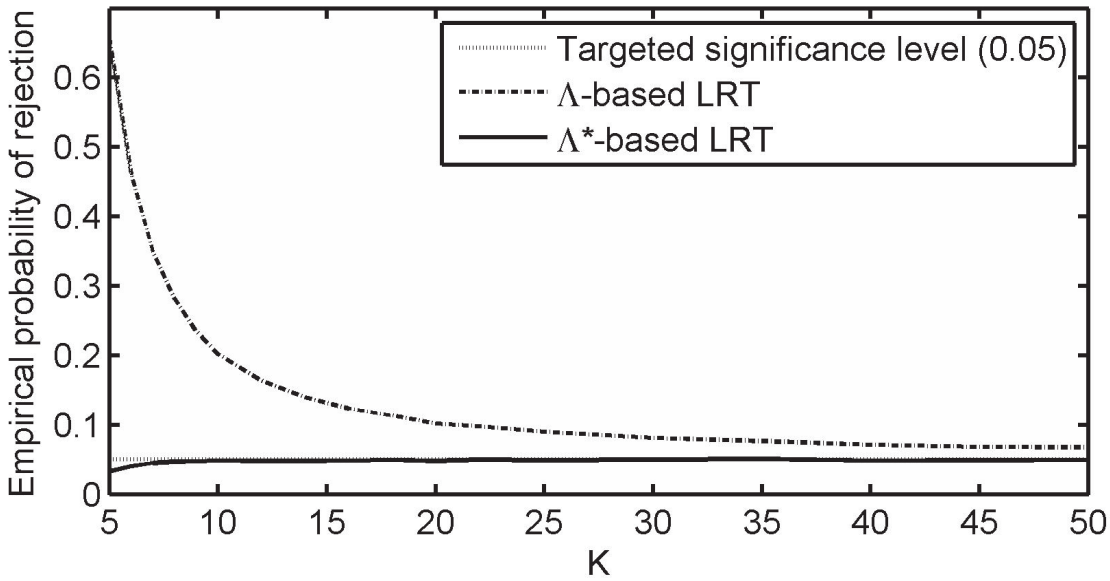


Figure 4.1: Empirical probability of rejecting H_0 (simple separability of a variance-covariance structure) when true, as a function of sample size K, using the critical value $\chi^2_{0.95}(5) = 11.0705$ with $n_1 = 2$, $n_2 = 2$, and Mean Model 3 ($\hat{\mathbf{M}} = \bar{\mathbf{X}}$). The empirical significance level of the unbiased modified LRT based on Λ^* is very close to the targeted significance level.

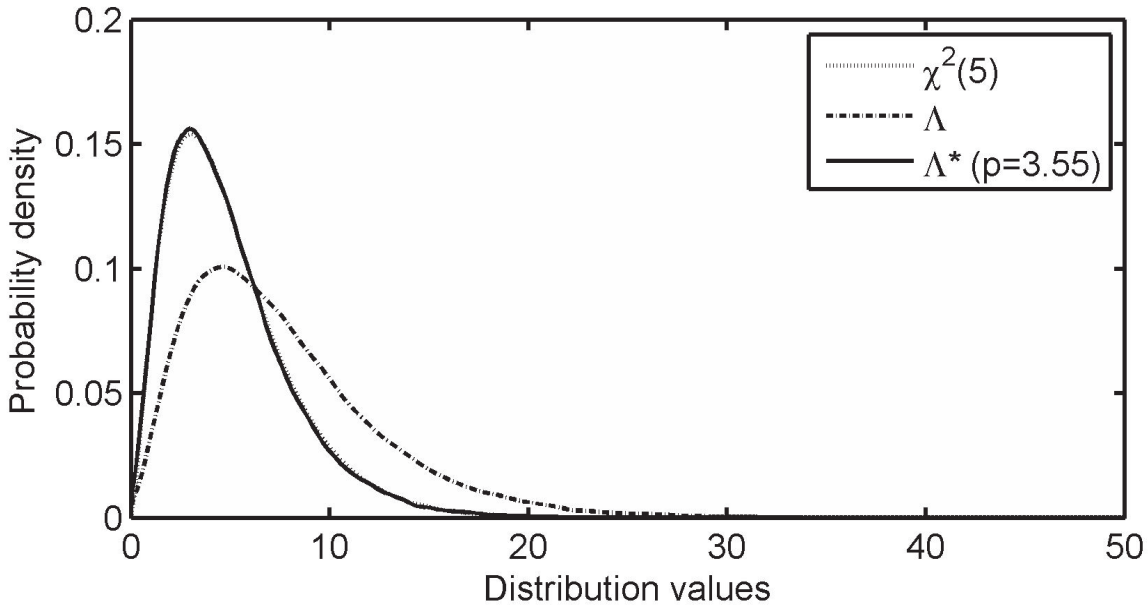


Figure 4.2: Empirical distribution of the LRT statistics Λ and Λ^* for $n_1 = 2$, $n_2 = 2$, Mean Model 3 ($\hat{\mathbf{M}} = \bar{\mathbf{X}}$), and $K = 10$, vs. the theoretical $\chi^2(5)$ distribution. The latter fits the empirical distribution of Λ^* so closely that they can barely be distinguished from each other.

Preface to Chapter 5

In Chapter 5, the matrix normal distribution model is extended to the tensor normal distribution model of order 3 and more. The object of study is the tensor normal distribution model for which a detailed definition is given, together with the MLE-3D algorithm (with the derivatives of the log-likelihood function for the tensor normal distribution model with respect to variance-covariance matrices in Appendix C). Thereafter, the bias and dispersion of the ML estimator of a doubly separable variance-covariance matrix are studied by simulation; the bias of the ML estimator of a doubly separable variance-covariance matrix is decomposed in Appendix D. An example using 3-D brain data completes the presentation. The unbiased modified LRT of double separability and the MLE-4D algorithm are presented in Appendix E and Appendix F, respectively. A manuscript (co-authored by Mr. Manceur and Prof. Dutilleul) based on this chapter was submitted in November 2011 to *Journal of Computational and Applied Mathematics* for publication. Mr. Manceur programmed the algorithm, performed the simulation studies, and wrote the initial manuscript. Both authors participated in writing the final manuscript.

Chapter 5. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion

5.1 Abstract

Recently, there has been a growing interest in the analysis of multi-dimensional data arrays (e.g. when a univariate response is sampled in 3-D space or when a multivariate response is sampled in time and 2-D space). In this chapter, we scrutinize the problem of ML estimation for the tensor normal distribution of order 3 or more, which is characterized by the separability of its variance-covariance structure; there is one variance-covariance matrix per dimension. In the 3-D case, the system of likelihood equations for the three variance-covariance matrices has no analytical solution, and therefore needs to be solved iteratively. We studied the convergence of an iterative three-stage algorithm (MLE-3D) that we propose for this, determined the minimum sample size required for matrix estimates to exist, and computed by simulation the empirical bias and dispersion of the Kronecker product of the three variance-covariance matrix estimators in eight scenarios. We found that the standardized bias and a matrix measure of dispersion decrease monotonically and tend to vanish with increasing sample size, so the Kronecker product estimator is consistent based on simulation results. An example with 3-D spatial measures of glucose content in the brain is also presented. Finally, results are discussed and the 4-D case is presented with simulation results in appendix.

5.2 Introduction

There is growing literature on the analysis of 2-D and 3-D data arrays, also called "multi-way data" (Dryden et al. 2008, Mardia and Goodall 1993, Bijma et al. 2005, Theobald and Wuttke 2008, Werner et al. 2008). Such data present correlations and heterogeneity of the variance, both within and among dimensions, through multiple responses and space-time levels. The variance-covariance structure is then often modeled to reduce the number of parameters and ensure the existence of parameter estimates. In a separable model (sometimes called "factorized" or "Kronecker structured"), the variance-covariance matrix of the vectorized multi-dimensional array is the Kronecker (direct) product of a number of variance-covariance matrices equal to the number of dimensions. The variance-covariance matrices used as factors in the Kronecker product define the

respective dependencies and variability among rows and columns in 2-D and among rows, columns and edges (or slices) in 3-D and beyond.

In 2-D, Dutilleul (1990, 1998, 1999) presented an iterative two-stage algorithm (MLE-2D) to estimate by ML the variance-covariance parameters of the matrix normal distribution $\mathbf{X} \sim N_{n_1,n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2)$, where the random matrix $\mathbf{X}$ is $n_1 \times n_2$, $\mathbf{M} = E(\mathbf{X})$, $\mathbf{U}_1$ is the $n_1 \times n_1$ variance-covariance matrix for the rows of $\mathbf{X}$ (e.g. repeated measures in space), and $\mathbf{U}_2$ is the $n_2 \times n_2$ variance-covariance matrix for the columns of $\mathbf{X}$ (e.g. repeated measures in time). The matrix normal distribution model implies a separable variance-covariance structure, defined by $\mathbf{U}_2 \otimes \mathbf{U}_1$. Other authors also studied the MLE-2D algorithm, and later nicknamed it "flip-flop" (Lu and Zimmerman 2005). Werner et al. (2008) compared it to four alternative estimation procedures, and found it was providing estimators with the lowest normalized root-MSE (mean square error), starting with very small sample sizes. The MLE-2D algorithm was found to be useful in brain science (Bijma et al. 2005), image analysis (Dryden et al. 2008), biochemistry (Theobald and Wuttke 2008), electrical engineering (Werner et al. 2008), and the environmental sciences (Dutilleul and Pinel-Alloul 1996), for example. Two unstructured variance-covariance matrices (with no other assumption than positive definiteness) are then estimated using a small number K of replicates, with $K \geq \max\left(\frac{n_1}{n_2}, \frac{n_2}{n_1}\right) + 1$, in order to ensure that the estimated variance-covariance matrices are positive definite.

Three-dimensional data arrays are obtained when a single response is sampled in 3-D space or in 2-D space and time or when multiple responses are recorded in 2-D space or in 1-D space and time. In the natural and life sciences, such data are provided by the measurement of wood density in given growth rings and directions at several heights in a tree trunk (Koga and Zhang 1996), the recording of a vector of air pollutants at a number of field stations over months (Mardia and Goodall 1993), and the monitoring of a vector of physiological variables in different organs over days (Roy and Leiva 2008). These data have rarely been analyzed by using a 3-D statistical methodology, apparently because it was not of easy access and the computational tools were not available. As the collection of 3-D data arrays is rising, it has become timely to fill in the gap. Therefore, we present the MLE-3D algorithm, define its conditions of application, and study by simulation the

properties of estimators in this chapter. In that non-trivial extension of the MLE-2D algorithm, the parameters are estimated by maximum likelihood under the relevant statistical distribution called "tensor normal distribution" and characterized by multivariate normality and a doubly separable variance-covariance structure, with no need to specify a variance-covariance matrix model at each dimension. Below, we summarize the approach followed and the results obtained in five earlier studies where a doubly separable variance-covariance structure was used for data analysis, prior to inserting our contribution in the developing field and explaining how this chapter is organized.

In 1993, Barton and Fuhrman explored the modeling of the variance-covariance structure of multi-dimensional data arrays that commonly arise in signal processing problems. They presented a notation system based on a "natural hierarchical block structure on the covariance data", and discussed the variance-covariance structures of block-circulant, block-Toeplitz type vs. unstructured type, with limited discussion of estimation algorithms. Corrections were provided by Fuhrman (1997).

Still in 1993, Mardia and Goodall (p. 358) presented an iterative three-stage estimation algorithm which resembles the MLE-3D algorithm that will be presented here, but the authors did not use it to analyze their multivariate spatio-temporal data due to insufficient replication. Eventually, they applied the MLE-2D algorithm by making the assumption of some temporal independence, and reported convergence in 10-14 iterations (Mardia and Goodall 1993, p. 357). Throughout, the authors assumed that the expected value of the random multi-dimensional array was constant along one of the three dimensions (i.e. time), and chose to use an isotropic variogram spherical model for the spatial variance-covariance matrix.

In the context of the analysis of doubly and triply repeated measures in the medical sciences, Galecki (1994) tried various types of variance-covariance structures, including autoregressive, compound symmetric, spherical (i.e. independence and homoscedasticity), and unstructured. He also developed the concept of covariance profile, and presented one application of an estimation algorithm without detailing it.

In 2008, an iterative ML algorithm for 3-D data arrays was proposed by Roy and Leiva, together with a LRT aimed to assess the adequacy of a doubly separable model for the variance-covariance structure. The estimation algorithm assumed an AR(1) or CS

structure for one variance-covariance matrix, and an intra-class correlation structure was assumed for the two others. The simulation study designed to verify the small-sample behavior of the LRT was limited to a reduced form of the tensor normal distribution of order 3, in which one of the three variance-covariance matrices was a scalar.

More recently, a Newton-Raphson type of algorithm (i.e. with no "flip-flop") was used by Richter et al. (2008) to estimate by ML the parameters of the tensor normal distribution model. This algorithm was developed in the context of tensor-valued signals in electrical engineering. In that Newton-Raphson ML algorithm, a score function is used and a variance parameter is estimated in addition to the three variance-covariance matrices. Furthermore, Richter et al. (2008) used a block-diagonal approximation for the Hessian matrix required by the Newton-Raphson algorithm.

To our knowledge, Dutilleul (1990) was the first to present the probability density function and the moment generating function of the tensor normal distribution, using tensor notations inspired from McCullagh (1987). Other notations for tensor operators such as the inner and outer products and the tensor multiplication appear more popular nowadays (Kolda and Bader 2009), so Dutilleul's original equations were re-written in Section 2 here. Still to our knowledge, Mardia and Goodall (1993), followed by Dutilleul (1998), first proposed the MLE-3D algorithm for the tensor normal distribution, which was recently presented by Hoff (2011, p. 185, in 3-D or more) and Ohlson et al. (2011, p. 15) in a theoretical framework and without numerical and simulation results concerning the convergence of the algorithm and the empirical properties of the estimators. The contribution lies in a detailed presentation of the algorithm with numerical and simulation results in addition to practical guidelines.

More specifically, we start by defining the tensor normal distribution of order 3 or more from the moment generating function in the general case and from the probability density function in the regular case (of particular interest for ML estimation) in Section 5.3; so doing, we will see that only the Kronecker product of variance-covariance matrices is defined uniquely. In Section 5.4, the system of likelihood equations for the mean and variance-covariance parameters of the tensor normal distribution of order 3 is derived; the complete MLE-3D algorithm is presented in details; and the minimum sample size required to ensure an output (i.e. positive-definite estimated variance-covariance matrices) is determined. In Section 5.5, simulations are used to study in small

and larger samples the empirical bias and dispersion of the Kronecker product of estimated variance-covariance matrices provided by the MLE-3D algorithm; special attention is paid to the questions of initial solutions and convergence criteria. In Section 5.6, an application with 3-D spatial data of brain glucose content is presented. In Section 5.7, other applications and extensions (e.g. MLE-4D) are discussed. Four appendices present the first derivatives of the log-likelihood function in the 3-D case (Appendix C), the bias decomposition for the ML estimator of a doubly separable variance-covariance matrix (Appendix D), the LRT of double separability (Appendix E), and the MLE-4D algorithm, with simulation results (Appendix F).

5.3 Definition of the tensor normal distribution

The random tensor of order J and of dimensions $n_1 \times ... \times n_J$, $\mathbb{X} = (x_{i_1 ... i_J})$ with $i_j = 1 ... n_j$ $(j = 1 ... J)$, is normally distributed with mean parameter $\mathbb{M}$ (same order and dimensions as $\mathbb{X}$) and variance-covariance parameters $\mathbf{U}_j$ ($n_j \times n_j$) positive semidefinite $(j = 1 ... J)$ if, by definition, its moment generating function is given by:

$$M_{\mathbb{X}}(\mathbb{T}) = \exp < \mathbb{M}, \mathbb{T} > \exp\{ \frac{1}{2} \mathbb{T} \times_{1...J} (\overset{J}{\underset{j=1}{\circ}} \mathbf{U}_j) \times_{1...J} \mathbb{T} \}, \qquad (5.1)$$

where $\mathbb{T} = (t_{i_1 ... i_J})$ is a non-random tensor of same order and dimensions as $\mathbb{X}$, "< , >" denotes the inner product, " $\circ$ " is the outer product and " $\times_{1...J}$ " is the tensor multiplication over all J dimensions, i.e. a generalization of the n-mode product (Kolda and Bader 2009):

$$< \mathbb{M}, \mathbb{T} > = \sum_{i_1=1}^{n_1} ... \sum_{i_J=1}^{n_J} m_{i_1 ... i_J} \, t_{i_1 ... i_J} \text{ and}$$

$$\mathbb{T} \times_{1...J} (\overset{J}{\underset{j=1}{\circ}} \mathbf{U}_j) \times_{1...J} \mathbb{T} = \sum_{i_1, i_1'=1}^{n_1} ... \sum_{i_J, i_J'=1}^{n_J} t_{i_1 ... i_J} \, u_{i_1 i_1'} ... u_{i_J i_J'} \, t_{i_1' ... i_J'} .$$

Setting $J = 3$ in equation (5.1) provides the moment generating function of the tensor normal distribution of order 3. This distribution is denoted:

$$\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3), \qquad (5.2)$$

where $\mathbb{M} = E(\mathbb{X})$.

An equivalent formulation of the tensor normal distribution involves the Kronecker product $\otimes$ and the vec operator (Kolda and Bader 2009). For a random tensor $\mathbb{X} = (x_{i_1 \ldots i_J})$ of order J and of dimensions $n_1 \times \ldots \times n_J$, $\text{vec}(\mathbb{X})$ is the $\prod_{j=1}^{J} n_j \times 1$ random vector:

$$\text{vec}(\mathbb{X}) = (x_{11\ldots11} \ldots x_{n_1 1\ldots11} x_{12\ldots11} \ldots x_{n_1 n_2 \ldots n_{J-1} 1} x_{11\ldots12} \ldots x_{n_1 n_2 \ldots n_{J-1} 2} \ldots x_{n_1 n_2 \ldots n_{J-1} n_J})^T \tag{5.3}$$

where the entries of $\mathbb{X}$ are aligned column after column, from edge to edge. Therefore, the tensor normal distribution of order J can also be defined as:

$$\mathbb{X} \sim N_{n_1, \ldots, n_J}(\mathbb{M}; \mathbf{U}_1, \ldots, \mathbf{U}_J) \tag{5.4}$$

if and only if

$$\text{vec}(\mathbb{X}) \sim N_{\prod_{j=1}^{J} n_j}(\text{vec}(\mathbb{M}), \bigotimes_{j=J}^{1} \mathbf{U}_j), \tag{5.5}$$

where $\text{vec}(\mathbb{M}) = E\{\text{vec}(\mathbb{X})\}$ and $\bigotimes_{j=J}^{1} \mathbf{U}_j = \text{var}\{\text{vec}(\mathbb{X})\}$ is positive-semidefinite. The tensor normal model is not adequate for any vector normal distribution of size $\prod_{j=1}^{J} n_j$, with $n_j > 1$ $(j = 1 \ldots J)$; only the vector normal distributions with the required separable variance-covariance structure can be modeled accordingly.

The 3-D distribution model defined above cannot be identified uniquely because for two scalars $a_1, a_2 > 0$ and for the same moment generating function, we have:

$$\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; a_1\mathbf{U}_1, a_2\mathbf{U}_2, (a_1 a_2)^{-1}\mathbf{U}_3). \tag{5.6}$$

A main advantage of the separable model is an important reduction in the number of variance-covariance parameters. An unstructured matrix $\mathbf{\Sigma}$ for the $n_1 n_2 n_3 \times 1$ random vector is composed of $\frac{n_1 n_2 n_3 (n_1 n_2 n_3 + 1)}{2}$ distinct variance-covariance parameters. By comparison, a separable matrix $\mathbf{\Sigma}$ for the random tensor of order 3 and dimensions $n_1 \times n_2 \times n_3$ is composed of $\frac{n_1(n_1+1)}{2} + \frac{n_2(n_2+1)}{2} + \frac{n_3(n_3+1)}{2}$ distinct variance-covariance parameters. This difference in the number of variance-covariance

parameters will be reflected in the minimum sample size (i.e. number of data tensors) required to obtain positive-definite estimated variance-covariance matrices under the separable normal distribution model (see Subsection 5.4.2), which is substantially smaller than the minimum sample size of $n_1 n_2 n_3 + 1$ (i.e. number of data vectors) required when separability is not assumed or does not hold. Furthermore, the properties of the Kronecker product help reduce the computational time of certain matrix operations (e.g. $(\mathbf{U}_1 \otimes \mathbf{U}_2 \otimes \mathbf{U}_3)^{-1} = (\mathbf{U}_1^{-1} \otimes \mathbf{U}_2^{-1} \otimes \mathbf{U}_3^{-1})$, assuming $\mathbf{U}_1, \mathbf{U}_2$ and $\mathbf{U}_3$ are positive definite).

Galecki (1994) suggested advantages for the interpretation of repeated factors under the separable variance-covariance structure. However, the parsimony of the separable model implies restrictions for cross-correlations. In fact, for $i_1, i_1' = 1 \ldots n_1$, $i_2, i_2' = 1 \ldots n_2$ and $i_3, i_3' = 1 \ldots n_3$, the cross-correlation function is restricted to be symmetrical:

$$\operatorname{cov}(x_{i_1 i_2 i_3}, x_{i_1' i_2' i_3'}) = u_{i_1 i_1'} u_{i_2 i_2'} u_{i_3 i_3'} = u_{i_1 i_1'} u_{i_2' i_2} u_{i_3 i_3'} = \operatorname{cov}(x_{i_1 i_2' i_3}, x_{i_1' i_2 i_3'}) \tag{5.7}$$

In particular, the variance is $\operatorname{var}[x_{i_1 i_2 i_3}] = u_{i_1 i_1} u_{i_2 i_2} u_{i_3 i_3}$.

If $\mathbb{X} \sim N_{n_1, \ldots, n_J}(\mathbb{M}; \mathbf{U}_1, \ldots, \mathbf{U}_J)$ and $\mathbf{U}_1 \ldots \mathbf{U}_J$ are positive definite, then the distribution of $\mathbb{X}$ is said to be regular, and is characterized by the probability density function:

$$f_{\mathbb{X}}(\mathbb{X}) = (2\pi)^{-\prod_{j=1}^{J} \frac{n_j}{2}} \prod_{j=1}^{J} |\mathbf{U}_j|^{-\prod_{\substack{j'=1 \\ j' \neq j}}^{J} \frac{n_{j'}}{2}} \exp\{-\frac{1}{2}(\mathbb{X} - \mathbb{M}) \times_{1 \ldots J} (\overset{J}{\underset{j=1}{\circ}} \mathbf{U}_j^{-1}) \times_{1 \ldots J} (\mathbb{X} - \mathbb{M})\}, \tag{5.8}$$

where $|\ .\ |$ denotes the determinant. For a regular normal tensor of order 3 (i.e. $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ are positive definite), the probability density function is:

$$f_{\mathbb{X}}(\mathbb{X}) = (2\pi)^{-\frac{n_1 n_2 n_3}{2}} |\mathbf{U}_1|^{-\frac{n_2 n_3}{2}} |\mathbf{U}_2|^{-\frac{n_1 n_3}{2}} |\mathbf{U}_3|^{-\frac{n_1 n_2}{2}} \exp\{-\frac{1}{2}(\mathbb{X} - \mathbb{M}) \times_{1 \ldots J} (\overset{3}{\underset{j=1}{\circ}} \mathbf{U}_j^{-1}) \times_{1 \ldots J} (\mathbb{X} - \mathbb{M})\} \tag{5.9}$$

In the development of the MLE-3D algorithm (see Section 5.4), data tensors, centered or not with respect to the sample mean tensor, will need to be reshaped as

matrices in various ways. The matricization operator presented in Kolda and Bader (2009) and the companion notations will be used for this. More specifically, if $\mathbb{A}$ is a tensor of third order and dimensions $n_1 \times n_2 \times n_3$, then $\mathbb{A}_{(1)}$ $\mathbb{A}_{(2)}$ and $\mathbb{A}_{(3)}$ will denote the matricized tensors with dimensions $n_1 \times n_2 n_3$, $n_2 \times n_1 n_3$ and $n_3 \times n_1 n_2$, respectively; see Kolda and Bader 2009, page 460 for details and an example.

5.4 Maximum likelihood estimation

5.4.1 The three-stage algorithm

The MLE-3D algorithm presented below requires an i.i.d. random sample of tensor data, which implies that the mean values and variance-covariance matrices are homogeneous within the sample; such practical implications will be discussed in the example with real data in Section 5.6. Let $\mathbb{X}_k$ $(k = 1 \ldots K)$ be an i.i.d. random sample from the regular tensor normal distribution $\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$. The log-likelihood function to be maximized can then be written as

$$l = -c - \frac{n_2 n_3 K}{2} \log |\mathbf{U}_1| - \frac{n_1 n_3 K}{2} \log |\mathbf{U}_2| - \frac{n_1 n_2 K}{2} \log |\mathbf{U}_3| \\ -\frac{1}{2} \sum_{k=1}^{K} \left[(\mathbb{X}_k - \mathbb{M}) \times_{1 \ldots J} (\overset{3}{\underset{j=1}{\circ}} \mathbf{U}_j^{-1}) \times_{1 \ldots J} (\mathbb{X}_k - \mathbb{M}) \right]. \tag{5.10}$$

First derivatives of the log-likelihood function are given in Appendix C.

Let $\hat{\mathbb{M}}$ and $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ denote the ML estimators of $\mathbb{M}$ and $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, respectively. Equalling to zero the first derivative of the log-likelihood function with respect to $\mathbb{M}$ provides that the sample mean is the ML estimator: $\hat{\mathbb{M}} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{X}_k = \bar{\mathbb{X}}$.

After replacing $\mathbb{M}$ by $\hat{\mathbb{M}} = \bar{\mathbb{X}}$ and using the property of positive definiteness of $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, equalling to zero the first derivatives of the log-likelihood function with respect to $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ provides the following system of three equations, for which $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ are the solutions:

$$\begin{cases} \hat{\mathbf{U}}_1 = \dfrac{1}{K n_2 n_3} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}{}^T; \\ \hat{\mathbf{U}}_2 = \dfrac{1}{K n_1 n_3} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)}{}^T; \\ \hat{\mathbf{U}}_3 = \dfrac{1}{K n_1 n_2} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)} (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)}{}^T. \end{cases} \quad (5.11)$$

There is no analytical solution to this system, but it can be solved iteratively by using the following three-stage algorithm, where $\|.\|$ denotes the Euclidean norm and $\hat{\mathbf{U}}_2^0$, $\hat{\mathbf{U}}_3^0$ are initial solutions for $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$.

Initialization: step = 0; $\mathbf{U}_2^* = \hat{\mathbf{U}}_2^0$; $\mathbf{U}_3^* = \hat{\mathbf{U}}_3^0$;

$$\mathbf{U}_1^* = \frac{1}{K n_2 n_3} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (\mathbf{U}_3^* \otimes \mathbf{U}_2^*)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}{}^T;$$

step = step+1;

$$\mathbf{U}_2^+ = \frac{1}{K n_1 n_3} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)} (\mathbf{U}_3^* \otimes \mathbf{U}_1^*)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)}{}^T; \quad (5.12)$$

$$\mathbf{U}_3^+ = \frac{1}{K n_1 n_2} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)} (\mathbf{U}_2^+ \otimes \mathbf{U}_1^*)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)}{}^T; \quad (5.13)$$

$$\mathbf{U}_1^+ = \frac{1}{K n_2 n_3} \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (\mathbf{U}_3^+ \otimes \mathbf{U}_2^+)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}{}^T; \quad (5.14)$$

While a given convergence criterion (based on separate rates of change in variance-covariance matrix estimates or on the rate of change in their Kronecker product) is not met,

Repeat:

step = step+1;

$\mathbf{U}_1^* = \mathbf{U}_1^+$; $\mathbf{U}_2^* = \mathbf{U}_2^+$; $\mathbf{U}_3^* = \mathbf{U}_3^+$;

Re-evaluate equations (5.12)–(5.14).

The convergence criterion is met. Maximum likelihood solutions are: $\hat{\mathbf{U}}_1 = \mathbf{U}_1^*$; $\hat{\mathbf{U}}_2 = \mathbf{U}_2^*$; $\hat{\mathbf{U}}_3 = \mathbf{U}_3^*$.

When available, prior information may direct the choice of $\hat{\mathbf{U}}_2^0$ and $\hat{\mathbf{U}}_3^0$. However, $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$ and $\hat{\mathbf{U}}_3^0 = \mathbf{I}_{n_3}$ will generally be suitable initial solutions for the search of $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$. Initial solutions $\hat{\mathbf{U}}_2^0$ and $\hat{\mathbf{U}}_3^0$ are also required to initiate the search for $\hat{\mathbf{U}}_1$, since estimates of $\mathbf{U}_2$ and $\mathbf{U}_3$ appear on the right-hand side of the first equation (i.e. for $\mathbf{U}_1$) in system (5.11). The choice of initial solutions for variance-covariance matrix estimates is further discussed in the frame of the simulation study (Section 5.5). The convergence criterion may be based on separate rates of change in variance-covariance matrix estimates (i.e. while $\| \mathbf{U}_1^+ - \mathbf{U}_1^* \| > \epsilon_1$ or $\| \mathbf{U}_2^+ - \mathbf{U}_2^* \| > \epsilon_2$ or $\| \mathbf{U}_3^+ - \mathbf{U}_3^* \| > \epsilon_3$, where ϵ_1, ϵ_2, ϵ_3 are infinitesimal positive quantities) or on the rate of change in their Kronecker product (i.e. while $\| \mathbf{U}_3^+ \otimes \mathbf{U}_2^+ \otimes \mathbf{U}_1^+ - \mathbf{U}_3^* \otimes \mathbf{U}_2^* \otimes \mathbf{U}_1^* \| > \epsilon$, where ϵ is an infinitesimal positive quantity). The former convergence criterion allows some flexibility, with the possible use of different infinitesimal positive quantities when the variance-covariance matrices are of very different sizes, and is not in contradiction with the fact that the variance-covariance matrix parameters are defined up to a multiplicative constant, since differences are not made relative to them; a similar criterion was used by Dutilleul (1999) for the MLE-2D algorithm. The latter convergence criterion is global and based on the fact that the Kronecker product of the three variance-covariance matrices is $\text{var}\{\text{vec}(\mathbb{X})\}$. We studied the use of the two convergence criteria with simulations (Section 5.5). With no other constraint than positive definiteness, the ML estimators $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ are defined up to two multiplicative constants, in the same way that the three variance-covariance matrices $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ of the tensor normal distribution of order 3 are defined up to two positive scalars a_1 and a_2; see equation (5.6). Only the Kronecker product $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ is defined uniquely and will therefore be studied for bias and dispersion in Section 5.5.

5.4.2 Existence of ML estimates for variance-covariance matrices

Hereafter, we prove that the ML estimators exist for the variance-covariance matrices of the regular tensor normal distribution of order 3 if and only if the sample size K of the i.i.d. random sample drawn from $\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ satisfies the condition:

$$K \geq \max\left(\frac{n_1}{n_2 n_3}, \frac{n_2}{n_1 n_3}, \frac{n_3}{n_1 n_2}\right) + 1. \qquad (5.15)$$

In particular, when $n_2 = 1$ and $n_3 = 1$, equation (5.15) reduces to the classical condition $K > n_1$ for the regular n_1-variate normal distribution; when $n_3 = 1$, it reduces to the condition in Dutilleul (1999) for the regular $n_1 \times n_2$ matrix normal distribution.

To prove equation (5.15), let us rewrite the system of equations (5.11) in matrix notation:

$$\begin{cases} \hat{\mathbf{U}}_1 = \dfrac{1}{K n_2 n_3}(\mathbb{X}_{1(1)} \ldots \mathbb{X}_{K(1)})\left\{\left(\mathbf{I}_K - \dfrac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1}\right\}(\mathbb{X}_{1(1)} \ldots \mathbb{X}_{K(1)})^T \\ \hat{\mathbf{U}}_2 = \dfrac{1}{K n_1 n_3}(\mathbb{X}_{1(2)} \ldots \mathbb{X}_{K(2)})\left\{\left(\mathbf{I}_K - \dfrac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1}\right\}(\mathbb{X}_{1(2)} \ldots \mathbb{X}_{K(2)})^T \\ \hat{\mathbf{U}}_3 = \dfrac{1}{K n_1 n_2}(\mathbb{X}_{1(3)} \ldots \mathbb{X}_{K(3)})\left\{\left(\mathbf{I}_K - \dfrac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1}\right\}(\mathbb{X}_{1(3)} \ldots \mathbb{X}_{K(3)})^T \end{cases} \qquad (5.16)$$

where $\mathbf{J}_K$ is the $K \times K$ matrix of ones.

Thus, the random matrices $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ are quadratic forms in $\mathbb{X}_{1(1)} \ldots \mathbb{X}_{K(1)}$, $\mathbb{X}_{1(2)} \ldots \mathbb{X}_{K(2)}$ and $\mathbb{X}_{1(3)} \ldots \mathbb{X}_{K(3)}$, respectively, and their ranks are given by

$$\begin{aligned} &\operatorname{rank}(\hat{\mathbf{U}}_1) = \operatorname{rank}\left\{\left(\mathbf{I}_K - \frac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1}\right\} = (K-1) n_2 n_3 \Leftrightarrow \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \text{ is positive definite;} \\ &\operatorname{rank}(\hat{\mathbf{U}}_2) = \operatorname{rank}\left\{\left(\mathbf{I}_K - \frac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1}\right\} = (K-1) n_1 n_3 \Leftrightarrow \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1 \text{ is positive definite;} \\ &\operatorname{rank}(\hat{\mathbf{U}}_3) = \operatorname{rank}\left\{\left(\mathbf{I}_K - \frac{1}{K}\mathbf{J}_K\right) \otimes (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1}\right\} = (K-1) n_1 n_2 \Leftrightarrow \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 \text{ is positive definite.} \end{aligned} \qquad (5.17)$$

It follows that the matrix estimators $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$ are positive definite, which includes that $\operatorname{rank}(\hat{\mathbf{U}}_1) = n_1$, $\operatorname{rank}(\hat{\mathbf{U}}_2) = n_2$ and $\operatorname{rank}(\hat{\mathbf{U}}_3) = n_3$, if and only if $(K-1) n_2 n_3 \geq n_1$, $(K-1) n_1 n_3 \geq n_2$ and $(K-1) n_1 n_2 \geq n_3$, or equivalently, $K \geq \dfrac{n_1}{n_2 n_3} + 1$, $K \geq \dfrac{n_2}{n_1 n_3} + 1$ and $K \geq \dfrac{n_3}{n_1 n_2} + 1$, which completes the proof.

5.5 Simulation studies

5.5.1 Description of scenarios

The general objective of our simulation studies was to determine empirically the bias and dispersion of $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ with small and large samples. Specific objectives concerned the choice of initial solutions for variance-covariance matrix estimates and the definition of the convergence criterion, and their effects on the performance of the algorithm and the final ML estimates. Two simulation studies were completed. Infinitesimal positive quantities (i.e. ϵ_1, ϵ_2, ϵ_3, ϵ) were set at 1.0*E*-6. Two 'extreme' sets of initial solutions for $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$ were tried for comparison purposes: very basic, $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$ and $\hat{\mathbf{U}}_3^0 = \mathbf{I}_{n_3}$, which may be far (initially) from the true parameter values in practice, vs. the true parameter values, $\hat{\mathbf{U}}_2^0 = \mathbf{U}_2$ and $\hat{\mathbf{U}}_3^0 = \mathbf{U}_3$.

In the first simulation study, n_1 = 5, n_2 = 3, n_3 = 4; the sample sizes considered are K = 5, 10, 15, 20, 100, and 500; and the number of scenarios is eight. In each scenario, variance-covariance matrices $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ were built so that they possess known theoretical properties measured by Box's 'epsilon' (Box 1954a, 1954b), which is a measure of the deviation of a variance-covariance matrix from the circularity condition (an $n \times n$ variance-covariance matrix $\mathbf{\Sigma}$ is said to be circular if $\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C} = \lambda\, \mathbf{I}_{n-1}$, with λ, a positive scalar, and $\mathbf{C}$, an $n \times (n-1)$ matrix of orthonormal contrasts). Box's 'epsilon' is given by $\dfrac{\{\mathrm{tr}(\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C})\}^2}{(n-1)\,\mathrm{tr}\{(\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C})^2\}}$. Its value varies from $\dfrac{1}{n-1}$ to 1.0 (i.e. $\mathbf{\Sigma}$ is circular). The smaller it is, the more complex the variance-covariance matrix usually is (see below and in Table 5.1).

Scenario T1: The three variance-covariance matrices satisfy the circularity condition, as they are identity matrices (i.e. there is no correlation and no heteroscedasticity): $\mathbf{U}_1 = \mathbf{I}_5$, $\mathbf{U}_2 = \mathbf{I}_3$, $\mathbf{U}_3 = \mathbf{I}_4$.

Scenario T2: $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ satisfy the circularity condition because they are compound symmetric; in other words, they are characterized by intra-class correlation (i.e. there is

constant correlation and no heteroscedasticity): $\mathbf{U}_1 = 3.0\,[(1-\frac{1}{6})\,\mathbf{I}_5 + \frac{1}{6}\,\mathbf{J}_5]$,

$\mathbf{U}_2 = 4.0\,[(1-\frac{1}{4})\,\mathbf{I}_3 + \frac{1}{4}\,\mathbf{J}_3]$, $\mathbf{U}_3 = 5.0\,[(1-\frac{2}{5})\,\mathbf{I}_4 + \frac{2}{5}\,\mathbf{J}_4]$.

Scenario T3: The three variance-covariance matrices are still circular but not of the compound symmetric type:

$$\mathbf{U}_1 = \begin{bmatrix} 5.00 & 3.00 & 2.50 & 2.00 & 1.50 \\ 3.00 & 4.00 & 2.00 & 1.50 & 1.00 \\ 2.50 & 2.00 & 3.00 & 1.00 & 0.50 \\ 2.00 & 1.50 & 1.00 & 2.00 & 0.00 \\ 1.50 & 1.00 & 0.50 & 0.00 & 1.00 \end{bmatrix}, \quad \mathbf{U}_2 = \begin{bmatrix} 3.00 & 1.50 & 1.00 \\ 1.50 & 2.00 & 0.50 \\ 1.00 & 0.50 & 1.00 \end{bmatrix},$$

$$\mathbf{U}_3 = \begin{bmatrix} 4.50 & 3.50 & 3.00 & 2.50 \\ 3.50 & 3.50 & 2.50 & 2.00 \\ 3.00 & 2.50 & 2.50 & 1.50 \\ 2.50 & 2.00 & 1.50 & 1.50 \end{bmatrix}.$$

Scenario T4: More complex, non-circular positive-definite variance-covariance matrices were generated artificially. Covariances do not necessarily decrease from the diagonal, and can be positive or negative. Heteroscedasticity (heterogeneity of the variance) is intermediate or stronger (i.e. ratio of 1 to 10 or 20):

$$\mathbf{U}_1 = \begin{bmatrix} 0.5528 & 0.8676 & 0.8938 & 0.1265 & -0.3505 \\ 0.8676 & 1.8519 & 1.5800 & 0.3881 & -0.4953 \\ 0.8938 & 1.5800 & 1.8656 & 0.3798 & -0.3307 \\ 0.1265 & 0.3881 & 0.3798 & 0.2375 & -0.0176 \\ -0.3505 & -0.4953 & -0.3307 & -0.0176 & 0.3947 \end{bmatrix},$$

$$\mathbf{U}_2 = \begin{bmatrix} 3.4397 & -1.0588 & -0.4014 \\ -1.0588 & 0.3455 & 0.0809 \\ -0.4014 & 0.0809 & 0.1598 \end{bmatrix}, \quad \mathbf{U}_3 = \begin{bmatrix} 0.5781 & -0.6113 & -0.0103 & 0.4031 \\ -0.6113 & 0.9624 & -0.2235 & -0.1695 \\ -0.0103 & -0.2235 & 0.2611 & -0.3606 \\ 0.4031 & -0.1695 & -0.3606 & 0.8857 \end{bmatrix}.$$

In the next four scenarios, combinations of variance-covariance matrices from Scenarios 1–4 are used, so Scenarios T5–T8 are not described in Table 5.1.

Scenario T5: $\mathbf{U}_1$ (Scenario T2), $\mathbf{U}_2$ (Scenario T1), $\mathbf{U}_3$ (Scenario T1). $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2 = 24.29$, Box's 'epsilon' = 0.9028.

Scenario T6: $\mathbf{U}_1$ (Scenario T1), $\mathbf{U}_2$ (Scenario T2), $\mathbf{U}_3$ (Scenario T4). $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2 = 31.19$, Box's 'epsilon' = 0.4532.

Scenario T7: $\mathbf{U}_1$ (Scenario T4), $\mathbf{U}_2$ (Scenario T3), $\mathbf{U}_3$ (Scenario T2). $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2 = 230.54$, Box's 'epsilon' = 0.1576.

Scenario T8: $\mathbf{U}_1$ (Scenario T3), $\mathbf{U}_2$ (Scenario T4), $\mathbf{U}_3$ (Scenario T3). $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2 = 450.02$, Box's 'epsilon' = 0.0444.

In Scenarios T1–T8 above, $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ are 5 × 5, 3 × 3, 4 × 4, respectively. Therefore, we completed a second simulation study for another group of scenarios, in which $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ are identity matrices but their size varies, with $n_1 = 2,...,5$, $n_2 = 2,...,5$ and $n_3 = 2,...,5$, for K = 5 and K = 10. The objective was to assess the effects of varying the numbers of levels for the three dimensions (e.g. responses, space-time) on the empirical bias and dispersion of $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$.

5.5.2 Procedures and methods

The number of simulation runs was set at 5.0*E*4, after preliminary trials showed that such a number was sufficient to obtain stable values of the empirical bias and measure of dispersion for the combinations of K and n_1, n_2, n_3 considered. Observations from $\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)$ were generated using $\text{vec}(\mathbb{X}) = \text{vec}(\mathbb{M}) + \mathbf{C}^T \text{vec}(\mathbb{Z})$, with $m_{i_1 i_2 i_3} = i_1 + i_2 + i_3 - 1 + i_1 i_2 i_3$ for the mean tensor $\mathbb{M}$; $\mathbf{C}$, the Cholesky root of $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$; and $\mathbb{Z} \sim N_{n_1, n_2, n_3}(\mathbb{O}; \mathbf{I}_{n_1}, \mathbf{I}_{n_2}, \mathbf{I}_{n_3})$, where $\mathbb{O}$ is the zero tensor. Simulations, array manipulation and matrix algebra calculations were performed with Matlab (The MathWorks, 2010). The existence of ML estimators $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ was ensured by the use of values of K satisfying equation (5.15), given n_1, n_2, n_3. The normal pseudo-random number generator was the Ziggurat algorithm (Marsaglia and Tsang 2000).

The standardized empirical bias was calculated as $\frac{\| \overline{\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1} - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}$, where the bar indicates that the Kronecker product

of ML estimates $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$, $\hat{\mathbf{U}}_3$ is averaged over the 5.0*E*4 simulation runs. Note that unlike the scalar case, such a measure of bias cannot be negative because it is based on the distance between two matrices, using the Euclidean norm. (For variance estimates (i.e. diagonal entries of $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$), we calculated the conventional empirical bias, which is not restricted to be non-negative, and this was found to be negative as expected in small samples for ML estimates of variances.) A standardized empirical measure of dispersion was defined as $\overline{\frac{\| \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}}$, where the bar indicates that the Euclidean norm of the difference between estimated and theoretical Kronecker products is averaged over the 5.0*E*4 simulation runs. The standardization ensures that results can be compared among scenarios, even though the norms of variance-covariance matrices vary by several orders of magnitude.

5.5.3 Results

Our simulation results are in accordance with the theoretical properties of ML estimators for variance-covariance matrices (Kendall and Stuart 1967), whether the initial solutions for the MLE-3D algorithm were identity matrices or the true parameter matrices and whether the convergence criterion was defined on each of the three variance-covariance matrices or on their Kronecker product. More specifically, the standardized empirical bias and measure of dispersion computed for $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ decreased monotonically towards zero as *K* increased (see Tables 5.2 and 5.3 for results obtained with identity matrices as initial solutions and the separate convergence criterion). The use of true parameter matrices as initial solutions decreased the number of iterations required to reach convergence by 10–20% when $K \leq 20$; this is the only difference worth reporting, as other differences regarding the ML estimates themselves and the empirical bias and dispersion were minor. In other words, we may conclude that $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ is a consistent estimator of $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$.

Standardized empirical bias values for *K* ranging from 5 to 500 were similar for all eight scenarios (Table 5.2), despite the differences in Box's 'epsilon' values. The mean standardized empirical measure of dispersion was the smallest, with a large standard error, for Scenarios T4 and T8 (Table 5.3). The average number of iterations

required to reach convergence decreased with increasing sample size; for example, it decreased from 25 for $K = 5$ to less than 5 when $K = 500$ in the simulations performed for Tables 5.2 and 5.3. Slightly more iterations were required to reach convergence in Scenarios T4 and T8, for which Box's 'epsilon' was close to the theoretical minimum. Overall, the choice of identity matrices as initial solutions for the MLE-3D algorithm can be considered appropriate in most future applications. Of course, the use of different initial solutions and the comparison of the corresponding final solutions remain a safe way to proceed.

Concerning the effects of size of the variance-covariance matrices on the properties of ML estimators (Figure 5.1), the standardized empirical bias and measure of dispersion are higher for $K = 5$ than for $K = 10$ as expected, and show a reverse relationship for a given sample size. Increasing n_1, n_2, n_3 results, in fact, in an increase of the bias and a decrease of the measure of dispersion; more specifically, the highest bias in our second simulation study is observed for $n_1 = n_2 = n_3$ 5, where the measure of dispersion is the lowest. Furthermore, the ranges of values of the empirical bias and dispersion are narrower with $K = 10$ than with $K = 5$. In summary, increasing the number of levels per dimension, without increasing the sample size, results in an increased bias combined with a decreased dispersion for the estimated Kronecker product in the case of the tensor normal distribution of order 3.

In Appendix D, the bias of the ML estimator of the doubly separable variance-covariance matrix is decomposed, as an extension of the bias analysis performed for the simply separable variance-covariance matrix in Chapter 3. The existence of a 'peak-trough' pattern is established therein.

5.6 Example with real dataset

We used 3-D spatial data of glucose measures in 15 regions of the brain, defined by their centroid coordinates (x, y, z) in cm, to illustrate the application of the MLE-3D algorithm; these data were collected by Tyler et al. (1988) and presented in Worsley et al. (1991). In our example, the glucose measure at position (x, y, z) in one of the two hemispheres of the brain is projected on the vertical y–z plane, which is divided into 2×3 sections. Six relatively well-aligned regions corresponding to these sections were selected: posttemporal, occipital, precentral, hippocampus, midfrontal, and caudate

(Worsley et al. 1991, Table 1). The same selection procedure was followed for the other hemisphere. Spatial heteroscedasticity and auto- and cross-correlations (auto: within a hemisphere in 2-D; cross: between hemispheres) are likely to exist, as different parts of the brain have specific roles and do not function independently. This provided one $2 \times 3 \times 2$ tensor of data per individual.

In the original study, there were two experimental groups, with 10 individuals each ($K_1 = K_2 = 10$), and brain glucose measures were made with slightly different methods in the two groups. Following Worsley et al. (1991), we used different means for the two groups but assumed their variance-covariance structure was the same, and we log-transformed the data. Then, the residuals of the log-transformed data were pooled, so $K = 20 > n_1 n_2 n_3 = 12$ and tests of multivariate normality became possible. Provided multivariate normality is accepted (see below), the minimum sample size required for the existence of solutions to the MLE-3D algorithm is two in this case; see equation (5.15) in Subsection 5.4.2.

With $n_1 = 2$, $n_2 = 3$, $n_3 = 2$, and $K = 20$, univariate and multivariate tests of normality were performed at $\alpha = 0.05$ on the residuals of the log-transformed data, with the SAS 9.1 macro multnorm. For each of the 12 glucose measure variables taken separately, the null hypothesis of normal distribution was not rejected by the univariate test. Multivariate normality tests were then performed on the 12×1 data vectors, using $\hat{\boldsymbol{\Sigma}} = \frac{1}{K} \sum_{k=1}^{K} \left\{ \mathrm{vec}(\mathbb{X}_k) - \mathrm{vec}(\bar{\mathbb{X}}) \right\} \left\{ \mathrm{vec}(\mathbb{X}_k) - \mathrm{vec}(\bar{\mathbb{X}}) \right\}^T$ as variance-covariance matrix estimate.

Mardia's (Mardia 1974) test based on multivariate skewness ($P = 0.6221$) and the Henze-Zirkler's (Henze and Zirkler 1990) test ($P = 0.1648$) both supported multivariate normality; only Mardia's (Mardia 1974) test based on multivariate kurtosis ($P = 0.0294$) rejected it at $\alpha = 0.05$. Therefore, we proceeded to ML estimation under multivariate normality, with vs. without separability of the variance-covariance structure (with: under the tensor normal distribution model; without: under the vector normal distribution model).

After 17 iterations, starting with $\hat{\mathbf{U}}_2^0 = \mathbf{I}_3$, $\hat{\mathbf{U}}_3^0 = \mathbf{I}_2$ and using $\epsilon_1 = \epsilon_2 = \epsilon_3 = 1.0E-6$, the MLE-3D algorithm converged, and the following estimates were obtained:

$$\hat{\mathbf{U}}_1 = \begin{bmatrix} 0.0136 & 0.0067 \\ 0.0067 & 0.0160 \end{bmatrix} \text{ (along } y\text{)}, \quad \hat{\mathbf{U}}_2 = \begin{bmatrix} 0.9681 & 0.4525 & 0.6306 \\ 0.4525 & 1.0946 & 0.4812 \\ 0.6306 & 0.4812 & 0.9995 \end{bmatrix} \text{ (along } z\text{)},$$

$$\hat{\mathbf{U}}_3 = \begin{bmatrix} 1.2623 & 0.9680 \\ 0.9680 & 1.1344 \end{bmatrix} \text{ (between hemispheres), and}$$

$\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 =$

$$\begin{bmatrix}
0.0166 & 0.0082 & 0.0078 & 0.0039 & 0.0108 & 0.0054 & 0.0128 & 0.0063 & 0.0060 & 0.0030 & 0.0083 & 0.0041 \\
0.0082 & 0.0195 & 0.0039 & 0.0091 & 0.0054 & 0.0127 & 0.0063 & 0.0150 & 0.0030 & 0.0070 & 0.0041 & 0.0097 \\
0.0078 & 0.0039 & 0.0188 & 0.0093 & 0.0083 & 0.0041 & 0.0060 & 0.0030 & 0.0144 & 0.0071 & 0.0063 & 0.0031 \\
0.0039 & 0.0091 & 0.0093 & 0.0220 & 0.0041 & 0.0097 & 0.0030 & 0.0070 & 0.0071 & 0.0169 & 0.0031 & 0.0074 \\
0.0108 & 0.0054 & 0.0083 & 0.0041 & 0.0172 & 0.0085 & 0.0083 & 0.0041 & 0.0063 & 0.0031 & 0.0132 & 0.0065 \\
0.0054 & 0.0127 & 0.0041 & 0.0097 & 0.0085 & 0.0201 & 0.0041 & 0.0097 & 0.0031 & 0.0074 & 0.0065 & 0.0154 \\
0.0128 & 0.0063 & 0.0060 & 0.0030 & 0.0083 & 0.0041 & 0.0150 & 0.0074 & 0.0070 & 0.0035 & 0.0097 & 0.0048 \\
0.0063 & 0.0150 & 0.0030 & 0.0070 & 0.0041 & 0.0097 & 0.0074 & 0.0175 & 0.0035 & 0.0082 & 0.0048 & 0.0114 \\
0.0060 & 0.0030 & 0.0144 & 0.0071 & 0.0063 & 0.0031 & 0.0070 & 0.0035 & 0.0169 & 0.0084 & 0.0074 & 0.0037 \\
0.0030 & 0.0070 & 0.0071 & 0.0169 & 0.0031 & 0.0074 & 0.0035 & 0.0082 & 0.0084 & 0.0198 & 0.0037 & 0.0087 \\
0.0083 & 0.0041 & 0.0063 & 0.0031 & 0.0132 & 0.0065 & 0.0097 & 0.0048 & 0.0074 & 0.0037 & 0.0154 & 0.0076 \\
0.0041 & 0.0097 & 0.0031 & 0.0074 & 0.0065 & 0.0154 & 0.0048 & 0.0114 & 0.0037 & 0.0087 & 0.0076 & 0.0181
\end{bmatrix}.$$

By comparison, the variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ estimated under the vector normal distribution model is equal to

$$\begin{bmatrix}
0.0311 & 0.0288 & 0.0252 & 0.0238 & 0.0261 & 0.0266 & 0.0269 & 0.0269 & 0.0242 & 0.0259 & 0.0262 & 0.0261 \\
0.0288 & 0.0337 & 0.0256 & 0.0220 & 0.0235 & 0.0257 & 0.0262 & 0.0318 & 0.0219 & 0.0257 & 0.0237 & 0.0245 \\
0.0252 & 0.0256 & 0.0282 & 0.0213 & 0.0256 & 0.0235 & 0.0231 & 0.0234 & 0.0239 & 0.0211 & 0.0245 & 0.0235 \\
0.0238 & 0.0220 & 0.0213 & 0.0261 & 0.0216 & 0.0219 & 0.0218 & 0.0212 & 0.0205 & 0.0247 & 0.0213 & 0.0224 \\
0.0261 & 0.0235 & 0.0256 & 0.0216 & 0.0339 & 0.0245 & 0.0225 & 0.0225 & 0.0228 & 0.0204 & 0.0297 & 0.0259 \\
0.0266 & 0.0257 & 0.0235 & 0.0219 & 0.0245 & 0.0265 & 0.0241 & 0.0238 & 0.0219 & 0.0229 & 0.0243 & 0.0254 \\
0.0269 & 0.0262 & 0.0231 & 0.0218 & 0.0225 & 0.0241 & 0.0261 & 0.0251 & 0.0211 & 0.0233 & 0.0243 & 0.0231 \\
0.0269 & 0.0318 & 0.0234 & 0.0212 & 0.0225 & 0.0238 & 0.0251 & 0.0326 & 0.0196 & 0.0243 & 0.0219 & 0.0221 \\
0.0242 & 0.0219 & 0.0239 & 0.0205 & 0.0228 & 0.0219 & 0.0211 & 0.0196 & 0.0237 & 0.0209 & 0.0219 & 0.0229 \\
0.0259 & 0.0257 & 0.0211 & 0.0247 & 0.0204 & 0.0229 & 0.0233 & 0.0243 & 0.0209 & 0.0297 & 0.0217 & 0.0235 \\
0.0262 & 0.0237 & 0.0245 & 0.0213 & 0.0297 & 0.0243 & 0.0243 & 0.0219 & 0.0219 & 0.0217 & 0.0294 & 0.0250 \\
0.0261 & 0.0245 & 0.0235 & 0.0224 & 0.0259 & 0.0254 & 0.0231 & 0.0221 & 0.0229 & 0.0235 & 0.0250 & 0.0267
\end{bmatrix}.$$

Thereafter, the null hypothesis of separability of the 3-D spatial variance-covariance structure, $\text{var}\{\text{vec}(\mathbb{X})\} = \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$ or

$\text{var}\{\text{vec}(\mathbb{X})\}^{-1}(\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) = \mathbf{I}_{12}$, was tested against the alternative, $\text{var}\{\text{vec}(\mathbb{X})\}$ is positive definite, using an unbiased modified likelihood ratio statistic of double separability, $\Lambda^* = (K-p)[n_1 n_2 \log|\mathbf{U}_3| + n_1 n_3 \log|\mathbf{U}_2| + n_2 n_3 \log|\mathbf{U}_1| - \log|\boldsymbol{\Sigma}_{UN}|]$, where p = 6.68 is a penalty, resulting in an observed value $\Lambda^* = 121.20$ here (Appendix E, Table E.1). The critical value is $\chi^2_{0.95}(67) = 87.1081$. In conclusion, the null hypothesis of separability is rejected, which confirms the visual inspection of the two 12×12 estimated variance-covariance matrices above; most covariances in the former are smaller than 0.01, whereas most covariances in the latter are greater than 0.02.

5.7 Discussion

We have given a detailed presentation of the MLE-3D algorithm for maximum likelihood estimation of the variance-covariance parameters of the tensor normal distribution of order 3, $\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$, with $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ unstructured. After determining the minimum sample size, we computed by simulation the empirical bias and dispersion of the estimated Kronecker product of variance-covariance matrices in a number of scenarios. As expected in the case of ML estimators, the standardized empirical bias and measure of dispersion were found to be decreasing with increasing sample size.

Recently, there has been a debate on the definition of ML estimators and their unicity for separable variance-covariance structures in 2-D. On the one hand, Werner et al. (2008) compared the MLE-2D (alias "flip-flop") algorithm to four alternative estimation procedures, and found it was providing estimators with the lowest normalized root-MSE already with very small samples. On the other hand, Srivastava et al. (2008) preferred to work by setting the bottom-right or top-left entry of one of the two variance-covariance matrices at 1.0. As pointed out by Lee et al. (2010) in their comment in response to the interpretation that had been made of their own results, Srivastava et al. (2008) worked in the curved exponential family instead of the regular exponential family, and presented theoretical results without simulations. This kind of discussion could be extended to the 3-D case, with a similar comment eventually.

In our example with a real dataset, the variance-covariance structure was assumed to be the same in both groups. It would have been better to test the assumption of homogeneity of the variance-covariance matrix between groups in a preliminary step, but

this was not possible because the sample sizes ($K_1 = K_2 = 10$) were smaller than $n_1 n_2 n_3 = 12$. Nevertheless, the MLE-3D algorithm can be used with datasets composed of tensors of order 3 collected for different levels of a treatment provided the number of replicates is sufficient. After centering each data tensor to the corresponding group mean, multivariate normality can be assessed, and if accepted, the homogeneity and separability of the variance-covariance structure can be tested in turn. When the three conditions are satisfied, the group means (mean tensors for the groups) can be compared with a modified *F*-test in which the autocorrelation and heteroscedasticity in 3-D are taken into account through an estimate of Box's 'epsilon' (Box 1954a, 1954b); see Dutilleul (1998) for examples in 2-D and 3-D.

The collection of data tensors of order 4 may become more common in a near future. For example, consider the non-destructive measurement by computed tomography scanning of several wood properties (e.g. density, fiber length) in several growth rings and directions at different heights in a tree trunk. Therefore, an extension of the MLE algorithm to four dimensions (MLE-4D), under multivariate normality and four-way separability of the variance-covariance structure, is presented in Appendix F, together with simulation results. Although the MLE-5D algorithm may be seen as the next obvious and straightforward extension technically speaking, one cannot do otherwise but question the scope of its applications, which seems to be, at least at this moment, more limited than the 2-D, 3-D and 4-D versions.

5.8 Acknowledgements

The financial assistance of the Natural Sciences and Engineering Research Council of Canada (NSERC), through an Alexander Graham Bell Canada Graduate Scholarship to Mr. Manceur and a Discovery Grant (Mathematics and Statistics) to Prof. Dutilleul, is gratefully acknowledged. Special thanks go to Dr. Guillaume Larocque and Mr. Liwen Han for help with simulations and general assistance.

Table 5.1: Characteristics of the variance-covariance matrices in simulation scenarios T1–T4.

Scenario		$\mathbf{U}_1(5\times5)$	$\mathbf{U}_2(3\times3)$	$\mathbf{U}_3(4\times4)$	$\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1$
T1	Euclidean norm	2.2361	1.7321	2.0	7.7560
	Box's 'epsilon' [a]	1.0	1.0	1.0	1.0
T2	Euclidean norm	7.0711	7.3485	12.1655	632.1392
	Box's 'epsilon'	1.0	1.0	1.0	0.5497
T3	Euclidean norm	10.7238	4.5826	11.0	540.5691
	Box's 'epsilon'	1.0	1.0	1.0	0.1439
T4	Euclidean norm	4.1354	3.8150	1.8982	29.9460
	Box's 'epsilon'	0.4825	0.5381	0.6436	0.0510

[a] A value of 1.0 means that the variance-covariance matrix is circular, while a value close to the minimum ($\frac{1}{n_j-1}$, j = 1, 2, 3, or $\frac{1}{n_1 n_2 n_3 - 1}$) indicates a strong deviation from circularity.

Table 5.2: Standardized empirical bias[a] as a function of the sample size (K) and the simulation scenario (T1–T8) (see text and Table 5.1 for the definition of the eight scenarios).

Sample size	Scenario							
K	T1	T2	T3	T4	T5	T6	T7	T8
5	0.1848	0.1847	0.1839	0.1838	0.1844	0.1849	0.1844	0.1836
10	0.09687	0.09733	0.09679	0.09648	0.09693	0.09680	0.09696	0.09675
15	0.06539	0.06592	0.06622	0.06558	0.06542	0.06584	0.06576	0.06608
20	0.04940	0.04934	0.04840	0.04843	0.04923	0.04943	0.04854	0.04841
100	0.01004	0.01008	0.01022	0.01022	0.01007	0.01010	0.01014	0.01025
500	0.00202	0.00206	0.00212	0.00213	0.00205	0.00205	0.00210	0.00213

[a] $\overline{\dfrac{\| \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}}$

Table 5.3: Standardized empirical measure of dispersion[a] with the standard error below (in parentheses), as a function of the sample size (K) and the simulation scenario (T1–T8) (see text and Table 5.1 for the definition of the eight scenarios).

Sample Size	Scenario							
K	T1	T2	T3	T4	T5	T6	T7	T8
5	0.4824 (0.00030)	0.4447 (0.00074)	0.3316 (0.00056)	0.3251 (0.00056)	0.4724 (0.00030)	0.4433 (0.00032)	0.3546 (0.00049)	0.3173 (0.00060)
10	0.3225 (0.00018)	0.2973 (0.00022)	0.2169 (0.00039)	0.2116 (0.00039)	0.3156 (0.00019)	0.2965 (0.00021)	0.2330 (0.00035)	0.2064 (0.00042)
15	0.2594 (0.00014)	0.2389 (0.00018)	0.1723 (0.00032)	0.1676 (0.00031)	0.2539 (0.00016)	0.2380 (0.00017)	0.1855 (0.00028)	0.1638 (0.00034)
20	0.2230 (0.00013)	0.2051 (0.00015)	0.1466 (0.00027)	0.1427 (0.00027)	0.2181 (0.00013)	0.2046 (0.00015)	0.1581 (0.00024)	0.1390 (0.00029)
100	0.09796 (5.75*E*-5)	0.08986 (6.89*E*-5)	0.06329 (0.00012)	0.06147 (0.00012)	0.09574 (5.98*E*-5)	0.08970 (6.67*E*-5)	0.06846 (0.00011)	0.05997 (0.00013)
500	0.04364 (2.58*E*-5)	0.04004 (3.10*E*-5)	0.02810 (5.36*E*-5)	0.02726 (5.31*E*-5)	0.04266 (2.67*E*-5)	0.03993 (2.97*E*-5)	0.03045 (4.83*E*-5)	0.0265 (5.69*E*-5)

[a] $\overline{\dfrac{\| \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}}$

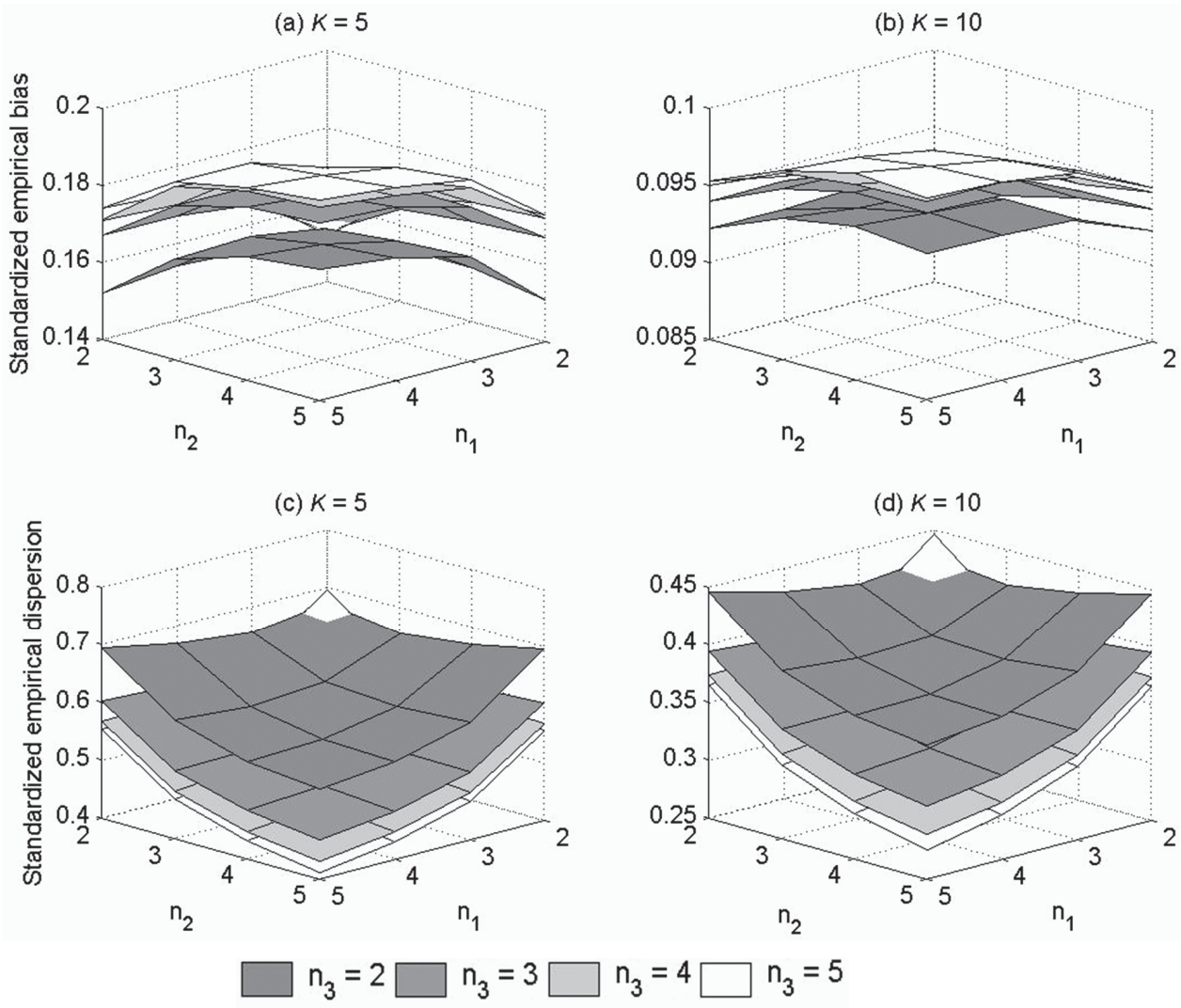


Figure 5.1: Standardized empirical bias [panels (a) and (b)] and measure of dispersion [panels (c) and (d)] for the Kronecker product of ML estimators, $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$, with $\mathbf{U}_1 = \mathbf{I}_{n_1}$, $\mathbf{U}_2 = \mathbf{I}_{n_2}$, $\mathbf{U}_3 = \mathbf{I}_{n_3}$, as a function of n_1, n_2, n_3, and sample size K (results of our second simulation study).

Preface to Chapter 6

The application of the tensor and matrix normal distribution models and the companion MLE-3D and MLE-2D algorithms and unbiased modified LRT tests of double and simple separability is the object of study in Chapter 6. It is performed on an original dataset of wood density estimates in two growth rings, two directions, and two heights, obtained from CT scanning data (with details of calibration process in Appendix G). The null hypothesis of double separability of the variance-covariance structure, tested with the unbiased modified LRT, is accepted. A triply repeated measures ANOVA shows significant variation in mean wood density with height and direction. This application presents the statistical toolbox developed for multi-dimensional data analysis in a less mathematical manner, while demonstrating the potential of the use of the statistical tools in the environmental sciences. As of May 2012, the manuscript (co-authored by Mr. Manceur, Dr. Beaulieu, Mr. Han, and Prof. Dutilleul) derived from this chapter is in press in the *Canadian Journal of Forest Research* (doi: 10.1139/x2012-053). Dr. Beaulieu furnished the prepared wood samples and contributed to the development of the approach. Mr. Han operated the computed tomography scanner and contributed to data analysis. Mr. Manceur wrote the initial manuscript. All authors participated in writing the final manuscript.

Chapter 6. A multi-dimensional statistical model for wood data analysis, with density estimated from CT scanning data as an example

6.1 Abstract

The trunk of a tree can be seen as a spatio-temporal sampling domain from the statistical perspective, where space is represented by direction horizontally and height vertically, and time through annual growth rings. In this framework, wood properties such as density can be the object of data collection for given estimation and testing purposes. We present a multi-dimensional statistical model, the tensor normal distribution, in which the variation (variance) of and dependency (covariance) between wood property measurements made for different years at various locations in a tree trunk can be inferred. Its application requires a smaller number of replicates (trees) than the traditional vector normal distribution because variances and covariances for directions and growth rings, for example, must be the same at all heights, up to a multiplicative constant. This assumption on the variance-covariance structure is called "separability", and we explain how to test it. An illustration with wood density estimates obtained from computed tomography scanning data for 11 white spruce (*Picea glauca* (Moench) Voss) trees is presented. This example is completed by assessing differences in mean wood density according to location in the trunk, with analysis-of-variance *F*-tests adjusted for the estimated variances and covariances obtained by fitting the model.

6.2 Introduction

The trunk of a tree may be conceptualized in a variety of ways, depending on the perspective. From the statistical perspective, it represents a spatio-temporal sampling domain, with height vertically and direction horizontally for space and annual growth rings for time (Figure 6.1a). As part of the tree growth process, some level of covariation (association) between measurements made for a given wood property may be expected inside the trunk, as cambial initials are always formed in the previous growing season (Fritts 1976). The analysis of variances (variation) and covariances (dependency) is a preliminary step for the analysis of autocorrelation, the property of correlation of a variable with itself at different times in different spatial locations. The variance-covariance structure must also be identified and the corresponding variance-covariance matrix must be estimated in order to perform valid tests of significance for mean values

in a multi-dimensional context, as the one of the spatio-temporal example of tree trunk. Modeling the variance-covariance structure reduces the number of parameters to be estimated and ensures that estimation can be carried out despite a limited sample size.

In forestry, and in dendrochronology in particular, autocorrelation in time series of tree-ring widths is often modeled with an autoregressive equation (Apiolaza et al. 2000, Apiolaza and Garrick 2001). Alternatively, shorter time series (also called longitudinal data) can be analyzed by repeated measures ANOVA (analysis of variance; Crowder and Hand 1990). In both cases, the observed time series is a vector of observations, or a one-dimensional statistical object. Multi-dimensional datasets such as matrix or tensor datasets can then be seen as extensions in 2-D and in 3-D or more dimensions (Figure 6.1b).

Assume, for example, that wood cores (from pith to bark) were collected at several heights on one side of a tree trunk, and a measure of density was made for each growth ring along the core available at each height. The resulting data constitute spatio-temporal repeated measures on the tree, with year as time factor and height as space factor. If this data collection procedure is reproduced for several trees with same age and approximately same total height, a variance and a covariance can be estimated for each spatio-temporal sampling site and each pair of sites inside the trunk, respectively. Spatio-temporal datasets are common in forestry, but are rarely analyzed in a multi-dimensional statistical framework based on a distributional model. Tian et al. (1995), Wilhelmsson et al. (2002), Molteberg and Høibø (2007), and Via et al. (2007), among others, collected 2-D data for wood properties on tree trunks but did not take into account the inherent spatio-temporal autocorrelation in their analyses. Jordan et al. (2005) and Antony et al. (2010) for their part did take autocorrelation into account but in one of the dimensions only, and Herman et al. (1999) accounted for autocorrelation in micro-fibril angles within a ring and among rings in different types of analysis, due to sample size limitation.

Three-dimensional datasets are less common, are often obtained with non-destructive techniques, and are very rarely analyzed with a statistical method that exploits the full dimensionality and very nature of the data. Gjerdum and Bernabei (2009) modeled grain angle in 3-D for Norway spruce wood, and stated that the interpretation of statistical analyses was limited by the lack of modeling of autocorrelation; Defo et al.

(2009) used SilviScan-3 and imaging algorithms to generate a 3-D image of the distribution of wood density values; and Pont et al. (2007) presented a literature review, and developed measurement tools and imaging algorithms to visualize wood properties in 3-D. To the best of our knowledge, only Dutilleul (1998) used a multi-dimensional normal distribution model to analyze wood properties in a tree trunk in 3-D.

Modern technologies allow the collection and visualization of multi-dimensional datasets in forestry, but appropriate statistical methods for their analysis are handful. As mentioned above, modeling the variance-covariance structure in multi-dimensional data should be a pre-requisite to testing hypotheses about mean values, while such modeling is of primary interest to understand key physiological processes such as tree growth. Furthermore, CT scanning and other non-destructive techniques that can be used to measure wood properties generate extremely large datasets. For all those reasons, we built up on the vector normal distribution model and the repeated measures ANOVA and their two-dimensional version, and extended the distributional model and the companion statistical method to three dimensions. The 2-D and 3-D extensions are not trivial mathematically speaking, and new estimation algorithms had to be developed (Dutilleul 1999; Chapter 5). This chapter is mainly aimed at providing a simple presentation of the multi-dimensional statistical model called "tensor normal distribution" in the context of forestry, to illustrate its use and perform statistical inference (estimation and testing) under this model using wood density estimates obtained from CT scanning data for white spruce (*Picea glauca* (Moench) Voss). A specific objective is to present a ML estimation algorithm and a LRT for the variance-covariance structures which characterize the matrix (2-D) and tensor (3-D) normal distribution models. Secondarily, a link and a brief comparison are made between wood density estimates obtained from CT scanning data and those provided by SilviScan (Evans et al. 1995; Keunecker et al. 2009).

To achieve these objectives in section 6.3, the tensor and matrix normal distribution models are first defined, together with the corresponding MLE algorithm and LRT procedure. Then, the ANOVAs for triply and doubly repeated measures are presented, including modified F-tests designed to assess differences among means in the presence of autocorrelation and heteroscedasticity (heterogeneity of the variance). In section 6.4, the white spruce wood samples and the CT scanning protocol are described.

In section 6.5, the results obtained in the new statistical framework for the white spruce 3-D data are summarized. In the section 6.6, results are interpreted and discussed in statistical and non-statistical terms, and future perspectives are outlined. In Appendix G, the calibration method used to determine the beginning and the end of annual growth rings in series of SilviScan and CT scan wood density estimates is explained.

6.3 Statistical models

6.3.1 The tensor normal distribution model

As the collection of data is upgraded from a core of wood to the tree trunk, the mathematical object describing the biological object has a larger number of dimensions. A tensor of order three (shortened into "tensor" hereafter) is a mathematical object composed of three dimensions corresponding to the row, the column and the edge; a data tensor is a 3-D array of observations or measurements. In our example, there is one data tensor per tree for a total of 11 trees (sample size $K = 11$), and each data tensor is composed of $n_1 = 2$ rows (annual growth rings: $r_1 = 1997$, $r_2 = 2003$), $n_2 = 2$ columns (direction: d_1 = North, d_2 = South) and $n_3 = 2$ edges (heights: h_1 = breast; h_2 = live crown). All those data tensors will be assumed to come from the same statistical distribution (the same tensor statistical distribution) defined below; they are observations of the same random tensor.

As the number of dimensions of the mathematical object increases, the statistical distribution model becomes multi-dimensional. The tensor normal distribution is a three-dimensional extension of the scalar normal distribution $X \sim N(m, u)$ parameterized like this in order to facilitate extensions (see below), where X is a random scalar, m is the mean parameter and u is the variance parameter. In the tensor normal distribution, each element of the random tensor $\mathbb{X}$ is characterized statistically by a mean, a variance and covariances (dependencies) with the other $n_1 n_2 n_3 - 1$ elements. This can be written as follows:

$$\mathbb{X} \sim N_{n_1, n_2, n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3), \tag{6.1}$$

where $\mathbb{X}$ is the $n_1 \times n_2 \times n_3$ random tensor, $\mathbb{M}$ is the $n_1 \times n_2 \times n_3$ mean tensor parameter, $\mathbf{U}_1$ is the $n_1 \times n_1$ variance-covariance matrix between the rows of $\mathbb{X}$ (the annual growth rings in our example), $\mathbf{U}_2$ is the $n_2 \times n_2$ variance-covariance matrix

between the columns of $\mathbb{X}$ (the directions), and $\mathbf{U}_3$ is the $n_3 \times n_3$ variance-covariance matrix between the edges of $\mathbb{X}$ (the heights); while $\mathbb{M}$ is the 3-D extension of m in the scalar case, each of the three variance-covariance matrices $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ is an extension of u and characterizes variation and dependencies in one dimension.

A tensor with n_1 rows, n_2 columns and n_3 edges can be transformed into a vector with $n_1 n_2 n_3$ components by using the "vec" operator which concatenates the $n_1 \times n_2 \times n_3$ tensor $\mathbb{X}$, column by column and edge by edge, resulting in an $n_1 n_2 n_3 \times 1$ vector. This provides a second, equivalent formulation of the tensor normal distribution model:

$$\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1), \tag{6.2}$$

where the variance-covariance matrix of $\text{vec}(\mathbb{X})$, $\text{var}[\text{vec}(\mathbb{X})]$, is doubly separable, i.e., it is constructed as the Kronecker product ($\otimes$; see Van Loan 2000) of the variance-covariance matrices for edges, columns and rows in this order, $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$.

The assumption of double separability on the variance-covariance structure may seem arcane at first glance, but the use of any model requires a number of assumptions. Here, the separability assumption means among other things, that the variation and dependencies for annual growth rings and directions remain the same, up to a multiplicative constant, from one height to the other. In other words, the interaction between the first two dimensions is maintained along the third dimension. However, it does not imply that in each dimension, variances are constant and covariances only depend on the vector linking the measurement points (stationarity). A major advantage of the constraints of separability on the variance-covariance structure is that the sample size (number of trees) required for estimating a doubly separable variance-covariance matrix is reduced.

Indeed, for ML estimation purposes, assume that (1) an i.i.d. random sample of size K from a tensor normal distribution, $\mathbb{X}_1$, …, $\mathbb{X}_K$, is available, and (2) the sample size K satisfies the condition

$$K \geq \max\left(\frac{n_1}{n_2 n_3}, \frac{n_2}{n_1 n_3}, \frac{n_3}{n_1 n_2}\right) + 1. \tag{6.3}$$

The minimum sample size for the $n_1 \times n_2 \times n_3$ tensor normal distribution is much smaller than the number of replicates required for the general $n_1 n_2 n_3 \times 1$ vector normal distribution model (see below), which is $n_1 n_2 n_3 + 1$. The ML estimation is performed with the "MLE-3D" iterative algorithm, which finds the solutions to the system of likelihood equations (i.e. a system of three equations with three unknowns: the three variance-covariance matrices, $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$; Mardia and Goodall 1993, p. 358; Chapter 5).

For testing purposes, the general vector normal distribution model, with no other constraint on the variance-covariance matrix than positive definitiveness, usually plays the role of alternative hypothesis, while the tensor normal distribution model and the doubly separable variance-covariance structure constitute the null hypothesis. Again, the vec operator can be used to write the vector normal distribution model:

$$\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \mathbf{\Sigma}_{UN}), \tag{6.4}$$

but this time, $\text{var}[\text{vec}(\mathbb{X})] = \mathbf{\Sigma}_{UN}$ is "unstructured" (UN) and hence, not separable.

If the sample size K of an i.i.d. random sample from a vector normal distribution is at least equal to the number of components plus one, then the ML estimator of the unstructured variance-covariance matrix $\mathbf{\Sigma}_{UN}$ is $(K - 1)/K$ times the sample variance-covariance matrix. If the raw data are $n_1 \times n_2 \times n_3$ tensor data and $K \geq n_1 n_2 n_3 + 1$, then the ML estimator of $\mathbf{\Sigma}_{UN}$ can be written as:

$$\hat{\mathbf{\Sigma}}_{UN} = \frac{1}{K} \sum_{k=1}^{K} \left\{\text{vec}(\mathbb{X}_k) - \text{vec}(\bar{\mathbb{X}})\right\} \left\{\text{vec}(\mathbb{X}_k) - \text{vec}(\bar{\mathbb{X}})\right\}^T \tag{6.5}$$

where $\bar{\mathbb{X}}$ denotes the arithmetic mean of $\mathbb{X}_k$ ($k = 1, \ldots, K$).

There are many other variance-covariance structures (Wolfinger 1996), besides the doubly separable and unstructured ones. At one extreme, the simplest variance-covariance structure is one with all variances being equal and all covariances being zero; there is only one variance parameter. Such variance-covariance matrices are unrealistic in forestry applications. Comparatively, two advantages of a separable variance-covariance structure are that in general, variances need not be the same everywhere in space or time, or both, and that covariances, although restricted by the model, can be different from

zero. At the other extreme is the unstructured variance-covariance structure, which has $n_1 n_2 n_3 (n_1 n_2 n_3 + 1)/2$ distinct variance-covariance parameters. The major advantage of using a separable variance-covariance structure with $n_1 (n_1 + 1)/2 + n_2 (n_2 + 1)/2 + n_3 (n_3 + 1)/2$ distinct parameters is a substantial reduction of the sample size required to estimate them (see above). Also, there are computational advantages, in the inversion of variance-covariance matrices in particular (Van Loan 2000). Prior to taking advantage of the tensor normal distribution model, the assumptions of double separability and multivariate normality need to be tested.

6.3.2 Testing model assumptions

To assess the assumptions of the tensor normal distribution model, the multivariate normality of vec($\mathbb{X}$) is tested first, using vec($\mathbb{X}_k$) (k = 1, …, K). This can be done with the "multnorm" macro in SAS 9.2 (SAS Institute Inc., Cary, USA), which combines Mardia's tests based on multivariate skewness and kurtosis (Mardia 1974) and the Henze-Zirkler test (Henze and Zirkler 1990). Secondly, after the multivariate normality of vec($\mathbb{X}$) has been accepted, the double separability of the variance-covariance matrix structure is tested:

$$H_0 : \boldsymbol{\Sigma}_{UN}^{-1} (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) = \mathbf{I}_{n_1 n_2 n_3} \text{ against } H_1 : \boldsymbol{\Sigma}_{UN}^{-1} (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) \neq \mathbf{I}_{n_1 n_2 n_3} \quad (6.6)$$

where $\mathbf{I}_{n_1 n_2 n_3}$ is the identity matrix (i.e. a square matrix with ones on the diagonal and zeroes off the diagonal). The equality under the null hypothesis H_0 is equivalent to $\boldsymbol{\Sigma}_{UN} = \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$, which is rewritten in order to have the identity matrix on the right-hand side of the equality in (6.6); this is in accordance with the usual way of writing tests of significance for variance-covariance matrices (Muirhead 1982). The test used for (6.6) is an unbiased modified LRT (Muirhead 1982, Chapter 4, Appendix E):

$$\Lambda^* = (K - p)[n_1 n_2 \log|\hat{\mathbf{U}}_3| + n_1 n_3 \log|\hat{\mathbf{U}}_2| + n_2 n_3 \log|\hat{\mathbf{U}}_1| - \log|\hat{\boldsymbol{\Sigma}}_{UN}|] \quad (6.7)$$

where p is a penalty that adjusts the distribution of the test statistic so that it fits to the theoretical chi-square distribution, log is the natural logarithm and | . | denotes the determinant of a matrix. If the sum of the three first terms is close to the last one in (6.7), then the value of the test statistic is close to zero and there is evidence for the structure of the variance-covariance matrix to be doubly separable. More formally, if the value of the

test statistic Λ^* is smaller than the critical value for a given significance level (e.g. α = 0.05), the null hypothesis H_0 is accepted and the variance-covariance matrix is considered doubly separable.

6.3.3 The matrix normal distribution model

Assume wood density is studied in several growth rings along two directions at a given height for a number of trees, resulting in 2-D datasets or data matrices; these could be 'slices' of initially 3-D datasets (Figure 6.1). In this case, there is potential for estimating the variance-covariance parameters of the matrix normal distribution model with an MLE algorithm and for testing the simple separability of the variance-covariance structure with an unbiased LRT. The matrix normal distribution model can be written as:

$$\mathbf{X} \sim N_{n_1, n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2), \tag{6.8}$$

where $\mathbf{X}$ is an $n_1 \times n_2$ random matrix, $\mathbf{M}$ ($n_1 \times n_2$) is the mean matrix, $\mathbf{U}_1$ ($n_1 \times n_1$) is the variance-covariance matrix for the n_1 rows of $\mathbf{X}$, and $\mathbf{U}_2$ ($n_2 \times n_2$) is the variance-covariance matrix for the n_2 columns of $\mathbf{X}$. The two variance-covariance matrices are estimated with an iterative algorithm called "MLE-2D", assuming a sufficient sample size: $K \geq \max(n_2/n_1, n_1/n_2) + 1$ (Dutilleul 1999).

Under the vector normal distribution model $\text{vec}(\mathbf{X}) \sim N_{n_1 n_2}(\text{vec}(\mathbf{M}), \mathbf{\Sigma}_{UN})$, where $\mathbf{\Sigma}_{UN}$ ($n_1 n_2 \times n_1 n_2$) is positive definite without further constraint, multivariate normality can be tested first, provided $K \geq n_1 n_2 + 1$. Following acceptance of multivariate normality, simple separability of the variance-covariance structure can then be tested with the modified unbiased LRT (Muirhead 1982, Chapter 4):

$$\Lambda^* = (K - p)[n_2 \log|\hat{\mathbf{U}}_1| + n_1 \log|\hat{\mathbf{U}}_2| - \log|\hat{\mathbf{\Sigma}}_{UN}|], \tag{6.9}$$

where p is a penalty that adjusts the distribution of the test statistic, and $\mathbf{\Sigma}_{UN}$ is estimated by $\frac{1}{K}\sum_{k=1}^{K}\{\text{vec}(\mathbf{X}_k) - \text{vec}(\bar{\mathbf{X}})\}\{\text{vec}(\mathbf{X}_k) - \text{vec}(\bar{\mathbf{X}})\}^T$. Critical LRT values are obtained by simulation when the sample size is moderate, and interpretation rules regarding acceptance and rejection of the null hypothesis of simple separability are the same as for double separability.

6.3.4 Analysis of variance

Estimates of modeled variance-covariance matrices, such as those of the tensor and matrix normal distribution models, can be used in a variety of statistical applications. Here, we focus on the repeated measures ANOVA to test differences among means on autocorrelated and heteroscedastic sample data as this method is commonly used in ecological studies (Potvin et al. 1990), and is an extension of the classical ANOVA method. With triply repeated measures, at least three types of hypotheses about the mean value of the random variable of interest can be tested. These hypotheses concern: the main effects of repeated measures factors considered one at a time (i.e. actual differences among means) – there are three of them; the interactions between repeated measures factors taken two by two (i.e. 'differences in differences' between means) – again, there are three of them; and the interaction for all three repeated measures factors (i.e. differences in the interaction between two factors depending on the level of the third factor) – there is only one. The triply repeated measures ANOVA model without treatment factor specifically has seven error terms for testing each of the seven main or interaction effects listed above, using trees as replicates:

$$X_{i_1 i_2 i_3 k} = m + T_k + y_{i_1} + \varepsilon^{(1)}_{i_1 k} + d_{i_2} + \varepsilon^{(2)}_{i_2 k} + h_{i_3} + \varepsilon^{(3)}_{i_3 k} + (y \times d)_{i_1 i_2} + \varepsilon^{(4)}_{i_1 i_2 k}$$
$$+ (y \times h)_{i_1 i_3} + \varepsilon^{(5)}_{i_1 i_3 k} + (d \times h)_{i_2 i_3} + \varepsilon^{(6)}_{i_2 i_3 k} + (y \times d \times h)_{i_1 i_2 i_3} + \varepsilon^{(7)}_{i_1 i_2 i_3 k} \quad (6.10)$$

where $X_{i_1 i_2 i_3 k}$ denotes the wood density for tree k = 1, …, 11 in growth ring i_1 = 1997, 2003 along direction i_2 = North, South at height i_3 = breast, live crown; y_{i_1}, d_{i_2} and h_{i_3} are the main effects of year, direction and height (all fixed); $(y \times d)_{i_1 i_2}$, $(y \times h)_{i_1 i_3}$, $(d \times h)_{i_2 i_3}$ and $(y \times d \times h)_{i_1 i_2 i_3}$ are the year-by-direction, year-by-height, direction-by-height and year-by-direction-by-height interaction effects (all fixed); T_k denotes the random tree effect; and the remaining terms (all random) are the seven error terms.

From the triply repeated measures ANOVA model (6.10), a number of ANOVA models for the analysis of doubly repeated measures can be built by removing the main

and interaction effects of the factor for which the number of levels is reduced to one. For example, the doubly repeated measures ANOVA model to be used at a given height is:

$$X_{i_1 i_2 k} = m + T_k + y_{i_1} + \varepsilon'^{(1)}_{i_1 k} + d_{i_2} + \varepsilon'^{(2)}_{i_2 k} + (y \times d)_{i_1 i_2} + \varepsilon'^{(3)}_{i_1 i_2 k} . \qquad (6.11)$$

Note that with simply repeated measures and no treatment factor, observations would be indexed by only two subscripts and the model would stop after the first error term.

Whether with simply, doubly or triply repeated measures, classical (unmodified) ANOVA *F*-tests assume statistical independence and homogeneity of the variance for the random variable of interest at different levels of the classification factor(s). This is unlikely to be the case in wood data analysis; it suffices to think of the temporal autocorrelation and heteroscedasticity of tree-ring width, wood density and fibre length (Dutilleul et al. 1998). It follows that the actual Type I error risk (i.e. rate of rejection of the null hypothesis while, in fact, it is true) will generally be different from the nominal significance level set by the experimenter, if the classical testing procedure is used.

Box's 'epsilon' (Box 1954a, 1954b) provides a solution to the problem described above, in the form of modified ANOVA *F*-tests; the modification is applied to the number of d.f. of the numerator and denominator of the *F*-ratio test statistic. Box's 'epsilon' is a measure of the deviation of a variance-covariance matrix from circularity, which is the most general necessary and sufficient condition for valid unmodified *F*-tests in the ANOVA method. This condition states that an $n \times n$ variance-covariance matrix $\mathbf{\Sigma}$ (separable or not) is circular if $\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C} = \lambda\, \mathbf{I}_{n-1}$, with λ a positive scalar and $\mathbf{C}$ an $n \times (n-1)$ matrix of orthonormal contrasts (i.e. particular linear transformations of raw data; the coefficients defining contrasts as linear combinations of raw data add up to zero). In other words, circular variance-covariance matrices present forms of heterogeneity of variance and non-null covariances for raw data such that the classical ANOVA *F*-tests remain valid, because independence and homoscedasticity for orthonormal contrasts really matter. Box's 'epsilon' is calculated as:

$$\text{Box's 'epsilon'} = \frac{\{\mathrm{tr}(\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C})\}^2}{(n-1)\,\mathrm{tr}\{(\mathbf{C}^T \mathbf{\Sigma}\, \mathbf{C})^2\}}, \qquad (6.12)$$

where tr(·) denotes the trace operator, which applies to any square matrix and returns the sum of values of its diagonal entries. The value of Box's 'epsilon' varies from $1/(n-1)$ to

1.0 (i.e. $\boldsymbol{\Sigma}$ is circular); it is exactly 1 for any 2×2 variance-covariance matrix. Box's 'epsilon' multiplies the numbers of d.f. of the numerator and denominator of the F-ratio test statistic, thus taking into account any discrepancy from circularity. In the doubly and triply repeated measures ANOVAs with simple and double separable variance-covariance structures, there are different Box's 'epsilons' for different main and interaction effects involving one or several repeated measures factors; see Dutilleul and Pinel-Alloul (1996) and Dutilleul (2011) for details about the doubly repeated measures ANOVA.

This completes the description of some of the statistical tools recommended for analyzing multi-dimensional data such as wood density measurements made at different locations in a tree trunk.

6.4 Trees and CT scanning

6.4.1 Trees

Trunk sections were obtained from white spruce trees harvested for an ecological survey conducted by the Ministère des Ressources naturelles et de la Faune du Québec in the southern Québec forest (between 48.18 and 49.91 degrees latitude North and between 62.77 and 67.77 degrees longitude West) in September-October 2006 and August 2008. The trees were sampled in different ecological regions. Our goal here is to apply the statistical models and companion data analyses in order to assess differences in mean wood density within the tree trunk; the number of sampled trees per region does not allow comparisons among ecological regions.

Samples consist of non-extracted wood sections oriented North to South, taken at two heights: breast (1.3 m) and live crown (average: 9.4 m). The average diameter of the trunk (including bark), width and thickness of the wood samples at breast height are 257 mm, 52 mm and 18 mm, and the average number of annual growth rings is 60. Corresponding values at live crown height were 157 mm, 52 mm and 18 mm, and 31. The wood samples were air-dried, and were free of apparent compression wood, cracks, and knots, where measurements with the CT scanner were made (Figure 6.2a).

6.4.2 CT scanning

The X-ray CT scanning technology is based on X-ray emission by a source in direction of an object and the measurement by detectors of the X-rays not absorbed by the object. The X-ray absorption coefficient (also known as the linear attenuation coefficient), denoted μ,

is first calculated for each voxel (i.e. a volumetric unit, 3-D extension of a pixel) within a circular slice of the CT scanned object. The X-ray absorption coefficient is then converted to a CT number (CTN) expressed in Hounsfield units (HU):

$$\mathrm{CTN} = \frac{\mu_{object} - \mu_{water}}{\mu_{object} - \mu_{air}} \times 1000 \tag{6.13}$$

where μ_{object} is the X-ray absorption coefficient for one voxel of the CT scanned object, μ_{water} is the X-ray absorption coefficient for pure water (i.e. 0 HU), and μ_{air} is the X-ray absorption coefficient for pure air (i.e. –1000 HU). Thus, negative CTN values correspond to voxels 'less dense than water', and positive CTN values, to voxels 'more dense than water'. The CTN values for each slice (with a certain thickness) of the CT scanned object are saved as a 512 × 512 matrix, although only the interior circle of the slice really contains a part of the CT scanned object and the rest (outside the circle) is considered as 'background' and is not used in further analyses.

CT scanning sessions took place at the CT Scanning Laboratory for agricultural and environmental research on Macdonald Campus of McGill University (Ste-Anne-de-Bellevue, Québec, Canada), where a Toshiba Xvision high-resolution CT scanner (Toshiba Corporation, Medical Systems Division, Tokyo, Japan) is installed. Configuration parameters in our study were: field of view of 18 cm in diameter (SS), combined with a zoom factor of 8; tube voltage, 120 kV; and tube current, 50 mA. The helical scanning mode was chosen because of its advantages in terms of image reconstruction and advanced interpolation, resulting in enhanced image precision (Han et al. 2008). The reconstruction interval length was 0.1 mm, and a total of 40 CT images were constructed sequentially each time, covering a distance of 4 mm in the central part of the wood sample (Figure 6.2a). Wood samples were placed flat on the CT scanner couch, the bark-to-bark axis being in the *x*-*y* plane which is perpendicular to the *z*-axis provided by the couch axis. Because of the use of a zoom factor in order to ensure a spatial resolution of less than 50 microns in the *x*-*y* plane (i.e. 0.044 mm), each wood sample was CT scanned more than one time in order to cover the total distance from bark to bark passing by the center of the trunk. Therefore, each wood sample was carefully moved by increments of 2 cm starting from the North direction (see black marks in Figure 2a). The diameter of CT images being 2.25 cm (i.e. 18 cm divided by 8, the zoom

factor), the overlap along the *x* axis (Figure 6.2b) allowed the merging of CT scanning datasets collected for successive portions of the same wood sample in order to produce a continuous curve of mean CTN values prior to transformation into wood density estimates (see below and Figure 6.3 in Results section). The 25 central rows in 512 × 512 CTN matrices were used for this.

The mean CTN values thus obtained every 44 microns from bark to bark for each wood sample were then converted to wood density estimates, using Lindgren's (1991a) calibration equation developed for dry, unextracted wood: $D = 1053 + 1.052$ CTN, where D is wood density in $kg.m^{-3}$. The slope of the regression line between wood density and CTN is close to 1, because the density variation of the cell wall material is small so that the variation in X-ray absorption in the sample is directly related to the relative proportion of cell wall in a voxel (Jungnikl et al. 2009). This is not true when there are knots and other defects, as the cell wall composition, and not only the quantity, would then change in each voxel (Jungnikl et al. 2009). This underlines the importance of using unblemished wood in such studies. Also, because of the use of a high level of energy (120 kV) and dry and relatively small wood samples and since wood has densities lower than pure water, limited noise and high accuracy of density estimated from CTN can be assumed (Freyburger et al. 2009).

The profiles of wood density estimates show a regular oscillating pattern which follows the annual growth rings. The transition from late wood to early wood is gradual, while the onset of growth in the spring should result in a more abrupt transition. This gradual transition is a consequence of the use of means calculated over the successive growth rings, through the collection of CT scanning data which are three-dimensional by nature. Calibration was thus performed with SilviScan data available for the same wood samples, in order to determine where in the gradual transition (Figure 6.3) a ring starts and finishes on average; Appendix G is devoted to this calibration. Once the beginning and end of growth rings were determined, an average wood density measure was computed per ring for each of the 2 directions, 2 heights and 11 trees, and these triply repeated measures data were saved as tensors and processed in Matlab 2010a (The Mathworks, Inc., Natick, MA, USA).

6.5 Results

The dataset was composed of $K = 11$ tensors with dimensions $2 \times 2 \times 2$, which were reshaped for model testing purposes into 11 vectors with 8 elements. The minimum sample size required to estimate the variance-covariance parameters of the tensor normal distribution model by ML in such a case is 3; see (6.3). By comparison, the minimum sample size necessary to estimate the variance-covariance parameters of the vector normal distribution model is 9.

Mardia's tests based on multivariate skewness ($P = 0.3122$) and multivariate kurtosis ($P = 0.1137$) as well as the Henze-Zirkler test ($P = 0.1967$) did not reject multivariate normality on the raw data (i.e. wood density estimates). Accordingly, we proceeded with maximum likelihood estimation without applying a preliminary data transformation.

Under the assumption of double separability for the variance-covariance structure, convergence was achieved in 13 iterations with the MLE-3D algorithm. The estimated mean tensor $\hat{\mathbb{M}}$ is:

$$\hat{\mathbb{M}}_{..1} = \begin{bmatrix} 402.3 & 403.8 \\ 403.8 & 401.5 \end{bmatrix} \quad \text{and} \quad \hat{\mathbb{M}}_{..2} = \begin{bmatrix} 404.9 & 383.4 \\ 402.4 & 384.4 \end{bmatrix}.$$

The estimated variance-covariance matrix for the tensor lines (i.e. the two growth rings) is:

$$\hat{\mathbf{U}}_1 = \begin{bmatrix} 1071.7 & 735.2 \\ 735.2 & 989.1 \end{bmatrix};$$

the estimated variance-covariance matrix for the tensor columns (i.e. the two directions) is:

$$\hat{\mathbf{U}}_2 = \begin{bmatrix} 1.058 & 0.284 \\ 0.284 & 0.953 \end{bmatrix};$$

and the estimated variance-covariance matrix for the tensor edges (i.e. the two heights) is:

$$\hat{\mathbf{U}}_3 = \begin{bmatrix} 1.130 & 0.020 \\ 0.020 & 0.873 \end{bmatrix}.$$

On the other hand, under the vector normal distribution model, the estimated 8×8 variance-covariance matrix is given by:

$$\hat{\boldsymbol{\Sigma}}_{UN} = \begin{bmatrix} 656.2 & 321.8 & 405.6 & 415.5 & 191.5 & 164.7 & 226.7 & 156.5 \\ 321.8 & 876.3 & 335.2 & 377.5 & 98.5 & 134.4 & 389.2 & 401.7 \\ 405.6 & 335.2 & 2115.5 & 1796.8 & -562.6 & -730.1 & -392.5 & 64.2 \\ 415.5 & 377.5 & 1796.8 & 1624.0 & -531.4 & -570.4 & -267.6 & 142.6 \\ 191.5 & 98.5 & -562.6 & -531.4 & 865.6 & 693.2 & 763.7 & 207.6 \\ 164.7 & 134.4 & -730.1 & -570.4 & 693.2 & 1143.8 & 609.1 & 154.1 \\ 226.7 & 389.2 & -392.5 & -267.6 & 763.7 & 609.1 & 983.6 & 393.1 \\ 156.5 & 401.7 & 64.2 & 142.6 & 207.6 & 154.1 & 393.1 & 383.1 \end{bmatrix}.$$

At the significance level α = 0.05, the critical value of the unbiased modified LRT for double separability is $\chi^2_{0.95}(28) = 41.3371$, while the observed value (6.9) is 34.85 (with p = 5.38, Table E.1). Thus, the null hypothesis of double separability is not rejected, and it is justified to perform the triply repeated measures ANOVA in these conditions (Table 6.1). So doing, a significant height-by-direction interaction was found. The inspection of the estimated mean tensor $\hat{\mathbb{M}}$ (see above) reveals that this is due to a lower wood density in the South direction at live crown height on average, compared with the three other height-direction combinations.

Triply repeated measures data can also be analyzed from 'slices' made in 2-D, by working with the data at a given level of one of the three repeated-measures factors (e.g. at breast height only; see below). Multivariate normality needs not be checked again, because once accepted for longer random vectors, it is accepted for shorter vectors (or marginal distributions; Muirhead 1982). For the six types of 'slices', the MLE-algorithm converged in 7 to 24 iterations, and the estimated matrices (mean and variance-covariance), together with the LRT results, are presented in Table 6.2. Simple separability was accepted for all the 2-D datasets but one. The doubly repeated measures ANOVA at breast height is presented as an example (Table 6.3), and the direction main effects are significant at α = 0.05, again because the mean wood density in South direction is smaller than in North direction at that height. Note that PROC MIXED of SAS 9.2 (SAS Institute Inc., Cary, USA), used with the UN @ UN variance-covariance structure (equivalent to the simple separable structure here) provided similar results; no UN @ UN @ UN (our doubly separable variance-covariance structure) is available in SAS PROC MIXED at this moment.

6.6 Discussion

6.6.1 Models and estimation

In previous sections, multi-dimensional (3-D and 2-D) normal distributions with a separable variance-covariance structure were described and applied to wood data for statistical analysis. As part of this, procedures for estimating the parameters, tests for assessing the assumptions of models, and modified ANOVA *F*-tests taking into account the dependency and heterogeneity of variance in data were presented. In a multi-dimensional context, the estimation of the variance-covariance matrix often requires modeling. Indeed, if we were to measure density in 40 growth rings at 10 heights, estimating the 79, 800 (400 times 399 divided by 2) distinct parameters of an unstructured variance-covariance matrix would require 401 (40 times 10 plus 1) trees. Assuming homogeneity of the variance and ignoring covariances in the space-time domain is unrealistic, but based on our preliminary example, using separable variance-covariance structures is reasonable while requiring much smaller sample sizes (only 5 in the 40×10 matrix normal distribution model) and thus represents a middle ground between an unstructured variance-covariance matrix and an over-simplified homoscedastic and independent structure (with one variance and no covariance different from zero) for raw data.

Several other variance-covariance structures are available (e.g. ante-dependence, banded or Toeplitz; see Wolfinger 1996 for examples with the vector normal distribution). Some of them require the assumption of homogeneity of the variance in the space-time domain (e.g. AR(1) and CS), whereas the separable structure in its general form does not require such an assumption. In addition to offering the simple separable variance-covariance structure UN @ UN, where both variance-covariance matrices are not modeled, SAS also offers UN @ AR(1) and UN @ CS. However, for small sample sizes, the estimation algorithm used in SAS PROC MIXED may not converge (Roy and Khattree 2005).

6.6.2 Testing the assumptions of models

In our example with wood density estimates obtained from CT scanning data for 11 white spruce trees, the hypothesis of double separability of the variance-covariance structure was accepted with the unbiased LRT, which is encouraging for future applications in

wood data analysis since these are 'real' data and not simulated data presenting known statistical properties. Also, simple separability was found to be an acceptable variance-covariance structure for most types of 2-D 'slices' taken from the $2 \times 2 \times 2$ data tensors, and triply and doubly repeated measures ANOVAs were performed accordingly. In case separability would be rejected and depending on sample size (i.e. whether $K \geq n_1 n_2 n_3 + 1$, or not), modified ANOVA F-tests or multivariate ANOVA tests available for simple repeated measures ANOVA (Crowder and Hand 1990) could be used over all three dimensions or only one or two dimensions.

Postulating separability when its testing is not possible is another option (see Dutilleul and Pinel-Alloul (1996) for a 2-D example in a hydrological application). Realistically, postulating a separable variance-covariance structure is reasonable, at least more than assuming homogeneity of the variance and that all covariances are zero. The effects of mistakenly assuming separability on hypothesis testing in linear models and on the estimation of Box's 'epsilon' have already been studied (Boik 1991, Njue 2001), but robustness to failure of the separability assumption is worthy of further studies, especially in the 3-D case.

One of the conditions of application of the tensor and matrix normal distribution models is that the variance-covariance matrix is the same for all trees (individuals or subjects in general terms) for which a tensor or matrix of data is available. When there are several experimental groups, this condition may not be satisfied, which can then prevent the MLE 3-D and 2-D algorithms from converging if they are used with all experimental groups combined (Dutilleul and Pinel-Alloul 1996). Here, the trees came from different regions but the algorithms converged without a problem, and the acceptance of double and simple separability was the rule instead of the exception.

6.6.3 Computed tomography scanning

The use of non-destructive methods to measure wood properties is receiving increased attention lately. Following Pont et al. (2007), more work is needed to improve "the correspondence between image greyscale values and wood properties". In our study, we did not have this problem, since we had access to CTN values and transformed them into wood density estimates, using a calibration equation (Lindgren 1991a). Thereafter, the

multi-dimensional normal distribution models helped take advantage of the very large dataset thus generated, with valid and interpretable tests of the mean eventually.

Slight inaccuracies may have arisen from the data collection and calibration, though. A calibration equation specific to the CT scanner would be more precise than the standard equation. Also, the determination of the beginning of annual growth rings (Appendix G) was based on an average (32% of the inter-maximum distance; Figure 6.3). However, the grand mean value of our wood density estimates (394.5 $kg.m^{-3}$) is within the boundaries of what has been reported in the literature for the same species in the same region (Corriveau et al. 1990). As in Corriveau et al. (1990), a substantial portion (75%) of the variation in our wood density estimates is associated with tree-related main effects and interactions (Table 6.1). In conclusion, imperfections are likely to be small in magnitude, and should be much smaller than variations among trees.

The results of the measures of the wood density in 3-D show that at live crown height in the South direction the density was significantly lower. Wang and Micko (1984) found that in general density increased with height in white spruce trees from north central Alberta. More work is needed to link environmental conditions with wood density in the tree trunk.

6.6.4 Implications for forestry

The statistical method was demonstrated with the simplest possible tensor (dimension 2 x 2 x 2) with an original dataset of wood density calculated by CT scanning, a method which requires multi-dimensional statistics for data analysis. CT scanning was chosen as it is a tool which allows the rapid assessment of response variables in tree trunks or wood products in 3-D, and contributes to fill a gap in instrumentation noted by Gartner et al. (2002). However, the statistical method can be used with any multi-dimensional dataset composed of continuous, multivariate normal data of sufficient sample size, not necessarily obtained by CT scanning and with tensors of any size.

An example of application in dendrochronology would consist in the collection of density in the early and late wood (space horizontal), for all rings (time) at several heights (space vertical). Such a dataset can be obtained by CT scanning, X-ray densitometry, or manually but can be analyzed for significant difference in the mean with the statistical method presented here, especially if the number of trees is limited. In wood engineering,

the humidity levels in wood beams can be obtained at different points (two spatial dimensions) on several wood beams using a sensor, and a test of the mean can be performed to ensure that drying was on average uniform. Finally, in the field, a core can be obtained at various heights and directions (two spatial dimensions) in the trunk of several trees and the thickness of the sapwood measured, and tested for significant differences in the mean.

In addition to their use to construct valid tests of significance about means, estimated variance-covariance matrices present a biological interest, as temporal autocorrelation has a physiological basis in the tree growth process (Fritts 1976). As shown with $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$ in the example, the variance (variation) can change more in one dimension than in another, whereas covariances can be large or close to zero. Notably, the covariance (dependency) is stronger in the horizontal spatial dimension than in the vertical spatial one, due to smaller distances between measurement points at the same height (in cm) relative to distances between measurement points at different heights (in m) and the biological implications for the production of new cambium and the density of wood in the end.

In closing, as measurement techniques allow the collection of an increased number of multi-dimensional datasets, including for multiple wood properties (e.g. wood density, fiber length and microfibril angle in 2-D), statistical models and methods have no other choice but to evolve. With the upgraded models and methods presented here, researchers in forestry and wood technology should be prepared to undertake the analysis of these new multi-dimensional wood data, with confidence and trust in their statistical and biological results.

6.7 Acknowledgements

The financial assistance of the Natural Sciences and Engineering Research Council of Canada, through an Alexander Graham Bell Canada Graduate Scholarship to Mr. Manceur and a Discovery Grant to Prof. Dutilleul, and of the Canadian Wood Fibre Centre to Dr. Beaulieu is gratefully acknowledged. The authors are also grateful to Patrick Laplante and Philippe Labrie (Canadian Forest Service), to Maurice Defo (FPInnovations) for wood density assessment using the SilviScan technology, and to Chun-Huor Ung (Canadian Wood Fibre Centre) for access to wood disks.

Table 6.1: ANOVA table and results of *F*-tests for fixed and random effects in the statistical analysis of triply repeated measures for wood density estimated from CT scanning data in two annual growth rings (*r*: 1997, 2003), in two directions (*d*: North, South) and at two heights (*h*: breast, live crown) in 11 white spruce trees (*T*)[a].

Effect	Number degrees freedom	Sum of squares	Mean square	Observed *F*-value	*P*-value	Estimated Box's 'epsilon'	Adjusted *P*-value
Fixed							
r	1	8.44	8.44	0.02	0.886	1.0	0.886
d	1	2227.18	2227.18	1.44	0.257	1.0	0.257
h	1	1814.62	1814.62	0.56	0.473	1.0	0.473
$r \times d$	1	0.14	0.14	<0.01	0.985	0.641	1.000
$r \times h$	1	0.66	0.66	<0.01	0.968	0.565	1.000
$d \times h$	1	2056.78	2056.78	5.07	0.029	0.917	0.032
$r \times d \times h$	1	70.40	70.40	0.17	0.679	0.585	0.546
Error	50	20297.50	405.95				
Random							
T	10	26709.80	2670.98	6.58	<.001		
$T \times d$	10	15445.80	1544.58	3.80	0.001		
$T \times h$	10	32675.18	3267.52	8.05	<.001		
Total	87	101306.50					

[a] The random interaction effects $T \times r$, $T \times r \times d$, $T \times r \times h$, and $T \times d \times h$ were not significant at 0.05 level and were removed from the model. d is tested against $T \times d$ (= $\varepsilon^{(2)}_{i_2 k}$ in equation 6.10), h against $T \times h$ (= $\varepsilon^{(3)}_{i_3 k}$), and the other effects against the Error (= $\varepsilon^{(7)}_{i_1 i_2 i_3 k}$).

Table 6.2: Mean and variance-covariance parameter estimates for matrix normal distribution models[a].

	Sample mean matrix	Estimated variance-covariance matrices		Observed value of Λ^* statistic (decision)
Breast height: Rings ($\mathbf{U}_1$) by directions ($\mathbf{U}_2$)	$\begin{bmatrix} 402.34 & 403.81 \\ 403.76 & 401.50 \end{bmatrix}$	$\begin{bmatrix} 1531.1 & 1191.9 \\ 1191.9 & 1325.1 \end{bmatrix}$	$\begin{bmatrix} 1.1873 & 0.1815 \\ 0.1815 & 0.8274 \end{bmatrix}$	11.3167 (reject separability)
Live crown height: Rings ($\mathbf{U}_1$) by directions ($\mathbf{U}_2$)	$\begin{bmatrix} 404.89 & 383.45 \\ 402.39 & 384.36 \end{bmatrix}$	$\begin{bmatrix} 721.23 & 418.40 \\ 418.40 & 797.79 \end{bmatrix}$	$\begin{bmatrix} 1.1684 & 0.4730 \\ 0.4730 & 0.9143 \end{bmatrix}$	9.8907 (accept separability)
North direction: Rings ($\mathbf{U}_1$) by heights ($\mathbf{U}_3$)	$\begin{bmatrix} 402.34 & 404.89 \\ 403.76 & 402.39 \end{bmatrix}$	$\begin{bmatrix} 747.27 & 495.21 \\ 495.21 & 1011.8 \end{bmatrix}$	$\begin{bmatrix} 0.9788 & 0.1604 \\ 0.1604 & 1.0218 \end{bmatrix}$	0.7726 (accept separability)
South direction: Rings ($\mathbf{U}_1$) by heights ($\mathbf{U}_2$)	$\begin{bmatrix} 403.81 & 383.45 \\ 401.50 & 384.36 \end{bmatrix}$	$\begin{bmatrix} 1601.3 & 1120.6 \\ 1120.6 & 1006.3 \end{bmatrix}$	$\begin{bmatrix} 0.9873 & 0.0861 \\ 0.0861 & 1.0151 \end{bmatrix}$	10.7668 (accept separability)
1997 Ring: Directions ($\mathbf{U}_2$) by heights ($\mathbf{U}_2$)	$\begin{bmatrix} 402.34 & 404.89 \\ 403.81 & 383.45 \end{bmatrix}$	$\begin{bmatrix} 882.56 & 764.46 \\ 764.46 & 1591.3 \end{bmatrix}$	$\begin{bmatrix} 1.3970 & 0.1312 \\ 0.1312 & 0.6572 \end{bmatrix}$	7.1883 (accept separability)
2003 Ring: Directions ($\mathbf{U}_2$) by heights ($\mathbf{U}_2$)	$\begin{bmatrix} 403.76 & 402.39 \\ 401.50 & 384.36 \end{bmatrix}$	$\begin{bmatrix} 1131.0 & 275.90 \\ 275.90 & 892.61 \end{bmatrix}$	$\begin{bmatrix} 1.2913 & 0.1756 \\ 0.1756 & 0.7334 \end{bmatrix}$	9.3230 (accept separability)

[a] Unbiased modified LRT for simple separability of the variance-covariance structure with $p = 3.57$ (Table 4.1). If the null hypothesis is 'accepted', then the structure is found to be separable. The decision to accept or reject the null hypothesis of simple separability was made at the 0.05 significance level here, with critical value $\chi^2_{0.95}(5) = 11.0705$.

Table 6.3: ANOVA table and results of *F*-tests for fixed and random effects in the statistical analysis of doubly repeated measures for wood density estimated from CT scanning data in two annual growth rings (*r*: 1997, 2003) and two directions (*d*: North, South) at live crown height in 11 white spruce trees (*T*)[a].

Effect	Number degrees freedom	Sum of squares	Mean square	Observed *F*-value	*P*-value	Estimated Box's 'epsilon'	Adjusted *P*-value
Fixed							
r	1	6.92	6.92	0.02	0.898	1.0	0.898
d	1	4282.26	4282.26	10.41	0.003	1.0	0.003
$r \times d$	1	32.09	32.09	0.08	0.782	0.838	0.733
Error	30	12337.83	411.26				
Random							
T	10	24798.86	2479.89	6.03	<.001		
Total	43	41457.96					

[a] The random interaction effects $T \times r$ ($= \varepsilon'^{(1)}_{i_1 k}$ in equation 6.11) and $T \times d$ ($= \varepsilon'^{(2)}_{i_2 k}$) were not significant at 0.05 significance level and were removed from the model. All the effects are tested against the Error.

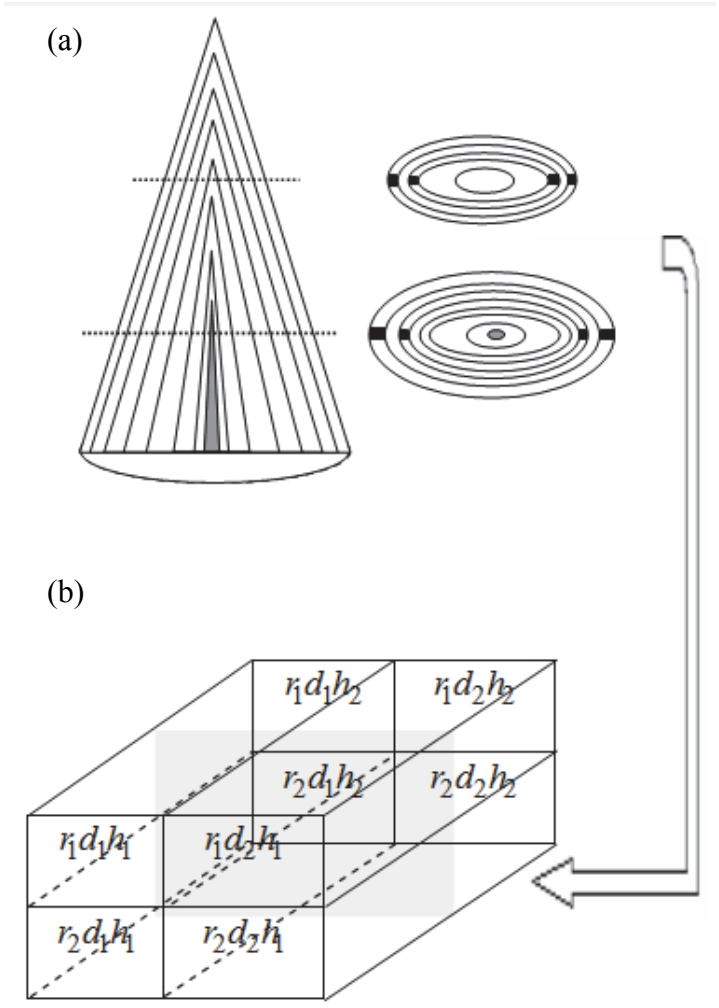


Figure 6.1: (a) A schematized 3-D representation of tree growth. The number of annual growth rings decreases with increasing height, while the cambium producing rings in higher position is younger for a given calendar year. Dark blocks in growth increments indicate sampling locations where wood density could be estimated for two rings, two directions and two heights. (b) The resulting triply repeated measures data are arranged in a $2 \times 2 \times 2$ third-order tensor (*r*: ring; *d*: direction; *h*: height).

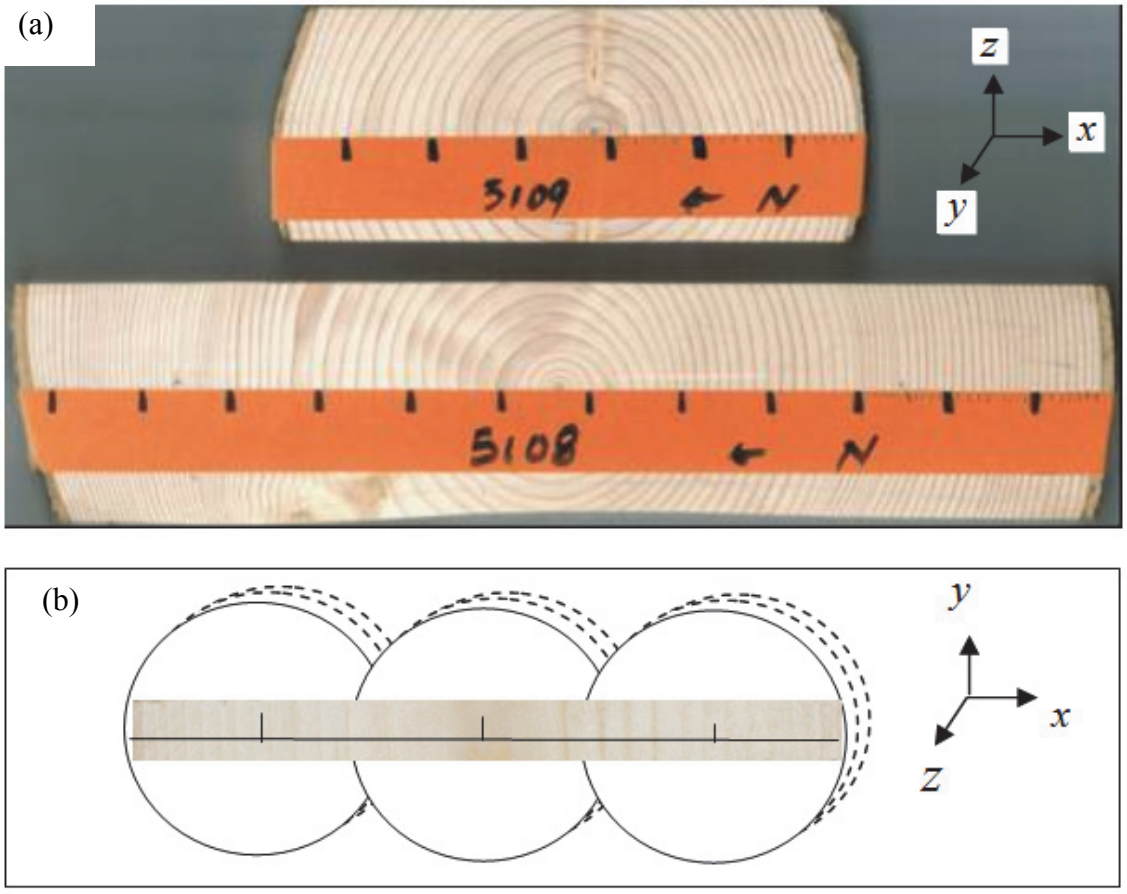


Figure 6.2: (a) Two wood samples collected at breast (below) and live crown (above) heights, viewed from the top while they lay on thinner side horizontally. They were CT scanned along the orange tape from right (North direction) to left (South). Black marks are separated by a distance of 2 cm. In this perspective, the *y*-axis points out of the figure, while the *x*- and *z*-axes are in the horizontal plane. (b) The view is transposed (90-degree rotation around the *x*-axis) here, compared to panel (a). CT images are constructed in the *x*-*y* plane, as schematized by the dashed circles. See text for CT scanning details.

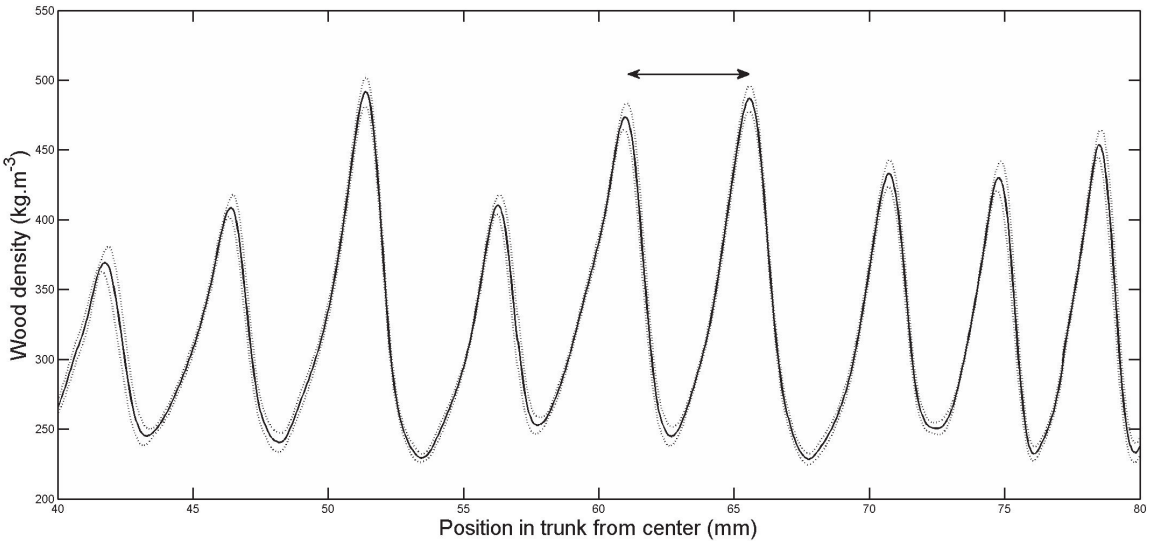


Figure 6.3: Illustration of the oscillating pattern in wood density as estimated from CT scanning data, with troughs and peaks corresponding to the succession of early and late wood within annual growth rings. This allows the determination of the beginning and end of rings, as 32% of the distance between peaks was found to be a good basis overall for the definition of ring limits (see Appendix G). Smoothness of the curve is the result of averaging wood density estimates, since CT scanning data are 3-D by nature. The two grey lines, above and below the black one which represents the mean wood density estimate at successive positions along the North-South axis, represent one standard error.

Chapter 7. General Discussion

Environmental datasets are often multi-dimensional and non-stationary. As extensions of the well-known scalar and vector normal distributions, the matrix and tensor normal distribution models do not require the assumption of stationarity in the general case. However, these multi-dimensional normal distributions imply a separable variance-covariance structure (simple, double or more). The main objectives of this Ph.D. thesis were to study the properties of the separable variance-covariance structure and contribute to the related estimation and testing procedures in order to present a new type of statistical applications.

The first hypothesis was that the bias of the ML estimator of a simply separable variance-covariance matrix structure, implied by the matrix normal distribution model, would decrease monotonically with increasing sample size. In Chapter 3, a 'peak-trough' pattern of the empirical bias was detected in a simulation study. To explain such unusual results, the empirical bias was renamed "ergodic" and decomposed into two components called "estimation" and "fluctuation", minus a non-orthonogonality factor. These results show that the use of numerical methods to find the solution to likelihood equations that do not have an analytical solution can lead to surprising discoveries. Nevertheless, the 'peak-trough' pattern of the bias of the ML estimator found here does not challenge the theory according to which the bias of a ML variance-covariance matrix estimator decreases with increasing sample size, since the pseudo-theoretical bias obtained without simulations decreased monotonically.

The second hypothesis was that the LRT presently available in the literature for testing the hypothesis of simple separability for a variance-covariance structure is biased, in that the rejection rate is not equal to the nominal Type I error risk when the null hypothesis is true. This second hypothesis of the thesis was verified in Chapter 4, and a simple and general modification of the LRT statistic corrected for the bias and improved the chi-square distribution of the LRT statistic in small to large samples, while theory classically predicts it only asymptotically. An EGLS algorithm, which allows modeling of the mean, was developed and the bias of the test was shown to be more severe with a more complex mean model.

The third hypothesis was that a 3-D extension of the MLE-2D algorithm, called "MLE-3D", would allow the estimation of the variance-covariance parameters of the tensor normal distribution model of order 3, with small samples. This hypothesis was verified in Chapter 5, where the algorithm converged when the sample size was small and the variance-covariance matrix was complex and it was shown that the initial solution provided to the algorithm did not influence the convergence properties. The "MLE-4D" algorithm was also developed as a further extension. Although these extensions do not constitute a conceptual leap after the 2-D case has been established, they are not straightforward since they require the use of advanced tensor operators whose properties and notations are not very familiar or common, and a limited number of statistical programs offer the possibility of estimating the parameters of a 3-D or 4-D normal distribution model from an i.i.d. random sample. Thereafter, specific objectives such as the development of an unbiased modified LRT for double separability of a variance-covariance matrix and the bias analysis for the ML estimator of a doubly separable variance-covariance matrix, showing a 'peak-trough' pattern with a slight but clear trough in the estimation bias as a function of sample size, were met.

The fourth hypothesis was that the variance-covariance structure for a data tensor made of wood density measures made from CT scan data on white spruce trees in two directions, at two heights and in two growth rings would be doubly separable. This hypothesis was verified in Chapter 6. In addition, simple separability was not rejected for almost all 2-D subsets of the data. Finally, ANOVA *F*-tests modified by the use of Box's 'epsilons' estimated from variance-covariance matrices themselves estimated under multi-dimensional normal distribution models showed significant variation in mean wood density at different heights and directions in the tree. Overall, a separable variance-covariance structure was found to be particularly appropriate for the tree data under study and the statistical models allowed the exploration of a very large and complex dataset, with valid statistical tests. Thus, multi-dimensional normal distribution models and separable variance-covariance structures together with wood density measures obtained by CT scanning form a coherent package. The relevant statistical software will be available at http://www.environmetrics.mcgill.ca soon.

The decomposition of the bias of a ML variance-covariance matrix estimator is an important conceptual contribution of this Ph.D. thesis. It followed from the focus on small-sample properties of estimators, in relation to the context in which multi-dimensional models are applied – the environmental sciences – where replication is possible and sample sizes are limited. It is difficult to tell if the same conceptual discovery would have occurred if the thesis had been realized in another context. However, the combination of the two disciplines was fruitful and reciprocally advantageous. On the one hand, statisticians are encouraged to pay attention to small-sample properties of estimators in addition to asymptotic properties, and on the other hand, environmental researchers are invited to model the often mentioned connectedness and interdependence of nature when performing statistical inference.

Chapter 8. Contributions to knowledge

The contributions to knowledge of this Ph.D. thesis, depending on the chapter or the appendix, read as follows.

The contributions from Chapter 3 and Appendices A and D are:

- The discovery of a non-monotonic 'peak-trough' pattern in the empirical bias of the ML estimator of a simply or doubly separable variance-covariance matrix in relation to sample size.
- The explanation of the discovered pattern through the decomposition of the ergodic (empirical) bias into estimation and fluctuation biases, minus a non-orthogonality factor, an important conceptual contribution.
- The verification at the same time that the fundamental theory of ML estimators was not challenged since the pseudo-theoretical bias showed a monotonic decreasing relationship with sample size.

The contributions from Chapter 4 and Appendices B and E are:

- The development of an unbiased modified LRT for simple separability of a variance-covariance matrix structure, through the introduction of an adjustment factor and the use in this context of an estimated generalized least-squares estimator of the mean matrix when modeled.
- The evaluation of the differences in terms of performance in small and large sample sizes between the proposed unbiased modified LRT and the biased LRT of the literature for simple separability of a variance-covariance matrix structure.
- The presentation of the unbiased modified LRT for double separability of a variance-covariance matrix structure.

The contribution from Chapter 5 and Appendices C and F is:

- The detailed exposition of (i) the tensor normal distribution model from its moment generating function and probability density function, (ii) the MLE-3D and MLE-4D algorithms, (iii) the empirical properties (bias, dispersion) of ML estimators with simulations, and (iv) a 3-D example with real data.

The contributions from Chapter 6 and Appendix G are:

- The application to an environmental dataset of the multi-dimensional normal distribution models and related estimation and testing procedures studied and

presented in other chapters and appendices (MLE-2D and MLE-3D algorithms, LRTs for simple and double separability, ANOVA with double and triply repeated measures).

- The development of a calibration procedure for tree ring determination, from wood density measures obtained by CT scanning.

Chapter 9. Future research directions

In this Ph.D. thesis, concepts and methods have been scrutinized, extended and developed. The work presented therein may be used as starting point for future research in the following directions:

1. The decomposition of the bias can be further explored. For example, the bias could be dissected for each of the component variance-covariance matrices rather than the Kronecker product, which assumes some preliminary adjustment to make estimated component variance-covariance matrices from different i.i.d. random samples comparable.
2. In addition to the bias, a detailed study of a measure of dispersion of the ML estimator of a simply or doubly separable variance-covariance matrix can be realized.
3. Since separability and multivariate normality may be assumed when it is not possible to test them in practice, the study of the robustness of algorithms and statistical methods (e.g. ANOVA) against departure from them will be useful.
4. ML estimators can be developed for separable multi-dimensional distributions other than normal, since not all environmental variables are distributed like multivariate normal (when data transformation is not a workable option).
5. Algorithmic extensions, such as 3-D and 4-D EGLS algorithms for modeling of the mean, are possible and would be of interest, and so would be a detailed numerical study of 2-D cases with very large numbers of iterations.
6. A comparison of the results of various statistical analyses for the same dataset (e.g. by assuming or not separability, stationarity, etc. with or without pre-whitening) could be performed, as a single approach may not be ideal for all datasets.
7. Exploration of the application of multi-dimensional normal distribution models to environmental datasets should be continued, with emphasis on the meaning and interpretation of variance-covariance parameters in fields such as limnology, with the lake as the individual under study instead of the tree in forestry/wood science.

References

Allen, G. I., and Tibshirani, R. 2010. Transposable regularized covariance models with an application to missing data imputation. Ann. Statist. 4 764–790.

Antony, F., Schimleck, L., Daniels, R., Clark III, A., and Hall, D. 2010. Modeling the longitudinal variation in wood specific gravity of planted Loblolly pine (*Pinus taeda*) in the United States. Can. J. For. Res. 40: 2439–2451.

Apiolaza, L., and Garrick, D. 2001. Analysis of longitudinal data from progeny test: some multivariate approaches. Can. J. For. Res. 47: 129–140.

Apiolaza, L., Gilmour, A., and Garrick, D. 2000. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. Can. J. For. Res. 30: 645–654.

Barndorff-Nielsen, O. E., and Cox, D. R. 1984. Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. J. Royal Stat. Soc. Series B (Methodological) 46: 483–495.

Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. Proc. Royal Soc. London. Series A (Mathematical and Physical Sciences) 160: 268–282.

Barton, T. A., and Fuhrman, D.R. 1993. Covariance structures for multidimensional data, Multidim. Syst. Sign. P. 4: 111–123.

Beaulieu, J., Doerksen, T., Boyle, B., Clément, S., Deslauriers, M., Beauseigle, S., Blais, S., Poulin, P.L., Lenz, P., Caron S., Rigault, P., Bicho, P., Bousquet, J., and Mackay, J. 2011. Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. Genetics 188: 197–214.

Bijma, F., de Munck, J. C., and Heethaar, R. M. 2005. The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. NeuroImage 27: 402–415.

Blum, J. R. 1981. Ergodic theorems. In: S. Kotz and N.L. Johnson, eds. Encyclopedia of Statistical Sciences. Wiley & Sons, New York, U.S.A.

Boik, R. J. 1991. Scheffé's mixed model for multivariate repeated measures: A relative efficiency evaluation. Commun. Stat. Theory Methods 20: 1233–1255.

Box, G. E. P. 1954a. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Ann. Math. Stat. 25: 290–302.

Box, G. E. P. 1954b. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. Ann. Math. Stat. 25: 484–498.

Bucur, V. 2003. Nondestructive Characterization and Imaging of Wood. Springer-Verlag, Dordrecht, The Netherlands.

Chen, H., Chen, J., and Kalbfleisch, J. D. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J. Royal Stat. Soc. Series B (Statistical Methodology) 63: 19–29.

Corriveau, A., Beaulieu, J., Mothe, F., Poliquin, J., and Doucet, J. 1990. Densité et largeur des cernes des populations d'épinettes blanches de la région forestière des Grands Lacs et du Saint Laurent. Can. J. For. Res. 20: 121–129.

Crowder, M. J., and Hand, D. J. 1990. Analysis of Repeated Measures. Chapman and Hall, London, U.K.

Dawid, A. P. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. Biometrika 68 265–274.

de Waal, D. J. 1988. Matrix-valued distributions. In: S. Kotz and N.L. Johnson, eds. Encyclopedia of Statistical Sciences. Wiley & Sons, New York, U.S.A.

Defo, M., Goodison, A., and Uy, N. 2009. A method to map within-tree distribution of fibre properties using SilviScan-3 data. Forest. Chron. 85: 409–414.

Downes, G. M., David, D., Battaglia, M., and Schulze, D. 2009. Measuring and modelling stem growth and wood formation: An overview. Dendrochronologia, 27: 147–157.

Dryden, I. L., Bai, L., Brignell, C. J., and Shen, L. 2008. Factored principal components analysis, with applications to face recognition, Statist. Comput. 19: 229–238.

Duff, N. J., and Nolan, G. H. 1953. Growth and morphogenesis in the Canadian forest species. i. the controls of cambial and apical activity in *Pinus resinosa* ait. Can. J. Bot. 13: 471–513.

Dutilleul, P. 1998. Tests of classification with simply, doubly and triply repeated measures. Invited talk at the Statistical Society of Canada Meeting, Sherbrooke, Canada, May 31-June 3.

Dutilleul, P. 1999. The MLE algorithm for the matrix normal distribution. J. Statist. Comput. Simul. 64: 105–123.

Dutilleul, P. 2011. Spatio-Temporal Heterogeneity: Concepts and Analyses. Cambridge University Press, Cambridge, U.K.

Dutilleul, P., and Pinel-Alloul, B. 1996. A doubly multivariate model for statistical analysis of spatio-temporal environmental data. Environmetrics 7: 551–565.

Dutilleul, P. 1990. Apport en Analyse Spectrale d'un Périodogramme Modifié et Modélisation des Séries Chronologiques avec Répétitions en vue de leur Comparaison en Fréquence. D.Sc. dissertation, Département de mathématique, Université catholique de Louvain, Louvain-la-Neuve, Belgique.

Dutilleul, P., Herman, M., and Avella-Shaw, T. 1998. Growth rate effects on correlations among ring width, wood density and mean tracheid length in Norway spruce (*Picea abies* (L.) Karst). Can. J. For. Res. 28: 56–68.

Evans, R., Gartside, G., and Downes, G. 1995. Present and prospective use of Silviscan for wood microstructure analysis. Appita J. 91–96.

Fey, M., Gojobori, P., Hannick, T., Hide, L., Hill, W., Kania, D. P., Schaeffer, R., St-Pierre, M., Twigger, S., White, S., Rhee, O., Howe, S.Y., and Costanzo, D. 2008. Big data: The future of biocuration. Nature 455: 47–50.

Freyburger, C., Longuetaud, F., Mothe, F., Constant, T., and Leban, J.-M. 2009. Measuring wood density by means of X-ray computer tomography. Ann. For. Sci. 66: 804–13.

Fritts, H. C. 1976. Tree rings and climate. Academic Press, London, U.K.

Fromm, J. H., Irina, S., Matthier, D., Kremer, J., Schumacher, P., and Ganter, C. 2001. Xylem water content and wood density in spruce and oak trees detected by high resolution computed tomography. Plant Phys. 127: 416–425.

Fuhrman, D. R. 1997. Correction to "On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices". IEEE Trans. Inform.Theory 43: 1094–1096.

Galecki, A. T. 1994. General class of covariance structures for two or more repeated factors in longitudinal data analysis. Commun. Stat. Theory Methods 23: 3105–3110.

Gartner, B. L., Aloni, R., Funada, R., Lichtfuss-Gautier, A. N., and Roig, F. A. 2002. Clues for dendrochronology from studies of wood structure and function. Dendrochronologia 20: 53–61.

Genton, M. G. 2007. Separable approximations of space-time covariance matrices. Environmetrics 18: 681–695.

Ghazal, G A. 2000. Recurrence formula for expectations of products of bilinear forms and expectations of bilinear forms and random matrices. Stat. Prob. Lett. 48:1–9.

Gjerdum, P., and Bernabei, M. 2009. Three-dimensional spiral grain pattern in five large Norway spruce stems. Silva Fenn. 43: 457–464.

Greenhouse, S. W., and Geisser, S. 1958. An extension of Box's results on the use of the F distribution in multivariate analysis. Ann. Math. Statist. 29:885-91.

Han, L., Dutilleul, P., Prasher, S., Beaulieu, C., and Smith, D. 2008. Assessment of common scab-inducing pathogen effects on potato underground organs via computed tomography scanning. Phytopath. 98: 118–1125.

Henze, N., and Zirkler, B. 1990. A class of invariant consistent tests for multivariate normality. Commun. Stat. Theory Methods 19: 3595–3617.

Herman, M., Dutilleul, P., and Avella-Shaw, T. 1999. Growth rate effects on intra-ring and inter-ring trajectories of microfibril angle in Norway spruce (*Picea abies*). IAWA J. 20: 3–21.

Hilts, M., and Duzenli, C. 2004. Image noise in x-ray ct polymer gel dosimetry. Journal of Physics: Conference Series 3: 252–256.

Hoff, P. 2011. Separable covariance arrays via the tucker product, with applications to multivariate relational data. Bayesian Anal. 6: 179–196.

Huang, H. C., Martinez, F., Mateu, J., and Montes, F. 2007. Model comparison and selection for stationary space–time models. Comp. Stat. Data Anal. 51: 4577–4596.

Huizenga, H. M., De Munck, J. C., Waldorp, L. J., and Grasman, R. 2002. Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. IEEE Trans. Biomed. Eng. 49: 533–539.

Huynh, H., and Feldt, L. S. 1976. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. J. Educ. Behavioral Stat. 1: 69–82.

Jordan, L., Daniels, R., Clark III, A., and He, R. 2005. Multilevel nonlinear mixed-effects models for the modeling of earlywood and latewood microfibril angle. For. Sci. 51: 357–371.

Jungnikl, K., Goebbels, J., Burgert, I., and Fratzl, P. 2009. The role of material properties for the mechanical adaptation at branch junctions. Trees 23: 605–610.

Kak, A. C., and Slaney, M. 2001. Principles of Computerised Tomographic Imaging. Society for Industrial and Applied Mathematics, Philadelphia, U.S.A.

Kendall, M. G., and Stuart, A. 1967. The Advanced Theory of Statistics. Hafner Publishing Company, New York, U.S.A.

Keunecke, D., Evans, R., and Niemz, P. 2009. Microstructural properties of common yew and Norway spruce determined with Silviscan. IAWA J. 30:165–178.

Koga, S., and Zhang, S. Y. 2004. Inter-tree and intra-tree variations in ring width and wood density components in Balsam fir (*Abies balsamea*). Wood Sci. Technol. 38: 149–162.

Kolda, T. G., and Bader, B. W., 2009, Tensor Decompositions and Applications. SIAM Review 51: 455-500.

Kotz, S., Johnson, N. L., and Read, C. B. 1982. Bias of an estimator. In: S. Kotz and N.L. Johnson, eds. Encyclopedia of Statistical Sciences. Wiley & Sons, New York, U.S.A.

Kouris, K., Tout, R. E., Gilboy, W. B., and Spyrou, N. M. 1981. Effect of constituent elements in wood on X-ray densitometry measurements. Archaeometry 23: 95–101.

Kroonenberg, P. M. 2008. Applied Multiway Data Analysis. Wiley Interscience, New York, U.S.A.

Larocque, G., Dutilleul, P., Pelletier, B., and Fyles, J. W. 2007. Characterization and quantification of uncertainty in coregionalization analysis. Math. Geol. 39 263–288.

Lecoutre, B. 1991. A correction for the e approximate test in repeated measures designs with two or more independent groups. J. Educ. Stat. 16: 371-372.

Lee, C. H., Dutilleul, P., and Roy, A. 2010. Comment on “Models with a Kronecker product covariance structure: Estimation and testing” by M.S. Srivastava, T. von Rosen, and D. von Rosen. Math. Methods Statist. 19 88–90.

Lindgren, L. O. 1991a. Medical CAT-scanning: X-ray absorption coefficients, CT-numbers and their relation to wood density. Wood Sci. Tech. 25: 341–349.

Lindgren. L. O. 1991b. The accuracy of medical CAT-scan images for non-destructive density measurements in small volume elements within solid wood. Wood Sci. Tech. 25: 425–432.

Lindstrom, H. 2002. Intra-tree models of juvenile wood in norway spruce as an input to simulation software. Silva Fenn. 36: 521–534.

Lu, N., and Zimmerman, D. L. 2004. On likelihood-based inference for a separable covariance matrix. Technical Report 337, University of Iowa, Iowa City, Iowa, U.S.A.

Lu, N., and Zimmerman, D. L. 2005. The likelihood ratio test for a separable covariance matrix. Stat. Prob. Lett. 73: 449–457.

Macedo, A., Vaz, C. M., Pereira, J. C. D., Naime, J. M, Cruvinel, P. E., and Crestana S. 2002. Wood density determination by X-ray and gamma-ray tomography. Holzforschung 56: 535–540.

Mardia, K. V. 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya B 36: 115–128.

Mardia, K.V., and Goodall, C. 1993. Spatial-temporal analysis of multivariate environmental monitoring data. In: G.P. Patil, C.R. Rao, Eds. Multivariate Environmental Statistics. North-Holland, New York, U.S.A.

Marsaglia, G., and Tsang, W. W. 2000. The Ziggurat method for generating random variables. J. Stat. Soft. 5: 1–7.

McCullagh, P. 2008. Marginal likelihood for parallel series. Bernoulli 14: 593–603.

McCullagh, P. 1987. Tensor Methods in Statistics. Chapman and Hall, London, U.K.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L., 2005. Testing for separability of space-time covariances. Environmetrics 16: 819–831.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L., 2006. A likelihood ratio test for separability of covariances. J. Mult. Anal. 97: 1025–1043.

Molteberg, D., and Høibø, O. 2007. Modelling of wood density and fibre dimensions in mature Norway spruce. Can. J. For. Res. 37: 1373–1389.

Muirhead, R. J. 1982. Aspects of Multivariate Statistical Theory. Wiley & Sons, New York, U.S.A.

Mull, R. T. 1994. Mass estimates by computed tomograph: physical density from CT numbers. Am. J. Roentgenology 143:1101–1104.

Naik, D. N., and Rao, S. S. 2001. Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. J. App. Stat. 28: 91–105.

National Forestry Database. 2009. Area planted by species. Canadian Forest Service, Natural Resources Canada, Government of Canada. Retrieved October 2011, from: http://nfdp.ccfm.org/index_e.php, 2011.

Njue, C. 2001. On the Efficiency of testing Procedures in the Linear Model for Multivariate Longitudinal Data. PhD thesis, Department of Statistics, University of Manitoba, Manitoba, Canada.

O'Hagan, A. 2002. Kronecker product. In: El-Shaarawi, A. H., and Piergorsch, W. W., Eds. Encyclopedia of Environmetrics. John Wiley & Sons, New York, U.S.A.

Ohlson, M., Ahmad, M. R., and von Rosen, D. 2011. The multilinear normal distribution: Introduction and some basic properties. Technical Report LiTH-MAT-R-2011/02-SE, Department of Mathematics, Linköping University, Linköping, Sweden.

Pliura, A., Zhang, S. Y., Bousquet, J., and MacKay, J. 2006. Age trends in genotypic variation of wood density and its intra-ring components in young poplar hybrid crosses. Ann. For. Sci. 63: 673–685.

Pont, D., Brownlie, R., and Grave, J. 2007. Disc image-processing software for three-dimensional mapping of stem ring width and compression wood. New Zeal. J. For. Sci. 37: 168–185.

Potvin, C., Lechowicz, M. J., and Tardif, S. 1990. The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. Ecology 71: 1389–1400.

Quintana, S. M., and Maxwell, S. E. 1994. A monte carlo comparison of seven epsilon-adjustment procedures in repeated measures designs with small sample sizes. J. Educ. Behavioral Stat. 19: 57-71.

Rencher, A. C. 1995. Methods of Multivariate Statistics. Wiley & Sons, New York, U.S.A.

Richter, A., Salmi, J., and Koivunen, V. 2008. ML estimation of covariance matrix for tensor valued signals in noise. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, U.S.A., March 30-April 14.

Roy, A. and Khattree, R. 2005. Testing the Hypothesis of a Kroneckar Product Covariance Matrix in Multivariate Repeated Measures Data. Paper 199-30, Statistics and Data Analysis, SUGI 30 Proceedings, Philadelphia, Pennsylvania, U.S.A.

Roy, A., 2006. Testing the Kronecker product structured mean vectors and covariance matrices. J. Stat. Theo. App. 5: 53–69.

Roy, A., and Khattree, R., 2005. On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. Stat. Meth. 2: 297–306.

Roy, A., and Leiva, R., 2008. Likelihood ratio tests for triply multivariate data with structured correlation on spatial repeated measurements. Stat. Prob. Lett. 78: 1971–1980.

Roy, A., and Leiva, R., 2011. Estimating and testing a structured covariance matrix for three-level multivariate data. Commun. Stat. Theory Methods 40: 1945–1963.

Schott, J. R. 1997. Matrix analysis for statistics. Wiley, New York, U.S.A.

Schweingruber, F. H. 2007. Wood Structure and Environment. Springer-Verlag, Dordrecht, Netherlands.

Searle, S. R., 1971. Linear models. Wiley, New York, U.S.A.

Simpson, S., 2010. An adjusted likelihood ratio test for separability in unbalanced multivariate repeated measures data. Stat. Meth. 7: 511–519.

Singh, T. 1984. Variation in the ovendry wood density of ten prairie tree species. For. Chron. 60: 217–220.

Smith, K. T. 2008. An organismal view of dendrochronology. Dendrochronologia 26: 185–193.

Srivastava, M. S., von Rosen, T., and von Rosen, D. 2008. Models with a Kronecker Product Covariance Structure: Estimation and Testing. Math. Methods Statist. 17: 357-370.

Steppe, K., Cnudde, V., Girard, C., Lemeur, R., Cnudde, J. P., and Jacobs, P. 2004. Use of X-ray computed microtomography for non-invasive determination of wood anatomical characteristics. J. Struct. Biol. 148: 11–21.

Sugiura, N., and Nagao, H. 1968. Unbiasedness of some test criteria for the equality of one or two covariance matrices. Ann. Math. Stat. 39: 1686–1692.

Svantesson, T., Wallace, J. W., 2003. Tests for assessing multivariate normality and the covariance structure of MIMO data. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China, 6-10 April.

Szeicz, J. M., and MacDonald, G. M. 1996. A 930-year ring-width chronology from moisture sensitive white spruce (*Picea glauca* moench) in Northwestern Canada. The Holocene 6: 345–351.

Theobald, D. L., and Wuttke, D. S. 2008. Accurate structural correlations from maximum likelihood superpositions. PLoS Comput. Biol. 4 1–8.

The Mathworks, Inc. 2010. MATLAB Version R2010b. The Mathworks, Inc., Natick, MA, U.S.A.

Tian, X., Cown, D., and McConchie, D. 1995. Modelling of Pinus radiata wood properties. Part II: Basic density. New Zeal. J. For Sci. 25: 214–30.

Tollner, E. W., Murphy, C., and Ramseur, E. L. 1994. Techniques and approaches for documenting plant root development with X-ray computed tomography. In: Anderson, S. H., Hopmans, J. W. , eds. Tomography of Soil–Water–Root Processes. SSSA Special Publication No. 36, Madison, WI, U.S.A.

Tyler, J. L., Strother, S. C., Zatorre, R. J., Alivisatos, B., Worsley, K. J., Diksic, M., and Yamamoto, Y. L. 1988. Stability of regional cerebral glucose metabolism in the normal brain measured by positron emission tomography. J. Nucl. Med. 29: 631–642.

Van Loan, C. F. 2000. The ubiquitous Kronecker product. J. Comput. App. Math. 123: 85–100.

Verbyla, A. P., and Cullis, B. R. 1992. The analysis of multistratum and spatially correlated repeated measures data. Biometrics 48: 1015–1032.

Via, B., So, C., Groom, L., Shupe, T. F. Stine, M., and Wikaira, J. 2007. Within tree variation of lignin, extractives and microfibril angle coupled with the theoretical and near infrared modeling of microfibril angle. IAWA J. 28: 189–209.

Wang, E. I. C., and Micko, M.M. 1984. Wood quality of white spruce from north central Alberta. Can. J. For. Res., 14: 181–185.

Wang, H., and West, M. 2009. Bayesian analysis of matrix normal graphical models. Biometrika 96 821–834.

Werner, K., Jansson, M., and Stoica, P. 2008. On estimation of covariance matrices with Kronecker product structure. IEEE Trans. Signal Process. 56: 479–491.

Werner, K., Jansson, M., and Stoica, P. 2007. Kronecker product covariance matrix estimation. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, Hawai'i, U.S.A.

Wilhelmsson, L., Arlinger, J., Spångberg, K., Lundqvist, S.-O., Grahn, T., Hedenberg, Ö., and Olsson, L. 2002. Models for predicting wood properties in stems of Picea abies and Pinus sylvestris in Sweden. Scand. J. For. Res. 17: 330–350.

Wimmer, R. 2002. Wood anatomical features in tree-rings as indicators of environmental change. Dendrochronologia, 20: 21–36.

Wolfinger, R. D. 1996. Heterogeneous variance: covariance structures for repeated measures. J. Agric. Biol. Envir. S. 1: 205–230.

Worsley, K. J., Evans, A. C., Strother, S. C., and Tyler, J. L. 1991. A linear spatial correlation model, with applications to positron emission tomography. J. Am. Stat. Assoc. 86: 55–67.

Zhang, S. Y., Yu, Q., and Beaulieu, J. 2004. Genetic variation in veneer quality and its correlation to growth in white spruce. Can. J. For. Res. 34: 1311–1318.

Appendix A. Matrix algebra and computational details for the evaluation of the bias of the ML estimator of a simply separable variance-covariance matrix

Prior to actually calculating $E(\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)$, each entry of the Kronecker product was re-expressed as a double sum of products of entries from two bilinear or quadratic forms in normal vectors, each of them defined by a random central matrix:

$$E(\hat{U}_{i_2 i_2'}\,\hat{U}_{i_1 i_1'}) = \sum_{k=1}^{K}\sum_{k'=1}^{K} E\left\{(\mathbf{x}_{i_2,k}{}^T (K n_1 \hat{\mathbf{U}}_1)^{-1} \mathbf{x}_{i_2',k})\,(\mathbf{x}_{i_1,k'}{}^T (K n_2 \hat{\mathbf{U}}_2)^{-1} \mathbf{x}_{i_1',k'})\right\}, \tag{A.1}$$

where $\hat{U}_{i_2 i_2'}$ is the entry $(i_2,\ i_2')$ of matrix $\hat{\mathbf{U}}_2$, idem for $\hat{U}_{i_1 i_1'}$, and

$$\begin{cases} \mathbf{x}_{i_2,k} \text{ and } \mathbf{x}_{i_2',k} \text{ are } n_1 \times 1 \text{ vectors made of columns } i_2,\ i_2' = 1, ..., n_2 \text{ of } \mathbf{X}_k - \bar{\mathbf{X}} \\ \mathbf{x}_{i_1,k'} \text{ and } \mathbf{x}_{i_1',k'} \text{ are } n_2 \times 1 \text{ vectors made of columns } i_1,\ i_1' = 1, ..., n_1 \text{ of } (\mathbf{X}_{k'} - \bar{\mathbf{X}})^T \end{cases} \tag{A.2}$$

for k, $k' = 1, \ldots, K$. Hereafter, the notations $\mathbf{x}_{i_1}$, $\mathbf{x}_{i_1'}$ and $\mathbf{x}_{i_2}$, $\mathbf{x}_{i_2'}$ are simplified to $\mathbf{x}_1$, $\mathbf{x}_1'$ and $\mathbf{x}_2$, $\mathbf{x}_2'$, respectively.

From Searle (1971, Chap. 2), the expected value of the product of two bilinear forms in normal vectors, defined by fixed central matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ and with zero means for the normal vectors, is:

$$E(\mathbf{x}_2{}^T \mathbf{A}_2\, \mathbf{x}_2'\, \mathbf{x}_1{}^T \mathbf{A}_1\, \mathbf{x}_1') =$$

$$tr[(\mathbf{A}_1 \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1} \mathbf{A}_2 \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2}) + (\mathbf{A}_1 \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1'} \mathbf{A}_2 \mathbf{C}_{\mathbf{x}_1 \mathbf{x}_2})] + tr(\mathbf{A}_2 \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_1})\, tr(\mathbf{A}_1 \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_2}), \tag{A.3}$$

where ^{T} denotes the transpose operator and $'$ is part of the notations of two of the normal vectors. The $\mathbf{C}$ matrices are covariance matrices between pairs of the normal vectors $\mathbf{x}_1$, $\mathbf{x}_1'$, $\mathbf{x}_2$, and $\mathbf{x}_2'$. Strictly speaking, the four random vectors are jointly normal:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_1' \\ \mathbf{x}_2 \\ \mathbf{x}_2' \end{pmatrix} \sim N_{n_2^2 n_1^2}\left(\begin{pmatrix} \mathbf{0} \\ \overline{\mathbf{0}} \\ \mathbf{0} \\ \overline{\mathbf{0}} \end{pmatrix}, \left(\begin{array}{cc|cc} \mathbf{C}_{\mathbf{x}_1 \mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_1 \mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_1 \mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_1 \mathbf{x}_2'} \\ \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2'} \\ \hline \mathbf{C}_{\mathbf{x}_2 \mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_2 \mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2 \mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_2 \mathbf{x}_2'} \\ \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_2' \mathbf{x}_2'} \end{array} \right) \right). \tag{A.4}$$

The application of (A.3) to calculate $E(\hat{U}_{i_2 i_2'}\, \hat{U}_{i_1 i_1'})$ (A.1) in the analysis of the ergodic (empirical) bias B_E (3.1) was made by using the observed values of $(K n_2 (\hat{\mathbf{U}}_2)_V)^{-1}$ and $(K n_1 (\hat{\mathbf{U}}_1)_V)^{-1}$ in place of $\mathbf{A}_1$ and $\mathbf{A}_2$ in (A.3). The computation of the **C** matrices is based on a property that can be found in Dutilleul (1999), among others. For an i.i.d. random sample $\mathbf{X}_1$, …, $\mathbf{X}_K \sim N_{n_1, n_2}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2)$, $\mathrm{cov}(X_{i_1 i_2, k}, X_{i_1' i_2', k}) = u_{i_1 i_1'}\, u_{i_2 i_2'}$ for any given k = 1, …, K and $\mathrm{cov}(\overline{X}_{i_1 i_2}, \overline{X}_{i_1' i_2'}) = \frac{1}{K}\, u_{i_1 i_1'}\, u_{i_2 i_2'}$. For $\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1}$, $\mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2}$, $\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1'}$, and $\mathbf{C}_{\mathbf{x}_1 \mathbf{x}_2}$, the calculations are different depending on whether $k \neq k'$ or $k = k'$ in (A.1). If $k \neq k'$, then

$$\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1} = -\frac{1}{K} (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1.})^T, \tag{A.5}$$

$$\mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2} = -\frac{1}{K} (\mathbf{u}_{.i_2}\, \mathbf{u}_{i_1'.}), \tag{A.6}$$

$$\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1'} = -\frac{1}{K} (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1'.})^T \text{ and} \tag{A.7}$$

$$\mathbf{C}_{\mathbf{x}_1 \mathbf{x}_2} = -\frac{1}{K} (\mathbf{u}_{.i_2}\, \mathbf{u}_{i_1.}), \tag{A.8}$$

where $\mathbf{u}_{.i_2}$ denotes the i_2-th column of $\mathbf{U}_2$, and $\mathbf{u}_{i_1.}$, the i_1-th row of $\mathbf{U}_1$, etc. On the other hand, if $k = k'$, then

$$\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1} = (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1.})^T - \frac{1}{K} (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1.})^T, \tag{A.9}$$

$$\mathbf{C}_{\mathbf{x}_1' \mathbf{x}_2} = (\mathbf{u}_{.i_2}\, \mathbf{u}_{i_1'.}) - \frac{1}{K} (\mathbf{u}_{.i_2}\, \mathbf{u}_{i_1'.}), \tag{A.10}$$

$$\mathbf{C}_{\mathbf{x}_2' \mathbf{x}_1'} = (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1'.})^T - \frac{1}{K} (\mathbf{u}_{.i_2'}\, \mathbf{u}_{i_1'.})^T \text{ and} \tag{A.11}$$

$$\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} = (\mathbf{u}_{.i_2}\,\mathbf{u}_{i_1.}) - \frac{1}{K}(\mathbf{u}_{.i_2}\,\mathbf{u}_{i_1.}). \tag{A.12}$$

Finally, the matrices $\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1}$ and $\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2}$ do not depend on whether k, k' are different or equal:

$$\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1} = u_{i_1 i_1'}\mathbf{U}_2 - \frac{1}{K}u_{i_1 i_1'}\mathbf{U}_2 \quad \text{and} \tag{A.13}$$

$$\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2} = u_{i_2 i_2'}\mathbf{U}_1 - \frac{1}{K}u_{i_2 i_2'}\mathbf{U}_1. \tag{A.14}$$

Appendix B. Estimated Generalized Least Squares (EGLS)

For Mean Model 2, to obtain $\text{vec}(\hat{\mathbf{M}}) = \mathbf{D}\hat{\boldsymbol{\beta}}_{EGLS}$, $\boldsymbol{\beta}_{EGLS}$ is estimated as

$$\hat{\boldsymbol{\beta}}_{EGLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\hat{\boldsymbol{\Sigma}}^{-1}\text{vec}(\mathbf{X}_k) \tag{B.1}$$

where $\mathbf{D}$ is a fixed design matrix with dimensions $n_1n_2 \times q$, and $\hat{\boldsymbol{\beta}}_{EGLS}$ is $q \times 1$. The estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ is given by $\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$ under the matrix normal distribution model, and by $\hat{\boldsymbol{\Sigma}}_{UN}$ under the vector normal distribution model. If $\hat{\boldsymbol{\Sigma}} = \mathbf{I}_{n_1n_2}$, then the GLS estimator reduces to the OLS estimator (Searle 1971, p. 87), used to initiate the algorithm below.

Computation under the matrix normal distribution model

Initialization: Step1 = step2 = 0; $\boldsymbol{\beta}^*_{EGLS} = \hat{\boldsymbol{\beta}}^0_{EGLS} = \hat{\boldsymbol{\beta}}_{OLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\text{vec}(\mathbf{X}_k)$

and $\text{vec}(\mathbf{M}^*) = \text{vec}(\hat{\mathbf{M}}^0) = \mathbf{D}\hat{\boldsymbol{\beta}}^0_{EGLS}$; $\mathbf{U}_2^* = \hat{\mathbf{U}}_2^0$;

$\mathbf{U}_1^* = \sum_{k=1}^{K}(\mathbf{X}_k - \mathbf{M}^*)(Kn_2\mathbf{U}_2^*)^{-1}(\mathbf{X}_k - \mathbf{M}^*)^T$; ϵ_1, ϵ_2 and ϵ_3 are set at 'infinitesimally small' values.

Step1 = 1;

$$\mathbf{U}_2^+ = \sum_{k=1}^{K}(\mathbf{X}_k - \mathbf{M}^*)^T(Kn_1\mathbf{U}_1^*)^{-1}(\mathbf{X}_k - \mathbf{M}^*); \tag{B.2}$$

$$\mathbf{U}_1^+ = \sum_{k=1}^{K}(\mathbf{X}_k - \mathbf{M}^*)(Kn_2\mathbf{U}_2^+)^{-1}(\mathbf{X}_k - \mathbf{M}^*)^T; \tag{B.3}$$

While: $\| \mathbf{U}_1^+ - \mathbf{U}_1^* \| > \epsilon_1$ or $\| \mathbf{U}_2^+ - \mathbf{U}_2^* \| > \epsilon_2$, repeat

Step1 = step1 + 1;

$\mathbf{U}_1^* = \mathbf{U}_1^+$;

$\mathbf{U}_2^* = \mathbf{U}_2^+$;

Recompute equations (B.2) and (B.3);

End

Step2 = 1;

$$\boldsymbol{\beta}^{+}_{EGLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^{T}(\mathbf{U}_2^{+}\otimes\mathbf{U}_1^{+})^{-1}\mathbf{D})^{-1}\mathbf{D}^{T}(\mathbf{U}_2^{+}\otimes\mathbf{U}_1^{+})^{-1}\mathrm{vec}(\mathbf{X}_k); \tag{B.4}$$

Calculate $\mathrm{vec}(\mathbf{M}^{+}) = \mathbf{D}\boldsymbol{\beta}^{+}_{EGLS}$;

While $\| \mathbf{M}^{+} - \mathbf{M}^{*} \| > \epsilon_3$, repeat

While: $\| \mathbf{U}_1^{+} - \mathbf{U}_1^{*} \| > \epsilon_1$ or $\| \mathbf{U}_2^{+} - \mathbf{U}_2^{*} \| > \epsilon_2$, repeat

Step1 = step1 + 1;

$\mathbf{U}_1^{*} = \mathbf{U}_1^{+}$;

$\mathbf{U}_2^{*} = \mathbf{U}_2^{+}$;

$$\mathbf{U}_2^{+} = \sum_{k=1}^{K}(\mathbf{X}_k - \mathbf{M}^{+})^{T}(Kn_1\mathbf{U}_1^{*})^{-1}(\mathbf{X}_k - \mathbf{M}^{+}); \tag{B.5}$$

$$\mathbf{U}_1^{+} = \sum_{k=1}^{K}(\mathbf{X}_k - \mathbf{M}^{+})(Kn_2\mathbf{U}_2^{+})^{-1}(\mathbf{X}_k - \mathbf{M}^{+})^{T}; \tag{B.6}$$

Recompute equations (B.5) and (B.6);

End

Step2 = step2 + 1;

$\mathbf{M}^{*} = \mathbf{M}^{+}$;

Recompute equation (B.4) and recalculate vec($\mathbf{M}^{+}$);

End

Solutions are $\hat{\mathbf{U}}_1 = \mathbf{U}_1^{+}$; $\hat{\mathbf{U}}_2 = \mathbf{U}_2^{+}$; $\hat{\mathbf{M}} = \mathbf{M}^{+}$.

Computation under the vector normal distribution model

Initialization: Step3 = 0; $\boldsymbol{\beta}^{*}_{GLS} = \hat{\boldsymbol{\beta}}^{0}_{GLS} = \hat{\boldsymbol{\beta}}_{OLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^{T}\mathbf{D})^{-1}\mathbf{D}^{T}\mathrm{vec}(\mathbf{X}_k)$ and $\mathrm{vec}(\mathbf{M}^{*})$

$= \mathrm{vec}(\hat{\mathbf{M}}^{0}) = \mathbf{D}\hat{\boldsymbol{\beta}}^{0}_{EGLS}$; $\boldsymbol{\Sigma}^{*}_{UN} = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{*})\right\}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{*})\right\}^{T}$; ϵ_4

and ϵ_5 are set at 'infinitesimally small' values.

Step3 = 1;

$$\boldsymbol{\beta}^{+}_{EGLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^{T}{\boldsymbol{\Sigma}^{*}_{UN}}^{-1}\mathbf{D})^{-1}\mathbf{D}^{T}{\boldsymbol{\Sigma}^{*}_{UN}}^{-1}\mathrm{vec}(\mathbf{X}_k);$$

Calculate $\mathrm{vec}(\mathbf{M}^{+}) = \mathbf{D}\boldsymbol{\beta}^{+}_{EGLS}$;

$$\boldsymbol{\Sigma}^{+}_{UN} = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{+})\right\}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{+})\right\}^{T};$$

While: $\| \boldsymbol{\Sigma}^{+}_{UN} - \boldsymbol{\Sigma}^{*}_{UN} \| > \epsilon_4$ or $\| \mathbf{M}^{+} - \mathbf{M}^{*} \| > \epsilon_5$, repeat

Step3 = step3 + 1;

$\boldsymbol{\Sigma}^{*} = \boldsymbol{\Sigma}^{+}$;

$\mathbf{M}^{*} = \mathbf{M}^{+}$;

$$\boldsymbol{\Sigma}^{+}_{UN} = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{*})\right\}\left\{\mathrm{vec}(\mathbf{X}_k) - \mathrm{vec}(\mathbf{M}^{*})\right\}^{T}; \quad \text{(B.7)}$$

$$\boldsymbol{\beta}^{+}_{EGLS} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}^{T}{\boldsymbol{\Sigma}^{+}_{UN}}^{-1}\mathbf{D})^{-1}\mathbf{D}^{T}{\boldsymbol{\Sigma}^{+}_{UN}}^{-1}\mathrm{vec}(\mathbf{X}_k); \quad \text{(B.8)}$$

Calculate $\mathrm{vec}(\mathbf{M}^{+}) = \mathbf{D}\boldsymbol{\beta}^{+}_{EGLS}$;

Recompute equations (B.7) and (B.8);

End

Solutions are $\hat{\boldsymbol{\Sigma}}_{UN} = \boldsymbol{\Sigma}^{+}$ and $\hat{\mathbf{M}} = \mathbf{M}^{+}$.

Matlab program

A customized Matlab program (The Mathworks Inc., 2010) was used to generate pseudo-random vectors from a vector normal distribution model with $n_1 = 4$ and $n_2 = 3$, using the Ziggurat method to generate pseudo-random values from a univariate normal distribution (Marsaglia and Tsang 2000). The mean models were of response surface type

quadratic (6 parameters) $\mathbf{M} = \{m_{i_1 i_2}\} = 1 + 9(i_1) + 10(i_2) + 4(i_1 i_2) + 2(i_1^2) + 3(i_2^2)$ (B.9)

and linear (3 parameters) $\mathbf{M} = \{m_{i_1 i_2}\} = 1 + 9(i_1) + 10(i_2)$, (B.10)

where $i_1 = 1...n_1$ and $i_2 = 1...n_2$. The mean parameters were estimated by EGLS, with $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4 = \epsilon_5 = 1E-6$, and $\hat{\mathbf{U}}_2^{0} = \mathbf{I}_{n_2}$ for the matrix normal distribution model.

Note that for $K = n_1 n_2 + 1$, the EGLS algorithm did not converge in a reasonable time (i.e., it had not converged even after several days) for a number of simulation runs ranging from 18 out of 1.0*E*5 when $n_1 = 2, n_2 = 2$ to 407 out of 1.0*E*5 when $n_1 = 5, n_2 = 5$. Supplementary simulations were then performed to obtain the needed results for 1.0*E*5 simulation runs.

Appendix C. Derivatives of the log-likelihood function for the tensor normal distribution model with respect to variance-covariance matrices

The first derivatives of the log-likelihood function l (equation (5.10)) with respect to $\mathbb{M}$ and $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ are:

$$\frac{\delta}{\delta \mathbb{M}} l = -\left\{ \sum_{k=1}^{K} (\mathbb{X}_k - \mathbb{M}) \right\} \times_{1\ldots J} \left(\overset{3}{\underset{j=1}{\circ}} \mathbf{U}_j^{-1} \right) \tag{C.1}$$

;

$$\frac{\delta}{\delta \mathbf{U}_1} l = Kn_2 n_3 \mathbf{U}_1^{-1} - \frac{Kn_2 n_3}{2} \operatorname{diag}(\mathbf{U}_1^{-1}) - \mathbf{U}_1^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}{}^T \right] \right\} \mathbf{U}_1^{-1}$$
$$+ \frac{1}{2} \operatorname{diag} \left[\mathbf{U}_1^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}{}^T \right] \right\} \mathbf{U}_1^{-1} \right] \tag{C.2}$$

;

$$\frac{\delta}{\delta \mathbf{U}_2} l = Kn_1 n_3 \mathbf{U}_2^{-1} - \frac{Kn_1 n_3}{2} \operatorname{diag}(\mathbf{U}_2^{-1}) - \mathbf{U}_2^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(2)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)}{}^T \right] \right\} \mathbf{U}_2^{-1}$$
$$+ \frac{1}{2} \operatorname{diag} \left[\mathbf{U}_2^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(2)} (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)}{}^T \right] \right\} \mathbf{U}_2^{-1} \right] \tag{C.3}$$

;

$$\frac{\delta}{\delta \mathbf{U}_3} l = Kn_1 n_2 \mathbf{U}_3^{-1} - \frac{Kn_1 n_2}{2} \operatorname{diag}(\mathbf{U}_3^{-1}) - \mathbf{U}_3^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(3)} (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)}{}^T \right] \right\} \mathbf{U}_3^{-1}$$
$$+ \frac{1}{2} \operatorname{diag} \left[\mathbf{U}_3^{-1} \left\{ \sum_{k=1}^{K} \left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(3)} (\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)}{}^T \right] \right\} \mathbf{U}_3^{-1} \right] \tag{C.4}$$

Note that equations (C.2)–(C.4) can be rewritten in a more general form as follows

$$\frac{\delta}{\delta \mathbf{U}_j} l = Kn_{(j)}\mathbf{U}_j^{-1} - \frac{Kn_{(j)}}{2}\mathrm{diag}(\mathbf{U}_j^{-1}) - \mathbf{U}_j^{-1}\left\{\sum_{k=1}^{K}\left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(j)}\,(\underset{\substack{j'=J\\ j'\neq j}}{\overset{1}{\otimes}}\mathbf{U}_{j'})^{-1}(\mathbb{X}_k - \overline{\mathbb{X}})_{(j)}{}^{T}\right]\right\}\mathbf{U}_j^{-1}$$

$$+\frac{1}{2}\mathrm{diag}\left[\mathbf{U}_j^{-1}\left\{\sum_{k=1}^{K}\left[(\mathbb{X}_k - \overline{\mathbb{X}})_{(j)}\,(\underset{\substack{j'=J\\ j'\neq j}}{\overset{1}{\otimes}}\mathbf{U}_{j'})^{-1}(\mathbb{X}_k - \overline{\mathbb{X}})_{(j)}{}^{T}\right]\right\}\mathbf{U}_j^{-1}\right] \tag{C.5}$$

with $j = 1, \ldots, J = 3$ and $n_{(j)} = \underset{\substack{j'=1\\ j'\neq j}}{\overset{J}{\Pi}} n_{j'}$.

Appendix D. Bias decomposition for the maximum likelihood estimator of a doubly separable variance-covariance matrix

D.1 Introduction

The tensor normal distribution model and the MLE-3D algorithm used for ML estimation were presented in Chapter 5. That chapter focused on a complete presentation of the estimation of the model parameters, with simulation results and an example with brain data. The objective of Appendix D is to decompose the bias of the ML estimator of a doubly separable variance-covariance matrix, under the tensor normal distribution model. Specifically, the ergodic, estimation and fluctuation biases and the non-orthogonality component were evaluated with simulations, in a way similar to Chapter 3 (Bias decomposition for the maximum likelihood estimator of a simply separable variance-covariance matrix). This appendix complements the simulation study presented in Chapter 5, by performing simulations starting from the minimum sample size K_{min}, for a larger number of values of *K*, and with a much larger number of simulation runs.

Such a study is necessary to answer the question: Is there a non-monotonic relationship between the empirical bias and the sample size in the doubly separable case also? To answer this question, the decomposition of the empirical bias in the doubly separable case is presented in Section D.2; the matrix algebra details are given in Section D.3; the components of the ergodic bias, together with a pseudo-theoretical bias, are evaluated in Section D.4; and the results are summarized and discussed in Section D.5.

D.2 Bias decomposition

The decomposition of the ergodic bias (B_E) into an estimation bias (B_S) and a fluctuation bias (B_F) minus a non-orthogonality factor (δ) is an extension from the case of a simply separable variance-covariance matrix to the case of a doubly separable variance-covariance matrix. The steps in the reasoning are the same as in Chapter 3, but with the object of estimation $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$ instead of $\mathbf{U}_2 \otimes \mathbf{U}_1$. The three biases are:

$$B_E = \| E(\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) \|, \tag{D.1}$$

$$B_S = \| E(\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - (\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V) \| \text{ and} \tag{D.2}$$

$$B_F = \| E((\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) \|, \tag{D.3}$$

where subscript V in $(\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)_V$ represents the average over V simulation runs with non-singular results (see the following section for simulation conditions). The decomposition reads as follows:

$$B_E = B_S + B_F - \delta. \tag{D.4}$$

As in the 2-D case, the pseudo-theoretical bias is calculated as:

$$B_T = \| E(\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) \|, \tag{D.5}$$

where the theoretical matrices $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ are used two by two, instead of the estimated matrices $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$, through the inverse of their Kronecker product (i.e. $(Kn_2n_1(\mathbf{U}_2 \otimes \mathbf{U}_1))^{-1}$, $(Kn_3n_1(\mathbf{U}_3 \otimes \mathbf{U}_1))^{-1}$, $(Kn_2n_3(\mathbf{U}_3 \otimes \mathbf{U}_2))^{-1}$) in the definition of the quadratic forms; see equation (D.8) below. No simulations are needed for the calculation of the pseudo-theoretical bias. Matrix algebra details concerning the calculation of biases are given in the next section.

D.3 Matrix algebra details

Prior to calculating $E(\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)$, each entry is re-expressed as a triple sum of expected values of products of entries from three bilinear or quadratic forms in normal vectors, each defined by a random matrix at the centre:

$$\begin{aligned} &E(\hat{U}_{i_3 i_3'} \hat{U}_{i_2 i_2'} \hat{U}_{i_1 i_1'}) \\ &= \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{k''=1}^{K} E\{(\mathbf{x}_{i_3,k}{}^T (Kn_1n_2\hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} \mathbf{x}_{i_3',k})(\mathbf{x}_{i_2,k'}{}^T (Kn_1n_3\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} \mathbf{x}_{i_2',k'}) \\ &(\mathbf{x}_{i_1,k''}{}^T (Kn_2n_3\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} \mathbf{x}_{i_1',k''})\} \end{aligned} \tag{D.6}$$

where $\hat{U}_{i_3 i_3'}$ is entry (i_3, i_3') of matrix $\hat{\mathbf{U}}_3$ (idem for $\hat{U}_{i_2 i_2'}$ and $\hat{U}_{i_1 i_1'}$), and

$$\begin{cases} \mathbf{x}_{i_3,k} \text{ and } \mathbf{x}_{i_3',k} \text{ are } n_1n_2 \times 1 \text{ vectors made of columns } i_3, i_3' = 1 \ldots n_3 \text{ of } (\mathbb{X}_k - \bar{\mathbb{X}})_{(3)}{}^T; \\ \mathbf{x}_{i_2,k'} \text{ and } \mathbf{x}_{i_2',k'} \text{ are } n_1n_3 \times 1 \text{ vectors made of columns } i_2, i_2' = 1 \ldots n_2 \text{ of } (\mathbb{X}_{k'} - \bar{\mathbb{X}})_{(2)}{}^T; \\ \mathbf{x}_{i_1,k''} \text{ and } \mathbf{x}_{i_1',k''} \text{ are } n_2n_3 \times 1 \text{ vectors made of columns } i_1, i_1' = 1 \ldots n_1 \text{ of } (\mathbb{X}_{k''} - \bar{\mathbb{X}})_{(1)}{}^T; \end{cases} \tag{D.7}$$

for $k,k',k''=1...K$. Hereafter, notations $\mathbf{x}_{i_1}$, $\mathbf{x}_{i_1'}$, $\mathbf{x}_{i_2}$, $\mathbf{x}_{i_2'}$ and $\mathbf{x}_{i_3}$, $\mathbf{x}_{i_3'}$ are simplified into $\mathbf{x}_1$, $\mathbf{x}_1'$, $\mathbf{x}_2$, $\mathbf{x}_2'$ and $\mathbf{x}_3$, $\mathbf{x}_3'$.

Based on equation (2.27) in Ghazal (2000), the expected value of the product of three bilinear forms in normal vectors, when each of the two vectors in each bilinear form has an expected value of zero, is:

$$
\begin{aligned}
&E(\hat{U}_{i_3i_3'}\hat{U}_{i_2i_2'}\hat{U}_{i_1i_1'}) \\
&=\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{k''=1}^{K}\{tr(\mathbf{A}_1\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1})[tr(\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2})tr(\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3})+tr(\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_2'})+tr(\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_2'})] \\
&+[tr(\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3})tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2}\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1'})+tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2}\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1'})+tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2}\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'})] \\
&+[tr(\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3})tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2'}\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_1'})+tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2'}\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1'})+tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2'}\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'})] \\
&+[tr(\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2})tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1'})+tr(\mathbf{A}_1\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2}\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1})+tr(\mathbf{A}_1\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2'}\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1})] \\
&+[tr(\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2})tr(\mathbf{A}_1{}^T\mathbf{C}_{\mathbf{x}_1\mathbf{x}_3'}\mathbf{A}_3{}^T\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'})+tr(\mathbf{A}_1\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2}\mathbf{A}_2\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1})+tr(\mathbf{A}_1\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2'}\mathbf{A}_2{}^T\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3}\mathbf{A}_3\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1})]\}
\end{aligned}
\tag{D.8}
$$

where ^{T} denotes the transpose operator and $'$ is part of the notation (subscript) of one of the two normal vectors. Equation (D.8) was used to calculate $E(\hat{U}_{i_3i_3'}\hat{U}_{i_2i_2'}\hat{U}_{i_1i_1'})$ in the analysis of the ergodic bias B_E, by substituting the observed values of $(Kn_2n_3(\hat{\mathbf{U}}_3\otimes\hat{\mathbf{U}}_2)_V)^{-1}$, $(Kn_3n_1(\hat{\mathbf{U}}_3\otimes\hat{\mathbf{U}}_1)_V)^{-1}$ and $(Kn_2n_1(\hat{\mathbf{U}}_2\otimes\hat{\mathbf{U}}_1)_V)^{-1}$ to $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$. (In the case of the theoretical bias B_T, $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ were replaced by the fixed values $(Kn_2n_1(\mathbf{U}_2\otimes\mathbf{U}_1))^{-1}$, $(Kn_3n_1(\mathbf{U}_3\otimes\mathbf{U}_1))^{-1}$, $(Kn_2n_3(\mathbf{U}_3\otimes\mathbf{U}_2))^{-1}$.

The $\mathbf{C}$ matrices are variance-covariance matrices or covariance matrices for the normal vectors $\mathbf{x}_1$, $\mathbf{x}_1'$, $\mathbf{x}_2$, $\mathbf{x}_2'$, $\mathbf{x}_3$ and $\mathbf{x}_3'$, taken two by two. The six random vectors are jointly normal:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_1' \\ \mathbf{x}_2 \\ \mathbf{x}_2' \\ \mathbf{x}_3 \\ \mathbf{x}_3' \end{pmatrix} \sim N_{n_1^4 n_2^4 n_3^4} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{C}_{\mathbf{x}_1\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_2'} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_3} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_3'} \\ \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_2'} & \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_3} & \mathbf{C}_{\mathbf{x}_1'\mathbf{x}_3'} \\ \mathbf{C}_{\mathbf{x}_2\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}_2'} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}_3} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}_3'} \\ \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_2'\mathbf{x}_3'} \\ \mathbf{C}_{\mathbf{x}_3\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_3\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_3\mathbf{x}_3'} \\ \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1} & \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_1'} & \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_2'} & \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3} & \mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3'} \end{pmatrix} \right), \tag{D.9}$$

where $\mathbf{C}_{\mathbf{x}_1\mathbf{x}_1}$ is the $n_2n_3 \times n_2n_3$ variance-covariance matrix of normal vector $\mathbf{x}_1$, $\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2}$ is the $n_2n_3 \times n_1n_3$ covariance matrix between normal vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, etc.

The computation of the $\mathbf{C}$ matrices is based on a property that can be found in Dutilleul (1999), among others. For an i.i.d. random sample $\mathbf{X}_1$, …, $\mathbf{X}_K \sim N_{n_1,n_2,n_3}(\mathbf{M}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$, $\text{cov}(X_{i_1i_2i_3}, X_{i_1'i_2'i_3'}) = u_{i_1i_1'}u_{i_2i_2'}u_{i_3i_3'}$ for any given k = 1, …, K, and $\text{cov}(\bar{X}_{i_1i_2i_3}, \bar{X}_{i_1'i_2'i_3'}) = \frac{1}{K}u_{i_1i_1'}u_{i_2i_2'}u_{i_3i_3'}$.

For the covariance matrices in equation (D.9), calculations are different depending on whether $k = k' = k''$; $k \neq k' \neq k''$; $k \neq k', k \neq k'', k' = k''$; $k = k', k \neq k'', k' \neq k''$; or $k \neq k', k = k'', k' \neq k''$. When $k = k' = k''$,

$$\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} = (\mathbf{U}_3 \otimes \mathbf{u}_{.i_2}\mathbf{u}_{i_1.}) - \frac{1}{K}(\mathbf{U}_3 \otimes \mathbf{u}_{.i_2}\mathbf{u}_{i_1.}) \tag{D.10}$$

$$\mathbf{C}_{\mathbf{x}_2\mathbf{x}_1} = (\mathbf{U}_3 \otimes \mathbf{u}_{.i_1}\mathbf{u}_{i_2.}) - \frac{1}{K}(\mathbf{U}_3 \otimes \mathbf{u}_{.i_1}\mathbf{u}_{i_2.}) \tag{D.11}$$

$$\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3} = (\mathbf{u}_{.i_3}\mathbf{u}_{i_2.} \otimes \mathbf{U}_1) - \frac{1}{K}(\mathbf{u}_{.i_3}\mathbf{u}_{i_2.} \otimes \mathbf{U}_1) \tag{D.12}$$

$$\mathbf{C}_{\mathbf{x}_3\mathbf{x}_2} = (\mathbf{u}_{.i_2}\mathbf{u}_{i_3.} \otimes \mathbf{U}_1) - \frac{1}{K}(\mathbf{u}_{.i_2}\mathbf{u}_{i_3.} \otimes \mathbf{U}_1) \tag{D.13}$$

$$\mathbf{C}_{\mathbf{x}_1\mathbf{x}_3} = (\mathbf{U}_2 \otimes \mathbf{u}_{i_1.} \otimes \mathbf{u}_{.i_3}) - \frac{1}{K}(\mathbf{U}_2 \otimes \mathbf{u}_{i_1.} \otimes \mathbf{u}_{.i_3}) \tag{D.14}$$

$$\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1} = (\mathbf{u}_{i_3.} \otimes \mathbf{U}_2 \otimes \mathbf{u}_{.i_1}) - \frac{1}{K}(\mathbf{u}_{i_3.} \otimes \mathbf{U}_2 \otimes \mathbf{u}_{.i_1}), \tag{D.15}$$

where $\mathbf{u}_{.i_2}$ denotes the i_2-th column of $\mathbf{U}2$, $\mathbf{u}_{i_1.}$ denotes the i_1-th row of $\mathbf{U}_1$, etc. Note that to compute $\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2'}$ for example, it suffices to replace i_2 by i_2' in the relevant equation above. Depending on the values of k, k' and k'', the second term on the right-hand side of the equations above is equal to zero, or not. For example, when $k \neq k'$, $k \neq k''$ and $k' = k''$, then $\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} = (\mathbf{U}_3 \otimes \mathbf{u}_{.i_2}\mathbf{u}_{i_1.})$, $\mathbf{C}_{\mathbf{x}_2\mathbf{x}_1} = (\mathbf{U}_3 \otimes \mathbf{u}_{.i_1}\mathbf{u}_{i_2.})$,

$\mathbf{C}_{\mathbf{x}_2\mathbf{x}_3} = (\mathbf{u}_{.i_3}\mathbf{u}_{i_2.} \otimes \mathbf{U}_1) - \frac{1}{K}(\mathbf{u}_{.i_3}\mathbf{u}_{i_2.} \otimes \mathbf{U}_1)$,

$\mathbf{C}_{\mathbf{x}_3\mathbf{x}_2} = (\mathbf{u}_{.i_2}\mathbf{u}_{i_3.} \otimes \mathbf{U}_1) - \frac{1}{K}(\mathbf{u}_{.i_2}\mathbf{u}_{i_3.} \otimes \mathbf{U}_1)$,

$\mathbf{C}_{\mathbf{x}_1\mathbf{x}_3} = (\mathbf{U}_2 \otimes \mathbf{u}_{i_1.} \otimes \mathbf{u}_{.i_3})$, and $\mathbf{C}_{\mathbf{x}_3\mathbf{x}_1} = (\mathbf{u}_{i_3.} \otimes \mathbf{U}_2 \otimes \mathbf{u}_{.i_1})$.

Finally, the second term on the right-hand side of the equations is always present for the matrices $\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1}$, $\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2}$ and $\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3}$, whether k, k' and k'' are equal or not:

$$\mathbf{C}_{\mathbf{x}_1'\mathbf{x}_1} = (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes u_{i_1'i_1}) - \frac{1}{K}(\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes u_{i_1'i_1}), \tag{D.16}$$

$$\mathbf{C}_{\mathbf{x}_2'\mathbf{x}_2} = (\mathbf{U}_3 \otimes \mathbf{U}_1 \otimes u_{i_2'i_2}) - \frac{1}{K}(\mathbf{U}_3 \otimes \mathbf{U}_1 \otimes u_{i_2'i_2}) \text{ and} \tag{D.17}$$

$$\mathbf{C}_{\mathbf{x}_3'\mathbf{x}_3} = (\mathbf{U}_2 \otimes \mathbf{U}_1 \otimes u_{i_3'i_3}) - \frac{1}{K}(\mathbf{U}_2 \otimes \mathbf{U}_1 \otimes u_{i_3'i_3}). \tag{D.18}$$

D.4 Simulation procedures

As in Chapter 5, a customized algorithm written in Matlab 2010a (The MathWorks Inc., Natick, MA) was used to generate observations from $\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)$, with $\text{vec}(\mathbb{X}) = \text{vec}(\mathbb{M}) + \mathbf{C}^T\text{vec}(\mathbb{Z})$, following the algorithm of Marsaglia and Tsang (2000). The mean tensor $\mathbb{M}$ was set at zero, and $\mathbf{U}_1 = \mathbf{I}_{n_1}$, $\mathbf{U}_2 = \mathbf{I}_{n_2}$ and $\mathbf{U}_3 = \mathbf{I}_{n_3}$. The infinitesimal quantities used to determine that the MLE-3D algorithm had converged were ϵ_1 = ϵ_2 = ϵ_3 = 1.0*E*-6, or the MLE-3D algorithm was stopped after 1.0*E*7 iterations. Cases where the number of iterations reached 1.0*E*7 were rare; in particular, there were seven of them when $n_1 = n_2 = n_3 = 2$. In all cases, the initial solutions were $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$ and $\hat{\mathbf{U}}_3^0 = \mathbf{I}_{n_3}$. A total of V = 1.0*E*6

estimates of $\hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1$, provided by as many simulation runs and permissible (non-singular) final solutions $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{U}}_3$ of the MLE algorithm at convergence, were used in the evaluation of the estimation and fluctuations biases B_S and B_F.

Simulations were performed with $K = K_{min}, \ldots, 25$ by steps of 1, in addition to 30, 35, 40, 45, 50, and 1000 (for one of the scenarios), where K_{min} is equal to INT[$\max(n_1 / n_2 n_3, n_2 / n_1 n_3, n_3 / n_1 n_2)$] + 1 if $\max(n_1 / n_2 n_3, n_2 / n_1 n_3, n_3 / n_1 n_2)$ is an integer number and INT[$\max(n_1 / n_2 n_3, n_2 / n_1 n_3, n_3 / n_1 n_2)$] + 2 otherwise. Finally, bias values were standardized relatively to the corresponding $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|$.

D.5 Results and discussion

For the simplest 3-D model possible (with all dimension sizes greater than 1), that is, $n_1 = n_2 = n_3 = 2$ so that $K_{min} = 2$, Figure D.1 shows that the ergodic bias decreases monotonically, while there is a clear 'trough' in the estimation and fluctuation biases. However, no 'peak' is observed in any of the three biases. When $K = 1000$, the ergodic, estimation, and fluctuation biases are equal to 3.717*E*-4, 6.316*E*-4, and 1.003*E*-4, respectively. These values are very small, thus the bias tends to vanish when K becomes larger.

In the second 3-D scenario considered, the six combinations of values of 3, 4 and 5 for n_1, n_2 and n_3 (so that $K_{min} = 2$ again) are explored (Figure D.2). The ergodic and fluctuations bias decrease monotonically; the ergodic bias has values close to those of the fluctuation bias and the two biases cannot be distinguished in some of the panels of Figure D.2. The estimation bias presents an elongated 'trough' pattern, and no 'peak' can be observed. Note that the case $n_1 = 5$, $n_2 = 3$ and $n_3 = 4$ was studied in Chapter 5, and the results presented for the ergodic bias here confirm those of the empirical bias in Chapter 5.

Finally, the pseudo-theoretical bias was calculated for $n_1 = n_2 = n_3 = 2$ and for $n_1 = 3$, $n_2 = 4$ and $n_3 = 5$ (Figure D.3). (Numerical results are almost identical for the other five combinations of values of 3, 4 and 5 for n_1, n_2 and n_3). The pseudo-theoretical bias decreases monotonically without exception (Figure D.3), as expected.

In summary, in both sets of scenarios (Figures D.1 and D.2), the 'peak' is absent; in Figure D.1, the 'trough' in the estimation and fluctuation biases is clearly present; and in Figure D.2, the 'trough' is present only in the estimation bias via a more subtle, elongated pattern. This is not very different from the 2-D case, where as the size of the dimensions increased, the 'peak-and-trough' pattern (always present in at least one of the three biases) was becoming more subtle. Also as in the 2-D case, the observed patterns are stable.

The results obtained in 3-D under the tensor normal distribution model do not preclude the use of the MLE-3D algorithm. There is no 'peak' in the bias of the ML estimator of a doubly separable variance-covariance matrix, but the bias is very large at values of K_{min} + 1 and K_{min} + 2. Thus, as in Chapter 3, a larger sample size than the minimum required for ML estimation can be useful to prevent a very large bias.

In conclusion, the 'peak-and-trough' pattern in the bias of the ML estimator of a separable variance-covariance matrix is not specific to the 2-D case, and is also found in the 3-D case. In both cases, since the pseudo-theoretical bias decreases monotonically with increasing sample size, the results do not challenge the theory of ML estimation. Instead, the results presented in this appendix as well as in Chapter 3 highlight the implications of the use of numerical methods for solving estimation equations that have no analytical solutions.

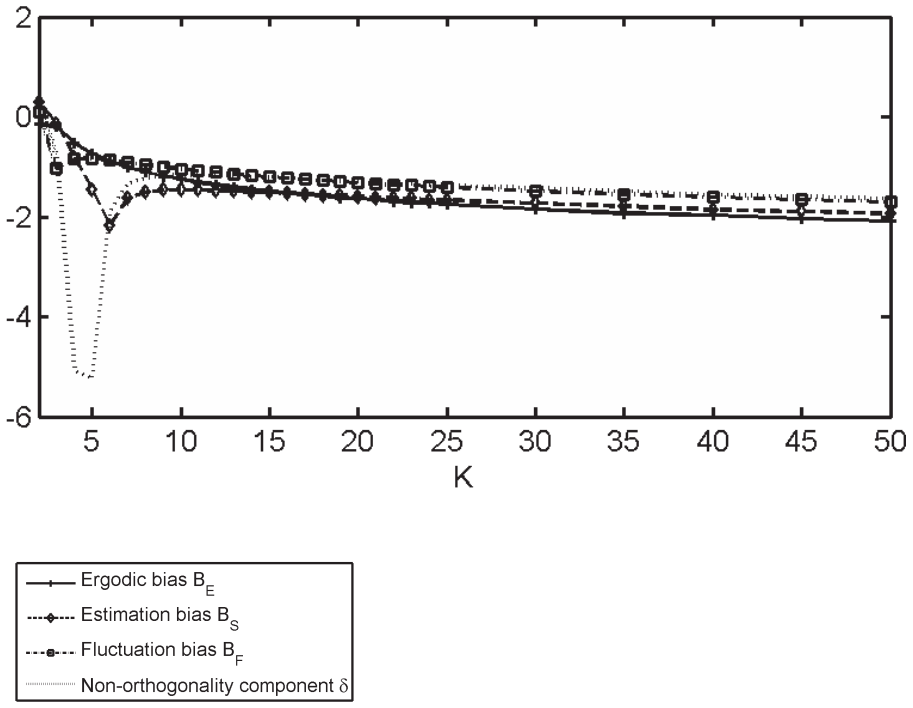


Figure D.1: Standardized values (i.e. divided by $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|$) of the ergodic, estimation, and fluctuation biases and the non-orthogonality component for the ML estimator of $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$, with $n_1 = n_2 = n_3 = 2$ and sample size $K = K_{min}, \ldots, 25$ by steps of 1 in addition to 30, 35, 40, 45, and 50. The vertical axis is in logarithmic scale, basis 10 ('0' means 1.0*E*0 = 1; '1' means 1.0*E*1 = 10, etc.). There is an atypical 'trough' in some of the biases and the non-orthogonality component as the sample size K increases.

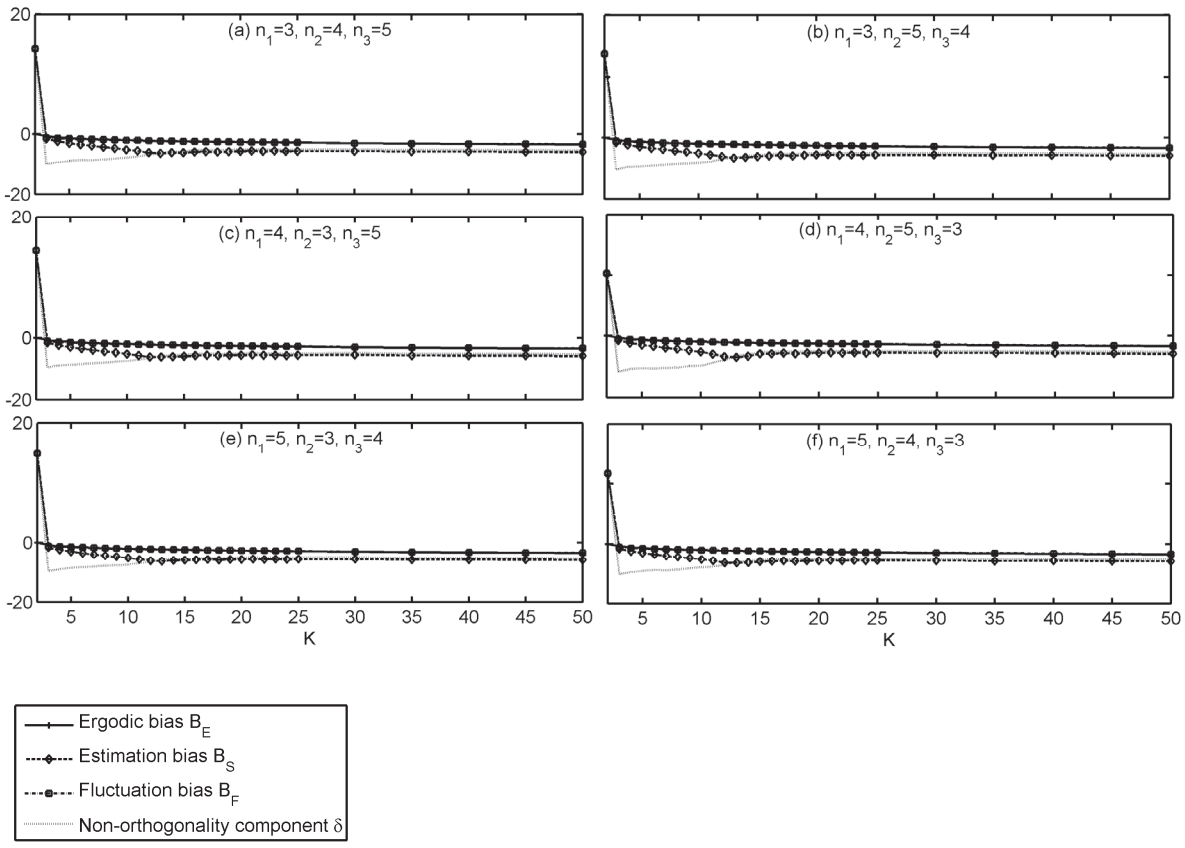


Figure D.2: Standardized values (i.e. divided by $\| \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|$) of the ergodic, estimation, and fluctuation biases and the non-orthogonality component for the ML estimator of $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$, with all six combinations of the values of 3, 4 and 5 for n_1, n_2 and n_3 and sample size $K = K_{min}$, …, 25 by steps of 1 in addition to 30, 35, 40, 45, and 50. The vertical axis is in logarithmic scale, basis 10. There is an atypical 'trough' in the estimation bias and the non-orthogonality component as the sample size K increases.

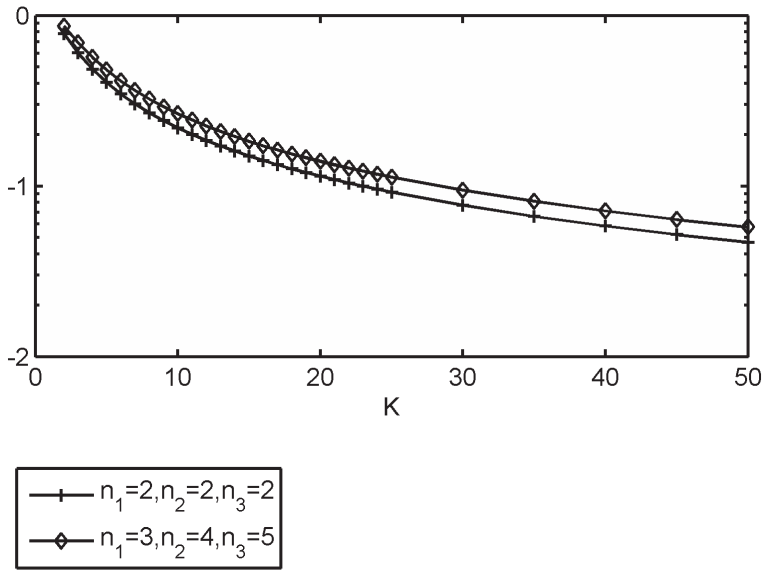


Figure D.3: Standardized pseudo-theoretical bias of the ML estimator of $\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1$, with $n_1 = n_2 = n_3 = 2$ vs. $n_1 = 3$, $n_2 = 4$, $n_3 = 5$ and sample size $K = K_{min}, \ldots, 25$ by steps of 1 in addition to 30, 35, 40, 45, and 50. The vertical axis is in logarithmic scale, basis 10. Unlike the patterns displayed in Figures D.1 and D.2, there is a monotonic decrease of the pseudo-theoretical bias with increasing sample size.

Appendix E. An unbiased likelihood ratio test for double separability of a variance-covariance structure

A biased LRT of double separability with one or two structured component variance-covariance matrices and one or two unstructured component variance-covariance matrices was presented in Roy and Leiva (2008), as well as Roy and Leiva (2011). To our knowledge, an unbiased modified LRT for double separability of a variance-covariance structure with three unstructured component variance-covariance matrices (referred to as the LRT-3D in this thesis) has never been presented.

Assuming an i.i.d. random sample of size $K \geq \max\left(n_1/n_2n_3, n_2/n_1n_3, n_3/n_1n_2\right) + 1$ is available for statistical inference, the ML equations for the component variance-covariance matrices under the tensor normal distribution model of order 3 $\mathbb{X} \sim N_{n_1,n_2,n_3}(\mathbb{M}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ (Chapter 5) are:

$$\begin{cases} \hat{\mathbf{U}}_1 = \sum_{k=1}^{K} (\mathbb{X}_k - \hat{\mathbb{M}})_{(1)} \, (K n_2 n_3 \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} (\mathbb{X}_k - \hat{\mathbb{M}})_{(1)}{}^T; \\ \hat{\mathbf{U}}_2 = \sum_{k=1}^{K} (\mathbb{X}_k - \hat{\mathbb{M}})_{(2)} \, (K n_1 n_3 \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \hat{\mathbb{M}})_{(2)}{}^T; \\ \hat{\mathbf{U}}_3 = \sum_{k=1}^{K} (\mathbb{X}_k - \hat{\mathbb{M}})_{(3)} \, (K n_1 n_2 \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \hat{\mathbb{M}})_{(3)}{}^T, \end{cases} \qquad \text{(E.1)}$$

where $\hat{\mathbb{M}} = \dfrac{1}{n_1 n_2 n_3 K} \left[\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} \sum_{k=1}^{K} \mathbb{X}_{i_1 i_2 i_3, k} \right] \mathbb{J}$ (Mean Model 1: stationary) or $\hat{\mathbb{M}} = \dfrac{1}{K} \sum_{k=1}^{K} \mathbb{X}_k = \bar{\bar{\mathbb{X}}}$ (Mean Model 2: unmodeled mean tensor); as in Chapter 5, the tensors are reshaped into matrices with the matricization operator (Kolda and Bader 2009). As in Chapter 5 also, an iterative algorithm is used to find solutions to the system of equations (E.1).

To test for double separability, the unstructured variance-covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{UN}$, needs to be estimated, as this was the case when testing for simple separability (Chapter 4). The vector normal distribution model is $\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3}(\text{vec}(\mathbb{M}), \boldsymbol{\Sigma}_{UN})$.

Provided $K \geq n_1 n_2 n_3 + 1$, the variance-covariance parameters of the vector normal distribution model can be estimated by ML. The hypotheses under testing are then:

$$H_0 : \boldsymbol{\Sigma}_{UN}^{-1}(\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) = \mathbf{I}_{n_1 n_2 n_3} \text{ against } H : \boldsymbol{\Sigma}_{UN}^{-1}(\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1) \neq \mathbf{I}_{n_1 n_2 n_3}. \quad \text{(E.2)}$$

The unbiased modified LRT of double separability is (by extension of the 2-D case):

$$\Lambda^* = (K - p)[n_1 n_2 \log|\mathbf{U}_3| + n_1 n_3 \log|\mathbf{U}_2| + n_2 n_3 \log|\mathbf{U}_1| - \log|\boldsymbol{\Sigma}_{UN}|]. \quad \text{(E.3)}$$

Asymptotically, the biased unmodified and unbiased modified LRT statistics follow a $\chi^2(f)$ distribution with

$$f = n_1 n_2 n_3 \left(\frac{n_1 n_2 n_3 + 1}{2}\right) - n_1 \left(\frac{n_1 + 1}{2}\right) - n_2 \left(\frac{n_2 + 1}{2}\right) - n_3 \left(\frac{n_3 + 1}{2}\right) + 1. \quad \text{(E.4)}$$

A simulation study with $n_1 = 2$, $n_2 = 2$, $n_3 = 2$ showed that using the biased unmodified LRT based on Λ (i.e., $p = 0$ in equation E.3), with $\chi^2_{0.95}(28) = 41.3371$ as critical value, may result in a rate of rejection of H_0, while it is true, of nearly 100% (Figure E.1). As for the unbiased modified LRT of simple separability, the optimal penalty p value was studied by simulation, and results (Table E.1) are similar. The optimal penalty p value (1) does not appear to be dependent on α (p = 5.6, 5.7, 5.8 for $\alpha = 0.1, 0.05, 0.01$); (2) increases with the product $n_1 n_2 n_3$ and with the number of mean parameters to be estimated; and (3) decreases with the sample size K. Figure E.1 shows that the optimal penalty p controls well the significance level of the unbiased modified LRT of double separability, which is just slightly conservative at values of $K = n_1 n_2 n_3 + 1$ (i.e., close to the minimum required for the ML estimation of $\boldsymbol{\Sigma}_{UN}$ and hence, for testing double separability), as in the unbiased modified LRT of simple separability.

Table E.1: Penalty (p) found to be optimal for modifying the LRT statistic for double separability of a variance-covariance structure, for two mean models (1: stationary; 2: unmodeled mean tensor), two triplets of dimension values (n_1, n_2, n_3), and sample sizes (K) starting from the minimum required for testing ($K = n_1 n_2 n_3 + 1$).

K	Mean Model 1	Mean Model 2
$n_1 = 2, n_2 = 2, n_3 = 2$		
9	4.58	6.00
10	4.40	5.58
11	4.30	5.38
12	4.25	5.28
13	4.23	5.25
15	4.23	5.18
20	4.25	5.15
$n_1 = 2, n_2 = 3, n_3 = 2$		
13	6.55	8.28
14	6.25	7.55
15	6.05	7.25
16	5.90	7.02
17	5.78	6.90
20	5.68	6.68
30	5.45	6.50

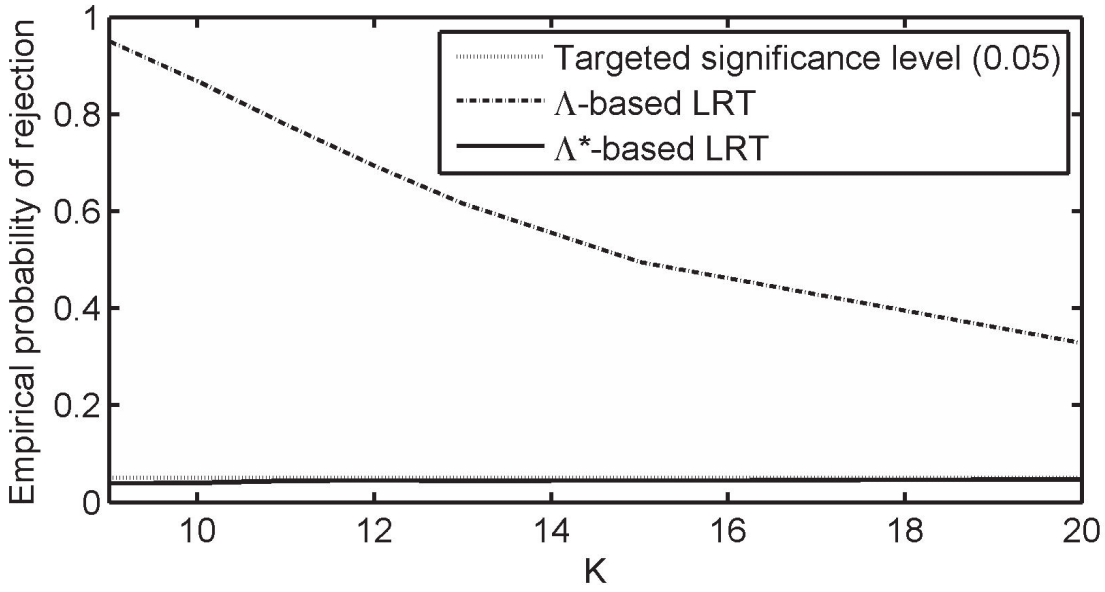


Figure E.1: Empirical probability of rejecting H_0 (double separability of a variance-covariance structure) when true, as a function of sample size K, using the critical value $\chi^2_{0.95}(28) = 41.3371$ with $n_1 = 2$, $n_2 = 2$, $n_3 = 2$ and Mean Model 2 ($\hat{\mathbb{M}} = \bar{\mathbb{X}}$).

Appendix F. The MLE-4D algorithm

Equations (5.1) and (5.8) provide the moment generating function and the probability density function for a tensor normal distribution of any order; the tensor normal distribution needs to be regular for the probability density function to exist. For a tensor normal distribution of order 4, that is, with $J = 4$, it follows that

$$\mathbb{X} \sim N_{n_1,n_2,n_3,n_4}(\mathbb{M}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4) \text{ if and only if}$$

$$\text{vec}(\mathbb{X}) \sim N_{n_1 n_2 n_3 n_4}(\text{vec}(\mathbb{M}), \mathbf{U}_4 \otimes \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1).$$

The four variance-covariance matrices can be estimated iteratively with the following 4-D version of the MLE algorithm:

$$\begin{cases} \hat{\mathbf{U}}_1 = \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)} (K n_2 n_3 n_4 \hat{\mathbf{U}}_4 \otimes \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(1)}^T \\ \hat{\mathbf{U}}_2 = \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)} (K n_1 n_3 n_4 \hat{\mathbf{U}}_4 \otimes \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(2)}^T \\ \hat{\mathbf{U}}_3 = \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)} (K n_1 n_2 n_4 \hat{\mathbf{U}}_4 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(3)}^T \\ \hat{\mathbf{U}}_4 = \sum_{k=1}^{K} (\mathbb{X}_k - \overline{\mathbb{X}})_{(4)} (K n_1 n_2 n_3 \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1)^{-1} (\mathbb{X}_k - \overline{\mathbb{X}})_{(4)}^T \end{cases} \quad \text{(F.1)}$$

As an extension of condition (5.15), the minimum sample size required for the existence of ML estimates for variance-covariance matrices is then

$$K \geq \max\left(\frac{n_1}{n_2 n_3 n_4}, \frac{n_2}{n_1 n_3 n_4}, \frac{n_3}{n_1 n_2 n_4}, \frac{n_4}{n_1 n_2 n_3} \right) + 1 \quad \text{(F.2)}$$

As we did for the tensor normal distribution of order 3, we performed simulations, this time with the mean tensor $\mathbb{M} = (m_{i_1 i_2 i_3 i_4}) = i_1 + i_2 + i_3 + i_4 - 1 + (i_1 i_2 i_3 i_4)$ ($i_1 = 1 \ldots n_1$, $i_2 = 1 \ldots n_2$, $i_3 = 1 \ldots n_3$, $i_4 = 1 \ldots n_4$) and using $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4 = 1.0E-6$ in the convergence criterion and $\hat{\mathbf{U}}_2^0 = \mathbf{I}_{n_2}$, $\hat{\mathbf{U}}_3^0 = \mathbf{I}_{n_3}$, $\hat{\mathbf{U}}_4^0 = \mathbf{I}_{n_4}$ as initial solutions. We considered two scenarios: one simple, in which the variance-covariance matrices were identity matrices: $\mathbf{U}_1 = \mathbf{I}_4$, $\mathbf{U}_2 = \mathbf{I}_2$, $\mathbf{U}_3 = \mathbf{I}_3$, $\mathbf{U}_4 = \mathbf{I}_2$; and one complex, with $\mathbf{U}_1$ and $\mathbf{U}_3$ being respectively $\mathbf{U}_3$ and $\mathbf{U}_2$ from Scenario T4 (Section 5.5), together with

$$\mathbf{U}_2 = \begin{bmatrix} 0.5528 & 0.8676 \\ 0.8676 & 1.8519 \end{bmatrix} \text{ and } \mathbf{U}_4 = \begin{bmatrix} 0.2375 & -0.0176 \\ -0.0176 & 0.3947 \end{bmatrix}.$$

Our results (Table F.1) show that the bias and dispersion properties expected for ML estimators are observed, since the standardized empirical bias and dispersion of the estimated Kronecker product decrease monotonically with increasing sample size and almost vanish at the largest sample size. The number of iterations needed to reach convergence was 21 for $K = 5$ and 6 for $K = 500$. As in the 3-D case, the measure of dispersion for the simple scenario is larger than for the complex scenario, but the value of the bias is very similar in both scenarios.

Table F.1: Standardized empirical bias[a] and standardized empirical measure of dispersion[b], with the standard error below (in parentheses), for the tensor normal distribution of order 4 (see text for $\mathbb{M}$ and $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, $\mathbf{U}_4$).

	Simple scenario		Complex scenario	
Sample Size	Bias	Measure of Dispersion	Bias	Measure of Dispersion
5	0.19003	0.43447 (0.00029)	0.19016	0.33578 (0.00050)
10	0.09789	0.29648 (0.00020)	0.09765	0.21947 (0.00035)
15	0.06582	0.23402 (0.00016)	0.06671	0.17395 (0.00028)
20	0.04960	0.20126 (0.00014)	0.05041	0.14850 (0.00024)
100	0.01010	0.08823 (6.35*E*-5)	0.01011	0.06410 (0.00011)
500	0.00201	0.03936 (2.86*E*-5)	0.00200	0.02853 (4.87*E*-5)

a $\overline{\frac{\| \hat{\mathbf{U}}_4 \otimes \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_4 \otimes \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_4 \otimes \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}}$

b $\overline{\frac{\| \hat{\mathbf{U}}_4 \otimes \hat{\mathbf{U}}_3 \otimes \hat{\mathbf{U}}_2 \otimes \hat{\mathbf{U}}_1 - \mathbf{U}_4 \otimes \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}{\| \mathbf{U}_4 \otimes \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 \|_2}}$

Appendix G. Calibration for ring determination

The objective was to find where in the gradual transition from late to early wood (i.e. from a peak to a trough in Figure 6.3), a ring is about to begin and end. Calibration was performed on wood samples for which both SilviScan (Evans et al. 1995) and CT scanning data were available and the oscillating pattern was clear. SilviScan data allow accurate determination of the beginning and end of rings (Figure G.1), thanks to rotations ensuring that the measurements made are tangent to the ring limits and because of the use of a very thin slice of wood. The smoothing of SilviScan data with two uniform moving averages of order 11 and 51 resulted in curves resembling the data obtained by CT scanning and showing a gradual transition of density from late to early wood in particular. Over the calibration dataset, it was calculated that, on average, the beginning of a ring (i.e. the end of the previous ring) was located at 32% of the inter-peak distance, from the center of the trunk to the bark.

For the dataset used here (i.e. $K = 11$; see Figure 6.3 for an illustration), the positions of maxima in the series of wood densities estimated from CT scanning data were determined and ring limits were putatively located at 32% of the inter-peak distances. More specifically, 32% of the distance between peaks was added in the South direction or removed in the North direction, to the position of the maximum. The ring limits (beginning and end) thus determined were then corrected if necessary, by inspecting the physical sample and making comparisons with the series of wood density estimates; for example, small dips in density caused by false rings and spurious rings were corrected. For the two calendar years selected (1997 and 2003), the oscillating pattern was clearly present in the North and South directions at both heights in the 11 wood samples.

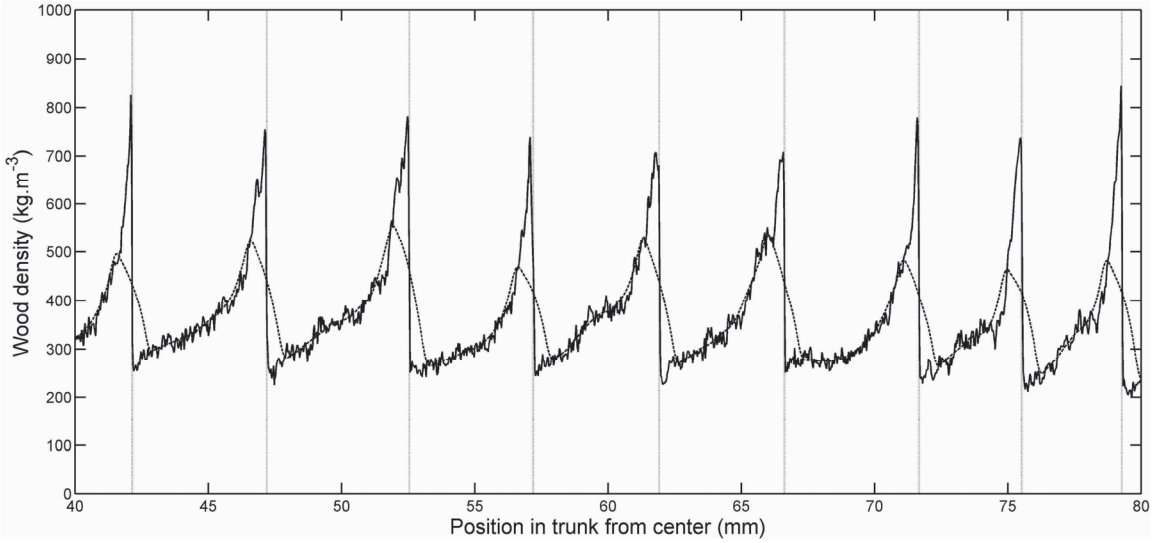


Figure G.1: Example of two series of wood density estimates for the same sample: ragged curve, made of raw measurements obtained with SilviScan, and smoothed curve, resulting from the application of two uniform moving averages of order 11 and 55. The beginning and end of annual growth rings are indicated by a dashed vertical line.