UNIVERSIDAD CATÓLICA DEL NORTE
FACULTAD DE CIENCIAS
Departamento de Matemáticas

Kriging for a Random Field of Positive-Definite Matrices

Tesis Para Optar al Grado de Doctor en Ciencias
Mención Matemáticas

Kjetil Brinchmann Halvorsen

Tutor: Dr. Victor Ayala Bravo

Antofagasta, Chile
2012

UNIVERSIDAD CATÓLICA DEL NORTE
FACULTAD DE CIENCIAS
Departamento de Matemáticas

Kriging for a Random Field of Positive-Definite Matrices[1]

Kjetil Brinchmann Halvorsen

COMISIÓN EXAMINADORA
Dr. Victor Leiva (Universidad de Valparaíso, Chile)
Dr. Xavier Emery (Universidad de Chile, Chile)
Dr. Eduardo Campos (Universidad Católica del Norte, Chile)
Dr. Eduardo Fierro (Universidad Católica del Norte, Chile)
Dr. Victor Ayala Bravo (Prof. Guía) (Universidad Católica del Norte, Chile)
Dra. Elva Ortega Torres (Universidad Católica del Norte, Chile)

Antofagasta, Chile
Diciembre de 2012

[1]Tesis de Doctorado parcialmente financiado mediante Proyecto Fondecyt Nr. 1100375 y Proyecto Mecesup UCN0711 Educación Superior

Tesis Doctoral: Kriging for a Random Field of Positive-Definite Matrices.
Escrita en el Departamento de Matemáticas de la Universidad Católica del Norte, Programa de Doctorado en Ciencias, Mención MAtemática. Parcialmente financiado mediante Proyecto Fondecyt Nr. 1100375 y Proyecto Mecesup UCN0711 Educación Superior.

Contents

List of Figures

Preface

Here is the result of too much time used to develop this ideas! My thanks go to everybody who have helped, and those who have suffered ... . First and foremost I must thank my advisors, Victor Ayala and Eduardo Fierro.

Abstract

The goal of this thesis is to develop methods for spatial interpolation of a tensor field, that is, a spatially correlated field of positive definite matrices. We develop a Wishart model for the field, and prove results needed for estimation of parameters using composite likelihood methods. We use this methodology to compute an interpolation in one example. The first chapter gives a summary of prior work and of some ideas we need, specifically likelihood and composite likelihood. The second chapter develops the Wishart random field model. The third chapter develop some distribution theory for the Wishart distribution that we need for the composite likelihood function, and the fourth chapter discusses estimation and prediction. The final chapter concludes and gives some challenges for further work.

Introduction

In this thesis we propose a solution to a particular case of a problem of spatial interpolation of tensors. In this work, "tensors" refer to symmetric positive-definite matrices, a misuse of terminology which is common in applications. We also use the abbreviation PDM for positive-definite matrix. The dimension of the tensor is always denoted by $m \times m$, where m is a positive integer, usually $m = 2$ or $m = 3$, although the theoretical development, as far as is possible, will be done for the general case.

We became interested in these problem because of an applied problem from Geology, interpolation of the anisotropy of magnetic susceptibility. This anisotropy can be represented as a tensor, that is, a PDM. For background see the paper [**59**], which also contains data which we will use as an example, referred to as the Pluton Example. This data is shown in figure (1). Another Data set we will be using as an example is anisotropy of magnetic susceptibility in the Sierra de Varas Plutón, see [**27**]. The data is shown as tensors in figure (2). For geological interpretation of such data it will be of interest

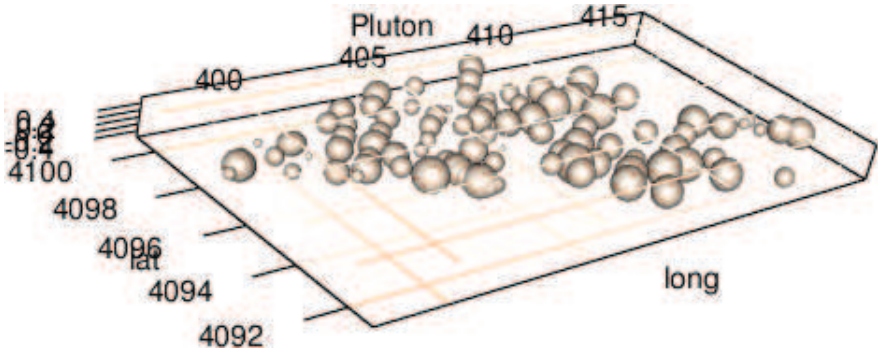

FIGURE 1. A real tensor field. The Pluton example data.

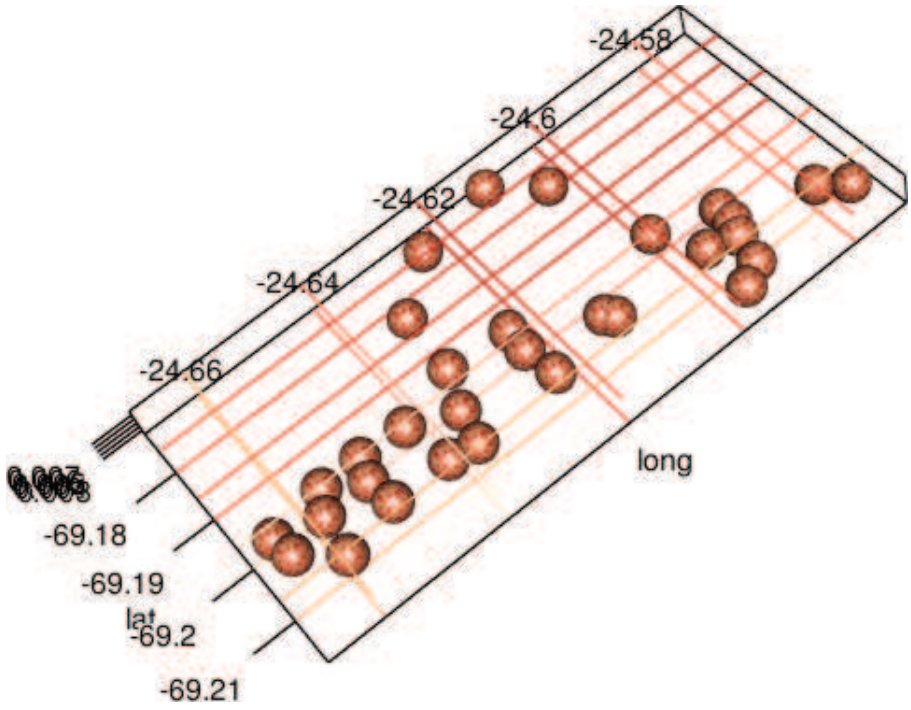

FIGURE 2. A real tensor field. The Sierra de Varas example data.

to present a geological map of the tensor, and that we will construct it using spatial interpolation. The problem of spatial interpolation is often referred to as "kriging", but the theory is mostly developed for a scalar variable. The tensors do not belong to a linear space, they form a non-linear manifold. There is not many publications for kriging applied to variables belonging to a non-linear manifold, a few examples are the papers [8] and [9]. So we need to develop a framework for working with data on non-linear manifolds.

We will present the usual framework for kriging, which we will be applying. Let $\mathcal{D}$ be a region (which we can take as meaning an open connected set) in $\mathbb{R}^d$, where d usually is 2 or 3. Within D is defined a field, that is, a continuous function

$$Y\colon \mathcal{D} \to F \tag{1}$$

where F usually is the field of real numbers, $\mathbb{R}$, but in our case will be the set of tensors, $\mathcal{P}^+(m)$. Let x denote the spatial referent, that is x

is a point within $\mathcal{D}$. Then our data consists of observations of the field Y at N spatial locations within $\mathcal{D}$, $Y_1 = Y(x_1), Y_2 = Y(x_2), \ldots, Y_N = Y(x_N)$. The goal now is to interpolate the field, that is, to construct a function $\hat{Y}\colon \mathcal{D} \to F$ such that $\hat{Y}(x_i) = Y_i, i = 1, \ldots, N$, and which is, elsewhere, a good approximation to the unknown function Y. We will need to specify in what sense it is a good approximation. In statistics, it is usual to give meaning to the interpolation problem by interpreting it as a "prediction" problem. That is, we model the field Y as a realization of a random function, and then define an optimal predictor as a function which minimizes the mean squared error of $\hat{Y}(x) - Y(x)$. For this to give meaning we need to define what we mean by the $-$ -sign (minus-sign) in the former sentence, what is a difference between tensors? One possibility is to use the Riemannian distance introduced in section (6), another possibility is to use the componentwise difference, but in this setting that does not have an intrinsic geometric meaning.

In this thesis, the first chapter is mostly a review of prior work and existing theory that we will need for the later work. Especially, here we give a rather long review of likelihood theory and its extensions. One point in this chapter might be new, or at least very little studied, we propose to base a predictive estimating equation on composite likelihood. In the second chapter we present the Wishart random field model on which the later work is based. In chapter three we calculate the marginal distribution of the two diagonal blocks in a blocked Wishart random matrix. And finally, in chapter four, we use the theory from the earlier chapters to develop methods for prediction of random tensor fields, based on the Wishart model from chapter two. The main contributions from this work are in chapters three and four. In the last fifth chapter we present some ideas for possible further work.

CHAPTER 1

Prior Work and Some Preliminary Ideas

In this chapter, we will lay the groundwork for what to come, that is, spatial interpolation of tensors. We will present concepts and methods we will use later, and illustrate its use in Kriging of scalar spatial fields, but its application to tensors will have to wait for later chapters.

There are not many published works on interpolation of tensor fields, and most seem to be pursuing heuristic approaches without a deep theoretical base. Here we will mention a few works that we are aware of. A book treatment of interpolation and visualization of tensor fields is [**38**]. A simple idea is to interpolate eigenvectors and eigenvalues separately, and construct the tensor field from there. This is pursued in [**32**]. Other works are [**75**] and [**76**].

1. A Conceptual Model for Spatial Interpolation

Conceptually, we have a spatial field of values defined on a domain (an open, connected set) $\mathcal{D} \subset \mathbb{R}^d$, where d usually is 2 or 3. In our applications, $\mathcal{D}$ will represent some region on a geological map. The values belong to the set F, which usually is $\mathbb{R}$ or $\mathbb{R}^m$ for some integer m, but later will be the space of tensors $\mathcal{P}^+(m)$. In this introductory part, we will use the symbol F. The spatial field is a function (usually continuous) $Y\colon \mathcal{D} \to F$, with value $Y(x)$ at the spatial referent $x \in \mathcal{D}$. At N locations within D, written $x_1, \ldots, x_N$, we have observed the value of the field Y, which we write $Y(x_1) = Y_1, \ldots, Y(x_N) = Y_N$. Otherwise the values of the field are unknown to us, but is knowable at some future time, for instance, with further exploration. The goal is to interpolate the field, that is, to construct a function $\hat{Y}\colon \mathcal{D} \to F$, which, in some sense, approximates the unknown function Y. Note that at this point we do not try to explicate in which sense $\hat{Y}$ is an

approximation of Y, since that will depend on our further interpretation of our conceptual model.

In the context of applications in geology, the field F might, for instance, represent a subsurface structure, so will only be known by perforation. It might be, for instance, the distance we have to drill before encountering the structure. As such, for each spatial location x, it is a given value, if unknown to us.

2. The conceptual Model in a Deterministic Framework

One possibility is to think about the function Y as simply given, although not known to us. This is similar to the status of an unknown parameter in statistics. We refer to this as the deterministic interpretation. If we use this interpretation, we understand the interpolation problem as a problem in approximation or interpolation of functions, as studied in numerical analysis. A recent book-length treatment from this point of view is [**74**]. Well-known methods in this framework are polynomial approximation, splines and radial basis functions.

Since we will not be following this lead, we do not have very much more to say about this subject, apart from giving some references. There exist some formal equivalences between spline approaches and the probabilistic approach we will look at in the next section. Two references for this development are [**45**] and [**57**]. For us to use splines to interpolate tensors, we would need to extend the theory of splines for splines having values in curved spaces. A few papers going in this direction are [**51**], [**69**] and [**70**].

There is also the possibility of using much simpler methods, like linear interpolation with weight as a function of distance. We could use the simple linear interpolator

$$\hat{Y}(x_0) = \sum_{i=1}^{N} \lambda_i Y(x_i), \tag{2}$$

with the weight constrained to sum to one. The weight should be a function $\lambda_i = \omega(\|x_i - x_0\|)$. For this to give a continuous surface, we will need that $\omega(h)$ approaches 1 when h goes to zero. Ripley [**62**, Section 4.2] gives a review and critique of such algorithms.

In section 6 we will look into how such ideas can be generalized to tensors.

3. The Conceptual Model in a Probabilistic Framework

We can think about the unknown function Y as a realization of a random function, usually a stationary random function. Note that we do not say that Y *is* a random function, we do not even know what saying such a thing would mean! In the context of application, Y is simply *given*, but representing it as a realization of a random function might be useful for the construction of methods.

Now we can explicate what we mean by approximation: the difference between $Y(x)$ and $\hat{Y}(x)$, which we can write as $\mathrm{dist}(\hat{Y}(x), Y(x))$, must be small in a probabilistic sense, for example, least squares distance. For this to make sense the space of values F must at least be a metric space, for instance, a Riemannian manifold. In the spatial context, this is often called "Kriging". In the following, we shall give a short presentation of Kriging, in the case when $F = \mathbb{R}$.

For the following, we refer to [**12**]. Suppose Y on $\mathcal{D}$ is a stationary (or intrinsic) random field, with constant expectation μ, and variogram function $2\gamma(h) = E(Y(x) - Y(x+h))^2$, the function $\gamma(h)$ itself is called a semivariogram. In the case that γ is a bounded function, the process is second-order stationary, with variance given by $\sigma^2 = \lim_{\|h\|\to\infty} \gamma(h)$[1]. If this limit does not exist, the process is called intrinsic, in which case the expectation of Y usually does not exist. The kriging predictor is given by a linear combination

$$\hat{Y}(x_0) = \sum_{i=1}^{N} \alpha_{i,0} Y_i + \beta_0, \tag{3}$$

[1]Note that this limit will never depend on in which direction h grows to infinity.

where the weights $\alpha_{1,0}, \ldots, \alpha_{N,0}, \beta_0$ are to be determined such that the mean squared error of prediction

$$\mathrm{E}(\hat{Y}(x_0) - Y(x_0))^2 \tag{4}$$

is minimized.

The solution is given by

$$\hat{Y}(x_0) = c^\top \Sigma^{-1}(Y - \mu 1) + \mu, \tag{5}$$

where $1 = (1, 1, \ldots, 1)^\top$ is a vector of N ones, μ is the constant expectation, $c = (C(x_0, x_1), \ldots, C(x_0, x_N))^\top$ is the vector of covariances between the prediction location and the data locations, and Σ is the $N \times N$ matrix with (i, j)-element given by $C(x_i, x_j)$, and C is the spatial covariance function, given by $C(x, x + h) = \sigma^2 - \gamma(h)$. Y is the vector of observations $(Y_1, \ldots, Y_N)^\top$.

Matheron, see [**12**] called this predictor for *simple kriging*, and it is based on an assumption of knowledge of the mean μ of the process. If the mean is unknown, which in most cases is a more palatable assumption, we could replace μ in the above expression with an estimate of it, $\hat{\mu}$. But mostly it is preferred to replace simple kriging with *ordinary kriging*, which we give below, in such cases.

It is of interest for us here, to contemplate simple kriging, and "estimated simple kriging", where μ is replaced by $\hat{\mu}$, since in more complex cases, as the case of tensors in a later chapter, that is maybe all we are able to do! This is somehow equivalent to ordinary kriging, provided that $\hat{\mu}$ is the kriging estimate of μ. This is called the "additivity relationship", see [**11**].

Now we will describe the changes in assumptions necessary to get to what is called *ordinary kriging*. Since we now do not assume that μ is known, we need a predictor which does not depend on μ. To obtain that we introduce the assumption that the predictor must be unbiased:

$$\mathrm{E}(\hat{Y}(x_0)) = \mathrm{E}(Y(x_0)), \tag{6}$$

for all values of μ. This leads to the requirement that $\mu = \sum_{i=1}^{N} \alpha_{i,0}\mu + \beta_0$, for all values of μ. This implies that $\beta_0 = 0$ and that $\sum_{i=1}^{N} \alpha_{i,0} = 1$. We can show the optimal coefficient vector is given by

$$\alpha_0^\top = (\gamma + \frac{1 - 1^\top \Gamma^{-1}\gamma}{1^\top \Gamma^{-1} 1})^\top \Gamma^{-1} \tag{7}$$

where $\gamma = (\gamma(x_0 - x_1), \ldots, \gamma(x_0 - x_N))^\top$ and Γ is an $N \times N$-matrix whose (i,j)th element is $\gamma(x_i - x_j)$.

Note that the meaning of the term "unbiased" here is different from that which is usual in parametric statistical estimation theory. Let $f(y;\theta)$ be a statistical model for an unobservable y, depending on an unknown parameter $\theta \in \Theta$, for some parameter space Θ. An estimator $\hat{\theta}$ of θ is *unbiased* if $E_\theta\, \hat{\theta} = \theta$ for all possible values of $\theta \in \Theta$. Comparing with the situation with ordinary kriging, the unknown quantity that we want to estimate or predict is $Y(x_0)$. But $Y(x_0)$ is a random variable, which is different from the situation in estimation theory where θ is a fixed, non-random, but unknown, quantity. Trying to use directly the estimation-theory definition of *unbiased*, we get

$$E(\hat{Y}(x_0)|Y(x_0)) = Y(x_0), \tag{8}$$

which is a very different criterion, and if used would lead to predictors with a larger variance. Plainly, in the parametric theory we require that the expectation of the estimator $\hat{\theta}$ be equal to θ for all values of θ, but in the prediction theory we do not require that the expectation of $\hat{Z}$ be equal to Z for all possible values of Z, that is, conditional on Z, but only that it be equal to the expectation of Z (for all possible values of that unknown expectation).

In reality, with ordinary kriging we will usually have that $E\,|\hat{Y}(x_0) - \mu| < |Y(x_0) - \mu|$, where here μ is the unknown true value of the constant expectation of the random field. Here we take the expectation of $\hat{Y}(x_0)$ conditional on the true value $Y(x_0)$. This shows that the predictor $\hat{Y}(x_0)$ is not unbiased in the sense of estimation theory, a

definition which really does not make sense for prediction of a random variable. For a general discussion of this phenomenon, see [**66**].

4. Likelihood, Predictive Likelihood and Extensions

In this section we will give a brief summary of the usual likelihood theory for parametric estimation, of the generalization of this theory to the theory of unbiased estimating equations, and the lesser known parallels of this theory for prediction of random variables.

4.1. The Ordinary Full Likelihood Function. Let $f(y;\theta)$ for $\theta \in \Theta \subset \mathbb{R}^p$ be a statistical model for a random observable $y \in \mathcal{Y}$, where θ is an unknown parameter belonging to some open subset Θ in $\mathbb{R}^p$. As our applications will be for a regular model, we suppose sufficient regularity for the theory to work as given, this includes that the support of the model function, that is, the closure of the set $\{y \in \mathcal{Y}\colon f(y;\theta) > 0\}$, do not depend on θ. Other regularity conditions is that the model function $f(y;\theta)$ can be differentiated with respect to θ a sufficient number of times, and that differentiation with respect to θ can be interchanged with integration with respect to y. The model function will be either a probability density function or a probability mass function, and in the density case we suppose the density is with respect to a common dominating measure. Here $\mathcal{Y}$ is the sample space for the observable y. A presentation with details of regularity conditions is [**41**].

This setting includes the case of an iid (independent and identically distributed) sample, in which case the model function is $\prod_{i=1}^{N} f(y_i;\theta)$. But it also includes cases where the observable y is a random vector with non-independent components.

The likelihood function is simply $L(\theta) = L_y(\theta) = f(y;\theta)$ considered as a function of θ, where the dependence on y is usually suppressed. Here y denotes an actually observed value of the observable y. Note that the equality in the definition is taken up to any

multiplicative constant that does not depend on the parameter vector θ, since any two functions which differ by a (positive) constant not depending on θ give the same values to all relative likelihoods, that is, all comparisons between two different values of θ. The likelihood principle says that, in the context of the given model function, all information in the sample y about the unknown parameter θ, is contained in the likelihood function, thus explaining the importance of the likelihood function in both theory and applications. For an extended discussion of the likelihood principle, see [**3**]. For a recent critique of the likelihood principle, see [**46**].

We will more often use the log-likelihood function $l_y(\theta) = l(\theta) = \log(f(y;\theta))$ where again, mostly we will let the dependence on y be implicit. The efficient score function is given by

$$U(\theta) = \frac{\partial}{\partial\theta} l(\theta) = \left(\frac{\partial l(\theta)}{\partial\theta_1}, \ldots, \frac{\partial l(\theta)}{\partial\theta_p}\right)^\top \tag{9}$$

where we remind the reader that the parameter space Θ is an open subset of $\mathbb{R}^p$. In the regular case, the maximum likelihood estimator of θ,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta) \tag{10}$$

can be found by solving the score equation $U(\theta) = 0$.

Note that the expectation of the score is zero: Calculate

$$\mathrm{E}_\theta\, U(\theta) = \int U(\theta) f(y;\theta)\, dy = \int \frac{\partial}{\partial\theta} l(\theta) f(y;\theta)\, dy \tag{11}$$

$$= \int \frac{\frac{\partial f(y;\theta)}{\partial\theta}}{f(y;\theta)} \cdot f(y;\theta)\, dy = \frac{\partial}{\partial\theta}\int f(y;\theta)\, dy = 0. \tag{12}$$

The covariance-matrix of the score function is called the (expected) Fisher information matrix. By manipulations similar to what used above, we can show that

$$i(\theta) = \mathrm{Cov}_\theta(U(\theta), U(\theta)) = \int \frac{\partial l(\theta)}{\partial\theta} \cdot \frac{\partial l(\theta)}{\partial\theta^\top} f(y;\theta)\, dy \tag{13}$$

also can be given by the expression

$$i(\theta) = E_\theta(-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}) \tag{14}$$

and the expression we take the expectation of above is called the observed information and denoted $j(\theta)$. To show this, first show that

$$\frac{\partial^2}{\partial\theta\partial\theta^\top}\log f(y;\theta) = \frac{\frac{\partial^2}{\partial\theta\partial\theta^\top}f(y;\theta)}{f(y;\theta)} - \frac{\frac{\partial}{\partial\theta}f(y;\theta)}{f(y;\theta)}\cdot(\frac{\frac{\partial}{\partial\theta}f(y;\theta)}{f(y;\theta)})^\top \tag{15}$$

and then take expectation of both sides.

A heuristic justification for this use of the concept "information" is that, if the likelihood function really contains much information about θ, then it should be peaked around its maximum, $\hat{\theta}$, and the observed information is then simply the negative of the Hessian matrix, and if this is large then the likelihood function is indeed very peaked, and θ is well defined. So the size of the observed information matrix is a measure of the peakedness of the loglikelihood function. Some authors ([**7**], [**19**]) call the inverse of the information matrix the "formation" matrix.

The Fisher information matrix gives information about with how large precision it is possible to estimate θ. The following result is called the Cramer-Rao inequality, or simply the information inequality. First, we give a simple result which can be seen as a multivariate Cauchy-Schwarz inequality.

LEMMA 1. *Let y and z be two square-integrable random vectors, not necessarily with the same number of components. Then we have that*

$$\mathrm{Var}(z) \geq \mathrm{Cov}(z,y)\,\mathrm{Var}(y)^{-1}\,\mathrm{Cov}(y,z) \tag{16}$$

where $\geq$ signifies the usual order in the cone of positive-definite matrices.

The proof of this result is simple: Note that the following matrix product is a positive-semi-definite matrix:

(17)

$$[I, -\operatorname{Cov}(z,y)\operatorname{Var}(y)^{-1}] \begin{pmatrix} \operatorname{Var}(z) & \operatorname{Cov}(z,y) \\ \operatorname{Cov}(y,z) & \operatorname{Var}(y) \end{pmatrix} \begin{pmatrix} I \\ -\operatorname{Var}(y)^{-1}\operatorname{Cov}(y,z) \end{pmatrix}$$

and multiply out. The information inequality is the following. Let g be a possibly vector-valued function of the parameter vector θ. Some possibilities are $g(\theta) = \theta$ and $g(\theta) = \theta_1$. Let T be a function of the data y with expectation $g(\theta)$, that is, an unbiased estimator of $g(\theta)$. Then we have, under regularity conditions as before, that

$$\operatorname{Var}_\theta(T) \geq g'(\theta) i(\theta)^{-1} (g'(\theta))^\top \tag{18}$$

where $g'(\theta)$ is the gradient of g, that is, the matrix with (i,j)-element given by $\frac{\partial}{\partial \theta_j} \operatorname{E}_\theta T_i$. As an example, for unbiased estimation of θ, that is $\operatorname{E}_\theta T(y) = \theta$, we find that $g'(\theta)$ is the identity matrix, so that $\operatorname{Var}_\theta T \geq i(\theta)^{-1}$. To show this result, use the above lemma to write

$$\operatorname{Var}_\theta(T) \geq \operatorname{Cov}_\theta(T, \frac{\partial \log f}{\partial \theta}) i(\theta)^{-1} \operatorname{Cov}(\frac{\partial \log f}{\partial \theta}, T) \tag{19}$$

and write the right-hand side as

$$\int (T - g(\theta)) \frac{\partial \log f}{\partial \theta^\top} f \, dy \cdot i(\theta)^{-1} \cdot \int \frac{\partial \log f}{\partial \theta} (T - g(\theta))^\top f \, dy \tag{20}$$

and, with manipulations similar to the ones used earlier, with interchange between differentiation and integration, this reduces to the given result (18).

4.2. Unbiased Estimating Equations. But for our application, we will not be able to calculate the full likelihood function, so we must search for some other approach. Godambe [**26**] had the idea to generalize the score equation to what he called an "unbiased estimating equation", that is, a function H of data y and parameter vector θ, $H(y,\theta)$ such that the unbiasedness property of the score function generalizes, that is,

$$\operatorname{E}_\theta H(y,\theta) = 0. \tag{21}$$

Note that this is connected to, but weaker than, the concept of a pivote: A pivote is a function of data and parameter with a distribution

which does not depend on the parameter. To fix ideas, let us give an elementary example of a pivote. If Y is a random variable with a normal distribution with mean θ and variance 1, then $H = Y - \theta$ is a pivote: H has a normal distribution with mean zero and variance 1, thus a completely known distribution. We can see that H is both a pivote and an estimating function, since its mean is free of the unknown parameter. If we change this example a little: let Y be normally distributed with mean θ and variance σ^2, then H defined in the same way is not a pivote, since its distribution is not completely known; it depends on the unknown variance σ^2. But H is still an estimating function for θ. We call H an estimating function. We could say that an estimating function is a pivote in expectation. Exactly as with the score function, we can obtain an estimator by solving the equation $H(y, \theta) = 0$, which is called an estimating equation.

Note that the very traditional estimation method, the *method of moments*, is really an example of unbiased estimating equations. If T is a function of the observed data Y with expectation θ, then the moment estimator is to solve $T(y) = \mathrm{E}\, T(y) = \theta$, and clearly this is the same as solving the estimating equation $H(y, \theta) = T(y) - \theta = 0$. Traditional least squares estimation can also be seen as unbiased estimating equations, so we see that this concept is an important unifying idea in estimation theory.

Now much of what we have said of the score function can be repeated. Clearly, the score function is a special case of an estimating function, in a sense, it is the ideal estimating function. We can define the information in the estimating function, sometimes called the Godambe information, by

$$\mathfrak{i}_H(\theta) = \mathrm{E}_\theta \frac{\partial H(\theta)}{\partial \theta} \operatorname{Var} H(\theta)^{-1} \mathrm{E}_\theta \frac{\partial H(\theta)}{\partial \theta^\top} \tag{22}$$

the idea being that, the variance of the estimator obtained by solving $H(\theta) = 0$ can often be approximated by the Godambe information. This can be shown by a similar argument as for the maximum likelihood estimator. Introduce the notation $V(\theta) = \mathrm{E}_\theta - \frac{\partial H(\theta)}{\partial \theta}$, which is called the sensitivity or Hessian matrix, and the variability matrix,

$J(\theta) = \mathrm{Var}_\theta H(\theta)$. The sensitivity matrix can be seen as the observed information based on the composite likelihood function. With this notation the Godambe information matrix is $i_H(\theta) = V(\theta)J(\theta)^{-1}V(\theta)^\top$.

Let us give an alternative form for $i_H(\theta)$. By differentiating the equation (21) we can show that

$$\mathrm{E}_\theta(\frac{\partial H}{\partial \theta} + UH^\top) = 0 \tag{23}$$

where again U is the efficient score. Using this we find that

$$i_H(\theta) = \mathrm{E}(UH^\top)\,\mathrm{Var}(H)^{-1}\,\mathrm{E}(HU^\top). \tag{24}$$

In particular, for the case where θ is a scalar, we find

$$i_H(\theta) = \mathrm{corr}(U, H)^2 i(\theta), \tag{25}$$

since $\mathrm{E}\left(UH^\top\right) = \mathrm{Cov}(U, H)$ and $\frac{\mathrm{Cov}(U,H)^2}{\mathrm{Var}\,H\,\mathrm{Var}\,U} = \mathrm{corr}(U, H)^2$. This result is given just after equation (2.3) in [**43**]. So full information is associated with a linear relationship between U and H. It is in this sense that the efficient score is an ideal estimating function.

We have an extension of the information inequality to this setting, which is given in [**43**]. It is

$$i(\theta) \geq i_H(\theta) \tag{26}$$

where $i(\theta) = i_U(\theta)$ is the expected Fisher information in the efficient score function. This can be proven using the multivariate Cauchy-Schwarz inequality, which gives

$$\mathrm{Var}\,U \geq \mathrm{E}(UH^\top)\,\mathrm{Var}(H)^{-1}\,\mathrm{E}(HU^\top) \tag{27}$$

where again $\geq$ is the usual order in the positive-definite cone. Here we use that both U and H are unbiased estimating functions.

4.3. Composite Likelihood. One method of obtaining an estimating equation is by using a composite likelihood function, as defined by [**43**]. To define this, we start with a set of marginal or

conditional events for which we can write log-likelihood functions $l_i(\theta)$, and then we sum them to obtain a composite log-likelihood

$$cl(\theta) = \sum_i l_i(\theta). \tag{28}$$

By differentiation we can define the composite score function,

$$CU(\theta) = \frac{\partial}{\partial\theta} Cl(\theta) \tag{29}$$

and since each component in the component likelihood function is a true likelihood, the expectation of the composite score function is zero. Therefore, $CU(\theta) = 0$ is an unbiased estimating equation and the composite score is an unbiased estimating function. This composite Likelihood function is a special case of the use of alternative or approximate likelihood, for which some references will be given later. They have been studied from many points of view, particularly from the view of a *misspecified model*. From this point of view, a composite likelihood is a misspecified likelihood. A book-length treatment from this point of view is [**22**].

In this thesis we will only use marginal events for the definition of a composite likelihood, in which case it is called a marginal composite likelihood, see [**71**]. An often used case is to use only the bivariate distributions of all pairs of random variables in the data. In this case it is called a pair (composite) likelihood. If the data vector is $y = (y_1, \ldots, y_N)^\top$, then the pair log likelihood is $cl(\theta) = \sum_i \sum_{j<i} \log f(y_i, y_j; \theta)$. Now let us ask what happens if θ has a value corresponding to independence between y_i and y_j (for $i \neq j$). Then the above becomes $cl(\theta) = (N-1)l(\theta)$, where $l(\theta)$ is the full log likelihood function. Observe that this is the log likelihood function we would observe if we had $N-1$ times more observations (or, more generally, quantity of information) than we actually have. So if we treat the composite log likelihood as a true likelihood, it will misrepresent the quantity of information we have, and lead to falsely precise inferences. This explains why the information in the composite likelihood cannot be measured by the Fisher information, or by the observed information, but must be calculated via the more involved Godambe information. It also leads to the idea to try to correct the

composite log likelihood, so as to behave more like an ordinary full likelihood. We will treat this only in the case of a pair likelihood, which is the case we will be using. Then define the corrected pair likelihood as

$$cl_c(\theta) = \sum_i \sum_{j<i} w_{ij} \log f(y_i, y_j; \theta) \tag{30}$$

where w_{ij} are some positive weights to be determined. These weights would indicate the weight to be given to each pair, in analogy with weighted least squares. For instance, in a spatial setting, they could depend somehow on the spatial distance between the two observations, maybe giving less weight to very distant pairs.

The maximizer of this corrected pair log likelihood will be called the maximum corrected composite likelihood estimator, denoted by $\hat{\theta}_c$. We will return later to how we can choose the weights. The question of consistency of the estimator maximizing the pair-likelihood is discussed in the paper [**61**], they note that in the case of the composite likelihood there is still no known complete characterization of the conditions for consistency. They say that the complete problem is too general, and that useful conditions probably only can be found for special cases, such as pair-likelihood. [**61**] also defines a more general version of the pair-likelihood as

$$\sum_{s<t} \log f(y_s, y_t; \theta) - a \sum_s \log f(y_s; \theta) \tag{31}$$

and asks the question about how to choose a in some optimal way.

Now we will try to justify the definitions of the Fisher and Godambe information matrices by giving some approximations for the maximization estimators. Again we assume sufficient regularity. Now let θ_0 indicate the "true" value of the parameter, that is, that parameter which actually governed the generation of the data. If θ is any other value, we get

$$E_{\theta_0} \log \frac{f(Y;\theta)}{f(Y;\theta_0)} \leq \log E_{\theta_0} \frac{f(Y;\theta)}{f(Y;\theta_0)} = 0, \tag{32}$$

by using Jensen's inequality. This shows that

$$\mathrm{E}_{\theta_0} \log f(Y;\theta) \leq \mathrm{E}_{\theta_0} \log f(Y;\theta_0) \tag{33}$$

that is, the expectation under the generating model of the log likelihood is maximized for the true parameter. This inequality is sometimes called the Kullback Leibler information inequality. It can be used to show that the maximization estimator is consistent, that is, when sample size (or some other measure of total information in the data) increases without bound, the estimator converges in probability to the true parameter. Note that since the composite log likelihood is a sum of true log likelihoods, that inequality is true for each of the terms in the sum and so also for the composite log likelihood.

Now let $\hat{\theta}$ be the maximum likelihood estimator, and let us use a Taylor expansion to approximate the log likelihood around θ_0. We assume the maximum is found by solving the score equation. Then we find that

$$-l'(\theta_0) = l'(\hat{\theta}) - l'(\theta_0) = l''(\theta^*)(\hat{\theta} - \theta_0) \tag{34}$$

where θ^* is some parameter value between θ_0 and $\hat{\theta}$. Here l' is the gradient (with respect to the parameter vector) and l'' is the Hessian matrix. Assuming the Hessian matrix is invertible, we find that

$$\hat{\theta} - \theta_0 = -l''(\theta^*)^{-1} l'(\theta_0) = j(\theta^*)^{-1} l'(\theta_0) \tag{35}$$

Where as before, j indicates the observed information matrix. When the quantity of information is large, the observed information will be close to the expected Fisher information matrix i, so the above becomes

$$\hat{\theta} - \theta_0 \approx i(\theta_0)^{-1} l'(\theta_0). \tag{36}$$

leading to $\mathrm{E}_{\theta_0}(\hat{\theta} - \theta_0) = \mathrm{E}_{\theta_0}(i(\theta_0)^{-1} l'(\theta_0) = 0$ and $\mathrm{Var}(\hat{\theta} - \theta_0) = \mathrm{Var}_{\theta_0}(i(\theta_0)^{-1} l'(\theta_0)) = i(\theta_0)^{-1} i(\theta_0) i(\theta_0)^{-1} = i(\theta_0)^{-1}$, explaining the name "information matrix". Now we will do the parallel approximations for the estimating equations case. Note that this includes composite likelihood.

As before, let $H(\theta)$ be an unbiased estimating function based on the data y. It plays the same role as $l'(\theta)$ in the former paragraph. The estimator solving $H(\theta) = 0$ will be denoted $\hat{\theta}$. We find

$$-H(\theta) = H(\hat{\theta}) - H(\theta_0) = H'(\theta^*)(\hat{\theta} - \theta_0) \tag{37}$$

which closely parallels (34), but note that the matrix on the right now is a Jacobian matrix, not a Hessian, so need not be symmetric. We can solve:

$$\hat{\theta} - \theta_0 = (-H'(\theta^*))^{-1} H(\theta_0). \tag{38}$$

Again, in the situation with much information, $\hat{\theta}$ will be close to θ_0, so $H'(\theta^*)$ will be close to $H(\theta_0)$, so we have the approximation

$$\hat{\theta} - \theta_0 \approx (-H'(\theta_0))^{-1} H(\theta_0) \tag{39}$$

which gives the approximations $E(\hat{\theta} - \theta_0) \approx 0$ and $\mathrm{Var}(\hat{\theta} - \theta_0) \approx i_H(\theta_0)^{-1}$, justifying the definition of the Godambe information. This formula for the approximation of the variance is also called the "sandwich formula".

4.4. Generalizations to obtain Predictive Likelihoods. Now, let us ask: How can this ideas around likelihood, unbiased estimating equations, and composite likelihood be used, not for estimating parameters, but for prediction of (future) random variables?

We will first review ideas for a definition of *predictive likelihood* in the contect of an ordinary, full likelihood function, and after that look into possibilities for generalizing that idea to unbiased estimating functions, and, in particular, component likelihood. In this generalized context predictive likelihood seem to be very little explored. In the first part we will mostly be following the paper [**5**] and also [**2**]. It is important to understand that in the usual case when the unknown quantities in a statistical problem are unknown but fixed *parameter*, the definition of the likelihood function is uncontroversial. But in a prediction problem we conceptualize the unknown quantity that we want to predict z, as a presently unobserved, but in the future possibly observable *random variable*, not as a fixed, but unknown *constant*. In this case it is not at all clear how a likelihood function

should be defined, if it is at all possible to define a likelihood function in this case! Bjørnstad [5], give 14 possible definitions! and it might possibly be more. Traditionally, prediction problems have been solved by other methods. A detailed discussion of the problems involved with extending the definition of a likelihood function to cover the case of unknown random variables is given in the paper [6].

We start with a review of prediction within a full Bayesian model, that is, a full probability model, where not only stochastic variables, but also fixed parameters, are given probability distributions. We use f as a generic symbol for a density function (or probability mass function) and indicate the random variables involved with the argument. We write y for the observed data, z for the unobserved future observable we want to predict, and θ for the parameter governing the distribution of y and z. The predictive distribution of z given y is

$$f(z|y) = \frac{f(y,z)}{f(y)} = \frac{\int f(y,z|\theta)f(\theta)\,d\theta}{\int\int f(y,z|\theta)f(\theta)\,d\theta\,dz} \tag{40}$$

Of course this depends on the choice of a suitable prior distribution $f(\theta)$. What can we do if we are not willing or able to use this Bayesian machinery? We can at least keep the idea that the *ideal* predictive distribution for z would be in the form of a probabilistic distribution for z. So this structures the prediction problem in two steps

(1) find a predictive distribution for z, $f(z|y)$.
(2) we use the predictive distribution to construct a prediction function $\hat{z} = \hat{z}(y)$.

But we emphasize that the first item here is an ideal, without the Bayesian method we will not necessarily be able to achieve that. The basis for constructing a predictive likelihood will always be the model function of (y,z), that is, $f(y,z|\theta) = f_\theta(y,z)$. Since the interest is not in the unknown parameter θ, but in the unknown random variable z, we will threat θ as a "nuisance parameter"[2], that is, we want to remove it, and by removing it, in some way, we obtain $f(z|y)$. In this way the concept of a predictive likelihood is rather vague. How can

[2]If this language sounds strange, note that in statistics , a nuisance parameter, or simply a nuisance, simply is a parameter without inferential interest.

we use the predictive likelihood to obtain a point prediction for z? A simple idea is to obtain the maximum likelihood predictor:

$$\hat{z}_{ml}(y) = \arg\max_z f(z|y), \tag{41}$$

but the usually good properties of the maximum likelihood estimator in parametric problems does not necessarily transfer to the maximum likelihood predictor. Bjørnstad [5] gives one example (his example 7) where the MLP always is zero, irrespective of the data y. But, that example uses the exponential distribution. As the mode of the exponential distribution is always zero, irrespective of the parameter, the result is maybe not that surprising. But it reminds us that the MLP always must be investigated, in each particular case, to see if it is reasonable.

An other idea, which can be used if the predictive likelihood is a probability distribution in z, is to use its mean: $\hat{z}_m(y)(= \int zf(z|y)\,dz$. This predictor would work well in the exponential distribution example above.

The simplest idea for removing the "nuisance" θ, is the profile predictive likelihood. Define

$$\hat{\theta}_z(y) = \arg\max_\theta f(y, z|\theta), \tag{42}$$

that is, the maximum likelihood estimator using (y, z) as data. Then the profile predictive likelihood is

$$f_{pr}(z|y) = f(z|y; \theta = \hat{\theta}_z(y)). \tag{43}$$

which in many examples is remarkably good. Leonard (1982) [**42**] shows close connections with the Bayesian posterior predictive density with flat prior. The profile predictive likelihood and its extensions seem to be the most practical in complicated situations. In complicated situations other, maybe theoretically superior methods, based on sufficiency and conditioning, are often not available. Remember, that in a parametric statistical problem, a sufficient statistic for θ, is a function $t(y)$ of data y such that the conditional distribution of y given $t(y)$ does not depend on the parameter θ. In this

sense, $t(y)$ contains all the information in the data about θ, of course, assuming the model is true. To interprete this, write

$$f(y;\theta) = f(y|t(y);\theta) \cdot f(t(y);\theta) = f(y|t(y)) \cdot f(t(y);\theta). \tag{44}$$

In the factorization above, we can see that the likelihood function really only depends on y via the function $t(y)$, so that a likelihood function based on this model will depend on the data only via the function $t(y)$. In this sense all the information in the data is contained in the *sufficient* statistic $t(y)$.

In the following we will be following Davison [**13**], which gives approximate predictive likelihoods based on the Laplace approximation. So first we will review the Laplace approximation to a multivariate integral.

Assume we want to approximate an integral of the form

$$\int e^{-ng(t)}\, dt. \tag{45}$$

Heuristically, the largest contribution should come from a region around the mode of the integrand, corresponding to a minimum of $g(t)$ (assuming n is positive). Denote this mode by t^*. Assuming we can Taylor-expand, write

$$\int e^{-ng(t)}\, dt = \int e^{-n(g(t^*)+g'(t^*)(t-t^*)+\frac{1}{2}g''(t^*)(t-t^*)^2+\cdots)}\, dt \tag{46}$$

$$\approx e^{-ng(t^*)} \int e^{-\frac{1}{2}g''(t^*)(t-t^*)^2}\, dt \tag{47}$$

Observe the integrand is a multiple of a multinormal density with covariance matrix given by $\frac{1}{n}g''(t^*)^{-1}$, which we can use to evaluate the integral and finally reach the approximation

$$\frac{e^{-ng(t^*)}(2\pi)^{p/2}}{n^{p/2}\det(g''(t^*))^{1/2}}, \tag{48}$$

where p is the dimension of the integral. If we are approximating a Bayesian posterior predictive distribution, we have a ratio of two integrals, so we want to approximate, say,

$$E_{p^*}(h(\theta)) = \frac{\int h(t)p^*(t)\, dt}{\int p^*(t)\, dt}, \tag{49}$$

we simply apply the Laplace approximation to both numerator and denominator separately, and then take the ratio:

$$\mathrm{E}_{p^*}(h(\theta)) = \frac{\det(H^{**})^{-1/2}\exp(\log(p^{**}(t^{**})))}{\det(H^*)^{-1/2}\exp(\log(p^*(t^*)))} \tag{50}$$

where $p^{**} = h(t)p^*(t)$, t^* is a mode of p^*, t^{**} is a mode of p^{**} and

$$H^* = -\nabla^2 \log p^*(t) \text{ at } t = t^* \tag{51}$$

$$H^{**} = -\nabla^2 \log p^{**}(t) \text{ at } t = t^{**} \tag{52}$$

Now we want to approximate integrals of the form

$$f_Y(y) = \int f_Y(y|\theta)h(\theta)\, d\theta \tag{53}$$

Write $l(\theta) = l(Y|\theta) = \log f_Y(y|\theta)$ for the log-likelihood. Write also $\rho(\theta) = \log h(\theta)$, and we assume l and ρ are twice continuously differentiable with respect to θ.

Laplace' method now gives

$$f_Y(y) = (2\pi)^{p/2}\exp\{l(y|\theta^*)\}h(\theta^*)\det(I(\theta^*))^{-1/2} \tag{54}$$

where $I(\theta^*)$ is minus the Hessian of $l(\theta)+\rho(\theta)$ evaluated at θ^* . If the prior $h(\theta)$ is flat then $I(\theta^*)$ is the observed information matrix. θ^* is the maximizer of $l(\theta) + \rho(\theta)$. This approximation does not apply to cases where the maximum likelihood estimator is not given as a solution of the score equation.

A similar expansion of the joint density of (Y, Z) leads to

$$f_{Y,Z}(y,z) = (2\pi)^{p/2}\exp\{l(y,z|\theta^*(z)\}h(\theta^*(z))\det(J(\theta^*(z)))^{-1/2} \tag{55}$$

where $J(\theta)$ is the observed information based on $f_{y,Z}(y,z)$. Substituting this approximations into 40 gives an approximate posterior predictive density

$$f_{Z|Y}(z|y) \approx \frac{f_{Y,Z}(y,z|\theta^*(z))h(\theta^*(z))\det(I(\theta^*))^{1/2}}{f_Y(y|\theta^*)h(\theta^*)\det(J(\theta^*(z)))^{1/2}} \tag{56}$$

If h is constant, we call the right-hand side of 56, regarded as a function of z, the approximate predictive likelihood of $Z = z$ given

$Y = y$. Leaving out the constant h factor, the above equation becomes

$$f_{Z|Y}(z|y) \approx \frac{f_{Y,Z}(y, z|\theta^*(z))\det(I(\theta^*))^{1/2}}{f_Y(y|\theta^*)\det(J(\theta^*(z)))^{1/2}}. \tag{57}$$

Note that this has the form of a (modified) profile likelihood, which we see by the appearance of $\theta^*(z)$ in the expression, not by decision, but by the logic of the Laplace approximation. A possible practical difficulty with using this, is that we need a new optimization for each value of z we want to consider. In that case a possible further approximation is to use a one-step Newton algorithm, to obtain a one-step estimate, starting from θ^*, the optimum based only on y.

Now we will look into the problem of prediction, using ideas from the theory of unbiased estimating equations. There seems to be few papers dedicated to this, so this is largely unexplored territory. First, we will be following the paper [**14**]. He asks how to get analogues of unbiased estimating equations, when the goal is not to estimate parameters, but to predict random variables. He defines a predictive estimating equation to be a function $H(z, y)$ of observed data y and unobserved random variable z, which we want to predict. The equation

$$H(z, y) = 0 \tag{58}$$

is referred to as a predictive estimating equation. Again, we assume that the expectation of the predictive estimating function is zero.

How can we define optimality criteria for prediction within this framework? The interest is in the future value z from a parametric family $f_Z(z;\theta)$ depending on current data y which is a sample from $f_Y(y;\theta)$. An unbiased prediction function now is a function H of data y and future data z such that

$$E_{Z,Y}(H(Z, Y) = 0. \tag{59}$$

The expectation being over the joint distribution of Z and of Y.

We could also consider conditional unbiasedness:

$$E_{Z|y}(H(Z, y)|y) = 0. \tag{60}$$

The point predictor $\hat{Z}(y)$, say, obtained as a solution of $H(z, y) = 0$, may or may not be unbiased, a departure from classical prediction. We will define optimality of our prediction function in terms of one of the two following criteria, both of which we want to minimize.

$$\frac{E_{Z,Y}(H^2)}{(E_{Z,Y}\frac{\partial H}{\partial z})^2} \tag{61}$$

or

$$\frac{E_{Z|Y}(H^2)}{(E_{Z|Y}\frac{\partial H}{\partial z})^2}. \tag{62}$$

Desmond [**14**] denote this criteria by **EFF** for efficiency, but we will avoid that, since in statistics we are usually maximizing efficiency, not minimizing it! This criteria can be justified, noting that we want to have H close to zero, so we should minimize its variance, and we want it to be as sensitive as possible, thus maximizing its derivative with respect to z, leading to the given criteria. There is much discussion as to should we use the conditional criterion (62) or the unconditional one (61). Desmond says that both may be of interest. We note that kriging ordinarily uses an unconditional criterion, indicating that we should also do so. There is an interesting discussion of this issue in [**24**, pages 248–249], which strongly indicates we should use the unconditional criterion. In the following we will do so. Compare also our earlier discussion of this issue, page 13.

Now let us see what is an optimal predicting function according to this criterion. Suppose Z is a scalar future observable from our model, and y is current data. The ideal object for predictive purposes is the conditional probability density of the future observable z conditional on current data y, $f(z|y)$, say. If we have a prior density on θ the Bayes solution is

$$f(z|y) = \int f(z|y;\theta) f(\theta|y)\, d\theta. \tag{63}$$

We argue here, informally, that if a predictive density is available, then the optimal prediction function is

$$H^* = \frac{\partial \log f(z|y)}{\partial z}. \tag{64}$$

We assume regularity conditions similar to those of [**21**] but involving conditions on the existence of certain derivatives with respect to z rather than with respect to θ, and that certain interchange of differentiation and integration be possible. Denote by $\mathcal{G}$ the class of all prediction functions $H : \mathbb{Z} \times \mathbb{Y} \to \mathbb{R}$, such that $E_{Z,Y}(H) = 0$ and $E_{Z,Y}(H^2) < \infty$. Then $\mathcal{G}$ is a real vector space, with addition of functions defined in the usual way. Now define an inner product $< H_1, H_2 >= E_{Z,Y}(H_1 H_2)$. This makes $\mathcal{G}$ into a Hilbert space. Consider an arbitrary element H of $\mathcal{G}$, denote by $\dot{H}$ its derivative with respect to z. Differentiating under the integral sign in the equation

$$E_{Z,Y}(H) = 0 \tag{65}$$

and arguing in a similar way as in the parametric case, leads to

$$E\,\dot{H} = - < H, H^* > \tag{66}$$

where $H^* = \frac{\partial \log f(z|y)}{\partial z}$. Cauchy-Schwarz inequality now yields

$$(E(\dot{H})^2) =< H, H^* > \leq \|H\|^2 \|H^*\|^2 \tag{67}$$

and optimality for H^* follows from the condition for equality in the Cauchy-Schwarz inequality.

In practice, the predictive density $f(z|y)$ will not be available and so we will need to restrict attention to certain subclasses of unbiased prediction functions. Desmond [**14**] now continues to show how his criteria applied to the spatial setting of kriging leads to simple kriging. We will not review this.

How can we get a predictive function based on the pair composite likelihood? The simplest idea is first to define a pair predictive composite likelihood, based on the joint density of Z and Y. We will now specialize the notation to our spatial setting, with N observations at locations $x_1, \ldots, x_N$ within the domain $\mathcal{D}$. We want to predict the value of the spatial field defined on $\mathcal{D}$, at location x_0 within $\mathcal{D}$. The joint density of current data $y_1 = y(x_1), \ldots, y_N = y(x_N)$ and future data $Z = y(x_0)$ is

$$f(z, y_1, \ldots, y_N; \theta) \tag{68}$$

and this is also the likelihood function. The corresponding pair likelihood function is

$$\mathrm{CL}(z,\theta) = \prod_{i=1}^{N} f(z,y_i;\theta) \cdot \prod_{i=1}^{N}\prod_{j<i} f(y_i,y_j;\theta) \tag{69}$$

and now we must face the problem of how to eliminate θ to get a predictive pair likelihood for prediction of z. The simplest idea is, again, substitute for θ the maximization estimator. We can also apply profiling to this pair likelihood. Define

$$\mathrm{cl}(z,\theta) = \sum_{i=1}^{N} \log f(z,y_i;\theta) + \sum_{i=1}^{N}\sum_{j<i} \log f(y_i,y_j;\theta) \tag{70}$$

and then

$$\widehat{\theta}_z = \arg\max_{\theta} \mathrm{cl}(z,\theta). \tag{71}$$

The profile pair likelihood for prediction can then be defined as

$$\mathrm{cl}_{\mathrm{pr}}(z) = \mathrm{cl}(z,\widehat{\theta}_z(y)) \tag{72}$$

But how can this be used? Earlier on page 20 we saw that the composite pair likelihood, if it is used as a true likelihood, will be falsely precise. So we try to correct it. Now we will see into how this can be done.

4.5. A Corrected Pair Likelihood. Here we will follow the paper [**52**] which discusses how marginal composite likelihood can be used in a Bayesian setting. The idea is to obtain a posterior distribution for θ by formally using the composite likelihood in Bayes theorem as if it was a true likelihood. This is not a new idea, there is an extensive literature on the use of alternative or approximate likelihoods in a Bayesian context, see [**23**], [**60**] and [**63**]. But as we have seen, the composite likelihood will be falsely precise if we use it as a true likelihood. Then this false precision will give rise to a falsely precise posterior distribution. Thus we need some method for correcting the composite likelihood.

We start with the definition given in equation (30) on page 21. The goal now is to find some way to choose the weights $w_{ij}, i, j =$

$1, \ldots, N$. We ask how we can correct the composite likelihood to obtain the correct variance in the normal approximation, as well as a correct shape of the posterior in its Laplace approximation. Now we write the theory for a parametric likelihood, following the paper. The modifications for the predictive case will be discussed afterwards. For inference about θ, usually tests and confidence intervals based on the likelihood ratio statistic are preferable, can we adjust this for the case of composite likelihood? We could define

$$W_c(\theta) = 2(cl(\hat{\theta}_c) - cl(\theta)), \tag{73}$$

with $cl(\theta) = \log CL(\theta)$. But this has a non-standard asymptotic null distribution. Indeed (under regularity conditions) the asymptotic distribution of $W_c(\theta)$ is a linear combination of independent chi-square variables, $W_c(\theta) \xrightarrow{d} \sum_{i=1}^{p} \lambda_i(\theta) Z_i^2$, where the Z_i are independent standard normal variables, and the $\lambda_i(\theta)$ are the eigenvalues of the matrix $V(\theta)^{-1}J(\theta)$, for $i = 1, \ldots, p$. Here the arrow denotes convergence in distribution, also called weak convergence[3]. Several adjustments based on this have been proposed, we will use a first-moment matching procedure. [**25**] proposes to use

$$W_c^{\dagger}(\theta) = W_c(\theta)/\tilde{\lambda} \tag{74}$$

with $\tilde{\lambda} = \frac{1}{p}\sum_{i=1}^{p} \lambda_i(\theta) = \frac{\operatorname{Tr} V(\hat{\theta}_c)^{-1} J(\hat{\theta}_c)}{p}$, which is one in case the Fisher and Godambe information coincides, when we obviously do not need any correction. This correction corresponds to use the same value for all the weights w_{ij}, namely $1/\tilde{\lambda}$.

Now, we can simply use this same correction for the case of the predictive composite likelihood, or we could possibly calculate $\tilde{\lambda}$ anew, based directly on the predictive pair likelihood. We could now use this corrected predictive composite likelihood in a Laplace approximation, in the same way as we did in equation (56. Ideas like this do not seem to have been much explored in the literature, at least I did not find references.

[3] The concept called weak convergence in probability and statistics, correspond to weak*-convergence in functional analysis.

4.6. Some Computational Aspects. We need to discuss some computational aspects. Here we will be mainly following [**72**]. When using the ordinary, full, likelihood, the precision of estimates is indicated by the information matrix. But should we use the observed information matrix, $j(\theta)$, or, its expectation $i(\theta)$ which is the Fisher information. The naive theoretical development as referred earlier in this chapter, seems to indicate use of the Fisher information matrix, but extensive experience and theory really gives the opposite answer, that the observed information $j(\theta)$ should be preferred. One key paper in this development is [**17**] and a book discussing this issue at length is [**7**]. In a particularly well-behaved family of models, the flat (or full) exponential family models, we have the identity $i(\theta) = j(\theta)$. If we are using numerical methods to maximize the likelihood, then it is easy to calculate $j(\theta)$: It is just the negative of the Hessian at the maximum! The issue with composite likelihood is more complicated. The curvature of the log composite likelihood function at the maximum is not enough to characterize the variability of the maximization estimator. To estimate the Godambe information matrix, we need to estimate its two factors, the sensitivity matrix $V(\theta)$ and the variability matrix $J(\theta)$. Suppose the composite likelihood is based on n independent observations, and that the number of components in the likelihood is m. We follow the paper in discussing the two cases:

(1) n large, m fixed.
(2) n small (possibly $n = 1$), m large, not necessarily fixed.

The first case is the easy one. The sensitivity matrix can be calculated via the Hessian matrix of the composite log likelihood, evaluated at its maximum:

$$\widehat{V(\theta)} = -\frac{1}{n}\sum_{i=1}^{n} cl''(\hat{\theta}, y_i) \tag{75}$$

or via the alternative expression,

$$\widehat{V(\theta)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m} cl'(\hat{\theta}; y_{ir}) cl'(\hat{\theta}; y_{ir})^{\top} \tag{76}$$

Validity follows from the identities (14), which are valid for each component likelihood, since each component is a true likelihood. Here y_{ir}

for $r = 1, \ldots m$ indexes the part of data used in component r of the component likelihood, for the ith observation. This formula can also be used in the case with small n.

To get a sample estimate of the variability matrix $J(\theta)$ is more difficult. In the large n case we can use the empirical variance of the composite score vectors, which is

$$\hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^{n} cl'(\hat{\theta}, y_i) cl'(\hat{\theta}, y_i)^\top \tag{77}$$

where cl' is the gradient of the composite log likelihood, that is, the composite score. This can be very imprecise if n is not large, as always for a variance estimate based on few observations.

An alternative approach is to use a jackknife estimate, that is,

$$\mathrm{Var}_{\mathrm{jack}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}^{(-i)} - \hat{\theta})(\hat{\theta}^{(-i)} - \hat{\theta})^\top \tag{78}$$

where $\hat{\theta}^{(-i)}$ is the maximum composite likelihood estimate with observation i omitted.

But what to do in the low-n case, especially in the extreme case $n = 1$, which is our applied case, since we have only one observation of our spatial random field! These proposals are not of much use, especially not the jackknife estimate, which is not even defined. For us to have any hope of success, we must have some internal replication within our observed random field. Assuming that, there are several possibilities. We could divide the data into several (possibly overlapping) groups, and then use the empirical variance estimate based on the composite score vectors for each group. This could possibly be approximated, using the full-data estimate as a starting point, and then use one step of a Newton algorithm within each group. Another possibility is to use parametric bootstrapping, or we could even simulate data from the estimated model, and use a Monte-Carlo estimate, given by

$$\hat{J}_{\mathrm{mc}}(\theta) = \frac{1}{B} \sum_{b=1}^{B} (cl'(\hat{\theta}; y^{(b)}))(cl'(\hat{\theta}; y^{(b)}))^\top$$

where $y^{(1)}, \ldots, y^{(B)}$ are B Monte-Carlo simulations from the fitted model, or we could even use direct estimation of the covariance matrix of $\hat{\theta}$ based on simulated replications from the fitted model. But note that such a simulation-driven estimation would be more dependent upon the assumed model, and less data-driven, countering the basic philosophy behind the use of sandwich-type variance estimation.

5. Applications to Spatial Predictin — Kriging

Our application to spatial prediction is for the case of tensors, and will need the results of the chapters (2) and (3), so will be presented in chapter (4).

6. Geometry of Tensors

The space of tensors, that is, $m \times m$ positive definite matrices, has an interesting geometry which might be useful. Let us write $\mathcal{P}^+(m)$ for the space of tensors. This is not a vector space, but it can be seen as a cone within the vector space $\mathcal{P}(m)$ of symmetric $m \times m$ matrices. The book [4] have a chapter dedicated to the geometry of positive-definite matrices. Moakher has published a lot of interest on the geometry of tensors, see [49], [48] and [47].

It is possible to make $\mathcal{P}^+(m)$ into a vector space, using a metric called the log-Euclidean metric. This theory is presented in the works [54], [55] and [56] and [1]. This way $\mathcal{P}^+(m)$ can be made into a Riemannian manifold with a flat metric. We will not review this theory here, since we will not have any use of it. The cited papers use it as a basis for simple interpolation algorithms in the context of image analysis. It could possibly have been used as a basis for kriging of tensors, but we have choosen not to follow that idea.

We will review another way of giving the tensor space a geometry. In the following we will be following [56]. Another related work is [30]. The tensor space $\mathcal{P}^+(m)$ can be seen to be a homogeneous space of the Lie group $\mathcal{GL}(m)$. $\mathcal{GL}(m)$ is the matrix group consisting of all real $m \times m$ invertible matrices. A homogeneous space is a set (or, as in this case, a manifold) with a transitive group action defined on it. The acting group in this case is the group of invertible $m \times m$ matrices,

$\mathcal{GL}(m)$. The action is by conjugation. Let us write the action as a centered dot $\cdot$, to distinguish it from matrix multiplication:

$$g \cdot p = gpg^{\top} \tag{79}$$

with $p \in \mathcal{P}^+(m)$, $g \in \mathcal{GL}(m)$. Observe that the isotropy group of the identity matrix $I \in \mathcal{P}^+(m)$ is the orthogonal group $\mathcal{O}(m)$. We could alternatively define the acting group as the subgroup of $\mathcal{GL}(m)$ consisting of matrices with positive determinant, which we write as $\mathcal{GL}^+(m)$. The action is still transitive in this case. In this case the isotropy group of the identity is the special orthogonal group, $\mathcal{SO}(m)$. By general results for group actions[4], we can write $\mathcal{P}^+(m)$ as a quotient of groups,

$$\mathcal{P}^+(m) = \mathcal{GL}(m)/\mathcal{O}(m) = \mathcal{GL}^+(m)/\mathcal{SO}(m). \tag{80}$$

We can give $\mathcal{P}^+(m)$ the structure of a Riemannian manifold by defining a metric on the tangent spaces. We can use the homogeneous space structure by simply defining a metric at the tangent space of the identity matrix, and then transporting it to the other tangent spaces by the group action. The tangent space at the identity is simply the vector space of symmetric matrices, $\mathcal{P}(m)$. At the identity, define an inner product by $< x, y >= \mathrm{Tr}(xy)$ ([**56**] gives some more possible definitions, which we will not explore). We can use the action of $\mathcal{GL}(m)$ to transport this inner product to other points of $\mathcal{P}^+(m)$. Let $p \in \mathcal{P}^+(m)$ and define $g = p^{-1/2}$, where we use the symmetric square root. It follows that $g \cdot p = I_m$. So we can define $< x, y >_p = < gxg^{\top}, gyg^{\top} >_I = \mathrm{Tr}(gxg^{\top}gyg^{\top}) = \mathrm{Tr}(p^{-1}xp^{-1}y)$. This construction is really an example of a Cartan Moving Frame. This way we can find the geodesics in $\mathcal{P}^+(m)$. One result from the differential geometry of homogeneous spaces establishes that, for an invariant metric, the geodesics are generated by the one-parameter subgroups of the acting group, see [**29**]. As the one-parameter subgroups of $\mathcal{GL}(m)$ simply are given by the matrix exponential, the geodesics emanating from the identity have the form $\Gamma_{(I,W)}(t) = \exp(tW)$, and the other geodesics can be obtained by translation. This way we can

[4] In these days the easy way out is simply to give Wikipedia as a reference for this general result!

finally obtain the Riemannian exponential map at each point, given by

$$\exp_\Sigma(W) = \Sigma^{1/2} \exp(\Sigma^{-1/2} W \Sigma^{-1/2}) \Sigma^{1/2} \tag{81}$$

Now we can obtain the Riemannian distance by integration, or more simply, by the norm of the initial tangent vector of the geodesic joining the two points (and reaching in time 1). This gives:

$$\mathrm{dist}^2(\Sigma, \Lambda) = \| \log_\Sigma(\Lambda) \|_\Sigma^2 = \mathrm{Tr}(\log(\Sigma^{-1/2} \Lambda \Sigma^{-1/2})^2) \tag{82}$$

where the Riemannian logarithmic function is given by

$$\log_\Sigma(\Lambda) = \Sigma^{1/2} \log(\Sigma^{-1/2} \Lambda \Sigma^{-1/2}) \Sigma^{1/2}. \tag{83}$$

where the Riemannian logarithmic map is the inverse of the Riemannian exponential map, and in the expression above we have used the (principal) matrix logarithm. In the following we will discuss means of tensors. But first we need to give some necessary background, before we can define a mean on a Riemannian manifold.

Now let M be a (connected) Riemannian manifold, x a point on M and the tangent space at x is $T_x M$. Let $x\vec{y}$ be a point in $T_x M$. The exponential map at x is the map sending the tangent vector $x\vec{y}$ to the endpoint of the geodesic emanating from x in direction $x\vec{y}$ and reaching at time 1. The maximum domain of definition of the exponential map at x is a star-shaped set delimited by a continuous curve in $T_x M$ denoted by C_x and denoted the tangential cut locus. The image by the exponential map of the tangential cut locus C_x is the *cut locus* $\mathcal{C}_x$. To fix ideas, take as a simple example the circle $\mathbb{S}^1$. For a point x on the circle, the cut locus is its antipodal point. In general, the cut locus $\mathcal{C}_x$ is the closure of the set of points where several different minimizing geodesics emanating from x meet. In the circle example, if we take the circle as of radius 1 with its metric as the metric as embedded in the Euclidean plane, the tangential cut locus will consist of the endpoints of the interval $[-\pi, \pi]$.

The exponential map within its domain realizes what is called the *exponential chart*. It covers all of the (connected) manifold except for the cut locus, which has null measure. In this chart, geodesics

emanating from x are straight lines, and the distances from the reference point are conserved. This chart is somehow the "most linear" chart of the manifold with respect to the reference point x. Again, to fix ideas, for the circle example, the exponential chart at x can be seen as obtained from "cutting" the circle at the cut locus, that is, the antipodal point of x, and then unrolling the circle so it becomes flat. The distance from x in this exponential chart corresponds to the angle between two points (where of course we take the arc connecting the two points and not containing the cut locus).

[**56**] gives the following table which can be used as a dictionary for "translating" algorithms from the vector space setting to the manifold setting:

Operation	Vector space	Manifold
subtraction	$x\vec{y} = y - x$	$x\vec{y} = \log_x y$
addition	$y = x + x\vec{y}$	$y = \exp_x(x\vec{y})$
distance	$\operatorname{dist}(x,y) = \|y - x\|$	$\operatorname{dist}(x,y) = \|x\vec{y}\|_x$
mean value (implicit)	$\sum_i x\vec{x}_i = 0$	$\sum_i \log_x x_i = 0$
gradient descent	$x_{t+\epsilon} = x_t - \epsilon\nabla C(x_t)$	$x_{t+\epsilon} = \exp_{x_t}(-\epsilon\nabla C(x_t))$
linear (geodesic) interpolation	$x_t = x_1 + tx_1\vec{x}_2$	$x_t = \exp_{x_1}(tx_1\vec{x}_2)$

Now, how can we define a concept of mean or average on a manifold? Remember that on the real line, the average of n points $x_1, \ldots, x_n$ can be obtained as the minimizer of the variance function $c \mapsto \sum_{i=1}^n (x_i - c)^2$. This same idea can be used on a Riemannian manifold, and is then called the Fréchet mean. It is defined as "a" minimizer of the variance functional

$$c \mapsto \sum_{i=1}^{n} \operatorname{dist}(c, x_i)^2. \tag{84}$$

We must say "a" mean, because on a general manifold there need not be uniqueness. As a simple example showing this, take again the circle, $n = 2$, and let the two points be the north and the south pole. Then, following Winnie Pooh, there are two minimizers, namely the east and the west pole! We can generalize this to a mean of order-$\alpha > 0$ by considering minimizers of $c \mapsto \sum_{i=1}^n |\operatorname{dist}(c, x_i)|^\alpha$. The case $\alpha = 1$ is called the geometric median. Take again the circle example with $n = 2$, the north and south pole. Now all the points on the circle

are minimizers!, showing a quite extreme non-uniqueness. Papers on the Fréchet mean are [**33**], [**65**] and [**37**].

Karcher [**34**] proposed to consider as means not only all the global minimizers, but all the local minimizers of the variance functional. This new set of means he called Riemannian centers of mass. Using this new definition, [**34**] and [**35**] were able to establish existence and uniqueness theorems for distributions on the manifold with a compact and small enough support.

The tensor space, with the geometry we have introduced, is a space of non-positive sectional curvature. Such a space is called a Hadamard space, the condition on the curvature has the effect that geodesics are diverging, thus are never meeting, and the cut locus is the empty set. For such spaces the variance functional is always convex, so has a unique minimum. A proof of this is contained in [**10**, Proposition 9.2.20]. The curvature of the tensor space can be calculated from information in [**49**].

The paper [**18**] proposes to use *exponential barycenters*, that is, the point at which the mean in the local exponential chart $\sum_{i=1}^{n} x\vec{x}_i = 0$. If the support of the distribution is contained in a strongly convex open set, he shows that the exponential barycenters were the critical points of the variance. They are then a superset of the Riemannian centers of mass that includes the Fréchet means. For the special case of the tensor space, by the results mentioned above, the exponential barycenter is unique and coincides with the Fréchet mean.

This idea can be used for interpolation. The linear kriging predictor of form $\sum_{i=1}^{N} \alpha_i Y_i$ can be replaced by a weighted Fréchet mean, that is, minimizer of the weighted variance functional

$$c \mapsto \sum_{i=1}^{N} \alpha_i(x_0)\, \mathrm{dist}(c, Y_i)^2. \tag{85}$$

Then the interpolated (or kriged) value at x_0 can be constructed as the minimizer of this functional, for some "optimal" weights $\alpha_i(x_0)$. However, it is not easy to see how such a program can be implemented,

and some approximations would probably be necessary. Therefore we decided to follow the other route, via construction of unbiased prediction functions as outlined in earlier sections.

CHAPTER 2

The Wishart Random Field Model

In this chapter we develop the stochastic model which will be used as a basis for construction of a methodology for interpolation of tensors. Let $\mathcal{D}$ be a domain (open, connected set) in $\mathbb{R}^d$, where d is typically 2 or 3. In $\mathcal{D}$ is defined a random field of tensors, that is, a random function $\mathcal{D} \to \mathcal{P}^+(m)$, so for each spatial referent $x \in \mathcal{D}$ there is associated a tensor $Y(x)$. In our applications, $\mathcal{D}$ will represent some region of interest on a geological map.

We will construct this random field from n independent and identically distributed Gaussian random fields of m-vectors, defined on $\mathcal{D}$, which we write as $S_i(x), x \in D, i = 1, \ldots, n$. These Gaussian random fields all have mean zero, covariance matrix Σ, an $m \times m$ positive definite matrix given as $\Sigma = E(S_i(x)S_i(x)^\top)$, for all i and all $x \in \mathcal{D}$. Finally, for the spatial correlation structure, we specify an intrinsic multivariate correlation model, see [**73**, Chapter 22]. This is also called an *separable* covariance model. This means that all the m components of the vector S_i have the same spatial correlation function, $\rho(h) = 1-\gamma(h)$ where $\gamma(h)$ is a variogram function, which is also the cross-correlation function between any pair of components. We suppose here that there are no delay effects, such that the correlation can be described by a variogram. Of course delay relationships are not consistent with an intrinsic correlation structure. The variogram is given by

$$2\gamma(h) = E((S_{ij}(x) - S_{ij}(x+h))^2) \tag{86}$$

for $x, x+h \in \mathcal{D}$, $j = 1, \ldots, m$. Here S_{ij} is the jth component of vector S_i. We will work with an isotropic variogram, so that we have $\gamma(h) = \gamma(\|h\|))$. But the construction will work with a general, not necessarily isotropic, variogram, only that we will not explore that

possibility in this work. So the specification is completed by

$$\mathrm{E}(S_i(x)S_i(x+h)^\top) = \Sigma\rho(h). \tag{87}$$

The random tensor field is then constructed by

$$Y(x) = \sum_{i=1}^{n} S_i(x)S_i(x)^\top. \tag{88}$$

By this construction the degrees of freedom parameter n is restricted to be a positive integer, but the model gives meaning also for fractional values of n. For the tensors to be non-singular it is clearly necessary that $n \geq m$. Note that the Gaussian random fields $S_i(x)$ introduced here are *latent* variables. They do not represent, by themselves, meaningful information, in particular, they do not represent observable variables. They are used here for the purpose of constructing the Wishart random field model, and will not play an individual role in the methodologies we are to develop based on this model.

We can easily simulate data from this model, on a set of spatial locations $x_1, \ldots, x_K$. Let $h_{jl} = x_j - x_l$, the spatial separation between locations x_j and x_l. Then define a $K \times K$ covariance matrix H by letting the element $H_{jl} = \rho(h_{jl})$, the spatial correlation between the sites. Then define the Gaussian random matrix S_i which is a $K \times m$ matrix, to have row k given by $S_i(x_k)^\top$, the random m-vector S_i at location k. S_i has the matrix normal distribution $N_{Km}(0, H \otimes \Sigma)$, see [**50**, Chapter 3]. This means that the random vector $\operatorname{vec} S_i^\top$ has a multinormal distribution with expectation zero and covariance matrix $H \otimes \Sigma$. Here $\otimes$ represents the Kronecker product of matrices. The way of thinking about this is that Σ represents the covariance within each sample (at each site), while H represents the correlation between sites. Using this representation we can sample the S_i using standard algorithms, and then form the Wishart distributed tensors at location $k = 1, \ldots, K$ by the sum $Y_k = \sum_{i=1}^{n} S_i(x_k)S_i(x_k)^\top$. Individually each of these tensors has a Wishart distribution with scale matrix Σ and degrees of freedom parameter n. Since the underlying Gaussian fields are spatially correlated, it is clear that the tensors will have a spatially correlated distribution.

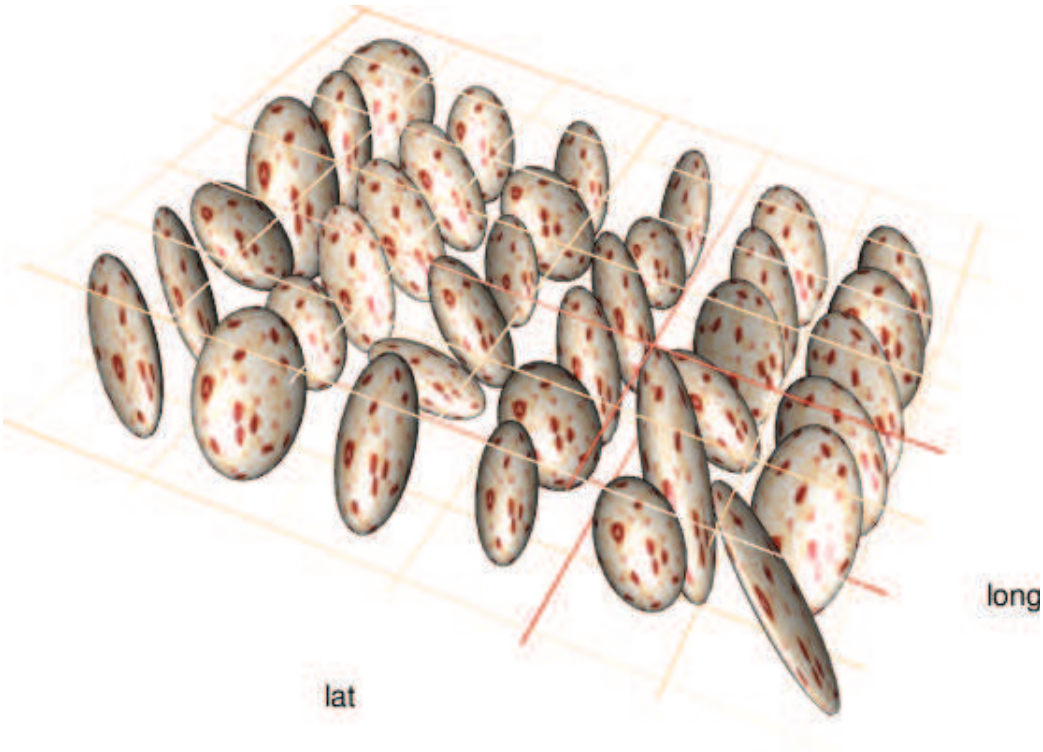

FIGURE 1. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by $\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$.

To investigate this distribution, we concentrate on the pair-distribution for any pair of two distinct sites. For simplicity we write them x_1 and x_2, with separation vector h. We write the vectors $S_i(x_1)$ and $S_i(x_2)$ at the two sites together as one vector of length $2m$, $(S_i(x_1)^\top, S_i(x_2)^\top)^\top = S_i'$. Then we can form the sum

$$\sum_{i=1}^{n} S_i' S_i'^\top \tag{89}$$

which can be seen to have the form

$$\begin{pmatrix} \sum_{i=1}^n S_i(x_1)S_i(x_1)^\top & \sum_{i=1}^n S_i(x_1)S_i(x_2)^\top \\ \sum_{i=1}^n S_i(x_2)S_i(x_1)^\top & \sum_{i=1}^n S_i(x_2)S_i(x_2)^\top \end{pmatrix} = \begin{pmatrix} Y_1 & Y_{12} \\ Y_{12}^\top & Y_2 \end{pmatrix} \tag{90}$$

and we can recognize the Wishart-distributed tensors Y_1 and Y_2 in the block diagonal position. This $(2m) \times (2m)$ random tensor has a Wishart distribution with n degrees of freedom and scale matrix

$\begin{pmatrix} \Sigma & \rho(h)\Sigma \\ \rho(h)\Sigma & \Sigma \end{pmatrix}$. The problem is that our observed data do not contain such $2m \times 2m$-tensors, we only observe the two diagonal blocks. The off-diagonal block is an artifact of our method of model construction, so to be able to use this model we need to decide what to do with the off-diagonal blocks. One idea is to eliminate it as part of the estimation process, for example by using an EM-algorithm, see [**53**]. We have experimented with this, and could not get it to function well, so decided to try another approach. That is to integrate out the off-diagonal blocks from the Wishart density, so as to obtain the marginal distribution of the two diagonal blocks. That will be reported in chapter 3.

Having obtained that marginal distribution, which effectively is a distribution of two correlated Wishart matrices, we can use the pair composite likelihood approach to estimate parameters, and further to construct predictive likelihoods. That will be the content of chapter 4.

We will show a few more simulated tensor fields. For the spatial correlation structure, we will use the exponential, spherical and gaussian variograms, see [**73**] for the definitions. We simulate over a 10×10 grid with grid spacing of 2, we use correlation functions with range 3 and without nugget effects. Finally, we use for all the simulations $n = 10$ degrees of freedom, and two different scale matrices, an identity matrix and an equicorrelation matrix with diagonal elements 1 and off-diagonal elements 0.5.

Note how the simulated fields with an equicorrelation scale matrix with correlation 0.5 visually looks more spatially continuous than the corresponding simulations with an identity scale matrix. This is a purely visual effect, due to the random changes in a round ellipsoid visually appearing as a larger change, it appears that the ellipsoid has changed form. The flatter ellipsoid resulting from the equicorrelation matrix (with one eigenvalue equal to 2 and the others 0.5) appears to have a more stable form.

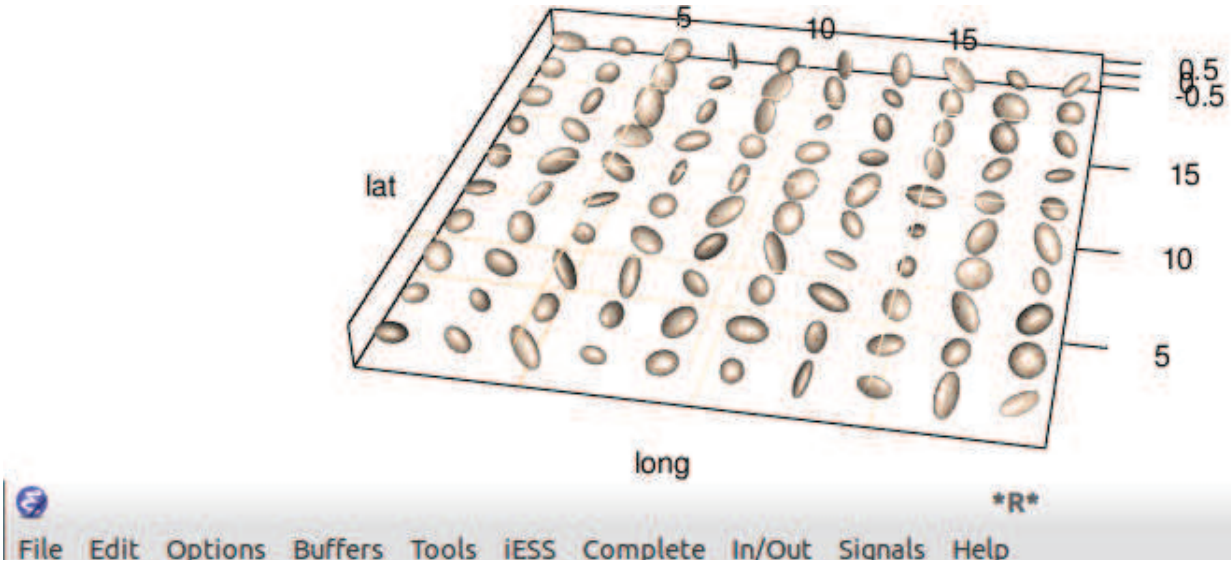

FIGURE 2. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by the identity matrix, and exponential variogram.

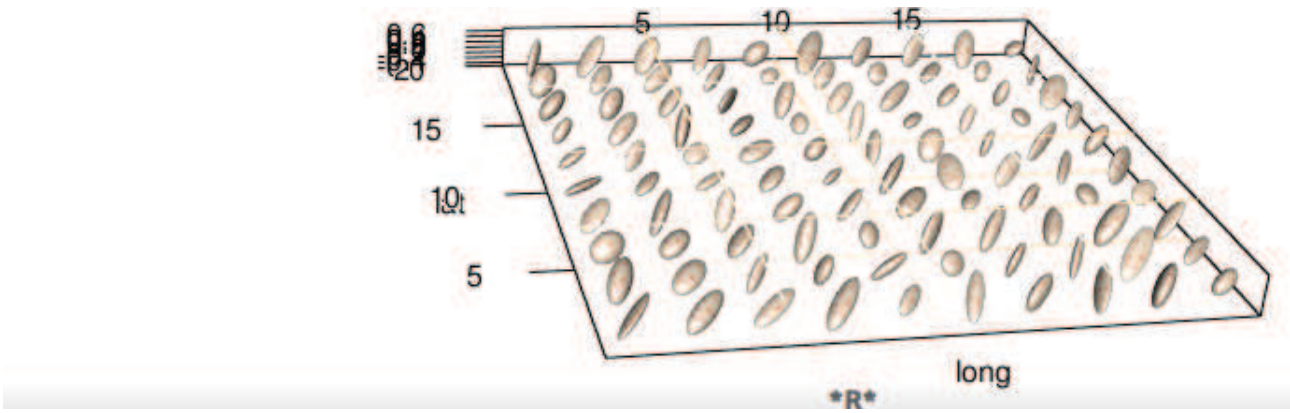

FIGURE 3. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by $\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$, and an exponential variogram.

We will show a series of simulated tensor fields where only one parameter is changed, the degrees of freedom parameter n. We do this to show how the tensor geometry becomes less variable when n increases. When $n = 1$ or $n = 2$ the tensors are singular, that is, the ellipsoid has volume zero. We do not plot this case, so we start

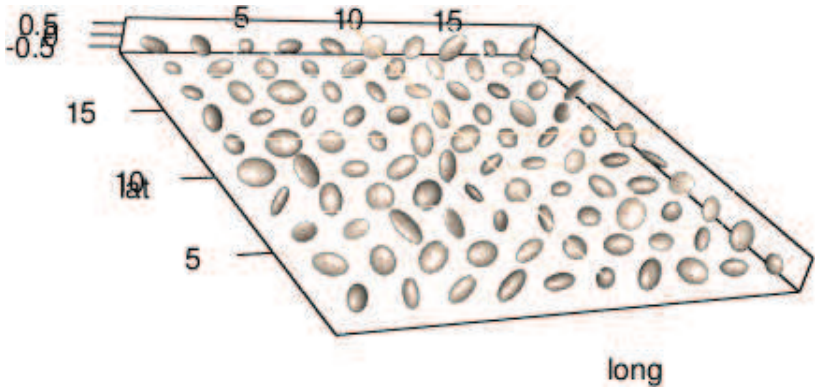

FIGURE 4. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by the identity, and a Gaussian variogram with (practical) range 3.

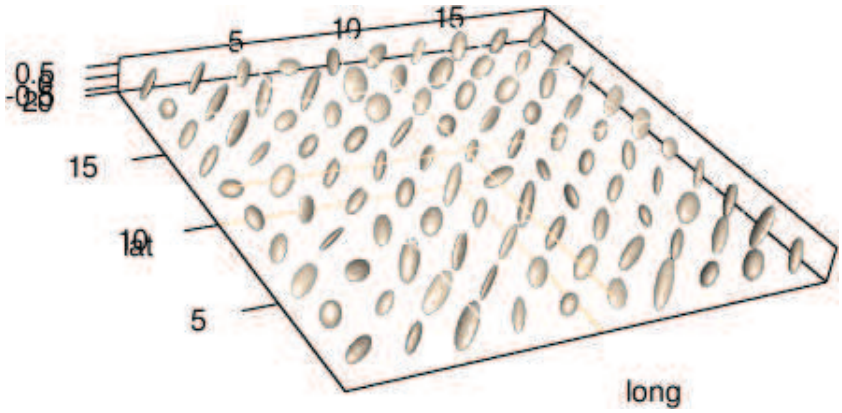

FIGURE 5. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by $\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$, and a Gaussian variogram with (practical)range 3.

with $n = 3$ and increases n up to 20. In all cases the scale matrix is the equicorrelation matrix used above, with off-diagonal elements all equal to 0.5. The simulations are done on a 20×20 regular grid, using a sperical covariance with range equal to 3, and a very small nugget effect (0.005).

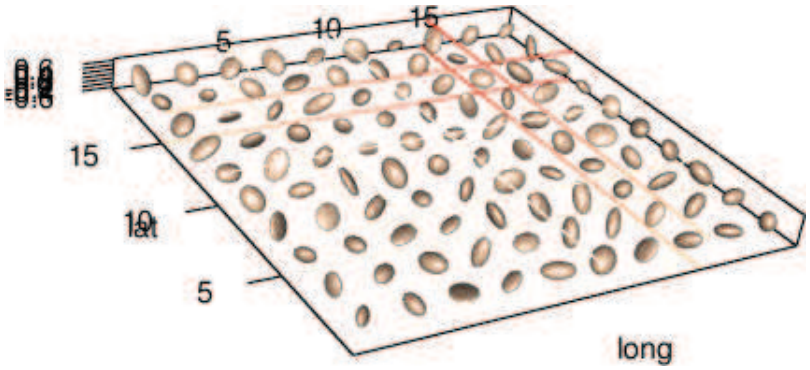

FIGURE 6. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by the identity, and a spherical variogram with range 3.

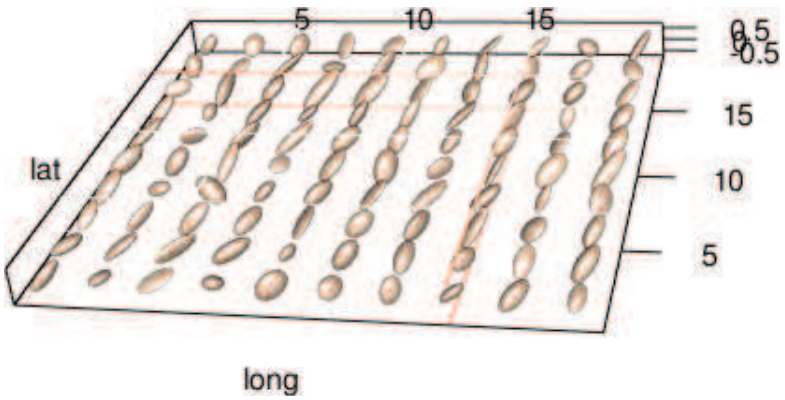

FIGURE 7. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by $\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$, and a spherical variogram with range 3.

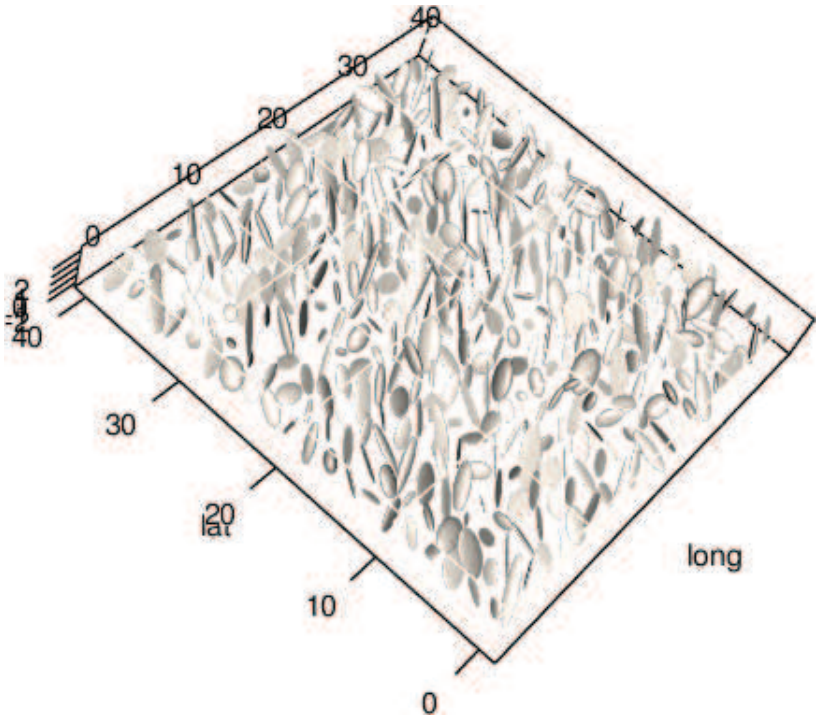

FIGURE 8. A simulated tensor field. The parameters used was $n = 3$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

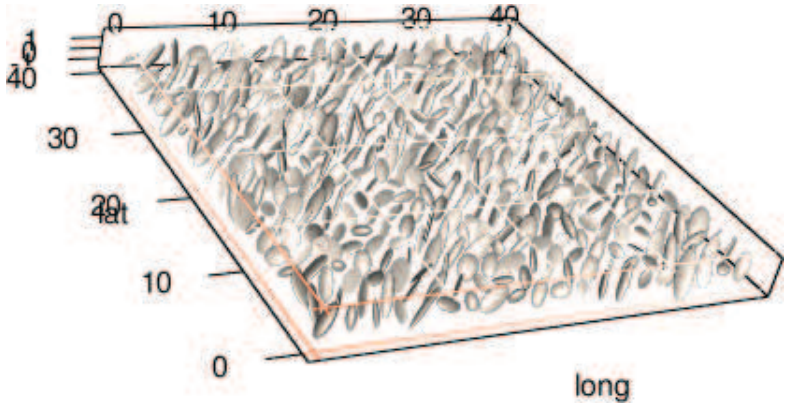

FIGURE 9. A simulated tensor field. The parameters used was $n = 4$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

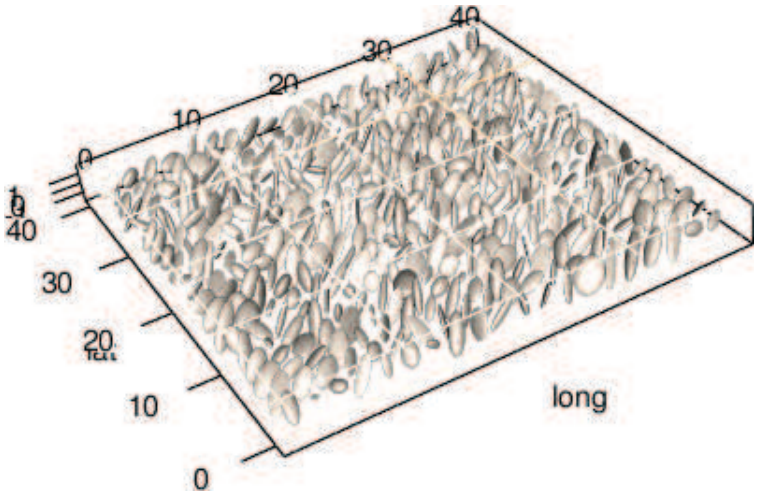

FIGURE 10. A simulated tensor field. The parameters used was $n = 5$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

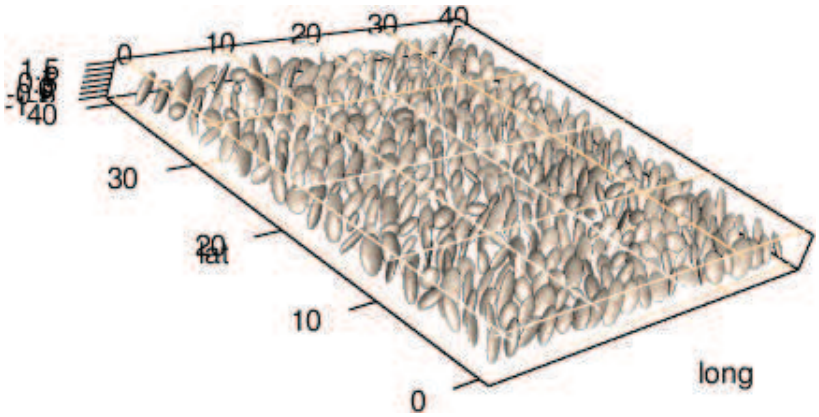

FIGURE 11. A simulated tensor field. The parameters used was $n = 7$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

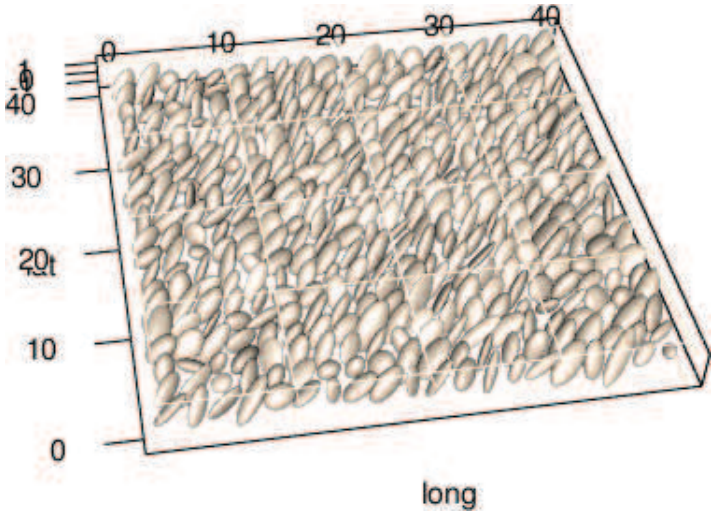

FIGURE 12. A simulated tensor field. The parameters used was $n = 10$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

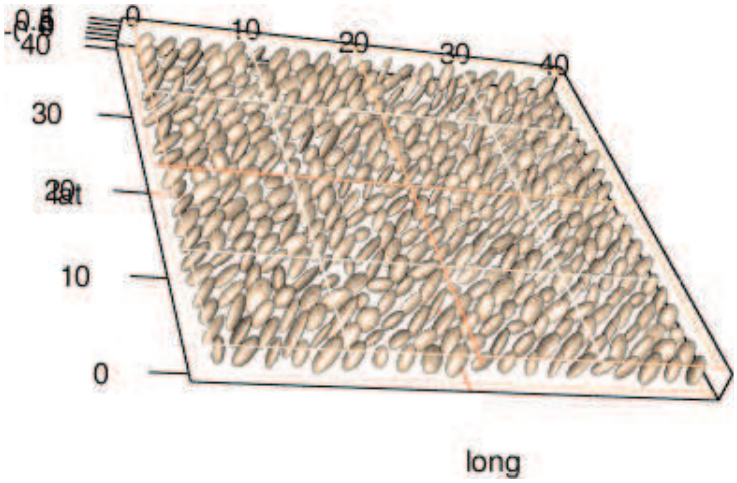

FIGURE 13. A simulated tensor field. The parameters used was $n = 15$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

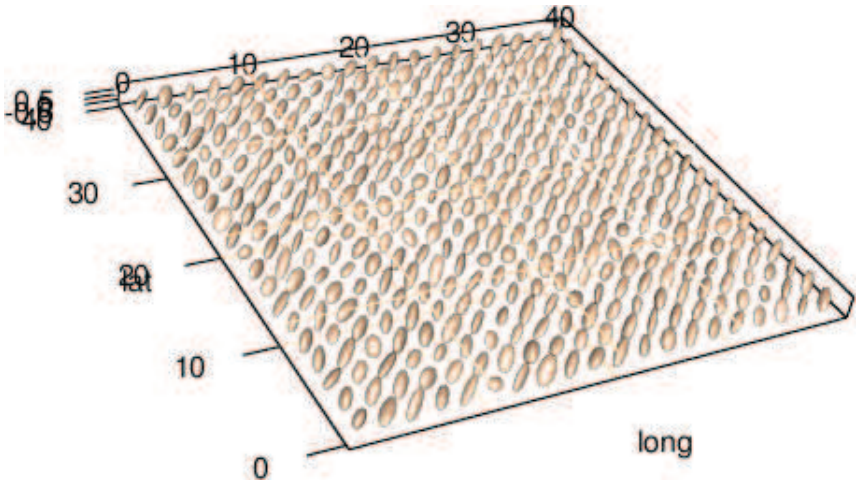

FIGURE 14. A simulated tensor field. The parameters used was $n = 20$ and a scale matrix given by an equicorrelation matrix, and spherical variogram.

CHAPTER 3

The Marginal Distribution of the Diagonal Blocks of a Blocked Wishart Random Matrix

1. Introduction

This chapter is based on the paper [**28**]. The goal of this paper is to find a useful form for the marginal distribution of the diagonal blocks of a 2×2 blocked Wishart random matrix. This problem arises in an applied problem, to estimate the parameters of a Wishart random field using composite likelihood methods, which will be reported elsewhere.

Let A be a $2m \times 2m$ Wishart random matrix, where each of the blocks are of size $m \times m$, where in our intended application m will be a small integer, typically $m = 3$. Write $A = \begin{pmatrix} A_1 & A_{12} \\ A_{12}^\top & A_2 \end{pmatrix}$.

Denote the number of degrees of freedom by n and the scale parameter, which is blocked in the same fashion as A, by $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{pmatrix}$. We are mostly interested in the special case $\Sigma = \begin{pmatrix} \Sigma_0 & \rho\Sigma_0 \\ \rho\Sigma_0 & \Sigma_0 \end{pmatrix}$ where the absolute value of ρ is less than one, but the general case dos not seem to be more difficult.

All matrices are real. Notation: we use $\mathrm{Tr}(A)$ for the trace of the square matrix A, and $\mathrm{etr}(A) = \exp(\mathrm{Tr}(A))$. We write $\mathcal{P}^+(m)$ for the cone of real $m \times m$ positive definite matrices, and we write $\mathcal{O}(m)$ for the orthogonal group, that is, the set of $m \times m$ orthogonal matrices. The Stiefel manifold, that is the set of $n \times m$ column orthogonal matrices, is written $\mathcal{V}_{m,n}$. Note that $\mathcal{V}_{m,m} = \mathcal{O}(m)$.

In the cone of positive definite matrices, we use the cone order (also called the Loewner order), defined by $A < B$ meaning that

$B - A$ is positive definite, written $B - A > 0$. Integrals over cones are written as $\int_0^I g(A)\ (dA)$ meaning the integral is taken over the cone $0 < A < I$.
The multivariate gamma function is defined by

$$\Gamma_m(a) = \int_{A>0} \operatorname{etr}(-A) \det(A)^a \det(A)^{-(m+1)/2}\,(dA) \tag{91}$$

for $\Re(a) > (m-1)/2$. We can show this is equal to

$$\Gamma_m(a) = \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma(a - \frac{i-1}{2})$$

for $\Re(a) > \frac{m-1}{2}$, which is proved in [**50**].

For more information on integrals, the notation for differentials and Jacobians, see the appendix to this chapter on page 65. The single most important reference for background material for this paper is [**50**].

Let us state our main result:

THEOREM 1. *(The Marginal Distribution of the Diagonal Blocks of a Blocked Wishart Random Matrix) Let* $A = \begin{pmatrix} A_1 & A_{12} \\ A_{12}^\top & A_2 \end{pmatrix}$ *be a* $2m \times 2m$ *blocked Wishart random matrix, where the blocks are* $m \times m$. *The Wishart distribution of* A *has* $n \geq 2m$ *degrees of freedom and positive definite scale matrix* $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{pmatrix}$ *blocked in the same way as* A. *The marginal distribution of the diagonal blocks* A_1 *and* A_2 *has density function given by*

$$\begin{aligned} &\frac{1}{2^{mn}\Gamma_m(n/2)\Gamma_m(n/2)\det(\Sigma)^{n/2}} \\ &\qquad \operatorname{etr}\left(-\frac{1}{2}(\Sigma_1^{-1}A_1 + F^\top C_2 F A_1 + C_2^{-1}A_2)\right) \\ &\qquad\qquad \det(A_1)^{\frac{n-m-1}{2}} \det(A_2)^{\frac{n-m-1}{2}}\, {}_0F_1\left(\left.\frac{n-m-1}{2}\right| G\right) \end{aligned} \tag{92}$$

where $C_2 = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}$, $F = C_2^{-1}\Sigma_{12}^\top\Sigma_1^{-1}$ *and* $G = \frac{1}{4}A_1^{1/2}F^\top A_2 F A_1^{1/2}$. ${}_0F_1$ *is the generalized matrix-variate hypergeometric function defined later.*

The rest of this chapter consists in a proof of this theorem. For completeness, we also state a version of the result for the case where the blocks are not necessarily of the same size:

THEOREM 2 (The Marginal Distribution of the Diagonal Blocks of a Blocked Wishart Random Matrix with Blocks of unequal sizes). *Let* $A = \begin{pmatrix} A_1 & A_{12} \\ A_{12}^\top & A_2 \end{pmatrix}$ *be a* $m \times m$ *blocked Wishart random matrix, where* $m = m_1 + m_2$ *and the diagonal blocks are of size* $m_1 \times m_1$ *and* $m_2 \times m_2$*, respectively. Write* $r = \min(m_1, m_2)$ *and* $s = \max(m_1, m_2)$*. The Wishart distribution of* A *has* $n \geq m$ *degrees of freedom and positive definite scale matrix* $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{pmatrix}$ *blocked in the same way as* A*. The marginal distribution of the two diagonal blocks* A_1 *and* A_2 *has density function given by*

$$\begin{aligned} (93)\quad & \frac{2^s \pi^{(r^2+s^2)/2} \Gamma_r((n-s)/2)}{2^{(r+s)n/2} \Gamma_{r+s}(n/2) \Gamma_s(s/2) \Gamma_r((n+r-s)/2) \det(\Sigma)^{n/2}} \\ & \cdot \operatorname{etr}\left\{ -\frac{1}{2}(\Sigma_1^{-1} A_1 + F^\top C_2 F A_1 + C^{-1} A_2) \right\} \det(A_1)^{(n-m_1-1)/2} \\ & \det(A_2)^{(n-m_2-1)/2} \, {}_1F_2\left(\begin{matrix} r/2 \\ s/2, (n+r-s)/2 \end{matrix} \middle| G \right) \end{aligned}$$

and where the notation not given here corresponds to the one in the first theorem.

2. A Wishart block matrix

The Wishart density function is, for our case,

$$(94)\qquad c \cdot \operatorname{etr}(-\frac{1}{2}\Sigma^{-1} A) \det(A)^\gamma$$

where $c = \left(2^{mn} \Gamma_{2m}(n/2) \det(\Sigma)^{n/2}\right)^{-1}$, $\gamma = \frac{n-2m-1}{2}$, $n \geq 2m$ and A is positive definite. (If A is not positive definite, the density is zero, a fact which will be understood hereafter).

The following simple facts will be useful. They are shown using standard formulas for block matrices. Introduce the following notation: The Schur complements of $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{pmatrix}$ are $C_1 = \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}$ and $C_2 = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}$. Then define $F = C_2^{-1}\Sigma_{12}^\top \Sigma_1^{-1}$.

a: $\det(\Sigma) = \det(\Sigma_1)\det(C_2) = \det(\Sigma_2)\det(C_1)$.

b:

$$\begin{pmatrix}\Sigma_1 & \Sigma_{12}\\ \Sigma_{21} & \Sigma_2\end{pmatrix}^{-1} = \begin{pmatrix}\Sigma^1 & \Sigma^{12}\\ \Sigma^{12\top} & \Sigma^2\end{pmatrix} = \begin{pmatrix}\Sigma_1^{-1} + F^\top C_2 F & -F^\top\\ -F & C_2^{-1}\end{pmatrix}. \tag{95}$$

c: $\operatorname{Tr}\Sigma^{-1}A = \operatorname{Tr}\left(\Sigma_1^{-1}A_1 + F^\top C_2 F A_1 - F^\top A_{12}^\top - F A_{12} + C_2^{-1}A_2\right) = \operatorname{Tr}\left(\Sigma^1 A_1 + \Sigma^{12}A_{12}^\top + \Sigma^{12\top}A_{12} + \Sigma^2 A_2\right)$.

d: $\det(A) = \det(A_1)\det(A_2 - A_{21}A_1^{-1}A_{12}) = \det(A_2)\det(A_1 - A_{12}A_2^{-1}A_{21})$

Using this we can rewrite the density from (94) as a function of the blocks:

$$c\cdot\exp\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma_1^{-1}A_1 + F^\top C_2 F A_1 - F^\top A_{12}^\top - F A_{12} + C_2^{-1}A_2\right)\right) \cdot \det(A_2)^\gamma \det(A_1 - A_{12}A_2^{-1}A_{21})^\gamma \tag{96}$$

In the following we will work with the density concentrating on the factors depending on A_{12}. To prove the theorem we need to integrate out the variable A_{12}. The other variables, which are constant under the integration, will be concentrated in one constant factor. So we repeat the formula (96) written as a differential form with the constants left out, indicating the intention to integrate over A_{12}:

$$K_1\cdot\operatorname{etr}(\frac{1}{2}(F^\top A_{12}^\top + F A_{12}))\det(A_1 - A_{12}A_2^{-1}A_{21})^\gamma\ \ (dA_{12}) \tag{97}$$

where $K_1 = c\cdot\operatorname{etr}\left(\frac{1}{2}(\Sigma_1^{-1}A_1 + F^\top C_2 F A_1 + C_2^{-1}A_2)\right)\det(A_2)^\gamma$. Now, to find the marginal distribution of the diagonal blocks, we need to integrate over the off-diagonal block A_{12}. Under this integration the value of the diagonal blocks A_1 and A_2 will remain fixed, and the region of integration will be a subset of $\mathbb{R}^{m\times m}$ consisting of the matrices A_{12} such that the block matrix $A = \left(\begin{smallmatrix}A_1 & A_{12}\\ A_{12}^\top & A_2\end{smallmatrix}\right)$ is positive definite. This seems like a complicated set, but we can give a simple description of it using the polar decomposition of a matrix. Note that this is one of the key observations for the proof, and this author has not seen any use of this observation earlier.

The region of integration is the set

$$\{A_{12} \in \mathbb{R}^{m\times m} \colon 0 < A_{12}A_2^{-1}A_{12}^\top < A_1\} \tag{98}$$

Introduce $D = A_{12}A_2^{-1/2}$ where we use the usual symmetric square root, given by $A_2^{1/2} = U\Lambda^{1/2}U^\top$ where $A_2 = U\Lambda U^\top$ is the spectral decomposition. Then in terms of the new variable D the region of integration becomes

$$\{D \in \mathbb{R}^{m\times m} \colon 0 < DD^\top < A_1\} \tag{99}$$

and with the polar decomposition $D = PU$ with $P \in \mathcal{P}^+(m), U \in \mathcal{O}(m)$, $DD^\top = PUU^\top P^\top = P^2$ so the region of integration can be written as

$$\{P \in \mathcal{P}^+(m), U \in \mathcal{O}(m) \colon 0 < P^2 < A_1\} \tag{100}$$

which is a Cartesian product of a cone interval with the orthogonal group.

One way to understand this result is as follows: We want to integrate over all matrices D such that $DD^\top$ belongs to the cone interval $(0, A_1)$. D is a square root of $DD^\top$, it is clear that any other square root will do as well. Let E be another square root of $DD^\top$, that is, $DD^\top = EE^\top$. It is clear that the product of D with an orthogonal matrix Q, DQ, will be another square root, and we ask: Do all square roots arise in this way? Indeed, they do: If D and E are two different square roots, define $Q = D^{-1}E$. Now a simple calculation confirms that Q is orthogonal: $QQ^\top = D^{-1}EE^\top D^{-\top} = D^{-1}DD^\top D^{-\top} = I$. This fact, that all square roots are orthogonally related, is sometimes described as unitary freedom of square roots. So, the region of integration consists of all the square roots of all the members of the cone interval $(0, A_1)$. That this is a Cartesian product as described above is a simple consequence of the unitary freedom of square roots.

In the following we will, by using successive transformations, get the density into a form we are able to integrate. The necessary Jacobians can be found in the appendix to this chapter, on page 65.

Write $D = A_{12}A_2^{-1/2}$ where here and elsewhere we use the symmetric square root. It follows that $A_{12}A_2^{-1}A_{12}^\top = DD^\top$ and the Jacobian is $(dD) = \det(A_2)^{-m/2}\,(dA_{12})$, see lemma (121) in appendix to this chapter. Using this the form (97) becomes

$$K_2 \cdot \operatorname{etr}(DA_2^{1/2}F)\det(A_1 - DD^\top)^\gamma \;\; (dD) \tag{101}$$

where $K_2 = K_1 \cdot \det(A_2)^{m/2}$. Now define $E = A_1^{-1/2}D$ with Jacobian $(dE) = \det(A_1)^{-m/2}\,(dD)$, see lemma (121). The form (101) becomes

$$K_3 \cdot \operatorname{etr}(A_1^{1/2}EA_2^{1/2}F)\det(I - EE^\top)^\gamma\,(dE) \tag{102}$$

where $K_3 = K_2 \cdot \det(A_1)^{\gamma+m/2}$.

Finally, we introduce the polar decomposition of E, $E = PU$, where $P \in \mathcal{P}^+(m)$ is a positive definite matrix and $U \in \mathcal{O}(m)$ is an orthogonal matrix. This transformation has a non-trivial Jacobian which again can be found in the appendix to this chapter (lemma (125)) as $(dE) = \prod_{i<j}^m (D_i + D_j)\,(dP)\left(U^\top dU\right)$ where $D_i, i = 1, \ldots, m$ are the eigenvalues of P. With this transformation the integral of (102) can be written as

$$K_3 \cdot \int_{\mathcal{O}(m)} \int_0^I \operatorname{etr}(A_1^{1/2}PUA_2^{1/2}F)\det(I - P^2)^\gamma \prod_{i<j}^m (D_i + D_j)\;\; (dP)\;\; \left(U^\top dU\right) \tag{103}$$

which we can write as an iterated integral as

$$K_3 \cdot \int_0^I \prod_{i<j}^m (D_i + D_j)\det(I - P^2)^\gamma \;\; (dP) \int_{\mathcal{O}(m)} \operatorname{etr}(A_1^{1/2}PUA_2^{1/2}F) \;\; \left(U^\top dU\right). \tag{104}$$

We are ready to perform the integration over the orthogonal group. For this purpose we need the following result from [**50**, Theorem 7.4.1, page 262], which we cite here.

Let X be an $m \times n$ real matrix with $m \leq n$ and $H = [H_1 : H_2]$ an $n \times n$ orthogonal matrix, where H_1 is $n \times m$. Then

$$\int_{\mathcal{O}(n)} \operatorname{etr}(XH_1) \;\; (dH) = {}_0F_1\left(\begin{matrix} \\ n/2 \end{matrix}\middle| \frac{1}{4}XX^\top\right). \tag{105}$$

The special case that we need for the main theorem is for $m = n$ and is given below:

$$\int_{\mathcal{O}(m)} \operatorname{etr}(XH)\ (dH) = {}_0F_1\left(\begin{matrix} \\ m/2 \end{matrix}\middle| \frac{1}{4}XX^\top\right) \tag{106}$$

where (dH) denotes Haar measure on $\mathcal{O}(m)$ normalized to have unit total mass.

Here ${}_0F_1$ denotes the generalized matrix-variate hypergeometric function, which is discussed in chapter 7 of [**50**]. We give the definition here:

Definition 1 (Matrix-variate hypergeometric function). The generalized hypergeometric function of symmetric matrix-variate argument X are given by

$${}_pF_q\left(\begin{matrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{matrix}\middle| X\right) = \sum_{k=0}^{\infty} \sum_{\kappa \vdash k} \frac{(a_1)_\kappa \dots (a_p)_\kappa}{(b_1)_\kappa \dots (b_q)_\kappa} \frac{C_\kappa(X)}{k!} \tag{107}$$

where $\sum_{\kappa \vdash k}$ denotes summation over all partitions of the integer k into no more than m parts, $\kappa = (\kappa_1, \kappa_2, \dots, \kappa_m)$ where $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m \geq 0$, $\kappa_1 + \dots + \kappa_m = k$. C_κ is the zonal polynomial of argument X corresponding to the partition κ of k, and the generalized hypergeometric coefficient $(a)_\kappa$ is defined as $(a)_\kappa = \prod_{i=1}^{m} (a - \frac{1}{2}(i-1))_{\kappa_i}$ where $(a)_{\kappa_i}$ is the rising factorial $(a)_k = a(a+1)(a+2)\dots(a+k-1)$ and $(a)_0 = 1$. The argument X is a symmetric $m \times m$ matrix and the zonal polynomial is a symmetric polynomial in the eigenvalues of X. Note that if the matrix argument X is a symmetric positive definite matrix, then the value of the zonal polynomial is positive. This is [**50**, corollary 7.2.4].

The matrix-variate hypergeometric function is a formal generalization from the classical hypergeometric function, based on simply using, in some sense, the same series expansion as a definition. It is not clear at all if other properties generalize. A good, modern introduction to the classical hypergeometric function is the book [**58**]. For efficient calculation of the matrix-variate hypergeometric function, see [**36**].

The zonal polynomial (and thus the hypergeometric function) can be extended to an argument which is a product of symmetric matrices (one of which is positive semidefinite) by defining $C_\kappa(XY) = C_\kappa(X^{1/2}YX^{1/2})$ and in this way we can extend the definition to non-symmetric matrices. This definition does not extend to all square matrices. As shown in [**67**], all square matrices can be written as a product of two symmetric matrices, but it is not possible to always choose one of the factors as positive-semidefinite.

Since the eigenvalues of products of matrices is invariant under cyclic permutations of the factors in the product, we have $C_\kappa(XYZ) = C_\kappa(YZX)$ etc, a fact which is very useful for manipulation of expressions. One basic property of the zonal polynomials is that $(\operatorname{Tr} X)^k = \sum_{\kappa\vdash k} C_\kappa(X)$, using this it is easy to see that

$$ {}_0F_0\left(\left| X\right.\right) = \operatorname{etr}(X). $$

Another special case is ${}_1F_0\left(a\middle| X\right) = \det(I-X)^{-a}$.

Convergence properties: The series converges absolutely for all X if $p \leq q$, and it converges for $\|X\| < 1$ if $p = q+1$. In all other cases it diverges, unless it terminates. It will converge for the cases we will be using. More information about zonal polynomials and, in particular, everything we need, can be found in [**50**]. A more comprehensive, but still readable, introduction to zonal polynomials is [**68**].

Returning to our integral, the integral over the orthogonal group occurring in (104) can now be written as

(108)

$$\begin{aligned}\int_{\mathcal{O}(m)} &\operatorname{etr}(A_1^{1/2}PUA_2^{1/2}F)\ \left(U^\top dU\right)\\ &= \operatorname{Vol}(\mathcal{O}(m))\int_{\mathcal{O}(m)} \operatorname{etr}(A_1^{1/2}PUA_2^{1/2}F)\ (dU)\\ &= \operatorname{Vol}(\mathcal{O}(m))\int_{\mathcal{O}(m)} \operatorname{etr}(A_2^{1/2}FA_1^{1/2}PU)\ (dU)\\ &= \operatorname{Vol}(\mathcal{O}(m))\,{}_0F_1\left(m/2\middle|\frac{1}{4}A_2^{1/2}FA_1^{1/2}P^2A_1^{1/2}F^\top A_2^{1/2}\right)\end{aligned}$$

where we did use (106). Here $\operatorname{Vol}(\mathcal{O}(m)) = \frac{2^m\pi^{m^2/2}}{\Gamma_m(m/2)}$ is the volume of the orthogonal group, that is, $\operatorname{Vol}(\mathcal{O}(m)) = \int_{\mathcal{O}(m)}\left(U^\top dU\right)$, a result which can be found in [**50**]. Note that the differential form $\left(U^\top dU\right)$ is an invariant differential form on the orthogonal group, invariant means that it is invariant under right and left translations defined by $U \mapsto UQ$ and $U \mapsto QU$ for $Q \in \mathcal{O}(m)$. This is in fact the famous Haar measure for the orthogonal group. To show invariance calculate $\left((QU)^\top(d(QU)\right) = \left(U^\top Q^\top Q(dU)\right) = \left(U^\top dU\right)$ since Q is orthogonal. (A somewhat more involved calculation works for translation on the right). The differential form (dU) denotes Haar measure normalized to total mass unity. Also note that since the orthogonal group is compact and we are integrating a continuous function, there are no problems with existence of the integral.

Now write $G = \frac{1}{4}A_1^{1/2}F^\top A_2FA_1^{1/2}$ then we can write (104) as

(109)

$$K_3\operatorname{Vol}(\mathcal{O}(m))\int_0^I\prod_{i<j}^m(D_i+D_j)\det(I-P^2)^\gamma\,{}_0F_1\left(m/2\middle|GP^2\right)\ (dP)$$

and to evaluate this integral we need another result from [**50**, theorem 7.2.10, page 254]:

If Y is a symmetric $m \times m$ matrix then we have

$$\int_0^I \det(X)^{a-(m+1)/2} \det(I-X)^{b-(m+1)/2} C_\kappa(XY)\ (dX) = \frac{(a)_\kappa}{(a+b)_\kappa} \frac{\Gamma_m(a)\Gamma_m(b)}{\Gamma_m(a+b)} C_\kappa(Y) \tag{110}$$

for $\Re(a) > \frac{m-1}{2}$, $\Re(b) > \frac{m-1}{2}$.

We can immediately use this to find a result we need for the integral of an hypergeometric function, by using the series expansion definition of the hypergeometric function and integrating term by term:

THEOREM 3. *If Y is a symmetric $m \times m$ matrix we have that*

$$\int_0^I \det(X)^{a-(m+1)/2} \det(I-X)^{b-(m+1)/2} {}_pF_q\left(\left.\begin{matrix}a_1,\dots,a_p\\ b_1,\dots,b_q\end{matrix}\right| XY\right)\ (dX) = \frac{\Gamma_m(a)\Gamma_m(b)}{\Gamma_m(a+b)} {}_{p+1}F_{q+1}\left(\left.\begin{matrix}a_1,\dots,a_p,a\\ b_1,\dots,b_q,a+b\end{matrix}\right| Y\right) \tag{111}$$

so both degrees of the hypergeometric function are raised by one.

Let us show the details of this calculation:

$$\int_0^I \det(X)^{a-(m+1)/2} \det(I-X)^{b-(m+1)/2} {}_pF_q\left(\left.\begin{matrix}a_1,\dots,a_p\\ b_1,\dots,b_q\end{matrix}\right| XY\right)\ (dX) = \tag{112}$$

$$\int_0^I \det(X)^{a-(m+1)/2} \det(I-X)^{b-(m+1)/2} \sum_{k=0}^{\infty}\sum_{\kappa\vdash k} \frac{(a_1)_\kappa \dots (a_p)_\kappa}{(b_1)_\kappa \dots (b_q)_\kappa} \frac{C_\kappa(X)}{k!}\ (dX) =$$

$$\sum_{k=0}^{\infty}\sum_{\kappa\vdash k} \frac{(a_1)_\kappa \dots (a_p)_\kappa}{(b_1)_\kappa \dots (b_q)_\kappa} \int_0^I \det(X)^{a-(m+1)/2} \det(I-X)^{b-(m+1)/2} \frac{C_\kappa(X)}{k!}\ (dX) =$$

$$\sum_{k=0}^{\infty}\sum_{\kappa\vdash k} \frac{(a_1)_\kappa \dots (a_p)_\kappa}{(b_1)_\kappa \dots (b_q)_\kappa} \frac{(a)_\kappa}{(a+b)_\kappa} \frac{\Gamma_m(a)\Gamma_m(b)}{\Gamma_m(a+b)} C_\kappa(Y) =$$

$$\frac{\Gamma_m(a)\Gamma_m(b)}{\Gamma_m(a+b)} {}_{p+1}F_{q+1}\left(\left.\begin{matrix}a_1,\dots,a_p,a\\ b_1,\dots,b_q,a+b\end{matrix}\right| Y\right)$$

We transform our integral one more time, writing $X = P^2$ with Jacobian $(dX) = \prod_{i\leq j}^{m}(D_i + D_j)\,(dP)$ where again the D_i are the eigenvalues of P, see lemma (124) in the appendix to this chapter. The integral (109) becomes

(113)

$$K_3 \cdot \mathrm{Vol}(\mathcal{O}(m)) \int_0^I \prod_{i<j}^{m}(D_i + D_j) \prod_{j=1}^{m}(1 - D_i^2)^\gamma \left(\prod_{i\leq j}^{m}(D_i + D_j)\right)^{-1} \cdot {}_0F_1\left(\begin{matrix} \\ m/2 \end{matrix}\middle| XG\right)\,(dX)\,.$$

Look at the product of eigenvalues in the integral, it simplifies to

$$2^{-m}\prod_{j=1}^{m}\frac{(1-D_j^2)^\gamma}{D_j}$$

which we can see to be the determinant of the diagonal matrix $(I - Q^2)^\gamma Q^{-1}$ where Q is the diagonal matrix with the eigenvalues of P on the diagonal. Since $X = P^2$, this is the same as $\det(I-X)^\gamma \det(X)^{-1/2}$, and the integral (113) becomes

(114)

$$K_3 \cdot 2^{-m}\,\mathrm{Vol}(\mathcal{O}(m)) \int_0^I \det(X)^{-1/2} \det(I-X)^\gamma\, {}_0F_1\left(\begin{matrix} \\ m/2 \end{matrix}\middle| XG\right)(dX)$$

and now using (111) we get, finally, the result

$$K_3 \cdot 2^{-m}\,\mathrm{Vol}(\mathcal{O}(m))\frac{\Gamma_m(m/2)\Gamma_m(\gamma + (m+1)/2)}{\Gamma_m(\gamma + m + 1/2)} {}_1F_2\left(\begin{matrix} m/2 \\ m/2, \gamma + m + 1/2 \end{matrix}\middle| G\right) \tag{115}$$

but note that one pair of upper and lower arguments to the hypergeometric function are equal, those clearly cancels. We can do a last

simplification, calculating

$$\frac{\Gamma_m((n-m)/2)}{\Gamma_{2m}(n/2)} = \frac{\Gamma_m(n/2-m/2)}{\Gamma_{2m}(n/2)} = \tag{116}$$
$$\frac{\pi^{m(m-1)/4}\Gamma(n/2-m/2)\Gamma(n/2-m/2-1/2)}{\pi^{2m(2m-1)/4}\Gamma(n/2)\Gamma(n/2-1/2)}$$
$$\cdot \frac{\dots\Gamma(n/2-1/2)}{\dots\Gamma(n/2-(m-1)/2)\Gamma(n/2-m/2)\dots\Gamma(n/2-1/2)} =$$
$$\frac{1}{\pi^{m^2/2}\Gamma_m(n/2)}.$$

This completes the proof of our main theorem.

We will not give the complete proof of the general version theorem 2, but only indicate the few differences. The main difference is related to the polar decomposition used in (103), which we must replace with the less known generalized polar decomposition for rectangular matrices. Let C be an $n \times m$ rectangular matrix with $n \leq m$. Define the positive semi-definite matrix $H = (CC^\top)^{1/2}$. Now define U by $C = HU$ where U is an $n \times m$ matrix. First, in the case that H is non-singular we can write $U = H^{-1}C$ and then calculate $UU^\top = H^{-1}CC^\top H^{-1} = H^{-1}H^2H^{-1} = I_n$, which shows that $U^\top$ is a column-orthogonal matrix. In the case that H is singular we can simply take the limit of non-singular matrices, using the fact that the non-singular matrices forms an open dense set. This is the generalized polar decomposition for rectangular matrices. See [**31**] for this, and some more generalizations.

Using this, we are led in the same way as before to an iterated integral, but now the integral over the orthogonal group is replaced by an integral over the Stiefel manifold, and we need now to generalize the result (105) to an integral over the Stiefel manifold. What we need is the following. Let $\mathcal{V}_{m,n}$ be the manifold of $n \times m$ column orthogonal matrices with $m \leq n$, and let f be a function defined on the Stiefel manifold. We can extend this function to a function defined on $\mathcal{O}(n)$ in the following way. Let U be an $n \times n$ orthogonal matrix, and write it in block form as $[U_1 : U_2]$ such that $U_1 \in \mathcal{V}_{m,n}$. How can we characterize the set of U_2 which is complementing U_1 to form an orthogonal

matrix? First, let U_2 be a fixed, but arbitrary such matrix. Then clearly any other $n \times (n-m)$ column orthogonal matrix with the same column space also works. The common column space is the orthogonal complement of the column space of U_1. The set of such matrices can be described as $\{V \in \mathcal{V}_{n-m,n} \colon V = U_2 Q \text{ for } Q \in \mathcal{O}(n-m)\}$. For this set we write $\mathcal{V}^{H_1}_{n-m,n}$. As a set we can identify this with $\mathcal{O}(n-m)$. Specifically, we can identify U_2 with the very special column orthogonal matrix $\begin{pmatrix} 0_{m \times n-m} \\ Q \end{pmatrix}$ where $Q \in \mathcal{O}(n-m)$ which clearly forms a proper submanifold of the Stiefel manifold $\mathcal{V}_{n-m,n}$. The function f can now be extended to the orthogonal group by defining $f(U) = f([U_1 \colon U_2]) = f(U_1)$ and for the integral we find that

$$\begin{aligned} (117) \quad \int_{\mathcal{O}(n)} f(H_1) \left(H^\top dH\right) = \\ \int_{\mathcal{V}_{m,n}} \int_{\mathcal{V}^{H_1}_{n-m,n}} f([H_1 \colon H_2]) \left(H^\top dH\right) = \\ \int_{\mathcal{V}_{m,n}} f(H_1) \left(H_1^\top dH_1\right) \int_{\mathcal{V}^{H_1}_{n-m,n}} \left(Q^\top dQ\right) = \\ \operatorname{Vol} \mathcal{O}(n-m) \int_{\mathcal{V}_{m,n}} f(H_1) \left(H_1^\top dH_1\right). \end{aligned}$$

With these changes the proof goes through with only minor notational changes.

3. Appendix: Jacobians

When doing change of variables in a multiple integral we need to know the Jacobian. In this appendix we will list the ones we need, most can be found in [**50**] or in [**44**]. We are following the notation of [**50**]. First a very brief summary.

For any matrix X, let dX denote the matrix of differentials dx_{ij}. For an arbitrary $n \times m$ matrix X, the symbol (dX) denotes the exterior product of the mn elements of dX:

$$(118) \qquad (dX) \equiv \wedge_{j=1}^{m} \wedge_{i=1}^{n} dx_{ij}.$$

If X is a symmetric $m \times m$ matrix, the symbol (dX) will denote the exterior product of the $\frac{m(m+1)}{2}$ distinct elements of dX:

$$(dX) \equiv \wedge_{1 \le i \le j \le m} dx_{ij}. \tag{119}$$

With similar definitions for other kinds of structured matrices. Note that since we are only interested in the absolute value of the determinants, the indeterminacy in the order of the wedge product in the formulas above are of no import.

Specifically, note that for the $2m \times 2m$ block matrix $A = \begin{pmatrix} A_1 & A_{12} \\ A_{12}^T & A_2 \end{pmatrix}$ we get

$$(dA) = \wedge_{1 \le i \le j \le m} da_{ij} \wedge \wedge_{m+1 \le i \le j \le 2m} da_{ij} \wedge \wedge_{\substack{i=1,\dots m \\ j=m+1\dots 2m}} da_{ij} = (dA_1)\,(dA_2)\,(dA_{12})\,. \tag{120}$$

We need to integrate over the orthogonal group, then we need the following invariant form representing the Haar measure, $\left(H^\top dH\right) = \wedge_{i=1}^m \wedge_{j=i+1}^m h_j^\top dh_i$. Here H represents an orthogonal matrix. The integral of this form over the orthogonal group gives its volume. This form normalized to have total mass unity is represented by (dH).

LEMMA 2. *If* $X = BY$ *where* X *and* Y *are* $n \times m$*-matrices and* B *is a (constant) nonsingular* $n \times n$*-matrix, then*

$$(dX) = (\det B)^m\,(dY) \tag{121}$$

so that $J(X \to Y) = (\det B)^m$.

A proof can be found in [**50**].

LEMMA 3. *If* $X = BYC$, *where* X *and* Y *are* $n \times m$*-matrices, and* B *and* C *are* $n \times n$ *and* $m \times m$ *nonsingular matrices, then*

$$(dX) = (\det B)^m (\det C)^n\,(dY) \tag{122}$$

so that $J(X \to Y) = (\det B)^m (\det C)^n$.

A proof can be found in [**50**]

LEMMA 4. *If* $X = BYB^T$, *where* X *and* Y *are* $m \times m$ *symmetric matrices and* B *is a nonsingular* $m \times m$*-matrix, then*

$$(dX) = (\det B)^{m+1} (dY). \tag{123}$$

A proof can be found in [**50**].

A few more Jacobians, the following ones are special cases of results from [**15**] and [**16**]. The three following results are the only ones not to be found in [**50**].

LEMMA 5. *(Jacobian of the symmetric square root of a positive definite matrix). Let* S *and* R *be in* $\mathcal{P}^+(m)$ *such that* $S = R^2$ *and let* D *be a diagonal matrix with the eigenvalues of* R *on the diagonal. Then,*

$$(dS) = 2^m |D| \prod_{i<j}^{m} (D_{ii} + D_{jj}) (dR) = \prod_{i\leq j}^{m} (D_{ii} + D_{jj}) (dR) \tag{124}$$

This result can also be found in [**44**]

LEMMA 6. *(Polar decomposition). Let* X *be an* $m \times m$*-matrix of rank* m*, that is,* $X \in GL(m)$*, with* m *distinct singular values, and write* $X = PR$*, with* $P \in \mathcal{O}(m), R \in \mathcal{P}^+(m)$*.* R *has* m *distinct eigenvalues. Also let* $S = X^TX = R^TR = R^2$ *where* R *is the symmetric square root of* S*. Let* $L = D^2$ *where* D *is a diagonal matrix with the eigenvalues of* R *on the diagonal. Then*

$$(dX) = \prod_{i<j}^{m} (D_{ii} + D_{jj}) (dR) \left(P^T dP\right) \tag{125}$$

and

$$(dX) = 2^{-m} |L|^{-1/2} (dS) \left(P^T dP\right) \tag{126}$$

where $L = D^2$ *and* D *is the diagonal matrix of eigenvalues of* R*, and* L *with those of* S*.*

LEMMA 7. *(Generalized Polar decomposition) Let* X *be an* $N \times m$ *matrix with* $N \geq m$ *and of rank* m*, with* m *distinct singular values. Write* $X = PR$*, with* $P \in \mathcal{V}_{N,m}$ *and* $R \in \mathcal{P}^+(m)$*. Then* R *has*

m *distinct eigenvalues. Also let* D *be the diagonal matrix with the eigenvalues of* R *on the diagonal. Then*

$$(dX) = \det(D)^{N-m} \prod_{i<j}^{m} (D_i + D_j) \, (dR) \left(P^\top dP\right). \tag{127}$$

Note that when we are using this result in an integral, the assumption that all eigenvalues are distinct do not really matter, since the subset where some eigenvalues coincide is of measure zero and does not contribute to the integral.

CHAPTER 4

Estimation and Prediction

In this chapter we will finally present methods for spatial interpolation of tensors. We start by analyzing the pair likelihood function, and take as a starting point the generalized version given by equation (31), and apply it to the Wishart random field model of chapter (2). Here we use the notation and results of chapter (3). Since we assume the random field is stationary, the scale matrix Σ of the Wishart random field does not depend on spatial location. So the $2m \times 2m$ blocked scale matrix of the two-dimensional density of the two tensors $A_i = A(x_i)$ and $A_j = A(x_j)$ have the form

$$\begin{pmatrix} \Sigma & \rho(h_{ij})\Sigma \\ \rho(h_{ij})\Sigma & \Sigma \end{pmatrix} \tag{128}$$

where $\rho(h)$ is the spatial correlation function. Note that in this model, a positive constant can be freely moved between the two factors in $\rho(h)\Sigma$, that is, variance can be modelled in the correlation function $\rho(h)$ or in the covariance matrix Σ. To eliminate this redundancy we require that $\rho(0) = 1$. In terms of the corresponding variogram this is to say that the sill is one. If we are using an intrinsic random field model, that is, a model with a variogram without a sill, then we must fix the redundancy in some other way.

This simple stationary model leeds to simplifications in the main result theorem (1). Let us look at that. First, we find that

$$C_2 = (1 - \rho(h)^2)\Sigma \tag{129}$$

$$F = \frac{\rho(h)}{1 - \rho(h)^2}\Sigma^{-1} \tag{130}$$

$$F^\top C_2 F = \frac{\rho(h)^2}{1 - \rho(h)^2}\Sigma^{-1} \tag{131}$$

$$G = \frac{1}{4}\left(\frac{\rho(h)}{1 - \rho(h)^2}\right)^2 \Sigma^{-1}A_1\Sigma^{-1}A_2 \tag{132}$$

Using this we can see that the log pair likelihood based on the two observed tensors A_1 and A_2, is given by

$$\begin{aligned} \text{logpairlik} = &-mn\log(2) - 2\log(\Gamma_m(n/2)) - (n/2)\log\det(\Sigma) \\ &- (1/2)\frac{1}{1 - \rho(h)^2}\left(\Sigma^{-1}A_1 + \Sigma^{-1}A_2\right) \\ &+ \frac{n - m - 1}{2}\left(\log\det(A_1) + \log\det(A_2)\right) \\ &+ \log\ {}_0F_1\left(\left.(n - m - 1)/2\right| G\right). \end{aligned} \tag{133}$$

Let us now combine this with the marginal likelihood contribution, using equation (31) with a new matching parameter a, to be chosen later. Introduce the notation G_{ij} for the G matrix above calculated using the tensors A_i and A_j, with spatial separation vector $h_{ij} =$

$x_i - x_j$. Then after some manipulation we get

(134)

$$\text{logpairlik}_a = \underbrace{\frac{mn}{2} N \log(2)(a-(N-1)) + N \log \Gamma_m(n/2)(a-(N-1))}_{\alpha}$$

(135)
$$+ \underbrace{-\frac{nm}{2} \sum_{i<j}^{N} \log(1-\rho_{ij}^2)}_{\beta}$$

(136)
$$+ \underbrace{\log\det \Sigma \frac{nN}{2} (a-(N-1))}_{\gamma}$$

(137)
$$+ \underbrace{\frac{1}{2} \sum_{i=1}^{N} (a-\omega_i) \cdot \mathrm{Tr}(\Sigma^{-1} A_i)}_{\delta}$$

(138)
$$+ \underbrace{\frac{n-m-1}{2} ((N-1)-a) \sum_{i=1}^{N} \log\det A_i}_{\epsilon}$$

(139)
$$+ \underbrace{\sum_{i<j}^{N} \log\ {}_0F_1\left((n-m-1)/2 \middle| G_{ij} \right)}_{\eta}.$$

Here $\rho_{ij} = \rho(h_{ij})$ and $\omega_i = \sum_{j \neq i} \frac{1}{1-\rho_{ij}^2}$. Note that in the independence case, all ω_i is equal to $N-1$. Choosing $a = N-1$ is matching the contributions from the two parts of the log pair likelihood, so is annulling most of the likelihood! In the independence case it will in fact only leave contributions from parts β and η of the likelihood, the two parts which only have contributions from the bivariate distributions. So probably, $a = N-1$ will be a poor choice and goods values of a can be expected to be in the interval $(0, N-1)$. To get an optimal choice of a we would need to know the Godambe information matrix as a function of a, but even if that were possible there is no reason to expect a common value of a to be good for all parameters.

Some preliminary numerical work, using programming in the statistical programming language R, indicates that the optimization problem which must be solved to estimate the parameters via maximization of the composite log likelihood, is difficult. So a detailed study of the properties of the composite log likelihood might be an aid in constructing good optimization routines. So in the following we will make a detailed study of properties of logparlik_a. Doing this analytically will be difficult, so this must mostly be based on simulations.

Now we will study this composite likelihood function, term by term, and discuss which information is contained in each term. We will also comment on possible choices for the matching constant a. A term contains information only if it can be possibly different for different possible data, in other words, it must depend on the A_i.

For prediction of the value of the tensor field at some location $x_0 \in D$ where it is not observed, we can construct a predictive likelihood based on the pairlikelihood (134). The modification consists in including the unknown unobserved tensor $A_0 = A(x_0)$ in the pairlikelihood function, in exactly the same way as if it had been observed.

That leads to the following definition:

(140)

$$\text{logpairlik.pred}_a = \underbrace{\frac{mn}{2}(N+1)\log(2)(a-N) + (N+1)\log\Gamma_m(n/2)(a-N)}_{\alpha}$$

(141)
$$+\underbrace{-\frac{nm}{2}\left(\sum_{i<j}^{N}\log(1-\rho_{ij}^2)+\sum_{i=1}^{N}\log(1-\rho_{0i}^2)\right)}_{\beta}$$

(142)
$$+\underbrace{\log\det\Sigma\cdot\frac{n(N+1)}{2}(a-N)}_{\gamma}$$

(143)
$$+\underbrace{\frac{1}{2}\sum_{i=0}^{N}(a-\omega_i)\cdot\operatorname{Tr}(\Sigma^{-1}A_i)}_{\delta}$$

(144)
$$+\underbrace{\frac{n-m-1}{2}(N-a)\sum_{i=0}^{N}\log\det A_i}_{\epsilon}$$

(145)
$$+\underbrace{\sum_{0\le i<j\le N}\log\ {}_0F_1\left(\left.(n-m-1)/2\right|G_{ij}\right)}_{\eta}.$$

where $\rho_{ij}=\rho(h_{ij})$ includes the new cases ρ_{0i} and ω must be redefined to use the index 0 in the summation, as $\omega_i=\sum_{\substack{0\le j\le N\\ j\neq i}}\frac{1}{1-\rho_{ij}^2}$.

Note that is in general hopeless to do a global maximization of the pairlikelihood function, since it is in general unbounded, except in very special cases, like choosing $a=N-1$ ($a=N$ in prediction case) which will zero a lot of terms. Observe that the α term is unbounded as a function of n, as is also the β, γ and ϵ terms. A practical choice could be to fix n, alternatively we could fix the parameters in the spatial correlation structure, then the β term will be maximized by choosing n as small as possible. In some cases, depending on the structure of the observations, it might be possible to do a global optimization, but this will be fragile. Probably a solution

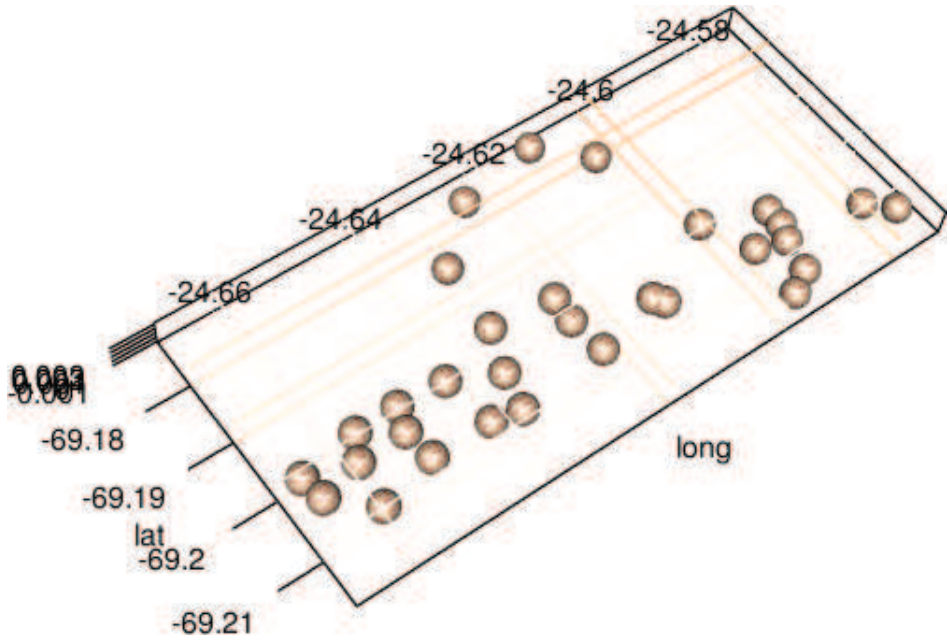

FIGURE 1. A real tensor field. The Sierra de Varas example data. The results of a cross-validation exercise.

to this problem must be developed on a case-by-case basis, maybe n could be estimated separately based on the variability of the size of the observed tensors. This phenomenon of a globally unbounded likelihood function is similar to what happens in the case of using the t-distribution as a data model. In that case also the likelihood function is unbounded. Solutions and pitfalls in this case is discussed in the papers [**20**] and [**39**].

An alternative would be to isolate some part of the likelihood function, for example only the marginal likelihood, and use that part to estimate n only, and then later fixing it.

We used the data shown in figure (2) to do prediction. In the following is shown the results of a cross-validation exercise based on this data. For each one of the 31 tensors of this dataset, we removed that one tensor, leaving 30, and use that 30 together with an estimated model to predict the tensor in that location. The results is shown in

figure (1), which should be compared to the data in figure (2). The geometry of the tensors is well reproduced, not so the volume.

CHAPTER 5

Conclusions and Challenges

We have developed a framework for spatial interpolation of tensors, that is, positive definite matrices. We have developed a stochastic model for a spatial field of tensors, and developed necessary theory for estimation of parameters, using composite likelihood methods. The basic functioning of the methodology is demonstrated, but there is a need for working out more examples, and constructing more stable software. The optimization of the composite likelihood function we constructed showed to be highly non-trivial, and time-consuming.

There is a lot more to do. We should try the new methods with more datasets, they should be extended to more general spatial covariance structures. We need to understand relationships with interpolation based on the Fréchet mean, see equation (84) and (85). The methods should also be studied in more extensive simulation exercises.

There is also a need to explore possibilities for better deterministic interpolation algorithms, like spline-based methods. The methods we developed should be programmed as software which can be published as an R-package on CRAN (the Comprehensive R Archival Network, `http://cran.r-project.org/`. The programs already developed as part of this thesis is a start in that direction.

Bibliography

[1] Vincent Arsigny, Pierre Fillard, Xaviet Pennec and Nicholas Ayache: "Fast and Simple Computations on Tensors with Log-Euclidean Metrics". INRIA, Rapport de Recherche N° 5584, May 2005.

[2] E. Bedrick and Joe E Hill: "Properties and applications of the generalized Likelihood as a summary function for prediction problems," Scandinavian Journal of Statistics, vol 26, No 4, 1999.

[3] J O Berger, R L Wolpert: "The Likelihood Principle, Second Edition". Institute of Mathemathical Statistics, Lecture Notes— Monograph Series, vol 6, 1988.

[4] Rajendra Bhatia: "Positive Definite Matrices". Princeton University Press. Princeton and Oxford, 2007.

[5] J. F. Bjørnstad: "Predictive Likelihood: A Review". In Statistical Science, 1990, vol 5, No 1, 242–265.

[6] J. F. Bjørnstad: "On the Generalization of the Likelihood Function and the Likelihood Principle". J. of the Amer. Statist. Assoc., vol 91, No 434, 791–806, 1996.

[7] O E Barndorff-Nielsen and D R Cox: "Inference and asymptotics". Chapman & Hall, 1994.

[8] K Gerald van den Boogaart and Helmut Schaeben: "Kriging of Regionalized Directions, Axes, and Orientations I: Directions and axes". Mathematical Geology, vol 34, No. 5, July 2002.

[9] K Gerald van den Boogaart and Helmut Schaeben: "Kriging of Regionalized Directions, Axes, and Orientations II: Orientations". Mathematical Geology, vol 34, No. 6, August 2002.

[10] D Burago, Y Burago and S Ivanov: "A course in metric geometry". Graduate Studies in Mathematics, vol 33, AMS.

[11] Jean-Paul Chilès and Pierre Delfiner: Geostatistics, Modelling Spatial Uncertainty. Wiley, 1999.

[12] Noel A C Cressie: "Statistics for Spatial Data, Revised Edition", Wiley, 1990.

[13] A C Davison: "Approximate Predictive Likelihood". Biometrika, Vol 72, No 2, 1986, 323–332.

[14] A F Desmond: "Prediction Functions and Geostatistics". In Institute of Mathematical Statistics, Lecture Notes — Monograph Series, Vol 32: Selected Proceedings of the Symposium on Estimating Equations. Ishwar V Basawa, V P Godambe and Robert L Taylor, editors, 1997.

[15] José A. Díaz-García, Graciela González-Farías: "Singular random matrix decompositions: Jacobians." Journal Of Multivariate Analysis, 2005.

[16] José A. Díaz-García, Ramon Gutierrez Jaimez, Kanti V. Mardia: "Wishart and Pseudo-Wishart Distributions and Some Applications to Shape Theory"

[17] B Efron: "The Geometry of Exponential Families". The Annals of Statistics, vol 6, No 2, 1978.

[18] Emery, M. and Mokobodzki, G. (1991). "Sur le barycentre d'une probabilité dans une variété". In J. Azema, P.A. Meyer, M. Y., editor, Séminaire de probabilités XXV, volume 1485 of Lect. Notes in Math., pages 220–233. Springer-Verlag.

[19] A W F Edwards: "Likelihood, Expanded Edition". The Johns Hopkins University Press, Baltimore and London, 1992.

[20] Carmen Fernández and Mark F J Steel: "Multivariate Student-t Regression Models: Pitfalls and Inference". Web document.

[21] Ferreira, P: "Estimating Equations in the Presence of Prior Knowledge". Biometrika 69, 1982, 667–669.

[22] T B Fomby and R C Hill: "Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later", Elsevier, 2003.

[23] Fraser, D A S, Reid, N, Wong, A and Yun Yi, G: "Direct Bayes for interest parameters". Valencia 7, 529–533, 2003.

[24] A Gelman, J B Carlin, H S Stern and D B Rubin: "Bayesian Data Analysis, second edition". Chapman & Hall/CRC, 2004.

[25] Geys, H, Molenberg, G and Ryan, L M: "Pseudolikelihood modelling of multivariate outcomes in developmental toxicology". J Amer. Statist. Assoc. 94, 734–745, 1990.

[26] V. P. Godambe: "An Optimum property of regular Maximum Likelihood Estimation". Ann. Math. Statist. 31, 1208–11, 1960.

[27] González, R. 2007: "Mecanismo de emplazamiento del plutón sierra de varas, cordillera de Domeyko, norte de Chile" / Rodrigo Iván González Tapia. Unpublished PhD thesis, Geology. Universidad Católica del Norte. 192 p. Antofagasta, Chile.

[28] K B Halvorsen, V Ayala and E Fierro: "On the Marginal Distribution of the Diagonal blocks in a 2×2 Blocked Wishart Random Matrix". Submitted for publication to Communications in Statistics, Theory and Methods.

[29] Sigurdur Helgason: "Differential Geometry, Lie Groups and Symmetric spaces". Academic Press, 1978.

[30] Fumio Hiai and Dénes Petz: "Riemannian Metrics on Positive Definite Matrices Related to Means". arXiv:0809.4974v3 [math-ph] 8 Nov 2008.

[31] Nicholas J. Higham, Christian Mehl , and Francçise Tisseur: "The Canonical Generalized Polar Decomposition". Web document, version of july 14, 2009.

[32] Ingrid Hotz, Jaya Sreevalsan-Nair, Hans Hagen and Bernd Hamann: "Tensor Field Reconstruction Based on Eigenvector and Eigenvalue Interpolation". Institute for Data Analysis and Visualization, (IDAV), Department of Computer Science, University of California, Davis CA 95616, USA. Se puede acceder desde `http://graphics.idav.ucdavis.edu/publications`.

[33] Justin Jacobs: "Fréchet Means on Manifolds". February 7th, 2007. Presentation, internet document.

[34] H Karcher: "Riemannian center of mass and mollifier smoothing". Communications in Pure and Applied Mathematics, 30, 509–541, 1977.

[35] W Kendall: "Probability, convexity and harmonic maps with small image. I: uniqueness and fine existence". Proc. London Math. Society 61(2), 371–406, 1990.

[36] Plamen Koev, Alan Edelman: "The Efficient Evaluation of the Hypergeometric Function of a Matrix Argument", in Mathematics of Computation, Volume 75, Number 254, 833–846, 2006.

[37] P. Thomas Fletcher, Suresh Venkatasubramanian, Sarang Joshi: "The Geometric Median on Riemannian Manifolds with Application to Robust Atlas Estimation". Neuroimage, 2009 March, (1 Suppl) S143–S152.

[38] David H Laidlaw, Joachim Weichert (Editores): "Visualization and Processing of Tensor Fields", Advances and Perspectives, Springer.

[39] Kenneth Lange, Roderick J A Little and Jeremy M G Taylor: "Robust Statistical Modeling Using the t Distribution". JASA, 1989, Vol 84, No 408, 881–896.

[40] C Lantuéjoul: "Geostatistical Simulation, Models and Algorithms". Springer, 2002.

[41] E L Lehmann and George Casella: "Theory of Point Estimation, Second Edition". Springer Texts in Statistics, Springer, 1998.

[42] T Leonard: Comment on "A simple predictive density function" by M Lejeune and G D Faulkenberry. J. Amer. Statist. Assoc. 81, 196–198, 1982.

[43] Bruce G. Lindsay: "Composite Likelihood Methods", en Contemporary Mathematics, AMS, Volume 80, 1988.

[44] A M Mathai: "Jacobians of Matrix Transformations and Functions of Matrix Arguments", 1997, World Scientific.

[45] Georges Matheron: "Splines and Kriging: Their Formal Equivalence". Internal Report, Centre. de Géostatistique: Ecole des Mines de Paris.

[46] Deborah G Mayo: An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle. Downloaded from `http://www.phil.vt.edu/dmayo/personal_website/ch%207%20mayo%20birnbaum%20proof.pdf`.

[47] Maher Moakher and Philipp G. Batchelor: "Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization".

[48] Maher Moakher: "On the Averaging of Symmetric Positive-Definite Tensors". Journal of Elasticity, (2006) 82: 273–296.

[49] Maher Moakher, Mourad Zevai: "the Riemannian geometry of the space of Positive-Definite Matrices and its application to the regularization of Positive-Definite Matrix-valued Data". (2011)

[50] Robb I Muirhead: "Aspects of Multivariate Statistical Theory", Wiley, 1982.

[51] Lyle Noakes, Greg Heinzinger , Brad Paden: "Cubic Splines on Curved Spaces". IMA Journal of Mathematical Control and Information, 1989.

[52] Francesco Pauli, Walter Racuguo, Laura Ventra: "Bayesian Composite Marginal Likelihood". In Statistica Sinica 21, 2011, 149–164.

[53] Yudi Pawitan: "In All Likelihood: Statistical Modelling and Inference using Likelihood", Oxford Science Publications.

[54] Xavier Pennec, Pierre Fillard, Nicholas Ayache: "A Riemannian Framework for Tensor Computing". International Journal of Computer Vision 66(1):41–66, January 2006.

[55] Xavier Pennec: "Probabilities and Statistics on Riemannian Manifolds: A Geometric Approach". INRIA, Rapport de recherche n° 5093 — January 2004.

[56] Xavier Pennec: "Statistical Computing on Manifolds for Computational Anatomy", Mémoire présenté pour l'obtention de l'habilitation à diriger des recherches.

[57] P Pluch: "Kriging and Splines: Theoretical Approach to Linking Spatial Prediction Methods". In Interfacing Geostatistics and GIS, Springer, Editor: Jürgen Pilz, Springer, 2009.

[58] Marko Petkovšek, Herbert S. Wilf , Doron Zeilberger: "A=B". Book downloadable from the internet.

[59] Michel de Saint-Blanquat, Richard D. Law, Jean-Luc Blouchez and Sven S. Morgan: "Internal structure and emplacement of the Papoose Flat pluton: An integrated structural, petrographic, and magnetic susceptibility study". GSA Bulletin, August 2001; v. 113, no (, 976–995.

[60] Reid, N: "Likelihood and Bayesian approximation methods. Bayesian Statistics", 5, 355–368, 1995.

[61] D R Cox & N Reid: "A note on pseudolikelihood constructed from marginal densities". Biometrika, Vol 91, No 3, 2004, 729–737.

[62] B D Ripley: "Spatial Statistics". John Wiley & Sons, 1981.

[63] Severini, T A: "On the Relationship between Bayesian and non-Bayesian elimination of nuisance parameters". Statist Sinica 9, 711–724, 1999.

[64] Xiaoyan Shi, Martin Styner, Jeffrey Lieberman, Joseph G. Ibrahim, Weili Lin and Hongtu Zhu: "Intrinsic Regression Models for Manifold-Valued Data". J Am Stat Assoc 2009, January 1, 5762, 192–199.

[65] Huiling Le and Alfred Kume: "The Fréchet Mean Shape and the Shape of the Means". Advances in Applied Probability, vol. 32, No 1 (March, 2000), 101–113.

[66] G K Robinson: "That BLUP is a good thing: The Estimation of Random Effects". In Statistical Science, 1991, Vol 6, No 1, 15–51.

[67] H Stenzel: "Über die Darstellbarkeit einer Matrix als Produkt von zwei symmetrischen Matrizen, als Produkt von zwei alternierenden Matrizen und als Produkt yon einer symmetrisehen und einer alternienden Matrix". mathematische zeitschrift, volume 15, Number 1, 1922.

[68] Akimichi Takemura: "Zonal Polynomials". Institute of Mathematical Statistics, Lecture Notes, monograph series, volume 4.

[69] Claire Tomlin: "Splining on Lie Groups". `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.9386`

[70] Florian Steinke, Matthias Hein, Jan Peters, Bernhard Schölkopf: Manifold-valued thin plate splines with applications in computer graphics.

[71] Varin, C: "On Composite Marginal Likelihood". Advances in Statistical Analysis 92, 1–28, 2008

[72] Varin, C, Reid, N and Firth, D: "An overview of composite Likelihood Methods". Statistica Sinica 21, 2011, 5–42.

[73] H Wackernagel: "Multivariate Geostatistics, An Introduction with Applications". Springer, 1995.

[74] Holger Wendland: "Scattered Data Approximation", in book series Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

[75] Burkhard Wünsche: "The visualization of 3D stress and strain tensor fields". Preprint, www.cs.auckland.ac.nz/~burkhard/GVRG/gradconf99.pdf

[76] Inas A Yassine, Tim McGraw: "A Subdivision Approach to Tensor Field Interpolation". Preprint, www.csee.wvu.edu/~tmcgraw/cdmri2008.pdf