Master Thesis

PREDICTION BASED ON PAIRWISE LIKELIHOOD

Supervisors

Professor Anthony C. Davison
Professor Alastair Young

Corina Grünenfelder

January 15, 2010

Contents

Introduction

In the context of global warming, extreme environmental phenomena such as heavy rainfall, avalanches, windstorms or high tides happen to occur more often and could cause major problems for human beings living in the affected regions. Scientists of various backgrounds want to be able to manage the risks induced by these rare events.

In statistics, spatial models are required to analyse extreme observations and to base predictions on them. Recently, max-stable processes (Smith (1990) and Schlather (2002)) have been considered for modelling extremal spatial data as these processes allow us to generalize extremal dependence structure in space. The analysis of a series of extremes observed at locations in a spatial domain can be based on different types of max-stable processes. Unfortunately, the multivariate density of these processes is intractable and different methods need to be considered. A possible approach is to consider composite likelihoods which are constructed by compounding marginal densities as suggested by Varin and Vidoni (2005) and many others. Padoan *et al.* (2009) and Davison and Gholamrezaee (2009) have lately applied this method to extreme data observed in a spatial domain. Their findings indicate that the composite likelihood methodology allows reliable model fitting and that models can be compared through modified likelihood criteria.

This project aims to give an approach to prediction, rather than model fitting, for extreme events in space based on composite likelihood methods. More specifically, pairwise likelihood is considered as obtaining higher-order joint densities for max-stable processes is non-trivial. Prediction methods for spatial extremes were already investigated by Cressie *et al.* (2005). In their article, predictors based on loss functions are compared to different types of kriging. More specifically, a predictor minimizing the integrated weighted quantile squared error loss is suggested whereas in this project, the predictors aim to minimize the mean square prediction error.

As traditional geostatistical tools for prediction, mostly based on multivariate normal theory, are not well established for extreme values, Gaussian processes are considered initially (section 3.4). On the Gaussian scale it is possible to compute the best linear predictor, that means the one minimizing the mean square prediction error, and to compare it to the suggested pairwise predictor. In addition, the influence of parameter estimation on the prediction is investigated. This is of particular interest as pairwise likelihood is then used for obtaining both the parameter estimates and the predictors and so there are two

potential sources of error. The results indicate that even though the parameter estimation does not influence the mean predictor it affects the prediction error. An illustration of the method is given by the analysis of mean precipitations in north-east Switzerland.

The pairwise method is adapted to simulated extremal data in a spatial domain. In this context, however, it is not possible to find the analytical form of the pairwise predictor and the composite likelihood needs to be optimized numerically. Moreover, the analysis of the prediction cannot be based on mean square prediction errors as on the one hand max-stable processes do not have finite moments and on the other the distribution of the predictors is not close to being normal. Therefore a more robust measure, the median square prediction error, is considered.

Both the 'random storm' and the 'random process' models, suggested by Smith (1990) and Schlather (2002), respectively, are considered in the simulation study. The properties of the predictors are poor if there is weak correlation or if few gauged locations are available. As the predictors are constantly underestimating the true values at the ungauged locations other methods are suggested. In particular, it turns out to be useful to transform the data to the Gaussian scale and then compute the best linear predictor. Back-transforming this predictor to the unit Fréchet scale yields then a predictor for the ungauged location. Even if the predictor is much more precise, there is a practical problem with this approach. In fact, for the bivariate Gaussian distribution, two extreme observations are asymptotically independent for a fixed correlation (Sibuya (1960)). This is not realistic in applications.

Seasonal maximum precipitations in Switzerland are considered to exemplify the method in the extremal spatial domain. Both the predictors on the unit Fréchet and the back-transformed predictors are considered. The back-transformed predictors allow more precise predictions but justifying the independence between the extrema is problematic.

Chapter 1

Composite marginal likelihood

In several applications it is hard to compute the likelihood as the datasets might be large or the model very complex, sometimes the likelihood is not even known analytically. In these cases it is useful to reduce the computational complexity by using composite likelihoods (see for example Varin and Vidoni (2005), Varin (2007) or Cox and Reid (2004)). Basically, composite marginal likelihoods correspond to likelihoods that are constructed by compounding marginal densities. Therefore, using composite marginal likelihoods can be seen as a particular case of model under-specification.

1.1 Maximum composite marginal likelihood estimator

Definition 1. For a parametric statistical model $\{f(y;\theta), y \in \mathcal{Y}, \theta \in \Theta\}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$, $\Theta \subseteq \mathbb{R}^p$, $d \geq 1$ and $p \geq 1$, let $\{\mathcal{A}_j : \mathcal{A}_j \subseteq \mathcal{F}, j \in J\}$ be a set of events, where $J \subseteq \mathbb{N}$ and $\mathcal{F}$ is some sigma algebra on $\mathcal{Y}$. Then, for a set of independent observations $(y_1, \ldots, y_n)$ a composite marginal likelihood is defined as

$$L_C(\theta) = \prod_{i=1}^{n} \prod_{j \in J} f(y_i \in \mathcal{A}_j; \theta)^{w_j},$$

where $f(y_i \in \mathcal{A}_j; \theta) = f(\{y_{i_l} \in \mathcal{Y} : y_{i_l} \in \mathcal{A}_j\}; \theta)$ and $\{w_j, j \in J\}$ is a set of suitable weights.

The corresponding composite log likelihood is thus

$$\ell_C(\theta) = \log L_C(\theta).$$

The full log likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$$

is a special case of a composite likelihood. Generally, two classes of composite likelihoods are distinguished; the subsetting methods and the omission methods. In the following, we will concentrate on the subsetting methods, which are based on taking for example only

pairs of observations into account (Cox and Reid (2004)). The omission methods, however, are constructed by omitting components of the full likelihood.

Under usual regularity conditions, the estimating equation

$$\frac{\partial \ell_C(\hat{\theta}_C)}{\partial \theta} = 0$$

is unbiased as each component of $L_C(\theta; y)$ is a likelihood object. In addition, the maximum composite likelihood estimator $\hat{\theta}_C$ is consistent and asymptotically normal with mean θ^0 and covariance matrix $H(\theta^0)^{-1}J(\theta^0)H(\theta^0)^{-1}$, which is the inverse of the sandwich information matrix, where

$$\begin{aligned} H(\theta^0) &= \mathrm{E}_f\{\nabla^2 \ell_C(\theta^0; Y)\}, \\ J(\theta^0) &= \mathrm{var}_f\{\nabla \ell_C(\theta^0; Y)\}. \end{aligned}$$

In a more compact form, we can write

$$\hat{\theta}_C \dot{\sim} N(\theta^0, H(\theta^0)^{-1}J(\theta^0)H(\theta^0)^{-1}).$$

If the additional condition that $k^{-1}J$ tends to a finite non-zero limit as $k \to \infty$ holds, the covariance matrix can be estimated consistently by

$$\mathcal{R} = \hat{H}(\hat{\theta}_C)^{-1}\hat{J}(\hat{\theta}_C)\hat{H}(\hat{\theta}_C)^{-1},$$

where $\hat{H}(\hat{\theta}_C)$ and $\hat{J}(\hat{\theta}_C)$ are estimators of the matrices $H(\theta^0)$ and $J(\theta^0)$, respectively. Under usual regularity conditions, a consistent estimator for $H(\theta^0)$ is obtained by taking

$$\hat{H}(\hat{\theta}_C) = \nabla^2 \ell_C(\hat{\theta}_C).$$

However, finding an appropriate estimator for $J(\theta^0)$ is much harder. Difficulties arise as the naive estimator

$$\hat{J}(\theta) = \nabla \ell_C(\theta; Y)\nabla \ell_C(\theta; Y)^T$$

vanishes at $\theta = \hat{\theta}_C(Y)$ as $\nabla \ell_C(\hat{\theta}_C; Y) = \mathbf{0}$.

Therefore a different estimator for $J(\theta^0)$ is considered. As θ^0 satisfies the estimating equation we have that

$$\begin{aligned} J(\theta^0) &= \mathrm{var}_f\{\nabla \ell_C(\theta^0; Y)\} \\ &= \sum_{i=1}^{n}\sum_{j \in J} \mathrm{var}_f\{\nabla \ell_C(\theta^0; y_i \in \mathcal{A}_j)\} \\ &= \sum_{i=1}^{n}\sum_{j \in J} \mathrm{E}\{\nabla \ell_C(\theta^0; y_i \in \mathcal{A}_j)\nabla \ell_C(\theta^0; y_i \in \mathcal{A}_j)^T\}. \end{aligned}$$

Therefore its empirical counterpart can be used to estimate $J(\theta^0)$. More explicitly, this means

$$\hat{J}(\hat{\theta}_C) = \sum_{i=1}^{n}\left[\left\{\sum_{j \in J} \nabla \ell_C(\hat{\theta}_C; y_i \in \mathcal{A}_j)\right\}\left\{\sum_{j \in J} \nabla \ell_C(\hat{\theta}_C; y_i \in \mathcal{A}_j)\right\}^T\right]$$

is an estimator for $J(\theta^0)$.

Chapter 2

Model selection

Once different models are fitted to a random sample $Y_1, \ldots, Y_n$ we are interested in making inference about the goodness of fit and in comparing them. The following approaches are described in Davison (2003).

2.1 Information criteria for full likelihoods

Assume that the random sample comes from an unknown true model $g(y)$ and that a candidate model $f(y;\theta)$ is fitted. The maximum likelihood estimator can be found for this model and is denoted as $\hat{\theta}$.

Moreover, it is possible to choose f based on the Kullback–Leibler divergence which is defined as

$$\mathrm{KL}(f_\theta, g) = \int \log\left(\frac{g(y)}{f(y;\theta)}\right) g(y)dy.$$

The property that $\mathrm{KL}(f, g) \geq 0$ with equality if $f = g$ indicates that f should be chosen such that $\mathrm{KL}(f_\theta, g)$ is minimal, yielding an estimate θ_g. However, there are infinitely many models such that $\mathrm{KL}(f_{\theta_g}, g) = 0$.

If we suppose that in addition the candidate model contains the true one, we must have $f(y;\theta_g) = g(y)$ for some θ_g. In this case, f_θ is called a **correct model**. Any correct model can be distinguished from the true one as g has fewer parameters than f. Among the correct models, we would like to find the one with fewest parameters having an equally good fit.

In order to compare g with f_θ at $\theta = \hat{\theta}$ for a different random sample $Y_1^*, \ldots, Y_n^*$ from g which is independent of $Y_1, \ldots, Y_n$ the expected likelihood ratio statistic can be computed

$$\mathrm{E}_g^*\left[\sum_{j=1}^{n} \log\left(\frac{g(Y_j^*)}{f(Y_j^*;\hat{\theta})}\right)\right] = n\mathrm{KL}(f_{\hat{\theta}}, g) \geq n\mathrm{KL}(f_{\theta_g}, g)$$

In order to remove the dependency on $\hat{\theta}$ of this quantity, its distribution is averaged, leading

to

$$\mathrm{E}_g\left\{\mathrm{E}_g^*\left[\sum_{j=1}^{n}\log\left(\frac{g(Y_j^*)}{f(Y_j^*;\hat{\theta})}\right)\right]\right\} = n\mathrm{E}_g(\mathrm{KL}(f_{\hat{\theta}},g)).$$

Expanding a Taylor series for $\log f(y;\hat{\theta})$ about θ_g yields

$$\begin{aligned}\log f(y;\hat{\theta}) &\doteq \log f(y;\theta_g) + (\hat{\theta}-\theta_g)^T\frac{\partial \log f(y;\theta_g)}{\partial\theta} \\ &\quad +\frac{1}{2}(\hat{\theta}-\theta_g)^T\frac{\partial^2 \log f(y;\theta_g)}{\partial\theta\partial\theta^T}(\hat{\theta}-\theta_g).\end{aligned}$$

Furthermore,

$$\int\frac{\partial \log f(y;\theta_g)}{\partial\theta}g(y)dy = 0$$

as θ_g minimizes $\mathrm{KL}(f_\theta,g)$. The previous two results allow one to compute

$$\begin{aligned}n\mathrm{KL}(f_{\hat{\theta}},g) &= n\int\log\left\{\frac{g(y)}{f(y;\hat{\theta})}\right\}g(y)dy \\ &\doteq n\mathrm{KL}(f_{\theta_g},g) + \frac{1}{2}\mathrm{tr}\{(\hat{\theta}-\theta_g)(\hat{\theta}-\theta_g)^T I_g(\theta_g)\}\end{aligned}$$

The results obtained for misspecified models imply that

$$n\mathrm{E}_g\{\mathrm{KL}(f_{\hat{\theta}},g)\} \doteq n\mathrm{KL}(f_{\theta_g},g) + \frac{1}{2}\mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}, \tag{2.1}$$

where the last term on the right penalizes for the dimension of θ. This trace is equal to p if f_θ is a correct and regular model because in this case $I_g(\theta_g) = K(\theta_g)$.

The aim is now to obtain an estimator for (2.1). For this, two facts are to be considered, first that $\int\log\{g(y)\}g(y)dy$ is constant and can therefore be ignored, second that we can write $\ell(\hat{\theta}) = \ell(\theta_g) + \{\ell(\hat{\theta}) - \ell(\theta_g)\}$. Thus, denoting the likelihood ratio statistic by W, we have

$$\begin{aligned}\mathrm{E}_g\{-\ell(\hat{\theta})\} &= -\mathrm{E}_g\{\ell(\theta_g) + \frac{1}{2}W(\theta_g)\} \\ &\doteq n\mathrm{KL}(f_{\theta_g},g) - n\int\log\{g(y)\}g(y)dy - \frac{1}{2}\mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}\end{aligned}$$

as under the wrong model we have that $\mathrm{E}_g\{W(\theta_g)\} = \mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}$.

Finally, an estimator of (2.1) is given by

$$-\ell(\hat{\theta}) + c,$$

where c is an estimator of $\mathrm{tr}(I_g(\theta_g)^{-1}K(\theta_g))$. In fact, there are two possible choices for c, either take it to be p, the model is supposed to be correct, or $\mathrm{tr}(J(\hat{\theta})^{-1}K(\hat{\theta}))$ where the model is misspecified.

This yields therefore directly two information criteria

$$\begin{aligned} \text{AIC} &= 2\{-\ell(\hat{\theta}) + p\} \\ \text{NIC} &= 2\left[-\ell(\hat{\theta}) + \text{tr}\{J(\hat{\theta})^{-1}K(\hat{\theta})\}\right], \end{aligned}$$

where the factor 2 is introduced in order to have the same scale as the likelihood ratio statistic. Model selection can then be based on choosing the one with minimal AIC, respectively NIC.

It is important to notice that model selection using AIC or NIC might be inconsistent. Consider the example where the true and a correct model are fitted yielding $\ell(\theta_g)$ and $\ell(\hat{\theta})$ with a number of parameters q and p respectively ($q < p$). The correct model f_θ will be preferred to g if $\ell(\theta_g) - q < \ell(\hat{\theta}) - p$. If g is nested within f_θ, the likelihood ratio statistic implies that

$$\text{pr}\{\ell(\theta_g) - q < \ell(\hat{\theta}) - p\} \doteq \text{pr}\{\chi^2_{p-q} > 2(p-q)\}.$$

In this case the model selection using AIC or NIC is not consistent as

$$\text{pr}(\text{true model is selected}) \nrightarrow 1 \quad \text{as } n \to \infty.$$

If the difference of penalty is however between $\mathcal{O}(1)$ and $\mathcal{O}(n)$ the criteria suggested above are consistent.

2.2 Information criterion for composite likelihoods

Based on composite likelihood methods, Varin and Vidoni (2005) suggest a new information criterion. The composite Kullback–Leibler divergence is their starting point to develop this information criterion.

For a random variable $Y = (Y_1, \ldots, Y_n)$ having density $f(y)$, the composite Kullback–Leibler divergence of a density $g(y)$ relative to $f(y)$ is defined as

$$\begin{aligned} \text{KL}_C(g, f) &= \text{E}_{g(y)}\left[\log\left\{\frac{L_C(g)}{L_C(f)}\right\}\right] \\ &= \sum_{i \in I} \text{E}_{g(y)}\{\log g(Y \in \mathcal{A}_i) - \log f(Y \in \mathcal{A}_i)\} w_i, \end{aligned}$$

where $L_C(f) = \prod_{i \in I} f(Y \in \mathcal{A}_i)^{w_i}$ and $L_C(g) = \prod_{i \in I} g(Y \in \mathcal{A}_i)^{w_i}$.

Notice that the composite Kullback–Leibler divergence is a linear combination of the ordinary Kullback–Leibler divergences.

For a given sample $Y = (Y_1, \ldots, Y_n)$, the objective is to find a model which predicts best another random variable Z having the same distribution. Let therefore $\{f(y;\theta), y \in \mathcal{Y}, \theta \in \Theta\}$ be a parametric statistical model specified by this family of density functions. The estimated density $\hat{f}(y) = f(y;\hat{\theta}_C)$ under this assumed model can then be used for model selection. The following information criterion, based on the expected composite

Kullback–Leibler information between the true density $g(y)$ and $\hat{f}(y)$ selects the model which minimizes

$$\mathrm{E}_{g(y)}\{\mathrm{KL}_C(g, \hat{f})\},$$

or equivalently, which maximizes

$$\varphi(g, f) = \sum_{i \in I} \mathrm{E}_{g(y)} \left[\mathrm{E}_{g(z)}\{\log f(Z \in (A)_i; \hat{\theta}_C)\} \right] w_i. \tag{2.2}$$

This selection criterion is based on the knowledge of the true density $g(y)$ which is unfortunately not often the case in practice. Therefore, we should rather maximize a statistic $\hat{\varphi}(g, f)$ which is an estimator of $\varphi(g, f)$ based on the sample Y. The counterpart of (2.2) is

$$\ell_C(\hat{\theta}_C) = \log L_C(\hat{\theta}_C) = \sum_{i \in I} \log f(Y \in \mathcal{A}_i; \hat{\theta}_C) w_i$$

and yields therefore an estimator.

Varin and Vidoni (2005) use the standard likelihood theory under misspecification for the composite likelihood. Under usual regularity assumptions, they have shown that the estimator above is biased and have introduced a modification in order to correct the first-order bias. They suggest the composite likelihood information criterion under the form stated below. If θ_g is a pseudo-true parameter value that minimizes the composite Kullback–Leibler divergence and Y a random sample, let $g(y) = f(y; \theta^0)$ be the correct model. For $J(\theta_g)$ and $H(\theta_g)$ which are defined as follows

$$J(\theta_g) = \mathrm{var}_{g(y)}\{\nabla \ell_C(\theta_g; Y)\}, \quad H(\theta_g) = \mathrm{E}_{g(y)}\{\nabla^2 \ell_C(\theta_g; Y)\},$$

the composite likelihood information criterion selects the model minimizing

$$-2 \left[\ell_C(\hat{\theta}_C) + \mathrm{tr}\{\hat{J}(\hat{\theta}_C) \hat{H}(\hat{\theta}_C)^{-1}\} \right], \tag{2.3}$$

where $\hat{J}(\hat{\theta}_C)$ and $\hat{H}(\hat{\theta}_C)$ are consistent and first-order unbiased estimators of $J(\theta_g)$, respectively $H(\theta_g)$.

Chapter 3

Spatial prediction

In this chapter the problem of predicting unobserved signals based on observed data is discussed. In geostatistics, the standard approach to obtain a predictor at ungauged locations is called kriging (see Diggle and Ribeiro (2007), Schabenberger and Gotway (2005), Wackernagel (2003) and Cressie (1993)). This approach is based on finding the minimum mean square error predictor which is equivalent to basing the predictor on the full conditional density. However, in more complex models, for example if the underlying process is a max-stable process, the full likelihood is not available. Only lower order joint densities can be obtained.

It is thus interesting to develop a predictor for $Z(x)$ using pairwise likelihood and apply it to processes for which the full likelihood is not available. In order to study the behavior of such a predictor, the Gaussian case is first considered as the full likelihood is known and the predictors can be compared directly through the mean square prediction error.

3.1 Ordinary kriging in the Gaussian case

Before discussing the kriging estimator, recall the definition of a stationary isotropic Gaussian process.

Definition 2. A stationary isotropic Gaussian process can be defined as a stochastic process $Z(x)$, where $x \in \mathcal{X}$ and $\mathcal{X} \subseteq \mathbb{R}^d$ and the joint distributions of $(Z(x_1), Z(x_2), \ldots)$ are multivariate normal with

$$\begin{aligned} &\mathrm{E}[Z(x_1)] = \mathrm{E}[Z(x_2)] = \cdots = \mu, \\ &\mathrm{cov}[Z(x_i), Z(x_j)] = \gamma(||x_i - x_j||), \end{aligned}$$

where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\gamma(\cdot)$ is a covariance function.

Consider now the problem of predicting the value of the random variable Z at an ungauged location $x \in \mathcal{X}$, $Z(x)$, based on observed data $Z(x_1), \ldots, Z(x_n)$ which are supposed to be generated by the Gaussian model defined above. In the following the covariance matrix of the process $\mathbf{Z} = \{Z(x_1), \ldots, Z(x_n)\}$ will be denoted as $V = \sigma^2 R$ where

$\sigma^2 = \text{var}\{Z(x_i)\}$ and $R_{ij} = \rho(||x_i - x_j||)$ corresponds to the correlation between $Z(x_i)$ and $Z(x_j)$.

The predictor $\hat{Z}(x)$ of $Z(x)$ which is of interest is that minimizing the mean square prediction error of $\hat{Z}(x)$

$$\text{MSE}\left\{\hat{Z}(x)\right\} = \text{E}\left[\{\hat{Z}(x) - Z(x)\}^2\right], \tag{3.1}$$

where the expectation is with respect to the joint distribution of $Z(x)$ and $\hat{Z}(x)$ or, equivalently, the joint distribution of $Z(x)$ and $Z(x_1), \ldots, Z(x_n)$. In prediction, the unconditional mean square error defined in (3.1) is distinguished from the conditional mean square error;

$$\text{MSE}\left\{\hat{Z}(x)\right\} = \text{E}\left[\{\hat{Z}(x) - Z(x)\}^2 | Z(x_1) = z(x_1), \ldots, Z(x_n) = z(x_n)\right].$$

In the following, mean square prediction error refers to the one defined in (3.1).

The following theorem gives the minimum mean square predictor.

Theorem 1. **(Diggle and Ribeiro (2007), page 135).** *The mean square prediction error of $\hat{Z}(x)$, MSE$\left\{\hat{Z}(x)\right\}$, is minimized at*

$$\hat{Z}(x) = \text{E}\left[Z(x) | \{Z(x_1), \ldots, Z(x_n)\}\right].$$

As in the Gaussian setting the multivariate conditional distribution is also multivariate normal, the minimum mean square error predictor $\hat{Z}(x)$ and its variance can be obtained immediately. More formally, the following well-known result applies.

Theorem 2. **(Diggle and Ribeiro (2007), page 136).** *If $X = (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$ are jointly multivariate normal, with $\mu = (\mu_1, \mu_2)$ and*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

then the conditional distribution of X_1 given X_2 is also multivariate normal

$$X_1 | X_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Denoting $\mathbf{r} = (r_1, \ldots, r_n)^T$ where $r_i = \rho(||x - x_i||)$, $i = 1, \ldots, n$, the minimum mean square error predictor $\hat{Z}(x)$ is readily obtained when applying the theorem to the joint distribution of $X_1 = Z(x)$ and $X_2 = \mathbf{Z}$ with mean vector $\mu\mathbf{1}$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}^T \\ \sigma^2\mathbf{r} & \sigma^2 R \end{bmatrix}.$$

This yields the minimum mean square error predictor for $Z(x)$

$$\hat{Z}(x) = \mu + \mathbf{r}^T R^{-1}\{\mathbf{Z} - \mu\mathbf{1}\}. \tag{3.2}$$

Notice that the predictor obtained in (3.2) is also obtained considering a full likelihood approach. The corresponding prediction variance is equal to the mean square prediction error in this particular setting

$$\text{MSE}\left\{\hat{Z}(x)\right\} = \text{var}\left\{\hat{Z}(x)\right\} = \sigma^2(1 - \mathbf{r}^T R^{-1}\mathbf{r}). \tag{3.3}$$

3.2 Extension to k observations

The kriging concept can be extended to datasets $Z_j(x_1), \ldots, Z_j(x_n)$, $j = 1, \ldots, k$, where k observations at each location are available.

Here a further assumption needs to be made. The observation to be predicted is supposed to be, say, $Z_1(x)$ and is therefore only correlated with the first observations at the other locations. The minimum mean square error predictor can thus be found using a cokriging approach (Cressie (1993), Section 3.2.3).

The estimator is thus a linear combination of all available observations and can be written as

$$\hat{Z}_1(x) = \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} Z_j(x_i).$$

This estimator is required to be uniformly unbiased, $\mathrm{E}\{\hat{Z}_1(x)\} = \mu$, yielding the condition

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} = 1.$$

The minimisation problem to be solved now can thus be written as

$$\begin{aligned} \min \quad & \mathrm{E}\left[\left\{ Z_1(x) - \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} Z_j(x_i) \right\}^2\right] \\ \text{such that} \quad & \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} = 1. \end{aligned}$$

Let m be the Lagrange multiplier of the problem and rewrite it as

$$\mathrm{var}\left\{ Z_1(x) - \hat{Z}_1(x) \right\} + \left[\mathrm{E}\left\{ Z_1(x) - \hat{Z}_1(x) \right\} \right]^2 - m \left(\sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} - 1 \right). \tag{3.4}$$

Notice that the second term will vanish as the estimator is chosen to be uniformly unbiased. Moreover, the first term can be expanded and, as $\mathbf{Z_j}$ is not correlated with $\mathbf{Z_p}$ for $j \neq p$, is equal to

$$\begin{aligned} & \mathrm{var}\{Z_1(x)\} + \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji}^2 \mathrm{var}\{Z_j(x_i)\} \\ & + 2 \sum_{i<l} \sum_{j=1}^{k} \lambda_{ji} \lambda_{jl} \mathrm{cov}\{Z_j(x_i), Z_j(x_l)\} - 2 \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} \mathrm{cov}\{Z_1(x), Z_j(x_i)\} \\ = \quad & \sigma^2 + \sigma^2 \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji}^2 + 2\sigma^2 \sum_{i<l} \sum_{j=1}^{k} \lambda_{ji} \lambda_{jl} R_{il} - 2\sigma^2 \sum_{i=1}^{n} \lambda_{1i} r_i \end{aligned}$$

Differentiating (3.4) with respect to m, λ_{ij} for $i = 1, \dots, n$ and $j = 1, \dots k$ in order to find the optimal values yields the following system of equations

$$2\sigma^2 \sum_{l=1}^{n} \lambda_{1l} R_{il} - 2\sigma^2 r_i - m = 0 \quad i = 1, \dots, n \tag{3.5}$$

$$2\sigma^2 \sum_{l=1}^{n} \lambda_{jl} R_{il} - m = 0 \qquad j = 2, \dots, n, i = 1, \dots, n. \tag{3.6}$$

In this case the mean square prediction error equals

$$\text{MSE}\left\{\hat{Z}_1(x)\right\} = \sigma^2 - \sigma^2 \sum_{i=1}^{n} \lambda_{1i} r_i + m. \tag{3.7}$$

Intuitively, as $\mathbf{Z}_j$ is not correlated with $\mathbf{Z}_p$ for $j \neq p$, one would assume that $\lambda_{ji} = 0$ for $j = 2, \dots, n$, $i = 1, \dots, n$ and thus that this predictor is equivalent to the one obtained in (3.2). However, under these assumptions, denoting $\lambda = (\lambda_{11}, \dots, \lambda_{1n})$, the optimal value of λ would equal

$$\lambda = \mathbf{r}^T R^{-1},$$

where $\mathbf{r} = (r_1, \dots, r_n)^T$ with $r_i = \rho(||x - x_i||)$, $i = 1, \dots, n$ and $R_{ij} = \rho(||x_i - x_j||)$ corresponds to the correlation between $Z(x_i)$ and $Z(x_j)$. Thus the optimal predictor would be equal to

$$\hat{Z}(x) = \mathbf{r}^T R^{-1} \mathbf{Z}_1.$$

This intuition is however misleading as the condition $\sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ji} = 1$ is not satisfied because $\sum_{i=1}^{n} \mathbf{r}^T R^{-1} \neq 1$. There is thus an influence of the observations $j = 2, \dots, n$ on the predictor $\hat{Z}_1(x)$ for the ungauged location x.

3.3 Kriging in the Gaussian case using pairwise likelihoods

Several problems may arise with the kriging approaches seen previously. It may not always be easy to solve the cokriging system of equations, especially for large datasets. As an alternative, the full likelihood approach yielding the predictor defined in (3.2) may be considered. However, for complicated models the full likelihood may not be known analytically or requires too much computational effort. A different approach to the kriging problem which relies on the knowledge of pairwise densities only is therefore considered.

As before, the observation to be predicted is $Z_1(x)$ and is thus only correlated with the first observations of the other locations. The corresponding pairwise likelihood can be written as

$$L_C\left\{Z_1(x)\right\} = \prod_{j=1}^{k} \prod_{i=1}^{n} f\{Z_1(x) | Z_j(x_i)\}^{w_{ji}},$$

and the corresponding log likelihood as

$$\begin{aligned}\ell_C\{Z_1(x)\} &= \sum_{j=1}^{k}\sum_{i=1}^{n} w_{ji}\log f\{Z_1(x)|Z_j(x_i)\} \\ &= \sum_{i=1}^{n}\left(w_{1i}\left[-\frac{1}{2}\log\{2\pi\sigma^2(1-r_i^2)\} - \frac{1}{2\sigma^2(1-r_i^2)}\left(Z_1(x)-[\mu+r_i\{Z_j(x_i)-\mu\}]\right)^2\right]\right. \\ &\quad \left. + \sum_{j=2}^{k} w_{ji}\left[-\frac{1}{2}\log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\{Z_1(x)-\mu\}^2\right]\right).\end{aligned}$$

The objective is to maximize this likelihood with respect to $Z_1(x)$. The score equation

$$\begin{aligned}\frac{\partial \ell_C\left\{\hat{Z}_C(x)\right\}}{\partial Z_1(x)} &= -\sum_{i=1}^{n}\left[\frac{w_{1i}}{\sigma^2(1-r_i^2)}\left(Z_1(x)-[\mu+r_i\{Z_1(x_i)-\mu\}]\right) + \sum_{j=2}^{k}\frac{w_{ji}}{\sigma^2}\{Z_1(x)-\mu\}\right] \\ &= 0\end{aligned}$$

allows us to obtain the maximiser $\hat{Z}_C(x)$. Solving this equation with respect to $\hat{Z}_C(x)$ gives the following composite likelihood predictor for $Z(x)$

$$\hat{Z}_C(x) = \frac{\sum_{i=1}^{n}\left([\mu+r_i\{Z_1(x_i)-\mu\}]\frac{w_{1i}}{\sigma^2(1-r_i^2)} + \sum_{j=2}^{k}\frac{w_{ji}}{\sigma^2}\mu\right)}{\sum_{i=1}^{n}\left(\frac{w_{1i}}{\sigma^2(1-r_i^2)} + \sum_{j=2}^{k}\frac{w_{ji}}{\sigma^2}\right)}.$$

This predictor only takes into account observation 1. Therefore the weights w_{ji}, $j = 2,\ldots,k$, $i = 1,\ldots,n$ can be set to be equal to 0. Consequently, the change of notation $w_{1i} \equiv w_i$, $i = 1,\ldots,n$ is introduced. This yields the following pairwise predictor

$$\hat{Z}_C(x) = \frac{\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\left[\mu+r_i\{Z_1(x_i)-\mu\}\right]}{\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}}. \tag{3.8}$$

This formula illustrates that if the parameters are known in advance, as for the first part of the simulation study, the additional observations are wasted. However, they may indirectly influence the predictor if the parameters of the Gaussian process are estimated.

The predictor obtained in (3.8) yields a sub-optimal linear predictor for $Z(x)$. As we are interested in finding the mean square prediction error of this predictor, its variance

needs to be calculated;

$$\begin{aligned}
\text{var}\left\{\hat{Z}_C(x)\right\} &= \left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-2}\text{var}\left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\left[\mu+r_i\{Z_1(x_i)-\mu\}\right]\right)\\
&= \left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-2}\left[2\sum_{j<i}\frac{w_jr_j}{1-r_j^2}\frac{w_ir_i}{1-r_i^2}\text{cov}\left\{Z_1(x_j),Z_1(x_i)\right\}\right.\\
&\quad\left.+\sum_{i=1}^{n}\frac{w_i^2r_i^2}{(1-r_i^2)^2}\text{var}\left\{Z_1(x_i)\right\}\right]\\
&= \sigma^2\left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-2}\left[2\sum_{j<i}\frac{w_i}{1-r_i^2}\frac{w_j}{1-r_j^2}r_jr_iR_{ji}\right.\\
&\quad\left.+\sum_{i=1}^{n}\frac{w_i^2r_i^2}{(1-r_i^2)^2}\right].
\end{aligned}$$

The mean square prediction error then equals

$$\begin{aligned}
\text{MSE}\left\{\hat{Z}_C(x)\right\} &= \text{E}\left[\left\{Z(x)-\hat{Z}_C(x)^2\right\}\right]\\
&= \text{var}\left\{Z(x)-\hat{Z}_C(x)\right\}+\text{E}\left\{Z(x)-\hat{Z}_C(x)\right\}^2\\
&= \text{var}\left\{Z(x)\right\}+\text{var}\left\{\hat{Z}_C(x)\right\}-2\text{cov}\left\{\hat{Z}_C(x),Z(x)\right\}\\
&= \sigma^2+\text{var}\left\{\hat{Z}_C(x)\right\}-2\left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-1}\\
&\quad\text{cov}\left\{Z(x),\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\left[\mu+r_i\{Z_1(x_i)-\mu\}\right]\right\}\\
&= \sigma^2+\text{var}\left\{\hat{Z}_C(x)\right\}-2\left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-1}\text{cov}\left\{Z(x),\sum_{i=1}^{n}\frac{w_ir_i}{1-r_i^2}Z_1(x_i)\right\}\\
&= \sigma^2+\text{var}\left\{\hat{Z}_C(x)\right\}-2\sigma^2\left(\sum_{i=1}^{n}\frac{w_i}{1-r_i^2}\right)^{-1}\sum_{i=1}^{n}\frac{w_ir_i^2}{1-r_i^2}
\end{aligned}$$

as the predictor is unbiased. The aim is to find weights such that the mean square prediction error is minimal. Different weights are considered for the simulation study because it is intractable to obtain the optimal weights analytically.

It would be interesting to compare the above mean square prediction error to the mean square prediction errors obtained in equations (3.3) and (3.7). As this is a non-trivial analytical problem, a numerical example is provided here. As described in section 3.4, a Gaussian process with underlying powered exponential covariance function observed at 6 locations as well as at the location to be kriged is simulated. The parameter values are

assumed to be known in order to calculate the predictors and their mean square prediction errors. Applying the different prediction methods, the following averages for the mean square prediction errors over $10,000$ simulations are obtained; the one corresponding to the best linear predictor based on one observation at each location (see equation (3.3)) equals 0.239, whereas the one based on all 25 observations (see equation (3.7)) is equal to 0.240. Finally the mean square prediction error for the pairwise likelihood approach described above equals 0.309. These results indicate that if all observations are taken into account, there is only little improvement in the mean square prediction error comparing to the full likelihood approach. However, there is the expected loss in precision if the pairwise predictor is considered.

3.4 Simulation study

The objective of this chapter is to investigate the behavior of the pairwise kriging predictor and compare it to the optimal kriging predictor. For the simulation study, let us consider three different covariance functions, the exponential, the Whittle-Matérn and the Cauchy covariance function. This allows to discover differences in the performance of the predictors due to the underlying covariance function.

Recall first the definition of the powered exponential covariance function

$$\gamma(h;\beta,\sigma^2) = \sigma^2 \exp\left\{\left(-\frac{h}{\beta}\right)^{\nu}\right\}, \quad h>0,$$

which depends on the parameters $\sigma^2 > 0$, $\beta > 0$, $\nu > 0$ and on the euclidean distance h between two positions $Z(x_i)$ and $Z(x_j)$. In this case, the parameter β determines how long the dependence range is.

Another possible choice is the Whittle-Matérn covariance function which is given by

$$\gamma(h;\sigma^2,\beta,\nu) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{h}{\beta}\right)^{\nu} K_\nu\left(\frac{h}{\beta}\right), \quad h>0,$$

where $K_\nu(\cdot)$ is the modified Bessel function of third kind of order $\nu > 0$ and $\Gamma(\cdot)$ is the gamma function.

Moreover, the Cauchy covariance function is given by

$$\gamma(h;\sigma^2,\beta,\nu) = \sigma^2\left\{1+\left(\frac{h}{\beta}\right)^2\right\}^{-\nu}, \quad h>0,$$

where $\beta > 0$ and $\nu > 0$ are the range and the smooth parameters, respectively.

The parameter values of μ and σ^2 are fixed to be equal to 0 and 1, respectively. Moreover, assume that there are 6 or 15 locations which are either uniformly distributed on $[0, 15]$ or form a regular grid on the same interval. Figure 3.1 shows the different parameter settings for covariance functions that are considered for the simulation studies. The code was implemented in `R` and, in order to reduce the computation time, the computation of

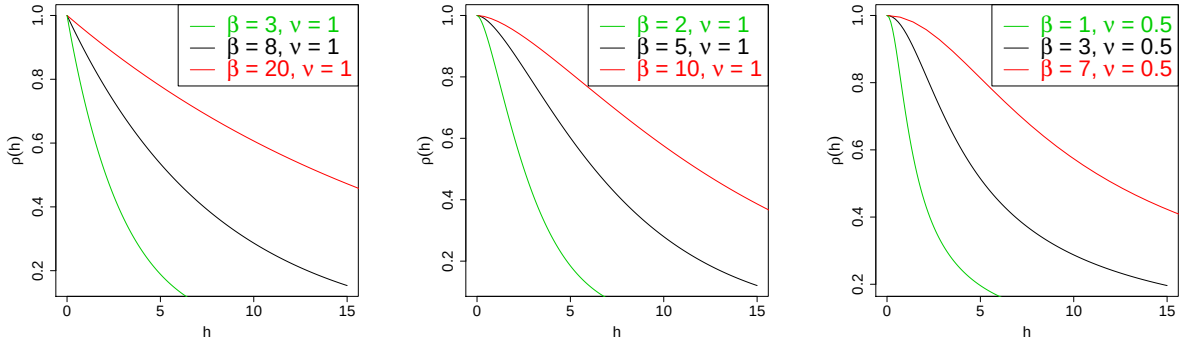

Figure 3.1: From left to right: Powered exponential, Whittle-Matérn and Cauchy covariance functions for different dependence configurations.

the pairwise likelihood, its gradient as well as the pairwise predictor and its mean square prediction error were implemented in `C`.

As already mentioned earlier, it is not obvious how to choose the weights for the pairwise predictors in order to minimize the mean square prediction error. Different choices of weights, all normalized such that they sum up to one, are therefore considered for the construction of the pairwise kriging predictor $\hat{Z}_C(x)$. The first choice of weights is the most basic one, setting

$$w_i = \frac{1}{n}, \quad i = 1, \dots, n, \tag{3.9}$$

and yielding the corresponding predictor denoted $\hat{Z}_{C_1}(x)$. A second pairwise predictor, $\hat{Z}_{C_2}(x)$, is obtained by taking weights that are monotonic decreasing in $||x_i - x||$;

$$w_i \propto \begin{cases} 1/||x_i - x|| & x_i \neq x \\ 1 & x_i = x \end{cases} \tag{3.10}$$

for $i = 1, \dots, n$, where $||\cdot||$ denotes the euclidean distance between the two locations. This choice seems plausible as stations that are located closer to the one being kriged should be given more weight than stations that are further apart from x. A third predictor, $\hat{Z}_{C_3}(x)$, can be computed considering a normalized version of the weights given by the formula

$$w_i = \begin{cases} 1 & \text{if } ||x_i - x|| \leq \frac{1}{n}\sum_{i=1}^{n} ||x_i - x|| \\ 0 & \text{if } ||x_i - x|| > \frac{1}{n}\sum_{i=1}^{n} ||x_i - x||. \end{cases} \tag{3.11}$$

In the following sections, different types of simulations are performed. In a first part, the predictors are calculated under the assumption that the true parameters of the Gaussian process are known. Later, they will be estimated and the effect of the estimation on the kriging predictor analysed.

3.4.1 Prediction of an observation at an ungauged location

This section investigates the simulation output for Gaussian processes where the underlying correlation structure as well as the corresponding parameters are supposed to be known.

Consider first the particular parameter setting where $n = 6$ locations, $x_1, \ldots, x_6$, are uniformly but randomly distributed on the interval $[0, 15]$. The location x to be interpolated is uniformly, randomly generated at each simulation on the same interval. Suppose moreover that the underlying correlation structure is powered exponential with parameter settings corresponding to short, medium and long range dependence on the interval $[0, 15]$ (Figure 3.1).

For each of the $10,000$ simulations, $k = 25$ observations of the Gaussian process are reported at each location including location x. Based on the observations at $x_1, \ldots, x_6$ it is possible to compute predictors at x. As discussed in chapter 3, different predictors are to be compared to the best linear predictor, $\hat{Z}(x)$, which can be obtained by solving the system of equations (3.5)–(3.6). One alternative predictor, denoted $\hat{Z}_\ell(x)$, is based on the full conditional density or, equivalently, on a full likelihood approach as defined in (3.2). In the particular case where the process is only observed once at each location, this predictor is equal to $\hat{Z}(x)$. Finally, the pairwise likelihood predictor as defined in (3.8) is calculated. For the simulation study, three different pairwise predictors, $\hat{Z}_{C_1}(x)$, $\hat{Z}_{C_2}(x)$ and $\hat{Z}_{C_3}(x)$, corresponding to the weight functions (3.9)–(3.11) are considered.

Figure 3.2 shows the boxplots of the true value of $Z(x)$ as well as the different predictors for the different strengths of correlation. It is immediate that if the observations are only weakly correlated, the predictions tend to be close to the mean of the process whereas the true values may not be. As the observations get more strongly correlated, the shapes of the boxplots become more similar, thus the distribution of the predictors is closer to the distribution of the true values observed at x.

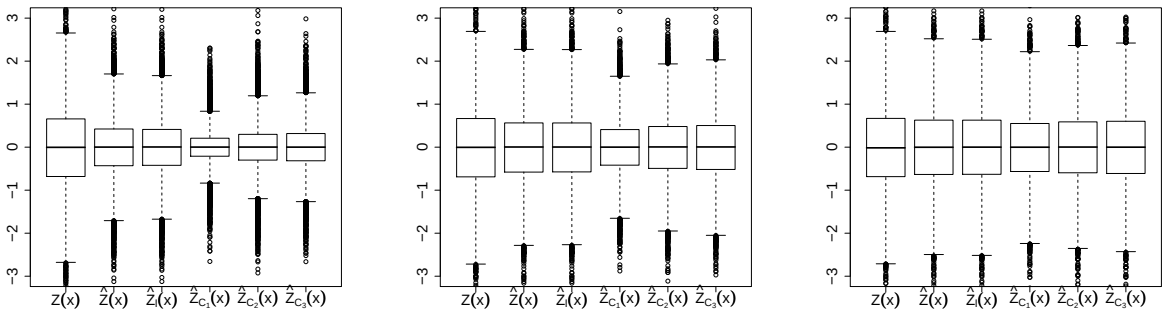

Figure 3.2: Boxplots of true values and predictors for $Z(x)$ based on 10,000 simulations of 25 observations at 6 locations which are uniformly distributed on $[0, 15]$. From left to right: powered exponential covariance function with short, medium and long range dependences.

In addition, there are differences in the behavior of the prediction methods. It is most obvious that the pairwise predictor $\hat{Z}_{C_1}(x)$ has always the smallest interquantile range.

The predicted values seem too much shrunk to the mean of the process. The other two pairwise predictors are similarly distributed, but slightly more spread, and get closer to the distribution of the true value as the dependence range increases. Similarly, there seems to be no distributional difference between the best linear predictor and the full likelihood predictor. There seems to be little improvement in the prediction when all observations are considered for the construction of the predictor.

For a more detailed analysis of the predictors, the performance of the predictors may be compared to the best linear predictor through the analysis of the ratio of the mean square prediction errors

$$\text{RMSE} = \frac{\text{MSE}\left\{\hat{Z}(x)\right\}}{\text{MSE}[p\{Z_j(x_i) : i = 1, \ldots, n, j = 1, \ldots, k\}]},$$

where $p\{\cdot\}$ stands for a specific predictor. The mean square prediction error for each predictor are calculated in two ways, either using the formulae obtained in sections 3.1, 3.2 and 3.3 or the empirical mean square prediction error $[Z(x) - p\{Z(x)\}]^2$. Tables 3.1–3.3 show the simulation results, where the gauged locations are also considered to be regularly spaced on $[0, 15]$. In addition, predictors are also calculated if the underlying covariance function is Cauchy or Whittle-Matérn.

In general, the empirical and the theoretical mean square error are equivalent, reflecting that using $10,000$ simulations is adequate. In addition, the average predictors are nearly always equal for the different approaches. However, there is a persistent difference in the minimum mean square prediction error between the two grids. The minimum mean square prediction error is greater if the locations are uniformly distributed.

For all correlation functions and strengths of correlation the full likelihood approach is equivalent to the best linear predictor. This shows that the influence of observations with which $Z(x)$ is not correlated is negligible. This was also confirmed when performing simulations with 50 observations per location.

Table 3.1: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 10,000 simulations of 25 observations at 6 locations. Short range dependence.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.01	0.00	0.01	0.00	0.00	0.00
$\hat{Z}(x)$	-0.01 [0.31, 0.30]	-0.01 [0.16, 0.16]	0.00 [0.28, 0.27]	0.00 [0.45, 0.45]	0.00 [0.33, 0.33]	0.00 [0.43, 0.43]
$\hat{Z}_\ell(x)$	-0.01 (1.0, 1.0)	-0.01 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)
$\hat{Z}_{C_1}(x)$	0.00 (1.7, 1.6)	0.00 (2.0, 1.9)	0.00 (1.7, 1.6)	0.00 (1.5, 1.3)	0.00 (1.7, 1.3)	0.00 (2.0, 1.3)
$\hat{Z}_{C_2}(x)$	0.00 (1.1, 1.2)	0.00 (1.3, 1.3)	0.00 (1.1, 1.2)	0.00 (1.1, 1.1)	0.00 (1.3, 1.2)	0.00 (1.5, 1.1)
$\hat{Z}_{C_3}(x)$	0.00 (1.3, 1.3)	0.00 (1.4, 1.4)	0.00 (1.3, 1.3)	0.00 (1.2, 1.1)	0.00 (1.4, 1.2)	0.00 (1.6, 1.1)

If the dependence range is small (Table 3.1), the inflation of the mean square error ratio is considerably low when considering composite likelihoods. Especially the second and the third composite likelihood approaches yield very reasonable results.

Table 3.2: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 10,000 simulations of 25 observations at 6 locations. Medium range dependence.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{Z}(x)$	0.00 [0.13, 0.13]	0.00 [0.04, 0.04]	0.00 [0.04, 0.04]	0.00 [0.24, 0.24]	0.00 [0.13, 0.13]	0.00 [0.12, 0.13]
$\hat{Z}_\ell(x)$	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)
$\hat{Z}_{C_1}(x)$	0.00 (1.7, 1.7)	0.00 (2.2, 2.2)	0.00 (3.4, 2.7)	0.00 (1.5, 1.3)	0.00 (1.9, 1.5)	0.00 (2.0, 1.8)
$\hat{Z}_{C_2}(x)$	0.00 (1.1, 1.1)	0.00 (1.5, 1.4)	0.00 (2.3, 1.7)	0.00 (1.1, 1.1)	0.00 (1.6, 1.3)	0.00 (1.7, 1.7)
$\hat{Z}_{C_3}(x)$	0.00 (1.2, 1.3)	0.00 (1.5, 1.6)	0.00 (2.4, 1.8)	0.00 (1.2, 1.1)	0.00 (1.6, 1.3)	0.00 (1.8, 1.7)

Table 3.3: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 10,000 simulations of 25 observations at 6 locations. Long range dependence.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.00	0.00	0.00	-0.01	-0.01	-0.01
$\hat{Z}(x)$	0.00 [0.05, 0.05]	0.00 [0.01, 0.01]	0.00 [0.002, 0.002]	0.00 [0.11, 0.12]	0.00 [0.05, 0.05]	0.00 [0.01, 0.01]
$\hat{Z}_\ell(x)$	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)
$\hat{Z}_{C_1}(x)$	0.00 (1.6, 1.6)	0.00 (2.1, 2.3)	0.00 (42.6, 7.7)	0.00 (1.4, 1.3)	0.00 (2.1, 1.8)	0.00 (51.1, 6.6)
$\hat{Z}_{C_2}(x)$	0.00 (1.1, 1.1)	0.00 (1.6, 1.6)	0.00 (37.3, 5.0)	0.00 (1.1, 1.1)	0.00 (1.8, 1.6)	0.00 (44.9, 5.9)
$\hat{Z}_{C_3}(x)$	0.00 (1.2, 1.2)	0.00 (1.6, 1.7)	0.00 (34.0, 5.7)	0.00 (1.2, 1.1)	0.00 (1.9, 1.5)	0.00 (45.3, 5.7)

Considering the medium correlation configuration (Table 3.2), the most striking observation is the decrease of the mean square prediction error of the optimal predictor. The ratios of the mean square prediction errors have nearly the same values as before meaning that the pairwise predictor improves the same way as the best linear predictors if the observations are more correlated. The only exception is when the underlying covariance function is Cauchy and the grid is regular.

Finally, when the observation are strongly correlated the optimal predictor becomes even more precise in terms of mean square prediction error (Table 3.3). In particular, with an underlying Cauchy covariance function the mean square error is very small and thus the ratios of mean square errors are much more inflated for the composite predictors. For the other covariance functions, the ratio of mean square prediction errors remain again nearly the same as for the medium covariance setting.

Increasing the number of sites on $[0, 15]$ to 15 for the medium range dependence (Table 3.4) yields a significant improvement of the mean square prediction error compared to the ones obtained for 6 sites shown in Table 3.2. Due to particularly small minimum mean square prediction errors in the case of an underlying Cauchy covariance function the theoretical mean square prediction error ratios are very large.

Moreover, it is interesting to observe that the inflation of the mean square error due to the use of pairwise predictors is greater if there are 15 locations except for $\hat{Z}_{C_2}$. This suggests that the other two pairwise predictors are not able to use the additional information as efficiently as the best linear predictor does.

The analyses for the short and long range dependences are not provided here as the results are similar, but scaled.

Table 3.4: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 10,000 simulations of 25 observations at 15 locations. Medium range dependence.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.00	0.00	0.00	0.00	-0.01	-0.01
$\hat{Z}(x)$	0.00 [0.05, 0.05]	0.00 [0.01, 0.01]	0.00 [0.001, 0.001]	-0.01 [0.08, 0.08]	-0.01 [0.01, 0.01]	-0.01 [0.001, 0.001]
$\hat{Z}_\ell(x)$	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	0.00 (1.0, 1.0)	-0.01 (1.0, 1.0)	-0.01 (1.0, 1.0)	-0.01 (1.0, 1.0)
$\hat{Z}_{C_1}(x)$	0.00 (3.0, 3.1)	0.00 (4.0, 4.7)	0.00 ($> 10^2$, 29)	0.00 (2.4, 2.3)	0.00 (3.6, 4.0)	0.00 ($> 10^2$, 64)
$\hat{Z}_{C_2}(x)$	0.00 (1.2, 1.2)	0.00 (1.8, 1.8)	0.00 ($> 10^2$, 9)	0.00 (1.2, 1.3)	0.00 (2.2, 2.3)	0.00 ($> 10^2$, 35)
$\hat{Z}_{C_3}(x)$	0.00 (1.9, 2.0)	0.00 (2.5, 2.9)	0.00 ($> 10^2$, 17)	0.00 (1.6, 1.6)	0.00 (2.5, 2.6)	0.00 ($> 10^2$, 41)

The previous results show that the first choice of weights is clearly the most inappropriate. The other two weight functions yield nearly equivalent predictors if there are only 6 locations. However, increasing the number of sites shows that the second weight function is yielding best results.

In order to try improving the third pairwise likelihood predictor, consider a more elaborated version of the third weight function related to the approach suggested by Bevilacqua *et al.* (2008) and Heyde (1997). In their approach the cut-off weights are defined as

$$w_i = \begin{cases} 1 & \text{if } ||x_i - x|| \leq \delta \\ 0 & \text{if } ||x_i - x|| > \delta, \end{cases}$$

such that δ is the value minimizing the mean square prediction error. Figure 3.3 shows how the mean square prediction error evolves as a function of δ for a fixed number of observations, replications and sites.

These plots show that there is nearly no change in the mean square prediction error as δ increases. There seems to be no point in considering different values of δ in order to choose the weights such that the mean square error is minimal.

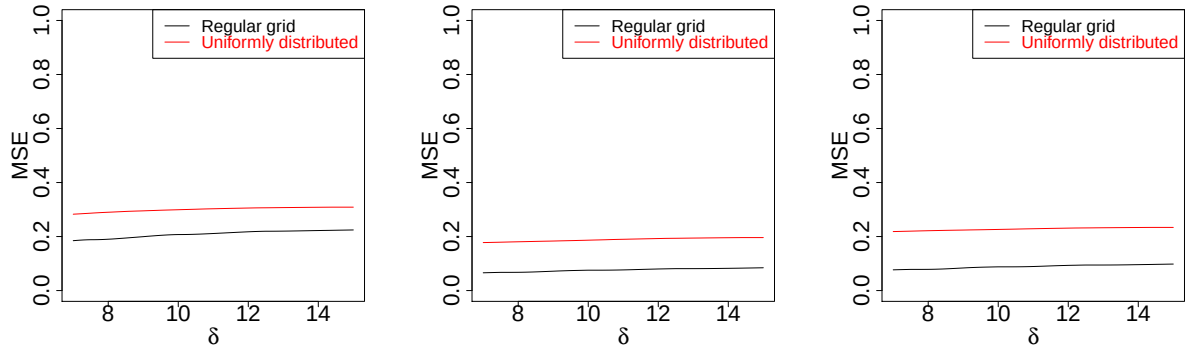

Figure 3.3: Mean square prediction errors computed for different values of δ for simulated samples based on 25 observations, 6 locations and 10,000 simulations. The parameters of the covariance functions are assumed to be known. From left to right: Exponential, Whittle-Matérn and Cauchy covariance functions.

3.4.2 Influence of parameter estimation on the prediction

This section investigates the situation where the parameters of the Gaussian process need to be estimated first. This can either be done by full likelihood estimation or by pairwise likelihood estimation where all the pairs are given a common constant weight. The number of times the process is observed at each location is likely to play a more important role as it improves the quality of the parameter estimates. Based on these estimates, it is then possible to establish predictors for ungauged locations. Hence, there are two potential sources of errors, the estimation error and the prediction error.

Consider first a process with underlying powered exponential covariance function. Only the results for the medium correlation are shown as parameter estimation for Gaussian processes by means of pairwise likelihood was already investigated by Grünenfelder (2009). Table 3.5 indicates the maximum likelihood and the maximum pairwise likelihood estimates as well as the corresponding normalized standard errors for the parameters of the Gaussian process. The standard errors are calculated by means of the inverse of the sandwich information matrix as described in section 1.1. They are therefore referred to as sandwich estimates. Detailed calculations of the gradients for the pairwise likelihoods can be found in the appendix.

It is interesting to observe in Table 3.5 that the estimates do not get significantly more precise as the number of observations increases. Even more observations would be required to improve the quality of the estimates (Grünenfelder (2009)). In addition, there is no remarkable difference in the estimation due to the distribution of the locations on $[0, 15]$.

Comparing the two methods, maximum likelihood estimation is, as expected, superior to pairwise. Nevertheless, the pairwise approach seems competitive as it yields reasonable estimates.

Consider now the parameter estimates obtained for a process with underlying Whittle-

Table 3.5: Average parameter estimation for powered exponential covariance function with medium range dependence using full and pairwise likelihood based on 1000 simulations of 25 and 50 observations, respectively, at 6 locations. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates. True values: $\mu = 0$, $\sigma^2 = 1$, $\beta = 8$ and $\nu = 1$.

k	grid	likelihood	$\hat{\mu}$ ($\sqrt{k}$·se($\hat{\mu}$))	$\hat{\sigma}^2$ ($\sqrt{k}$·se($\hat{\sigma}^2$))	$\hat{\beta}$ ($\sqrt{k}$·se($\hat{\beta}$))	$\hat{\nu}$ ($\sqrt{k}$·se($\hat{\nu}$))
25	regular	full	0.00 (0.73)	0.98 (0.88)	8.47 (16.0)	1.02 (0.99)
		pairwise	0.00 (0.76)	0.98 (0.90)	9.22 (54.3)	0.99 (2.43)
	uniform	full	0.00 (0.78)	0.98 (0.96)	8.41 (15.7)	1.01 (0.63)
		pairwise	0.00 (0.81)	0.98 (0.94)	8.84 (22.8)	1.01 (0.90)
50	regular	full	0.00 (0.74)	0.99 (0.88)	8.14 (13.8)	1.01 (0.98)
		pairwise	0.00 (0.76)	0.99 (0.90)	8.85 (33.6)	0.97 (1.88)
	uniform	full	0.00 (0.79)	0.98 (0.97)	8.09 (14.3)	1.01 (0.63)
		pairwise	0.00 (0.81)	0.98 (0.98)	8.35 (19.8)	1.00 (0.89)

Matérn covariance function (Table 3.6). It is not possible to compute a standard error for the pairwise estimates as the gradient of the log likelihood is not known analytically.

If there are only 25 observations, the estimation of β is not very precise for neither of the methods when the gauged locations are on a uniform grid. Furthermore, the pairwise likelihood does not yield precise parameter estimates for ν unless there are 50 observations.

A general issue with the Whittle-Mattérn covariance function is the identification problem of ν (Diggle and Ribeiro (2007)). As the considered sample is of size 1000, however, the identification of the parameter ν is good. Nevertheless, for smaller samples it may be preferable to choose the shape parameter ν from several values, in this case for example from the discrete set $\{0.75, 1, 1.25\}$, such that the likelihood is maximal.

Table 3.6: Average parameter estimation for Whittle-Matérn covariance function with medium range dependence using full and pairwise likelihood based on 1000 simulations of 25 and 50 observations, respectively, at 6 locations. True values: $\mu = 0$, $\sigma^2 = 1$, $\beta = 5$ and $\nu = 1$.

k	grid	likelihood	$\hat{\mu}$ ($\sqrt{k}$·se($\hat{\mu}$))	$\hat{\sigma}^2$ ($\sqrt{k}$·se($\hat{\sigma}^2$))	$\hat{\beta}$ ($\sqrt{k}$·se($\hat{\beta}$))	$\hat{\nu}$ ($\sqrt{k}$·se($\hat{\nu}$))
25	regular	full	0.00 (0.79)	0.92 (0.52)	4.97 (9.03)	1.08 (1.66)
		pairwise	0.00 (-)	0.91 (-)	5.14 (-)	1.63 (-)
	uniform	full	-0.01 (0.85)	0.92 (0.55)	5.45 (9.90)	0.99 (0.81)
		pairwise	-0.01 (-)	0.90 (-)	4.51 (-)	1.92 (-)
50	regular	full	0.00 (0.75)	0.94 (0.55)	4.95 (8.48)	1.03 (1.40)
		pairwise	0.00 (-)	0.93 (-)	5.01 (-)	1.19 (-)
	uniform	full	-0.01 (0.81)	0.93 (0.59)	4.82 (5.56)	1.01 (0.61)
		pairwise	-0.01 (-)	0.93 (-)	4.91 (-)	1.06 (-)

Similar analyses can be carried out when investigating the estimates (Table 3.7) for the process with underlying Cauchy covariance function. The parameters μ and σ^2 are well estimated by both the full and the pairwise likelihood approaches even if there are only 25 observations. The shape parameter ν gains precision as the number of observations at each location increases. The same is true for the range parameter β if the full likelihood estimation is considered. In contrast, the estimation of β by pairwise likelihood yields much larger values than the true one.

Table 3.7: Average parameter estimation for Cauchy covariance function with medium range dependence using full and pairwise likelihood based on 1000 simulations of 25 and 50 observations, respectively, at 6 locations. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates. True values: $\mu = 0$, $\sigma^2 = 1$, $\beta = 3$ and $\nu = 0 - 5$.

k	grid	likelihood	$\hat{\mu}$ ($\sqrt{k}$·se($\hat{\mu}$))	$\hat{\sigma}^2$ ($\sqrt{k}$·se($\hat{\sigma}^2$))	$\hat{\beta}$ ($\sqrt{k}$·se($\hat{\beta}$))	$\hat{\nu}$ ($\sqrt{k}$·se($\hat{\nu}$))
25	regular	full	-0.01 (0.73)	0.93 (0.87)	3.31 (5.20)	0.90 (8.16)
		pairwise	-0.01 (0.76)	0.98 (0.85)	7.78 (29.2)	0.68 (43.8)
	uniform	full	0.00 (0.77)	0.98 (0.95)	3.32 (3.30)	0.70 (2.92)
		pairwise	0.00 (0.81)	0.98 (0.95)	8.35 (36.9)	0.53 (74.5)
50	regular	full	0.00 (0.74)	0.95 (0.89)	3.05 (3.92)	0.55 (1.46)
		pairwise	0.00 (0.77)	0.98 (0.90)	7.29 (15.4)	0.53 (13.2)
	uniform	full	0.00 (0.78)	0.99 (0.98)	3.09 (2.66)	0.56 (1.30)
		pairwise	0.00 (0.81)	0.99 (0.99)	7.04 (26.9)	0.51 (28.2)

As the objective is to predict an observation at an ungauged location, the obtained estimates can be used to compute different predictors. The best linear predictor is calculated in two ways, using the estimates obtained through full and pairwise likelihood methods yielding $\hat{Z}_{fe}(x)$ and $\hat{Z}_{pe}(x)$, respectively. Moreover, the full likelihood predictor and the three pairwise predictors introduced previously are again considered. Naturally, the corresponding estimates are used for the calculations. The theoretical mean square prediction error becomes a plug-in mean square prediction error using the estimates.

The results for 25 observations with medium correlation configuration are indicated in Table 3.8. Comparing $\hat{Z}_{fe}(x)$ and $\hat{Z}_{pe}(x)$, there is no significant difference between the best linear predictors obtained through replacing the true values by the different estimates. The pairwise estimates imply however a loss in efficiency.

The full likelihood predictor is equivalent to the best linear predictor. Even if the estimation of the parameters may not be very precise in the pairwise setting, its influence on the predictors is negligible. In contrast, the plug-in mean square prediction errors are very sensitive to the estimated values of parameters. Plugging-in the pairwise estimates turns out to be very inappropriate. The ratios of mean square prediction errors are not reasonable at all for most of the considered settings. Therefore inference should not be based on the pairwise plug-in mean square prediction error.

In this context, the Whittle-Matérn and the Cauchy covariance functions are the ones yielding most precise results in terms of mean square prediction errors. In addition, the

pairwise predictor calculated with the second weight function is most efficient.

Increasing the number of observations to 50 changes the obtained predictors only slightly (Table 3.9). The only remarkable improvement is in the ratios of the empirical mean square prediction errors for uniformly distributed locations where a process with underlying exponential or Whittle-Matérn covariance function is observed.

A comparison of the predictors for a random location of a Gaussian process can now be provided. For all considered correlation structures, the predictors are reasonable even if the estimates of the covariance parameters may not be very close to the true values. However, the plug-in mean square prediction errors are not reliable.

As already discussed in section 3.4.1, the empirical mean square prediction errors for the pairwise predictors are, as expected, inflated. The mean square prediction error is minimal if the weights are inversely proportional to the distance to the kriging location. In addition, the most competitive results in terms of the mean square prediction are obtained using the Whittle-Matérn or the Cauchy covariance function.

Table 3.8: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 1000 simulations of 25 observations at 6 locations. Medium correlation function.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.06	0.05	0.06	0.02	0.02	0.01
$\hat{Z}_{fe}(x)$	0.04 [0.13, 0.13]	0.05 [0.04, 0.04]	0.04 [0.04, 0.04]	0.01 [0.24, 0.24]	0.01 [0.13, 0.13]	0.01 [0.12, 0.12]
$\hat{Z}_{pe}(x)$	0.04 (1.21, 1.05)	0.05 (1.39, 1.03)	0.04 (0.83, 1.23)	0.01 (1.01, 1.01)	0.02 (0.93, 2.85)	0.02 (0.99, 1.76)
$\hat{Z}_{\ell}(x)$	0.04 (1.00, 1.00)	0.05 (1.00, 1.00)	0.04 (1.00, 1.00)	0.01 (1.00, 1.00)	0.01 (1.00, 1.00)	0.01 (1.00, 1.00)
$\hat{Z}_{C_1}(x)$	0.03 (<0, 1.82)	0.04 (<0, 2.25)	0.04 ($>10^2$, 2.62)	0.00 (2.37, 1.28)	0.00 (<0, 1.51)	0.00 ($>10^2$, 1.77)
$\hat{Z}_{C_2}(x)$	0.04 (<0, 1.18)	0.05 (<0, 1.39)	0.04 ($>10^2$, 1.67)	0.01 (1.89, 1.12)	0.00 (<0, 1.37)	0.01 ($>10^2$, 1.63)
$\hat{Z}_{C_3}(x)$	0.03 (<0, 1.30)	0.04 (<0, 1.56)	0.04 ($>10^2$, 1.83)	0.00 (1.98, 1.15)	0.00 (<0, 1.38)	0.01 ($>10^2$, 1.65)

Table 3.9: Average predictors for uniformly, randomly generated locations, [MSE, empirical MSE] and (RMSE, empirical RMSE) respectively, based on 1000 simulations of 50 observations at 6 locations. Medium correlation function.

	Regular			Uniform		
	Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
$Z(x)$	0.00	0.00	0.02	0.00	0.00	-0.02
$\hat{Z}_{fe}(x)$	0.00 [0.13, 0.14]	0.00 [0.04, 0.04]	0.02 [0.04, 0.04]	-0.01 [0.23, 0.24]	-0.01 [0.12, 0.13]	-0.01 [0.12, 0.13]
$\hat{Z}_{pe}(x)$	0.00 (1.21, 1.02)	0.00 (1.14, 1.04)	0.03 (0.75, 1.46)	-0.01 (1.01, 1.00)	-0.01 (1.05, 1.03)	-0.01 (0.83, 1.04)
$\hat{Z}_{\ell}(x)$	0.00 (1.00, 1.00)	0.00 (1.00, 1.00)	0.02 (1.00, 1.00)	-0.01 (1.00, 1.00)	-0.01 (1.00, 1.00)	-0.01 (0.99, 1.00)
$\hat{Z}_{C_1}(x)$	0.01 (2.71, 1.86)	0.01 (<0, 2.58)	0.02 ($>10^2$, 2.86)	0.00 (<0, 1.24)	0.00 (<0, 1.39)	-0.02 ($>10^2$, 2.94)
$\hat{Z}_{C_2}(x)$	0.01 (1.91, 1.20)	0.01 (<0, 1.56)	0.02 ($>10^2$, 1.84)	-0.01 (<0, 1.09)	-0.01 (<0, 1.26)	-0.02 ($>10^2$, 2.14)
$\hat{Z}_{C_3}(x)$	0.01 (2.09, 1.35)	0.01 (<0, 1.77)	0.02 ($>10^2$, 2.05)	-0.01 (<0, 1.10)	-0.01 (<0, 1.27)	-0.02 ($>10^2$, 2.09)

The evolution of the mean square prediction error as a function of δ for a fixed number of observations, replications and sites is again investigated (Figure 3.4). As already observed in Figure 3.3, where the parameters were known, there is nearly no change in the behavior of the mean square prediction error if the parameters are estimated first. This approach of finding minimal weights will therefore no longer be pursued.

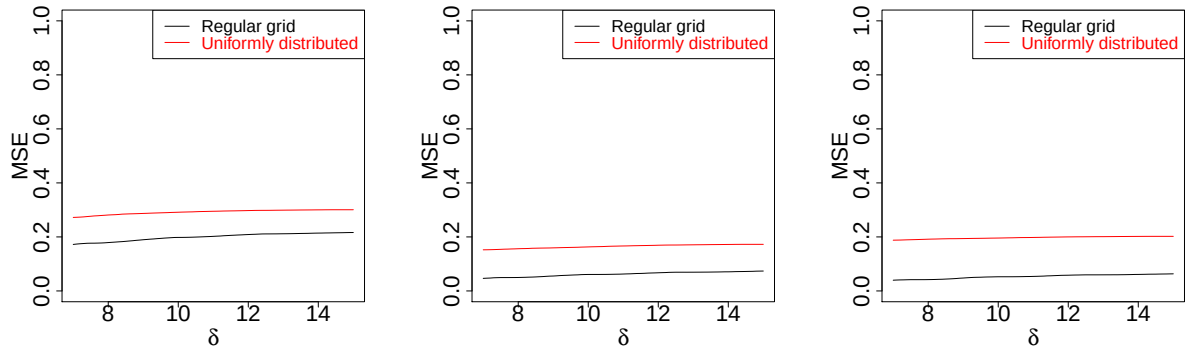

Figure 3.4: Mean square prediction error computed for different values of δ for simulated samples based on 25 observations, 6 locations and 1'000 simulations. The parameters of the covariance functions are estimated. From left to right: Exponential, Whittle-Matérn and Cauchy covariance functions.

3.5 Parameter estimation and prediction when assuming a wrong underlying correlation structure

As the underlying correlation structure is not known explicitly in applications, it is interesting to check whether the predictors are sensitive to this. A single simulation setting, 6 sites, 25 observations and 1000 replications, is considered as the behavior of the predictors is known as the values change.

Table 3.10 indicates the predictors, the empirical mean square prediction error and the ratios of empirical mean square prediction errors, respectively, for Gaussian processes observed at uniformly distributed locations. As the previous analyses have shown that the plug-in estimates for the mean square prediction errors are not reliable due to the estimated parameters, only the empirical mean square prediction errors are indicated.

All combinations of true and fitted models work well. The predictors are very close to the true values. The empirical ratios of mean square prediction errors are about the same for most of combinations of fitted and true models. The least precise predictors are obtained when fitting a model with Cauchy covariance function to data that is generated by a model with a different covariance function. As before, the pairwise predictor with the second choice of weight does best among the three pairwise predictors that are investigated.

To conclude, assuming an underlying correlation structure that might not be the true one does not cause particular problems, the predictors and the mean square prediction errors still look reasonable.

Table 3.10: Average predictors for a random site [empirical MSE] and (empirical RMSE) based on 6 uniformly distributed sites, 25 observations, 1000 replications, model fitted with an incorrect covariance function.

Fitted model:	Exponential		Whittle-Matérn		Cauchy	
True model:	Cauchy	Whittle-Matérn	Exponential	Cauchy	Exponential	Whittle-Matérn
$Z(x)$	0.02	0.02	0.02	0.02	-0.03	-0.01
$\hat{Z}_{fe}(x)$	0.01 [0.14]	0.01 [0.13]	0.01 [0.24]	0.02 [0.13]	-0.02 [0.26]	-0.02 [0.27]
$\hat{Z}_{pe}(x)$	0.01 (1.46)	0.01 (1.08)	0.01 (1.01)	0.02 (1.26)	-0.03 (1.30)	-0.03 (1.23)
$\hat{Z}_{\ell}(x)$	0.01 (1.00)	0.01 (1.00)	0.01 (1.00)	0.02 (1.00)	-0.02 (1.00)	-0.02 (1.00)
$\hat{Z}_{C_1}(x)$	0.00 (1.72)	0.00 (1.50)	0.00 (1.32)	0.00 (1.88)	-0.03 (1.27)	-0.01 (1.24)
$\hat{Z}_{C_2}(x)$	0.00 (1.60)	0.00 (1.38)	0.01 (1.14)	0.00 (1.72)	-0.03 (1.10)	-0.01 (1.08)
$\hat{Z}_{C_3}(x)$	0.00 (1.65)	0.00 (1.40)	0.00 (1.16)	0.00 (1.77)	-0.03 (1.08)	-0.00 (1.07)

3.6 Analysis of the mean precipitation data

The considered data are mean annual summer precipitations based on daily measurements in the region of Zürich (Figure 3.5). The distances between the 51 respective measurement stations are given in kilometres. The range of distances reaches from 2 kilometres up to nearly 60 kilometres. For each of these stations 47 annual means are reported. The mean observations can therefore be described as $Y_{j,i}$, where $j \in 1, \ldots, 47$ and $i \in 1, \ldots, 51$.

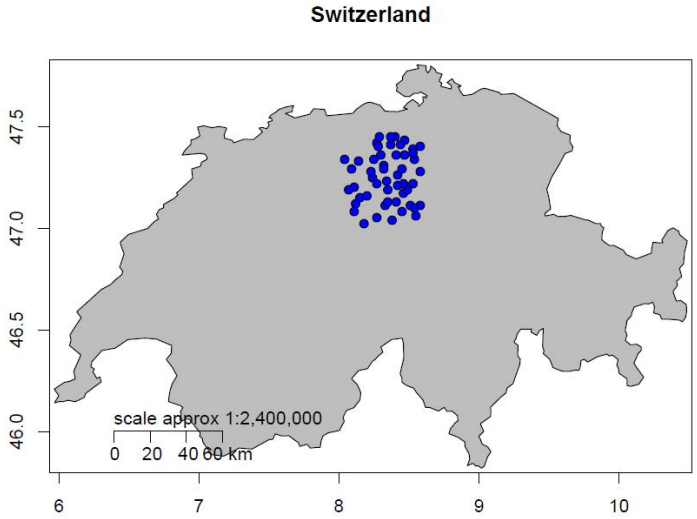

Figure 3.5: Locations of the measurement stations.

The hypothesis that the mean precipitation data is normally distributed is checked by graphical means (Figure 3.6). The histogram shows that the data are skewed and this is confirmed by the normal QQ-plot. This may be due to strong dependences of the daily rainfall and in addition, there may be another source of dependence through the spatial component as all the locations are investigated simultaneously. Nevertheless, a Gaussian process is fitted to the data and predictions at random locations calculated.

The parameter estimation is done by full and pairwise likelihood for different covariance functions (Table 3.11). For a given covariance function, the different approaches yield unexpectedly different parameter estimates. The shape of the profile likelihoods was investigated to make sure that the algorithm converged to the global maximum. The model fit is thus assessed by means of the information criteria for full and pairwise likelihood (Table 3.12) described in sections 2.1 and 2.2, respectively.

As for the Whittle-Matérn covariance function it is not possible to calculate the gradient for the shape parameter ν, this parameter is first estimated using the profile likelihood. In a second step, the other parameters are estimated and their gradients calculated. Considering a fixed value for ν allows thus to compute the information criterion.

If the models are fitted by full likelihood estimation, the AIC is minimal for the Whittle-Matérn covariance function, indicating the best fit. According to the composite likelihood

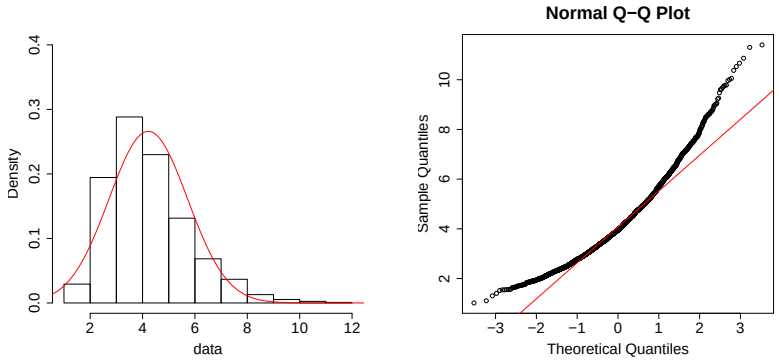

Figure 3.6: Histogram and QQ-plot of mean summer precipitation data.

information criterion, the powered exponential covariance function yields the most appropriate pairwise fitted model.

Table 3.11: Parameter estimation for Gaussian processes using full and pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates.

Covariance function		$\hat{\mu}$ (se($\hat{\mu}$))	$\hat{\sigma}^2$ (se($\hat{\sigma}^2$))	$\hat{\beta}$ (se($\hat{\beta}$))	$\hat{\nu}$ (se($\hat{\nu}$))
Exponential	full	4.72 (0.28)	5.19 (0.88)	123.79 (28.28)	0.88 (0.03)
	pairwise	4.23 (0.11)	2.35 (0.16)	22.09 (1.41)	1.49 (0.04)
Whittle-Matérn	full	4.71 (0.26)	4.70 (0.74)	131.0 (27.02)	0.41 (0.01)
	pairwise	4.23 (-)	2.35 (-)	12.9 (-)	1.03 (-)
Cauchy	full	4.49 (0.31)	5.62 (0.97)	2.38 (0.13)	0.04 (0.01)
	pairwise	4.23 (0.11)	2.36 (0.16)	22.72 (3.84)	1.80 (0.57)

Predictors can thus be obtained for random ungauged locations (Table 3.13). Only the full likelihood and the pairwise predictors are calculated. The best linear predictor is not considered as the computational effort is too large compared to the improvement on the full likelihood predictor.

Comparing to the full likelihood predictor, the second weight function yields the most reasonable predictors among the pairwise predictors, despite the fact that the ratio of mean square prediction errors is greater than the ones of the other pairwise predictors. Generally, there is a massive increase, much more pronounced than in the simulation study, of the ratio of mean square prediction errors when considering the pairwise setting. A probable reason for this may be that the data are not properly normally distributed as discussed before.

In order to check whether these prediction results are sensible, a cross-validation is used. In this context, the parameters of the process are estimated using two-thirds of the stations, meaning 34 stations, only. Based on these same stations, the observations at the 17 remaining stations are predicted. Only the best linear predictor and the pairwise

Table 3.12: Information criterion (AIC and CLIC) scores for models fitted by full and pairwise likelihood, respectively.

Covariance function	AIC	CLIC
Exponential	5289	**2003**
Whittle-Matérn	**5274**	2067
Cauchy	5738	2059

Table 3.13: Average predictors [MSE] and (RMSE) for 10,000 uniformly randomly generated ungauged locations, parameters of the Gaussian process estimated.

	Exponential	Whittle-Matérn	Cauchy
$\hat{Z}_\ell(x)$	3.85 [0.29]	3.86 [0.30]	3.86 [0.04]
$\hat{Z}_{C_1}(x)$	4.01 (6.96)	4.02 (6.77)	4.01 (51.25)
$\hat{Z}_{C_2}(x)$	3.93 (7.52)	3.94 (7.23)	3.93 (56.25)
$\hat{Z}_{C_3}(x)$	3.94 (7.14)	3.95 (6.87)	3.04 (52.75)

predictor based on the second weight function are considered. Figure 3.7 shows that there is a considerable difference in the distributions of the best linear predictors and the pairwise predictors. The distributions of the predictors indicate that they are not very close to the true values for both approaches, but less in the pairwise setting. Besides, there is no particular difference in the predictors due to the assumed correlation structure.

The prediction for the mean precipitation data turns out to be problematic. The mean square prediction errors are very large if a pairwise predictor is considered. Moreover, the pairwise predictors are often inappropriate. A reason for this may be that the data are not exactly normally distributed and that the method is very sensitive to this hypothesis.

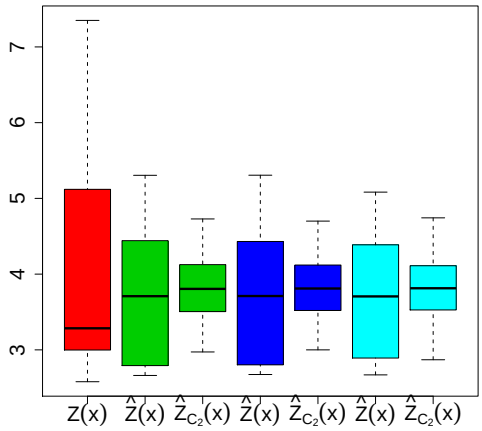

Figure 3.7: Boxplots of true mean precipitation values (red), best linear and pairwise predictors, $\hat{Z}_{fe}(x)$ and $\hat{Z}_{C_2}(x)$, respectively, for 17 randomly chosen stations based on observations at remaining 34 locations. Green: exponential covariance function, blue: Whittle-Matérn covariance function, light blue: Cauchy covariance function.

Chapter 4

Max-stable processes

Extreme values observed at locations in a spatial domain may be modelled by max-stable processes. Actually, these processes are the spatial analogues of multivariate extreme value models. However, inference can only be based on composite likelihood methods as the multivariate density function is intractable. Two special cases, the 'random storm' and the 'random process' formulations, for which the bivariate densities are available are considered here.

Definition 3. Let $\{\tilde{Z}_k(x)\}_{x\in\mathcal{X}}$ $k = 1, \ldots, K$ be independent replications of a continuous stochastic process for an index set $\mathcal{X}$ and assume that there are sequences of continuous functions $a_k(x) > 0$ and $b_k(x) \in \mathbb{R}$ such that

$$Z(x) = \lim_{k\to\infty} \frac{\max_{k=1}^{K} \tilde{Z}_k(x) - b_k(x)}{a_k(x)}, \quad x \in \mathcal{X}.$$

Provided that this limit exists, the limit process $Z(x)$ is a max-stable process (de Haan (1984)).

Without loss of generality, $Z(x)$ is assumed to be stationary and if $a_k(x) = k$ and $b_k(x) = 0$ the margins of $Z(x)$ are unit Fréchet with distribution function

$$\text{pr}(Z(x) \leq z) = \exp\left(-\frac{1}{z}\right), \quad x \in \mathcal{X} \quad z > 0.$$

Consider the particular canonical representation of a max-stable process where Π is a Poisson process on $\mathbb{R}_+$ of intensity $s^{-2}ds$ and $\{Y_s(x) : x \in \mathcal{X}, s \in \mathbb{R}_+\}$ a collection of independent identically distributed non-negative random processes with measure ν on $\mathcal{Y} = \mathbb{R}_+^{\mathcal{X}}$ satisfying $\text{E}\{Y_s(x)\} = 1$ for all x. Schlather (2002) showed that

$$Z(x) = \max_{s\in\Pi} sY_s(x), \quad x \in \mathcal{X},$$

is max-stable with unit Fréchet marginal distributions.

The 'random storm' model was first considered by Smith (1990). More recently, Padoan *et al.* (2009) applied composite likelihood approaches to this formulation in order to model

spatial extremes. In this setting, let $\mathcal{X} = \mathbb{R}^d$ and $Y_s(x) = f(x - X_s)$, where f is a density function on $\mathcal{X}$ and X_s a point of a Poisson process of unit rate in $\mathcal{X}$. The physical interpretation of this process is as follows. If s is the magnitude of the storm at location x, centred at X_s and of shape f, then $sY_s(x)$ is the impact of the storm at x. Therefore, $Z(x)$ is the impact of the largest storm observed at x.

Smith (1990) showed that if f is chosen to be multivariate normal with covariance matrix Σ, the bivariate marginal distribution equals

$$\text{pr}\{Z(x_i) \leq z_i, Z(x_j) \leq z_j\} = \exp\left\{-\frac{1}{z_i}\Phi\left(\frac{a(\mathbf{h})}{2} + \frac{1}{a(\mathbf{h})}\log\frac{z_j}{z_i}\right) - \frac{1}{z_j}\Phi\left(\frac{a(\mathbf{h})}{2} + \frac{1}{a(\mathbf{h})}\log\frac{z_i}{z_j}\right)\right\}, \tag{4.1}$$

where $\mathbf{h} = x_i - x_j$, $a(\mathbf{h})^2 = \mathbf{h}^T\Sigma^{-1}\mathbf{h}$ and $\Phi(\cdot)$ the standard normal distribution function.

By taking second-order partial derivatives of (4.1) the bivariate density function can be obtained

$$\begin{aligned} f(z_i, z_j) \quad = \quad & \exp\left[-\frac{\Phi\{w(\mathbf{h})\}}{z_i} - \frac{\Phi\{v(\mathbf{h})\}}{z_j}\right]\left[\left(\frac{\Phi\{w(\mathbf{h})\}}{z_i^2} + \frac{\phi\{w(\mathbf{h})\}}{a(\mathbf{h})z_i^2} - \frac{\phi\{v(\mathbf{h})\}}{a(\mathbf{h})z_i z_j}\right)\right. \\ & \left.\left(\frac{\Phi\{v(\mathbf{h})\}}{z_j^2} + \frac{\phi\{v(\mathbf{h})\}}{a(\mathbf{h})z_j^2} - \frac{\phi\{w(\mathbf{h})\}}{a(\mathbf{h})z_i z_j}\right) + \left(\frac{v(\mathbf{h})\phi\{w(\mathbf{h})\}}{a(\mathbf{h})^2 z_i^2 z_j} + \frac{w(\mathbf{h})\phi\{v(\mathbf{h})\}}{a(\mathbf{h})^2 z_i z_j^2}\right)\right], \end{aligned}$$

where $w(\mathbf{h}) = a(\mathbf{h})/2 + \log(z_j/z_i)/a(\mathbf{h})$, $v(\mathbf{h}) = a(\mathbf{h}) - w(\mathbf{h})$ and $\phi(\cdot)$ is the standard normal density function.

Let us now investigate the 'random process' formulation suggested by Schlather (2002) considering the particular setting $Y(x) = \max\{0, \sqrt{2\pi}\varepsilon(x)\}$, where $\{\varepsilon(x)\}$ is a stationary Gaussian process with zero mean, unit variance and correlation function $\rho(x)$. It is then possible to obtain the marginal bivariate distribution function

$$\text{pr}\{Z(x_i) \leq z_i, Z(x_j) \leq z_j\} = \exp\left[-\frac{1}{2}\left(\frac{1}{z_i} + \frac{1}{z_j}\right)\left\{1 + \sqrt{1 - 2\frac{\{\rho(h)+1\}z_i z_j}{(z_i + z_j)^2}}\right\}\right], \tag{4.2}$$

where $h = ||x_i - x_j||$.

Again, the pairwise density of $Z(x_i), Z(x_j)$ is computed by differentiating partially with respect to z_i, z_j,

$$f(z_i, z_j) = \left\{\frac{\partial V(z_i, z_j)}{\partial z_i}\frac{\partial V(z_i, z_j)}{\partial z_j} - \frac{\partial^2 V(z_i, z_j)}{\partial z_i \partial z_j}\right\}\exp\{-V(z_i, z_j)\}, \quad z_i, z_j > 0, \tag{4.3}$$

where

$$V(z_i, z_j) = \frac{1}{2}\left(\frac{1}{z_i} + \frac{1}{z_j}\right)\left[1 + \sqrt{1 - 2\frac{\{\rho(h)+1\}z_i z_j}{(z_i + z_j)^2}}\right].$$

4.1 Prediction based on pairwise likelihood

Similar to chapter 3, the pairwise likelihood can be considered in order to obtain a predictor for an ungauged location of a max-stable process

$$\begin{aligned}\ell_C\left\{Z_1(x)\right\} &= \sum_{i=1}^{n} w_i \log f\{Z_1(x)|Z_1(x_i)\} \\ &= \sum_{i=1}^{n} w_i \left[2\log Z_1(x_i) + \frac{1}{Z_1(x_i)} + \log f\{Z_1(x), Z_1(x_i)\}\right],\end{aligned}$$

where the second equality follows from the fact that the marginal bivariate distributions are unit Fréchet.

The predictor $\hat{Z}_C(x)$ cannot be computed as an analytical solution of the composite score equation. Composite maximum likelihood estimates can however be obtained through numerical maximization routines.

Due to non-existence of moments of max-stable processes, median predictors are considered instead of average predictors. Similarly, a useful quantity to describe the precision of the predictor is the median square prediction error

$$\text{MDSE}\left\{\hat{Z}(x)\right\} = \text{med}\left[\{\hat{Z}(x) - Z(x)\}^2\right],$$

where $\hat{Z}(x)$ is a predictor for $Z(x)$.

4.2 Simulation study

As the simulation study for Gaussian processes indicates that the pairwise predictors are of reasonable quality, they are considered for a simulation study in the case of max-stable processes. Model fitting through pairwise likelihood for both the 'random process' and the 'random storm' models is possible using the `R` package `SpatialExtremes` by Mathieu Ribatet.

Similar to the simulation study on the Gaussian scale, different strengths of correlations between the observations are considered.

4.2.1 'Random process' model

As previously, the initial parameter setting is the following, there are $n = 6$ uniformly but randomly distributed locations $x_1, \ldots, x_6$ on the interval $[0, 15]$. The underlying correlation structure is again assumed to be powered exponential with different strengths of dependence as depicted in Figure 3.1.

The corresponding max-stable process is simulated $10,000$ times at each location including x. Predictors at x can now only be based on composite likelihood as the full density is intractable. Thus, as previously, the pairwise predictors corresponding to the weight functions (3.9)–(3.11) are considered. Only one observation is generated at each location as the pairwise predictors do not take into account the other observations anyway.

The boxplots of the true value of $Z(x)$ and the pairwise predictors for $Z(x)$ for different correlation configurations are shown in Figure 4.1. If there is short range dependence, the predictors are very imprecise and tend to be too small compared to the true values. Increasing the dependence improves the predictors, but still, for medium range dependent observations, the predictors are clearly underestimating the true values. A possible reason for this behavior is the small number of pairs available to predict the observation at the ungauged location.

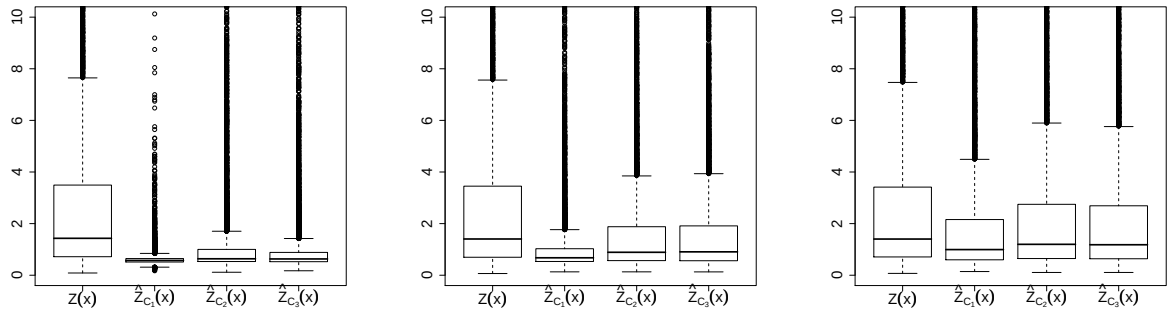

Figure 4.1: Boxplots of the true values and the pairwise predictors for $Z(x)$ based on $10,000$ simulations of one observation at 6 locations which are uniformly distributed on $[0, 15]$. From left to right: powered exponential covariance function with short, medium and long range dependences.

Consider therefore a max-stable process on $[0, 15]$ observed at 15, 25 or 50 uniformly, randomly distributed locations. Only the medium range dependence is considered to analyze the predictors as the number of observations increases.

The distributions of the predictors are shown in Figure 4.2. Compared to the middle panel of Figure 4.1 where there were 6 observations at each location for the same correlation structure, there is no noticeable change of the boxplots for $\hat{Z}_{C_1}(x)$ and $\hat{Z}_{C_3}(x)$. In contrast, an improvement in the precision of the pairwise predictor $\hat{Z}_{C_2}(x)$ is observed. However, as the number of observations is greater than 15, the gain in precision is not very remarkable. The additional information provided by the additional locations cannot be used very efficiently by the pairwise predictors in this case.

The previous analyses were performed for the Whittle-Matérn and Cauchy covariance functions, too. The details are not shown as the same conclusions can be drawn.

However, a detailed analysis of the prediction results based on 6 or 15 observed, medium range dependent extremes is performed for all the covariance functions. The median predictors and the median square prediction errors, which are displayed in Table 4.1, indicate again that the pairwise predictors are of bad quality if there are only 6 observations although the median square prediction errors are small. It is interesting that in the context of Gaussian processes, reasonable results were obtained for the same parameter settings.

If there are 15 locations, the predictors $\hat{Z}_{C_1}(x)$ and $\hat{Z}_{C_3}(x)$ do not become more precise.

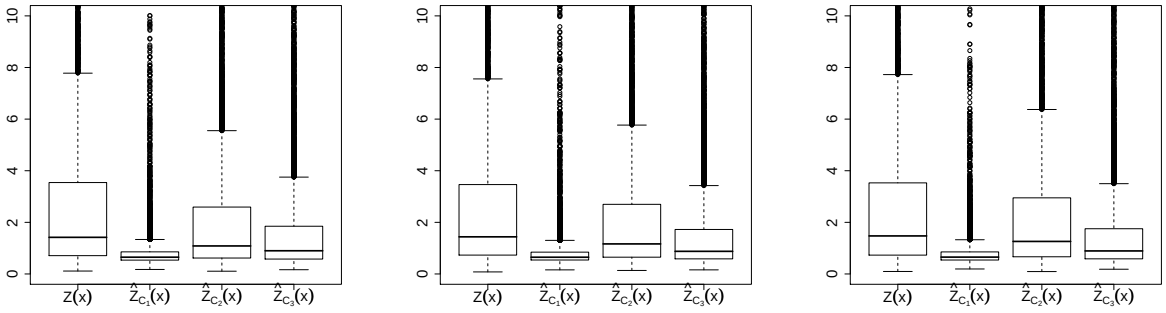

Figure 4.2: Boxplots of true values and pairwise predictors for $Z(x)$ based on 10,000 simulations of one observation at 15, 25 and 50 locations, respectively, which are uniformly distributed on $[0, 15]$, exponential covariance function with medium range dependence.

In fact, only $\hat{Z}_{C_2}(x)$ improves in terms of the median prediction as well as the median square prediction error. Nevertheless, this best pairwise predictor still underestimates the true value.

Table 4.1: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 10,000 simulations of one observation at 6 or 15 locations, respectively. Medium range dependence.

		Regular			Uniform		
n		Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
6	$Z(x)$	1.46	1.45	1.41	1.40	1.45	1.47
	$\hat{Z}_{C_1}(x)$	0.66 (0.26)	0.78 (0.06)	0.73 (0.06)	0.68 (0.20)	0.81 (0.07)	0.75 (0.07)
	$\hat{Z}_{C_2}(x)$	1.05 (0.09)	1.28 (0.02)	1.20 (0.02)	0.89 (0.10)	1.12 (0.04)	1.03 (0.04)
	$\hat{Z}_{C_3}(x)$	0.91 (0.11)	1.17 (0.03)	1.08 (0.03)	0.91 (0.12)	1.13 (0.04)	1.04 (0.04)
15	$Z(x)$	1.46	1.43	1.43	1.42	1.42	1.48
	$\hat{Z}_{C_1}(x)$	0.66 (0.28)	0.78 (0.03)	0.73 (0.04)	0.65 (0.26)	0.76 (0.04)	0.72 (0.05)
	$\hat{Z}_{C_2}(x)$	1.26 (0.03)	1.38 (0.004)	1.37 (0.004)	1.09 (0.04)	1.29 (0.01)	1.30 (0.01)
	$\hat{Z}_{C_3}(x)$	0.87 (0.09)	1.14 (0.01)	1.09 (0.01)	0.90 (0.08)	1.14 (0.01)	1.12 (0.01)

Again, it is worthwhile to investigate the influence of parameter estimation in the context of pairwise prediction. The parameter σ^2 is not estimated as the 'random process' is not measurable if σ^2 is strictly smaller than 1. Only the process consisting of observations at 6 locations with a medium range dependence is considered. The quality of the estimates is however improved if the process is observed at more locations or if there is a stronger correlation structure.

Consider the parameter estimates obtained for the medium range dependence (Table 4.2). As expected, the number of observations influences the accuracy of the estimates. Moreover, the grid type is also an important factor, more precise estimates are obtained if the locations are uniformly distributed on $[0, 15]$. For example, the shape parameter of

the Whittle-Matérn covariance function is poorly identified if the grid is regular.

Particular attention needs again to be paid to the Cauchy covariance function as the parameters of the process are not close to the true values. The estimated covariance functions tend to have a shorter dependence range than the true one has.

Table 4.2: Average parameter estimation based on $k = 25, 50$ observations at each of the 6 locations using full and pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates.

Covariance function	k	grid	$\hat{\beta}$ (se($\hat{\beta}$))	$\hat{\nu}$ (se($\hat{\nu}$))
Exponential	25	Regular	12.4 (32.4)	1.05 (0.36)
		Uniform	9.86 (7.2)	1.01 (0.22)
	50	Regular	8.99 (3.9)	1.02 (0.25)
		Uniform	8.72 (3.2)	1.00 (0.15)
Whittle-Matérn	25	Regular	6.78 (-)	2.14 (-)
		Uniform	5.45 (-)	1.36 (-)
	50	Regular	5.59 (-)	1.60 (-)
		Uniform	5.27 (-)	1.12 (-)
Cauchy	25	Regular	7.94 (16.6)	18.0 (88.9)
		Uniform	9.35 (20.8)	22.0 (105.7)
	50	Regular	4.11 (6.0)	2.65 (17.0)
		Uniform	5.59 (11.9)	7.5 (56.2)

Based on these parameter estimates, pairwise predictors can be calculated (Table 4.3). The results reveal that the parameter estimation has little effect on the predictors. There is no significant difference between the predictors obtained through the different parameter estimates. The predictors remain nearly the same as when the parameters are known (Table 4.1).

Table 4.3: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 1000 simulations of $k = 25, 50$ observations at 6 locations. Medium range dependence.

		Regular			Uniform		
k		Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
25	$Z(x)$	1.37	1.53	1.51	1.43	1.42	1.45
	$\hat{Z}_{C_1}(x)$	0.66 (0.19)	0.77 (0.06)	0.74 (0.09)	0.69 (0.18)	0.82 (0.06)	0.80 (0.07)
	$\hat{Z}_{C_2}(x)$	1.01 (0.07)	1.29 (0.02)	1.28 (0.03)	0.91 (0.09)	1.12 (0.04)	1.05 (0.04)
	$\hat{Z}_{C_3}(x)$	0.88 (0.08)	1.22 (0.03)	1.15 (0.03)	0.90 (0.10)	1.13 (0.04)	1.06 (0.05)
50	$Z(x)$	1.36	1.49	1.40	1.44	1.22	1.49
	$\hat{Z}_{C_1}(x)$	0.65 (0.18)	0.78 (0.07)	0.72 (0.06)	0.67 (0.21)	0.76 (0.04)	0.73 (0.09)
	$\hat{Z}_{C_2}(x)$	1.04 (0.08)	1.25 (0.02)	1.14 (0.03)	0.89 (0.09)	1.02 (0.03)	1.01 (0.04)
	$\hat{Z}_{C_3}(x)$	0.87 (0.10)	1.14 (0.03)	1.08 (0.03)	0.87 (0.12)	0.99 (0.03)	1.04 (0.05)

As suggested by the analysis where the parameters are known, it is advantageous to observe the process at more locations. The number of observations is now fixed to be 25

as there seems to be no improvement in the prediction even if the parameter estimates get more precise. Figure 4.3 shows the boxplots of the true values of the max-stable process with underlying medium range powered exponential covariance function and the pairwise predictor $\hat{Z}_{C_2}(x)$ for 6, 15, 25 and 50 locations, respectively. The other pairwise predictors are not considered here as they are of lower quality.

These boxplots indicate that the pairwise predictors constantly underestimate the true value of the process and that a lot of locations are required to improve the quality of the predictor. Similar results are provided if the underlying covariance functions are either Whittle-Matérn or Cauchy. Naturally, the predictors are also becoming more precise if the correlation between the extremes is stronger.

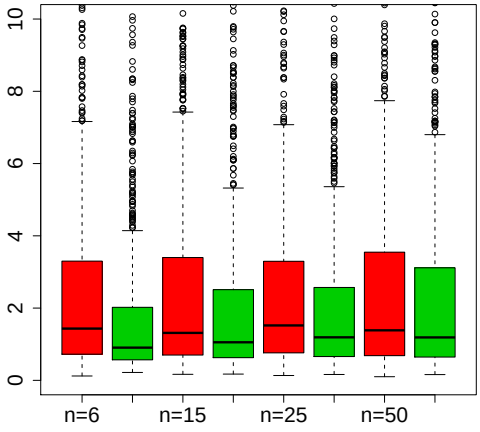

Figure 4.3: Boxplots of the true values (red) and the pairwise predictors $\hat{Z}_{C_2}(x)$ (green) for $Z(x)$ based on 1000 simulations of 25 observations at 6, 15, 25 and 50 locations, respectively, which are uniformly distributed on $[0, 15]$, powered exponential covariance function with medium range dependence.

Nonetheless, it would be preferable to have a different method which improves the predictors. As the prediction on the Gaussian scale results in more precise predictors, observations of max-stable processes may be transformed to the Gaussian scale and predictions calculated based on the formulae described in chapter 3. This approach is discussed in section 4.3.

4.2.2 'Random storm' model

A one dimensional simulation study is also carried out for the 'random storm' model where the parameter σ^2 of the normal distribution is chosen from the discrete set $\{2.5, 5, 10\}$. Initially, a simulated process of size $10,000$ observed a single time at 6 locations on $[0, 15]$ is considered. The predictors are again calculated corresponding to the three weight functions defined in (3.9)–(3.11). Figure 4.4 depicts the distributions of the true values of the

processes as well as the predictors for the different values of σ^2. As σ^2 increases, the quality of the predictors does, too. In accordance to the previous conclusions, the second weight functions yields predictors that mimic best the true distribution of the process.

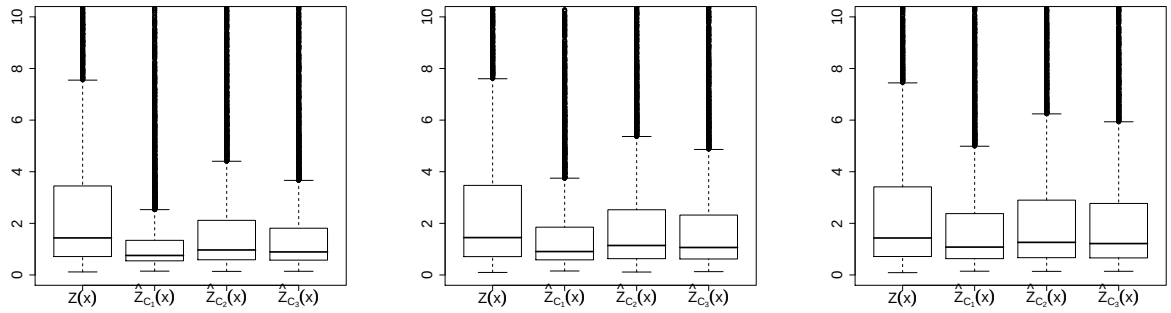

Figure 4.4: Boxplots of true values and pairwise predictors for $Z(x)$ based on 10,000 simulations of one observation at 6 locations which are uniformly distributed on $[0, 15]$. From left to right: $\sigma^2 = 2.5$, $\sigma^2 = 5$ and $\sigma^2 = 10$.

For a more profound analysis, the medians and the median square prediction errors are shown for both predictions based on 6 and 15 observed extreme values (Table 4.4). As for the 'random process' model, there is a general underestimation of the true value and the predictors are preciser if the gauged locations are regularly distributed.

Moreover, the increase of the number of locations yields an improvement in the median prediction as well as in the median square prediction errors. Given that the 15 gauged locations are on a regular grid, the pairwise approach using the second weight function gives a very precise predictor if the variance of the underlying normal distribution is greater or equal to 5.

Table 4.4: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 10,000 simulations of one observation at 6 and 15 locations.

		Regular			Uniform		
n		$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$	$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$
6	$Z(x)$	1.44	1.45	1.43	1.41	1.44	1.39
	$\hat{Z}_{C_1}(x)$	0.76 (0.18)	0.91 (0.10)	1.08 (0.05)	0.65 (0.19)	0.78 (0.12)	0.94 (0.05)
	$\hat{Z}_{C_2}(x)$	0.97 (0.08)	1.14 (0.05)	1.26 (0.02)	0.77 (0.10)	0.92 (0.07)	1.09 (0.03)
	$\hat{Z}_{C_3}(x)$	0.89 (0.10)	1.06 (0.06)	1.22 (0.03)	0.75 (0.12)	0.92 (0.07)	1.08 (0.04)
15	$Z(x)$	1.43	1.46	1.40	1.50	1.41	1.43
	$\hat{Z}_{C_1}(x)$	1.03 (0.07)	1.20 (0.03)	1.28 (0.01)	0.86 (0.10)	1.00 (0.04)	1.19 (0.02)
	$\hat{Z}_{C_2}(x)$	1.32 (0.01)	1.42 (0.01)	1.40 (0.01)	1.14 (0.03)	1.22 (0.01)	1.34 (0.01)
	$\hat{Z}_{C_3}(x)$	1.17 (0.03)	1.32 (0.01)	1.35 (0.01)	1.04 (0.05)	1.16 (0.02)	1.31 (0.01)

The quality of the prediction based on uniformly distributed locations is likely to improve if there are more gauged locations. The boxplots for the max-stable process with

$\sigma^2 = 5$ and different numbers of locations are shown in Figure 4.5 and indicate that the distributions of the predictors get very similar, especially for the second weight function. In the extreme setting where there are 50 gauged locations, $\hat{Z}_{C_2}(x)$ even tends to overestimate the value of $Z(x)$.

In contrast to the Gaussian processes and the 'random process' model, all the weight functions improve significantly as the number of locations increases. Still, the second one yields the most appropriate results.

If the number of locations is changed for other values of σ^2 similar conclusions can be drawn; if σ^2 is smaller, even more locations are required to be able to predict an observation at an ungauged location more precisely.

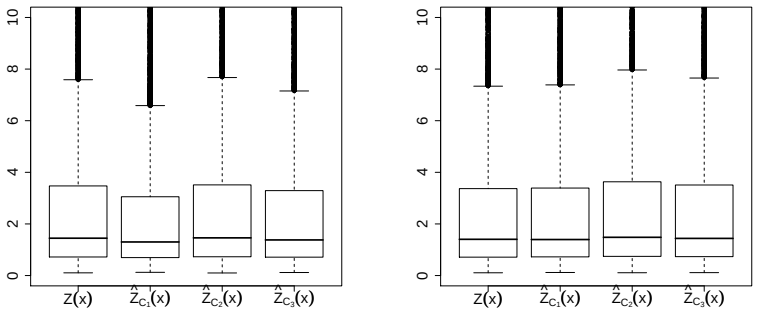

Figure 4.5: Boxplots of the true values and the pairwise predictors for $Z(x)$ based on 1000 simulations of 25 observations at 25 and 50 locations, respectively, which are uniformly distributed on $[0, 15]$, $\sigma^2 = 5$.

As the parameter σ^2 is estimated through pairwise likelihood, the estimation error may also affect the quality of the pairwise predictors. The parameter estimates obtained based on max-stable processes observed $k = 25, 50$ times at $n = 6, 15$ locations are displayed in Table 4.5. Independent of the number of observations, the parameter estimates are very imprecise if there are only 6 locations. Increasing the number of gauged locations to 15, the estimation is improved but still not very precise. More locations are required to improve the quality of the predictors.

Predictors for an ungauged location x can then be calculated using the obtained parameter estimates. The resulting median predictors based on 1000 simulations are shown in Table 4.6.

Compared to the results where the parameter value is known, the median predictions are equivalent. Even the median square prediction errors are nearly the same as before. This shows that the pairwise predictors are robust and give reliable results even if the parameter estimation is not very precise. As expected, increasing the value of σ^2 or the number of gauged locations implies a gain of precision.

These considerations give evidence that the pairwise methods require a lot of data in order to yield precise predictors. Different prediction approaches are therefore considered in the following section and compared to the pairwise approach given above.

Table 4.5: Average parameter estimation based on $k = 25, 50$ observations at each of the $n = 6, 15$ locations using full and pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates. True values: $\sigma_1^2 = 2.5$, $\sigma_2^2 = 5$ and $\sigma_3^2 = 10$.

n	k	grid	$\hat{\sigma}_1^2$ (se($\hat{\sigma}_1^2$))	$\hat{\sigma}_2^2$ (se($\hat{\sigma}_2^2$))	$\hat{\sigma}_3^2$ (se($\hat{\sigma}_3^2$))
6	25	Regular	4.5 (0.97)	9.2 (1.84)	18.3 (3.73)
		Uniform	4.6 (0.91)	8.9 (2.02)	18.5 (3.50)
	50	Regular	4.4 (1.04)	9.0 (2.00)	18.6 (3.43)
		Uniform	4.6 (0.86)	9.1 (1.89)	17.9 (4.07)
15	25	Regular	3.1 (0.96)	6.0 (1.85)	11.7 (3.72)
		Uniform	2.9 (0.96)	5.9 (1.78)	11.5 (3.76)
	50	Regular	3.0 (0.90)	5.7 (1.87)	11.6 (3.68)
		Uniform	2.9 (0.93)	5.7 (1.98)	11.6 (3.67)

In addition, the second weight function is most efficient according to the obtained results for the 'random process' and the 'random storm' storm models. This confirms the observation already made for Gaussian processes.

4.3 Gaussian anamorphosis

As kriging on the unit Fréchet scale does not yield predictions that are as precise as the ones obtained for Gaussian processes, max-stable processes may be transformed to the Gaussian scale, where standard multivariate normal approaches apply. It is possible to calculate predictors based on classical Gaussian geostatistics and back-transform them to the unit Fréchet scale in order to get predictors for max-stable processes.

More specifically, for every observation $Z(x_i) \equiv z$, $x_i \in \mathcal{X}$ the marginal cumulative distribution function is

$$F(z) = \exp\left(-\frac{1}{z}\right),$$

and therefore the data can be transformed to the standard Gaussian scale by taking

$$Z^*(x_i) = G^{-1}\{F(z)\}, \tag{4.4}$$

where G is the cumulative distribution function of a standard normal random variable. Hence, a predictor $\hat{Z}^*(x)$ for this transformed process can be obtained the same way as in chapter 3. Applying the inverse transformation of (4.4) to $\hat{Z}^*(x)$, a predictor $\hat{Z}^\dagger(x)$ for the max-stable process can be obtained. Like that, it even is possible to base predictors on the full conditional likelihood as it is known for Gaussian processes.

In the case of the Schlather model, the quantities required for predicting on the Gaussian scale are implicitly given through the definition of the model. For the Smith model however, the correlation function needs to be calculated. Consider therefore the geometric Gaussian model where $Y_s(x) = \exp\{t\varepsilon(x) - t^2/2\}$ and $\varepsilon(\cdot)$ is a standard Gaussian process

Table 4.6: Median predictors for uniformly, randomly generated locations (empirical median square prediction error) based on 1000 simulations of one observation at 6 locations.

			Regular			Uniform		
n	k		$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$	$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$
6	25	$Z(x)$	1.58	1.50	1.40	1.24	1.39	1.43
		$\hat{Z}_{C_1}(x)$	0.91 (0.14)	1.08 (0.06)	1.22 (0.03)	0.75 (0.10)	0.91 (0.08)	1.07 (0.05)
		$\hat{Z}_{C_2}(x)$	1.15 (0.10)	1.31 (0.04)	1.34 (0.02)	0.87 (0.07)	1.07 (0.05)	1.20 (0.04)
		$\hat{Z}_{C_3}(x)$	1.08 (0.11)	1.24 (0.04)	1.31 (0.02)	0.86 (0.08)	1.06 (0.06)	1.22 (0.04)
	50	$Z(x)$	1.39	1.52	1.41	1.48	1.44	1.41
		$\hat{Z}_{C_1}(x)$	0.85 (0.11)	1.05 (0.07)	1.18 (0.03)	0.76 (0.13)	0.88 (0.09)	1.07 (0.05)
		$\hat{Z}_{C_2}(x)$	1.05 (0.07)	1.25 (0.04)	1.30 (0.02)	0.90 (0.09)	1.03 (0.07)	1.19 (0.03)
		$\hat{Z}_{C_3}(x)$	0.89 (0.08)	1.21 (0.05)	1.27 (0.02)	0.91 (0.09)	1.03 (0.07)	1.20 (0.04)
15	25	$Z(x)$	1.32	1.40	1.39	1.52	1.43	1.48
		$\hat{Z}_{C_1}(x)$	1.02 (0.05)	1.23 (0.03)	1.31 (0.01)	0.90 (0.09)	1.03 (0.04)	1.19 (0.02)
		$\hat{Z}_{C_2}(x)$	1.27 (0.01)	1.41 (0.01)	1.43 (0.01)	1.18 (0.03)	1.25 (0.01)	1.38 (0.01)
		$\hat{Z}_{C_3}(x)$	1.15 (0.02)	1.34 (0.01)	1.37 (0.01)	1.10 (0.05)	1.18 (0.02)	1.32 (0.01)
	50	$Z(x)$	1.37	1.46	1.46	1.46	1.44	1.55
		$\hat{Z}_{C_1}(x)$	1.04 (0.06)	1.20 (0.03)	1.31 (0.01)	0.86 (0.07)	1.03 (0.03)	1.26 (0.02)
		$\hat{Z}_{C_2}(x)$	1.29 (0.02)	1.40 (0.01)	1.44 (0.01)	1.10 (0.03)	1.28 (0.01)	1.44 (0.01)
		$\hat{Z}_{C_3}(x)$	1.18 (0.03)	1.31 (0.01)	1.39 (0.01)	1.00 (0.04)	1.20 (0.02)	1.38 (0.01)

(personal communication from Anthony C. Davison). The bivariate distribution of this process is equal to the one of the Smith model and the following identity holds

$$a(\mathbf{h})^2 = 2t^2\{1 - \rho(||\mathbf{h}||)\}.$$

This allows to find the correlation function for the Smith process as a function of the euclidean distance between two locations and the parameter t;

$$\rho(||\mathbf{h}||, t) = 1 - \frac{a(\mathbf{h})^2}{2t^2}. \tag{4.5}$$

4.3.1 Transformed Gaussian predictors for the 'random process' model

For the same parameter setting as in section 4.2.1 pairwise kriging predictors directly computed on the unit Fréchet scale are compared to transformed kriging predictors calculated on the Gaussian scale. More precisely, the best linear predictor as well as the pairwise predictor based on the second weight function (see (3.10)) are calculated based on extreme data that is transformed to the Gaussian scale. These two predictors are back-transformed to the unit Fréchet scale and compared to $\hat{Z}_{C_2}(x)$ which is calculated directly from the data on the unit Fréchet scale.

The comparison is performed for $10,000$ simulated processes observed at 6 and 15 gauged locations as well as at the ungauged location (Table 4.7). As previously, different distributions of the locations as well as covariance functions, all corresponding to medium range dependence, are considered.

In general, there is a large discrepancy between the predictors calculated on the different scales. The transformed predictors $\hat{Z}^{\dagger}(x)$ and $\hat{Z}^{\dagger}_{C_2}(x)$ are both very close to the true values

observed at the ungauged location x. As expected, the median square prediction error is much smaller if the best linear predictor is considered.

Observe that the values of the transformed predictors are not more precise if there are more gauged locations, however, there is a reduction of the corresponding median square prediction error for the back-transformed pairwise predictions. In contrast, the pairwise predictors calculated on the unit Fréchet scale improve a lot as already seen in section 4.2.1. As well, if the locations are on a regular grid, the predictors at x are more appropriate.

Table 4.7: Median predictors for uniformly, randomly generated locations (median square prediction error) based on 10,000 simulations of one observation at 6 and 15 locations. Medium range dependence.

		Regular			Uniform		
n		Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
6	$Z(x)$	1.46	1.45	1.41	1.40	1.45	1.46
	$\hat{Z}^{\dagger}(x)$	1.43 ($< 10^{-3}$)	1.44 ($< 10^{-3}$)	1.43 ($< 10^{-3}$)	1.41 ($< 10^{-3}$)	1.44 ($< 10^{-3}$)	1.46 ($< 10^{-3}$)
	$\hat{Z}^{\dagger}_{C_2}(x)$	1.43 (0.10)	1.44 (0.02)	1.42 (0.02)	1.40 (0.11)	1.43 (0.04)	1.45 (0.04)
	$\hat{Z}_{C_2}(x)$	1.05 (0.09)	1.28 (0.02)	1.20 (0.02)	0.89 (0.10)	1.12 (0.04)	1.03 (0.04)
15	$Z(x)$	1.46	1.43	1.43	1.43	1.42	1.48
	$\hat{Z}^{\dagger}(x)$	1.43 ($< 10^{-3}$)	1.43 ($< 10^{-3}$)	1.43 ($< 10^{-3}$)	1.42 ($< 10^{-3}$)	1.41 ($< 10^{-3}$)	1.48 ($< 10^{-3}$)
	$\hat{Z}^{\dagger}_{C_2}(x)$	1.43 (0.04)	1.42 (0.004)	1.41 (0.004)	1.42 (0.05)	1.41 (0.01)	1.46 (0.01)
	$\hat{Z}_{C_2}(x)$	1.26 (0.03)	1.38 (0.004)	1.37 (0.01)	1.09 (0.04)	1.29 (0.01)	1.30 (0.01)

The details of the analyses for other strengths of correlations are not discussed as the behavior of the transformed predictors is very similar to the one of the predictors on the unit Fréchet scale which was described in section 4.2.1.

Similarly, confirming what was already observed in the previous sections, the pairwise kriging predictors are robust to imprecise parameter estimates. Therefore the results where the parameters of the process are estimated are not shown here.

4.3.2 Transformed Gaussian predictors for the 'random storm' model

The same approach is applied to the 'random storm' model. However, for this model another parameter, t, needs to be fixed in order to be able to obtain the correlation (see equation (4.5)) between two extremes.

Figure 4.6 shows the correlation functions for different values of the parameter σ^2. Furthermore, different values for t are considered in order to have different correlation ranges.

For none of the parameter settings, it was possible to invert the correlation matrix as the determinant equals 0. Therefore a generalized inverse, calculated using the singular value decomposition, is considered for calculating the best linear predictor. Only the results for the medium correlation range are shown here as the predictors are equivalent if the value of t changes according to the ones shown in Figure 4.6.

The results for $10,000$ simulations (Table 4.8) suggest that the transformed predictors are very powerful and yield reasonable predictors even as there are only 6 gauged locations.

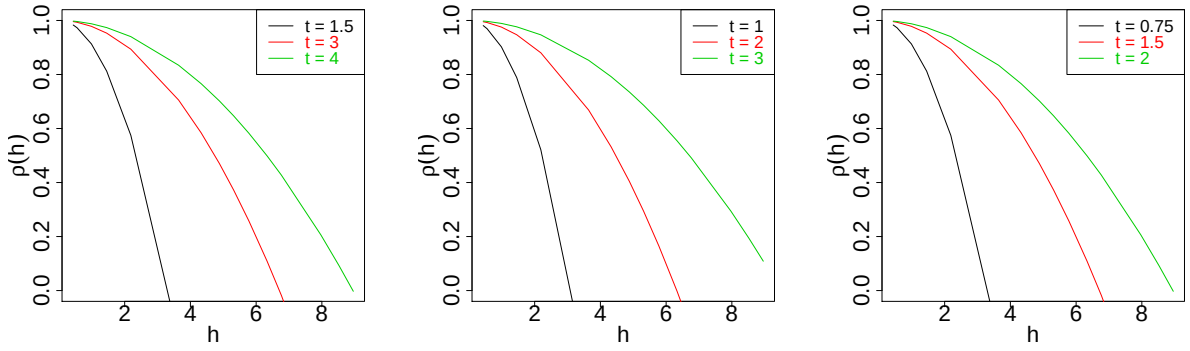

Figure 4.6: Correlation functions for the 'random storm' process for $\sigma^2 = 2.5$, $\sigma^2 = 5$ and $\sigma^2 = 10$ and for different values of t.

Interestingly, the transformed pairwise predictors yield median results that are much closer to the median true values of the process than the best linear predictors.

The direct pairwise approach is not competitive at all, even if the number of gauged locations is increased to 15. Naturally, there is a decrease in the median square prediction errors related to the increase of locations for both approaches.

It is also important to notice that the transformed predictions are very precise for small values of σ^2 whereas the quality of the pairwise predictors on the unit Fréchet scale are affected by this.

Table 4.8: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 10,000 simulations of one observation at 6 and 15 locations.

		Regular			Uniform		
n		$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$	$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$
6	$Z(x)$	1.44	1.45	1.43	1.41	1.44	1.39
	$\hat{Z}^{\dagger}(x)$	1.40 (0.004)	1.40 (0.003)	1.39 (0.001)	1.38 (0.001)	1.43 ($< 10^{-3}$)	1.42 ($< 10^{-3}$)
	$\hat{Z}^{\dagger}_{C_2}(x)$	1.44 (0.07)	1.44 (0.03)	1.43 (0.02)	1.42 (0.10)	1.42 (0.06)	1.41 (0.03)
	$\hat{Z}_{C_2}(x)$	0.97 (0.08)	1.14 (0.05)	1.26 (0.02)	0.77 (0.10)	0.92 (0.07)	1.09 (0.03)
15	$Z(x)$	1.43	1.46	1.40	1.50	1.41	1.43
	$\hat{Z}^{\dagger}(x)$	1.38 (0.003)	1.41 (0.003)	1.36 (0.002)	1.42 (0.004)	1.37 (0.002)	1.38 (0.001)
	$\hat{Z}^{\dagger}_{C_2}(x)$	1.43 (0.01)	1.46 (0.01)	1.40 (0.002)	1.48 (0.02)	1.40 (0.01)	1.42 (0.004)
	$\hat{Z}_{C_2}(x)$	1.32 (0.01)	1.42 (0.01)	1.40 (0.01)	1.14 (0.03)	1.22 (0.01)	1.34 (0.01)

As a conclusion, this method yields for both the 'random process' and the 'random storm' models, as expected, very precise predictors for ungauged locations. In fact, they perform much better than the pairwise predictors on the unit Fréchet scale. This approach has however the disadvantage that extremes of bivariate Gaussian distributions occur asymptotically independently for any fixed correlation (Sibuya (1960)). In applications such as the rainfall data this might not be realistic.

4.3.3 Alternative predictor for max-stable processes

It might be worth looking at another more intuitive predictor that may improve on the pairwise predictors on the unit Fréchet scale. The approach, analogue to the way max-stable processes are constructed, is to consider a predictor at x of the form

$$\hat{Z}^{\ddagger}(x) = \max_{i=1,\dots,n} \{w(x, x_i)Z(x_i)\},$$

where the weights are set to be monotonic increasing in $||x_i - x||$ in order to give more weight to observations made closer to the ungauged location. An appropriate choice of weights may therefore be the second weight function defined in (3.10).

This approach was tested on both the 'random process' and the 'random storm' model, $10,000$ simulations were performed for different parameter settings (Tables 4.9 and 4.10).

For the 'random process' model, the predictors $\hat{Z}^{\ddagger}(x)$ are of very low quality for all covariance functions and both distributions of the gauged locations. In particular, observe that, counterintuitively, the median predictors and thus the median square prediction errors get less precise as the number of gauged locations increases.

Table 4.9: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 10,000 simulations of one observation at 6, 15, 25 and 50 locations, 'random process' model.

		Regular			Uniform		
n		Exponential	Whittle-Matérn	Cauchy	Exponential	Whittle-Matérn	Cauchy
6	$Z(x)$	1.46	1.42	1.47	1.46	1.48	1.42
	$\hat{Z}^{\ddagger}(x)$	0.94 (0.30)	0.91 (0.31)	0.93 (0.30)	0.91 (0.32)	0.86 (0.39)	0.91 (0.27)
15	$Z(x)$	1.48	1.45	1.47	1.45	1.47	1.47
	$\hat{Z}^{\ddagger}(x)$	0.77 (0.45)	0.77 (0.43)	0.73 (0.45)	0.63 (0.57)	0.60 (0.68)	0.58 (0.68)
25	$Z(x)$	1.43	1.42	1.46	1.42	1.39	1.47
	$\hat{Z}^{\ddagger}(x)$	0.66 (0.50)	0.65 (0.50)	0.66 (0.52)	0.56 (0.65)	0.50 (0.67)	0.51 (0.73)
50	$Z(x)$	1.43	1.43	1.46	1.47	1.40	1.43
	$\hat{Z}^{\ddagger}(x)$	0.58 (0.62)	0.58 (0.60)	0.59 (0.63)	0.43 (0.90)	0.44 (0.74)	0.43 (0.81)

The obtained predictors for an ungauged location of the 'random storm' have the same properties; the predictors are constantly underestimating the true values and are even less precise if there are more gauged locations. It is interesting to notice that unlike the pairwise predictors on the unit Fréchet scale, there is only little difference between the predictors for processes with different values of σ^2, the variance of the underlying normal distribution.

This analysis indicates thus that the ad-hoc predictor $\hat{Z}^{\ddagger}(x)$ of $Z(x)$ is not powerful. The pairwise predictor on the unit Fréchet scale seems to be more appropriate.

Table 4.10: Median predictors for uniformly, randomly generated locations (empirical MDSE) based on 10,000 simulations of one observation at 6, 15, 25 and 50 locations, 'random storm' model.

		Regular			Uniform		
n		$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$	$\sigma^2 = 2.5$	$\sigma^2 = 5$	$\sigma^2 = 10$
6	$Z(x)$	1.43	1.45	1.39	1.43	1.43	1.44
	$\hat{Z}^{\ddagger}(x)$	1.07 (0.23)	0.97 (0.26)	0.86 (0.25)	0.95 (0.34)	0.95 (0.22)	0.74 (0.37)
15	$Z(x)$	1.44	1.41	1.44	1.42	1.44	1.41
	$\hat{Z}^{\ddagger}(x)$	0.77 (0.44)	0.71 (0.43)	0.70 (0.44)	0.69 (0.51)	0.57 (0.63)	0.48 (0.70)
25	$Z(x)$	1.45	1.42	1.44	1.42	1.43	1.40
	$\hat{Z}^{\ddagger}(x)$	0.66 (0.53)	0.64 (0.53)	0.63 (0.52)	0.52 (0.73)	0.55 (0.65)	0.52 (0.65)
50	$Z(x)$	1.43	1.41	1.43	1.43	1.46	1.43
	$\hat{Z}^{\ddagger}(x)$	0.58 (0.62)	0.55 (0.59)	0.57 (0.64)	0.46 (0.82)	0.45 (0.85)	0.39 (0.88)

4.4 Analysis of the extreme precipitation data

The observations of extreme rainfall are reported at the same stations as the means precipitations discussed in section 3.6. There are again 47 observations measured at each location. In this context, annual summer and winter maxima are observed. These spatial data can thus be modelled by the means of max-stable processes.

Both the 'random process' and the 'random storm' models are considered. Before a model can be fitted to the data, they need to be transformed to the unit Fréchet scale. This is done by considering the use of generalized extreme value marginals. More specifically, consider the bijection $(Y_{ji}, Y_{jk}) = g(Z_{ji}, Z_{jk})$, where the inverse function is

$$Z_{ji} = \left\{1 + \frac{\xi_i(Y_{ji} - \mu_i)}{\lambda_i}\right\}_+^{1/\xi_i} \quad Z_{jk} = \left\{1 + \frac{\xi_k(Y_{jk} - \mu_k)}{\lambda_k}\right\}_+^{1/\xi_k}.$$

The constants μ, ξ and $\lambda > 0$ ensure that Z_j is unit Fréchet distributed. Denoting $f_{Z_{ji},Z_{jk}}(z_{ji}, z_{jk})$ the density of the given model, the bivariate density can be written as

$$f_{Y_{ji},Y_{jk}}(y_{ji}, y_{jk}) = f_{Z_{ji},Z_{jk}}\{g^{-1}(y_{ji}, y_{jk})\}|J(y_{ji}, y_{jk})|,$$

where the Jacobian of the transformation equals

$$|J(y_{ji}, y_{jk})| = \frac{1}{\lambda_i\lambda_k}\left\{1 + \frac{\xi_i(Y_{ji} - \mu_i)}{\lambda_i}\right\}_+^{1/\xi_i - 1}\left\{1 + \frac{\xi_k(Y_{jk} - \mu_k)}{\lambda_k}\right\}_+^{1/\xi_k - 1}.$$

This configuration allows the pairwise likelihood to estimate the location, scale and shape parameters of the marginal distribution at each location as well as the unknown quantities of the underlying model.

Tables 4.11 and 4.12 indicate the parameter estimates and their estimated standard errors for the different underlying processes. In contrast to the simulation study, it is now possible to fit a two-dimensional 'random storm' model to the data.

In case of the 'random process' model, for a fixed covariance function the parameter estimates are very similar for the two seasons. However, for the 'random storm' model,

there is an interesting difference between the estimates obtained. Actually, the estimated values of the covariance matrix are smaller if the winter extremes are considered. This means that the diffusion in space is different in summer than it is in winter.

Table 4.11: Parameter estimation for the 'random process' model using pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates.

Season	Covariance function	$\hat{\beta}$ (se($\hat{\beta}$))	$\hat{\nu}$ (se($\hat{\nu}$))
Summer	Exponential	26.1 (5.7)	0.89 (0.12)
	Whittle-Matérn	31.9 (-)	0.40 (-)
	Cauchy	4.94 (1.10)	0.30 (0.09)
Winter	Exponential	21.00 (6.8)	0.83 (0.14)
	Whittle-Matérn	28.6 (-)	0.36 (-)
	Cauchy	4.33 (1.10)	0.33 (0.14)

Table 4.12: Parameter estimation for the 'random storm' model using pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates.

Season	$\hat{\sigma}_{11}$ (se($\hat{\sigma}_{11}$))	$\hat{\sigma}_{12}$ (se($\hat{\sigma}_{12}$))	$\hat{\sigma}_{22}$ (se($\hat{\sigma}_{22}$))
Summer	233.6 (2.8)	14.3 (1.9)	101.2 (3.8)
Winter	160.2 (4.8)	-3.9 (1.8)	54.2 (1.9)

The goodness of fit for the different models may be compared through the composite likelihood information criterion defined in (2.3). Again, if the Whittle-Matérn covariance function is considered, the information criterion is obtained by a model fitting in two steps, estimating first the parameter ν and then estimating the range parameter β along with its standard error.

The information criterion scores are shown in Table 4.13 and indicate that the best model for both seasons is a 'random process' with underlying Whittle-Matérn covariance function although the other 'random process' models give appropriate fit, too. In contrast, the 'random storm' model, having larger values for the composite likelihood criterion, does not fit the data as well as the 'random process' model.

Once the model is fitted, the objective is to check the prediction properties of the different models. Therefore 10,000 ungauged locations are uniformly but randomly generated and the unobserved values at these locations are predicted based on the given observations. As for the simulation study, the predictions are based on the most recent measurements, meaning on observation 47 only. For a fixed season, the median pairwise predictors (Table 4.14) are very similar for all the 'random process' models. In agreement with the simulation study, the second weight function yield the largest predictor. Generally, the predictors based on the 'random storm' model are larger than the ones yielded by the

'random process' model.

Table 4.13: Composite likelihood information criterion scores for the 'random process' and the 'random storm' models.

Season	Model	Covariance function	CLIC
Summer	'random process'	Exponential	489908
		Whittle-Matérn	**489861**
		Cauchy	490124
	'random storm'		496961
Winter	'random process'	Exponential	497324
		Whittle-Matérn	**497281**
		Cauchy	497423
	'random storm'		505124

Table 4.14: Median predictors for 10,000 uniformly randomly generated ungauged locations, parameters of the max-stable process estimated through pairwise likelihood.

		'random process'			'random storm'
		Exponential	Whittle-Matérn	Cauchy	
Summer	$\hat{Z}_{C_1}(x)$	19.2	19.2	19.2	21.7
	$\hat{Z}_{C_2}(x)$	19.8	19.7	19.8	22.8
	$\hat{Z}_{C_3}(x)$	19.5	19.5	19.4	22.2
Winter	$\hat{Z}_{C_1}(x)$	17.3	17.3	17.3	19.0
	$\hat{Z}_{C_2}(x)$	18.0	17.8	17.8	20.7
	$\hat{Z}_{C_3}(x)$	17.9	17.8	17.8	20.0

In order to judge the quality of the predictors, a cross validation is performed. The predictions are based on the same 34 stations as for the mean data cross-validation. The predictors, along with the true values of the process at the left-out stations are displayed in Figure 4.7. These Figures depict that the properties of the pairwise predictors are very poor. Especially for the 'random process' model, the predictions stick to a single value. Even if the 'random storm' predictors behave in a less erratic way than the 'random process' ones do, they are still far for being appropriate.

The cross-validation shows that the pairwise predictors for max-stable processes are not reliable. They constantly underestimate the true values at the locations. In addition, the values are strongly shrunk to a common value.

The application of the pairwise prediction approach to the rainfall datasets shows that the predictions on the unit Fréchet scale are not precise at all. As the simulation study has shown that the properties of the pairwise predictors are better on the Gaussian scale, it may be essential to transform the data, calculate predictors on this scale and then

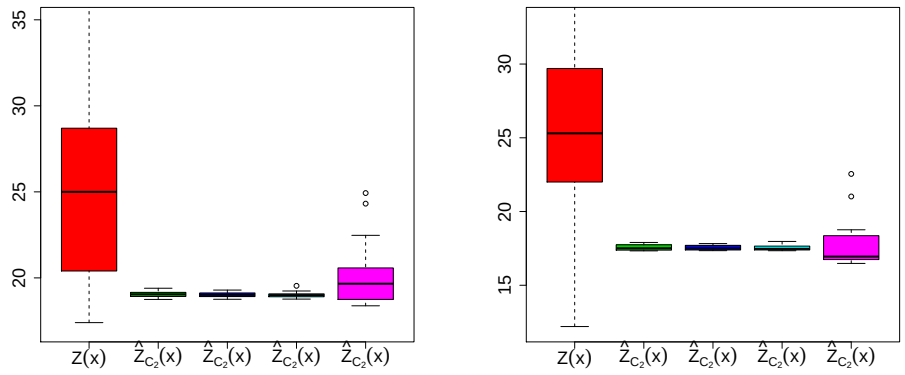

Figure 4.7: Boxplots of true extreme precipitation values (red) and pairwise predictors, $\hat{Z}_{C_2}(x)$, respectively, for 17 randomly chosen stations based on observations at remaining 34 locations. Left panel: summer data, right panel: winter data. Green: exponential covariance function, blue: Whittle-Matérn covariance function, light blue: Cauchy covariance function, purple: 'random storm' model.

back-transform them to the unit Fréchet scale in order to get predictors for max-stable processes.

As then the aim is to model the data as a Gaussian process, the observed maxima need to be transformed. Initially, a generalized extreme value (GEV) model is fitted to the data for each station (Coles (2001)). Recall that the generalized extreme value distribution corresponds to a family of models where the distribution functions are of the form

$$G(z) = \exp\left[-\left\{1+\xi\left(\frac{z-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right] \tag{4.6}$$

defined on the set $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. This model depends on three parameters, the location, scale and shape parameter, μ, σ and ξ, respectively.

Thus, for each station i, a generalized extreme value distribution is fitted to the observations $\{Y_{j,i}\}_{j=1}^{47}$. This yields maximum likelihood estimates $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$ for each station $i \in 1, \ldots, 51$.

Finally, the transformed values of $Y_{j,i}$ can be computed as follows

$$Y_{j,i}^* = F^{-1}\{G(Y_{j,i}; \hat{\theta}_i)\} \dot{\sim} \mathcal{N}(0,1),$$

where G is the cumulative distribution function of a GEV random variable and F is the cumulative distribution function of a standard normal random variable.

Figure 4.8 shows the histogram as well as the QQ-plot of the transformed summer and winter rainfall data. The histograms match well with the theoretical distribution whereas

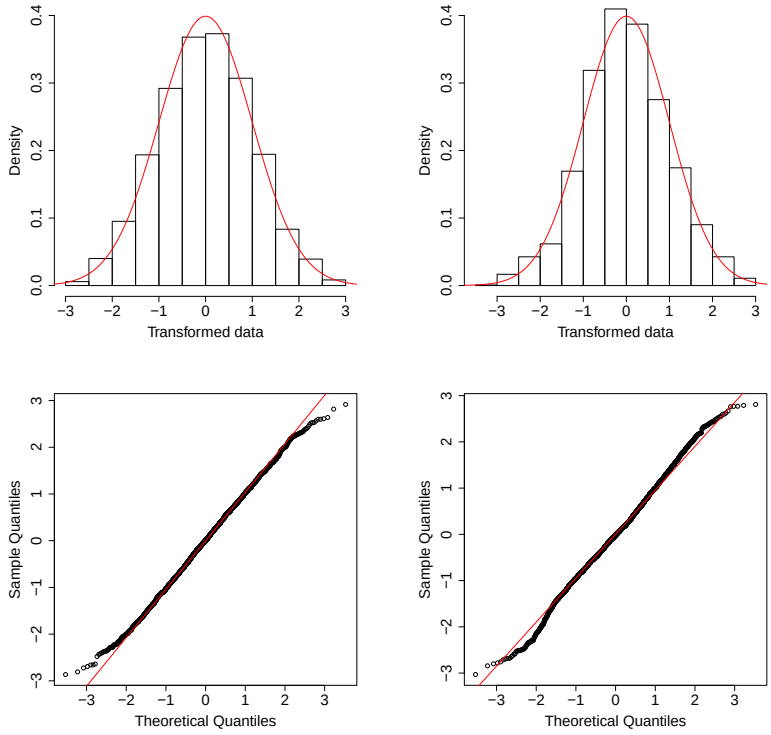

Figure 4.8: Histograms and QQ-plots of transformed summer (left) and winter data (right).

the QQ-plots indicate that the transformed data are slightly heavy-tailed for both data sets.

The objective is now to apply different models to the transformed data in order to obtain predictors for a random station located between the other stations. Initially, assume that the underlying model is the 'random process' model. In this case, the parameters of the Gaussian process are estimated by full and pairwise likelihood for different covariance functions (Table 4.15).

In order to be able to judge to model fit, information criteria are calculated the same way as for the mean precipitation data (see Table 4.16). Concerning the full likelihood approach, both the Whittle-Matérn and the powered exponential covariance functions give good fit. If the models are fitted by pairwise likelihood estimation, the Whittle-Matérn covariance function gives rise to the smallest value of the information criterion and thus the best fit.

For the 'random storm' model, it is not possible to obtain parameter estimates on the Gaussian scale. Therefore the model is fitted on the unit Fréchet scale (for parameter estimates see Table 4.12). As described in section 4.3, the additional parameter t needs to be fixed in order to calculate the best linear predictor on the Gaussian scale. For the extreme value data, two values for t are chosen according to the empirical correlation (Figure 4.9). However, the corresponding correlation functions do not fit the empirical one

Table 4.15: Parameter estimation for the transformed data using full and pairwise likelihood. Standard errors for the pairwise likelihood estimates are obtained through sandwich estimates.

Season			$\hat{\mu}$ (se($\hat{\mu}$))	$\hat{\sigma}^2$ (se($\hat{\sigma}^2$))	$\hat{\beta}$ (se($\hat{\beta}$))	$\hat{\nu}$ (se($\hat{\nu}$))
Summer	Exponential	full	0.0015 (0.11)	1.00 (0.12)	83.0 (26.8)	0.54 (0.03)
		pairwise	0.0013 (0.11)	1.00 (0.12)	73.4 (25.6)	0.59 (0.06)
	Whittle-Matérn	full	0.0014 (0.11)	0.98 (0.11)	141.4 (42.8)	0.235 (0.015)
		pairwise	0.0013 (-)	1.00 (-)	135.5 (-)	0.246 (-)
	Cauchy	full	0.0014 (0.12)	1.10 (0.15)	1.16 (0.13)	0.07 (0.013)
		pairwise	0.0013 (0.11)	1.00 (0.12)	2.34 (0.42)	0.12 (0.025)
Winter	Exponential	full	0.00 (0.11)	1.00 (0.12)	83.0 (29.28)	0.51 (0.03)
		pairwise	0.00 (0.11)	1.00 (0.15)	89.48 (45.8)	0.50 (0.05)
	Whittle-Matérn	full	0.00 (0.10)	0.91 (0.08)	111.5 (26.18)	0.23 (0.014)
		pairwise	0.00 (-)	1.00 (-)	207.0 (-)	0.20 (-)
	Cauchy	full	0.00 (0.12)	1.08 (0.14)	0.99 (0.13)	0.07 (0.013)
		pairwise	0.00 (0.11)	1.00 (0.15)	1.75 (0.33)	0.10 (0.026)

very well as its shape is not close to being quadratic.

Based on the parameter estimates, predictors for uniformly randomly generated ungauged locations can be computed for both models. Table 4.17 summarizes the back-transformed predictors. The full likelihood predictors are generally different than the pairwise predictors. For the 'random storm' model, they are constantly lower than the pairwise predictors. This is unexpected as the simulation study indicated that the pairwise predictors usually yield smaller estimates. The pairwise predictors do not depend on the value of t in the 'random storm' model and are thus equal. It is however interesting how much the value of the back-transformed best linear predictor changes as a function of t. If the dependence range is shorter ($t = 5$), the predictor is very large whereas for longer range dependence it is even smaller than the ones obtained based on the 'random process' model.

In order to confirm these results, a cross validation is carried out. Again, only the second choice of weights is considered, as it has so far been suggested to be the most reasonable one.

Figure 4.10 indicates that the back-transformed predictors based on the 'random process' model are much more precise than the ones obtained on the unit Fréchet scale, but compared to the true values they are still not very precise for neither of the assumed correlation structures. Moreover, the best linear predictor is preferable to the pairwise predictor in terms of the spread of the values. The pairwise predictors are still shrunk to the median, but less than when predicting on the unit Fréchet scale. The best linear predictor for the 'random storm' model is very sensitive to the value of t, the interquartile range is too large for $t = 10$ comparing to the true values. The predictors stemming from this model based on pairwise likelihood behave however well. Their distribution is nearly the same as the

Table 4.16: Information criterion (AIC and CLIC) scores for models fitted by full and pairwise likelihood, respectively.

Season	Covariance function	AIC	CLIC
Summer	Exponential	3949	3837
	Whittle-Matérn	**3944**	**3768**
	Cauchy	3985	3822
Winter	Exponential	4111	4737
	Whittle-Matérn	**4110**	**4675**
	Cauchy	4136	4723

distribution of the back-transformed best linear predictors obtained based on the 'random process' model.

The data analysis leads to the conclusion that pairwise predictors calculated on the unit Fréchet scale have poor properties. They do not manage to mimic the behavior of the true values. More precisely, the prediction is constantly underestimating the true value and sticks to a common value. When the data are transformed to the Gaussian scale and predictors computed, results become more appropriate. For the 'random process' model, the best linear predictor is most precise. In contrast, the best linear predictors based on 'random storm' model are strongly influenced by the value of t. Astonishingly, better results are obtained based on pairwise likelihood approaches for this model.

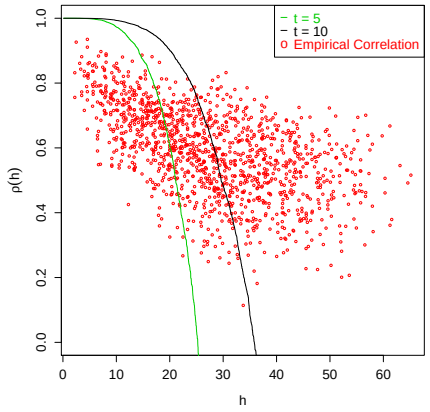

Figure 4.9: Correlation functions for $t = 5$, $t = 10$ and empirical correlation for the 'random storm' model.

Table 4.17: Back-transformed median predictors for 10,000 uniformly randomly generated ungauged locations, parameters of the Gaussian process estimated.

		'random process'			'random storm'	
		Exponential	Whittle-Matérn	Cauchy	$t = 5$	$t = 10$
Summer	$\hat{Z}^{\dagger}_{\ell}(x)$	21.4	21.4	21.4	24.9	20.8
	$\hat{Z}^{\dagger}_{C_1}(x)$	23.6	23.6	23.6	21.6	21.6
	$\hat{Z}^{\dagger}_{C_2}(x)$	23.1	23.1	23.1	21.6	21.6
	$\hat{Z}^{\dagger}_{C_3}(x)$	23.0	23.0	23.1	21.6	21.6
Winter	$\hat{Z}^{\dagger}_{\ell}(x)$	21.9	21.9	21.8	24.5	21.2
	$\hat{Z}^{\dagger}_{C_1}(x)$	22.4	22.4	22.4	21.4	21.4
	$\hat{Z}^{\dagger}_{C_2}(x)$	22.5	22.5	22.5	21.4	21.4
	$\hat{Z}^{\dagger}_{C_3}(x)$	22.8	22.8	22.8	21.4	21.4

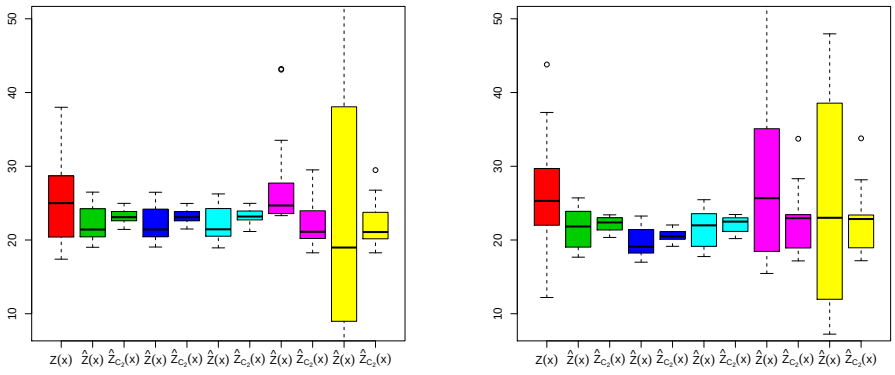

Figure 4.10: Boxplots of true extreme precipitation values (red), best linear and pairwise predictors, $\hat{Z}^{\dagger}_{fe}(x)$ and $\hat{Z}^{\dagger}_{C_2}(x)$, respectively, for 17 randomly chosen stations based on observations at remaining 34 locations. Left panel: summer data, right panel: winter data. Green: exponential covariance function, blue: Whittle-Matérn covariance function, light blue: Cauchy covariance function, purple: 'random storm' model with $t = 5$, yellow: 'random storm' model with $t = 10$.

Conclusion

As it is not possible to obtain the full conditional density for max-stable processes, predictors for ungauged locations of such processes cannot be calculated by means of standard Geostatistics procedures. Moreover, as standard approaches in spatial statistics are all based on Gaussian distributions, predictions for extreme observations need to be computed in a different way.

It is already known that the composite likelihood approach allows reliable model fitting for max-stable processes (Padoan *et al.* (2009)). Therefore, the performance of this approach is explored when it comes the predict observations at ungauged locations. The behavior of such predictors can be investigated when looking at Gaussian processes as the full conditional density is available for such processes. Based on traditional Geostatistics, it is possible to calculate the best linear predictor as well as pairwise predictors which depend on a weight function. These quantities can be compared through mean square prediction errors. A simulation study allows to conclude that for Gaussian processes, pairwise predictors induce a loss in precision but that they build nevertheless a reliable alternative to the best linear predictors in cases where the full conditional density is not available. Furthermore, the quality of the pairwise predictors depends on the given weight function. The weight function, of those studied, yielding the best pairwise predictor is the one which is inversely proportional to the distance between the ungauged location and the considered location.

The properties of the pairwise predictors can also be compared to the ones of the best linear predictor in a data analysis. The considered dataset consists of annual mean summer precipitations observed at 51 location in Switzerland for a period of 47 years. Predicting $10,000$ observations at ungauged locations shows that the mean square prediction error is likely to be very inflated if the pairwise predictors are considered. In addition, a cross-validation indicates that the pairwise predictors are slightly shrunk to a common value compared to the true values. However, this analysis needs to be considered carefully as the data do not seem to be properly normally distributed. In fact, there may be different factors causing the non-normality of the means; there may be strong dependence in the daily rainfall as well as dependence caused by considering the different locations simultaneously.

In a second part, prediction for max-stable processes based on pairwise approaches is investigated. Both the 'random process' and the 'random storm' models are fitted

to simulated data and predictions based on it. The quality of predictors based on this approach is analyzed through median square prediction errors as the Fréchet distribution does not have finite moments. Considering thus the median predictors and the median square prediction errors, it appears that the loss in precision is considerably higher than for Gaussian processes. The true values are constantly underestimated. The predictors can be improved by having more data, that means more locations where the process is observed. However, as extreme data are scarce this may be problematic. An alternative is to predict on the Gaussian scale using transformed data. Back-transforming the predictors obtained, yields very precise results although this approach has the major drawback that, for any fixed correlation, extremes of bivariate Gaussian distributions are asymptotically independent (Sibuya (1960)).

The approaches investigated by the simulation study are again applied to a dataset. Extreme observations are observed at the same locations as the mean summer precipitation data. Two datasets are available, the summer and winter extremes over a period of 47 years. Once the models are fitted it is possible to calculate predictors for ungauged locations. The results can be checked by performing a cross-validation. Especially for the 'random process' model, the predictors are concentrated to a common value and do not correspond to the true value. The transformation of the data to the Gaussian scale yields a significant improvement in the predictors. As it is possible to obtain the best linear predictor, the back-transformed predictors for the 'random process' model are of better quality than the pairwise predictors. Both predictors allow more precise prediction than on the unit Fréchet scale. Due to an additional parameter in the 'random storm' model, the best linear predictors give not very precise predictions but vary a lot. In this case the pairwise predictors are preferable as they give results that are nearly equivalent to back-transformed linear predictors for the 'random process' model.

It would be very interesting to pursue the composite likelihood approach and improve it. In this context, it may be reasonable to consider the triplewise density. Obtaining the closed form for the triplewise density of a max-stable process is a non-trivial problem. Hüsler and Reiss (1989) suggest a way to calculate higher order densities for max-stable processes.

In addition, it may be worthwhile to compare the pairwise likelihood approach to Bayesian approaches such as predictive likelihood (Davison (1986)). More specifically, approximate conditional prediction may lead to reasonable predictors for max-stable processes. There may be a link between these procedures as predictive likelihood approaches are based on repeating likelihood maximization of a posterior likelihood.

Acknowledgements

I would like to thank Professor Anthony C. Davison for giving me the opportunity to do my Master Thesis at Imperial College London. Moreover, he was always disposed to answer my questions. I also really appreciated the support by Professor Alastair Young. During my time in London he answered all my questions and gave me useful advice.

Thanks also to Mathieu Ribatet, who always manages to help me patiently with programming issues. Finally, I would also like to thank Simone Padoan who provided the details about the datasets.

Bibliography

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2008) Estimating space and space-time covariance functions: a weighted composite likelihood approach. Technical report, Universitat Jaume I 117.

Coles, S. G. (2001) *An Introduction to Statistical Modelling of Extreme Values*. New York: Springer.

Cox, D. and Reid, N. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91, 3**, 729–737.

Cressie, N., Zhang, J. and Craigmile, P. (2005) Geostatistical prediction of spatial extremes and their extent. *Geostatistics for environmental applications, Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications* pp. 27–37.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*. New York: Wiley.

Davison, A. C. (1986) Approximate predictive likelihood. *Biometrika* **73**, 323–332.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A. C. and Gholamrezaee, M. M. (2009) Geostatistics of extremes. *EPFL* **Preprint**, stat.epfl.ch.

Diggle, P. J. and Ribeiro, P. J. (2007) *Model-based Geostatistics*. New York: Springer.

Grünenfelder, C. (2009) Composite marginal likelihoods. Technical report, École polytéchnique fédérale de Lausanne.

de Haan, L. (1984) A spectral representation for max-stable processes. *Annals of Probability* **12**, 1194–1204.

Heyde, C. (1997) *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimatio*. New York: Springer.

Hüsler, J. and Reiss, R.-D. (1989) Maxima of normal random vectors: between independence and complete dependence. *Statistics and Probability Letters* **7**, 283–286.

Padoan, S., Ribatet, M. and Sisson, S. (2009) Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association* **to appear**.

Schabenberger, O. and Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis*. Chapman and Hall / CRC.

Schlather, M. (2002) Models for stationary max-stable random fields. *Extremes* **5**, 33–44.

Sibuya, M. (1960) Bivariate Extreme Statistics, I. *Annals of the Institute of Statistical Mathematics* **11**, 195–210.

Smith, R. L. (1990) Max-stable processes and spatial extremes. *Unpublished* .

Varin, C. (2007) On composite marginal likelihoods. Technical report, University Ca' Foscari San Giobbe.

Varin, C. and Vidoni, P. (2005) A note on composite likelihood inference and model selection. *Biometrika* **92, 3**, 519–528.

Wackernagel, H. (2003) *Multivariate Geostatistics: An Introduction with Applications*. New York: Springer.

Appendix

A.1 Gradient of pairwise log likelihood for Gaussian processes

The covariance matrix of the bivariate normal distribution and its inverse are denoted as

$$\Sigma = \begin{bmatrix} \sigma^2 & \gamma(h) \\ \gamma(h) & \sigma^2 \end{bmatrix} \qquad \Sigma^{-1} = \frac{1}{|\Sigma|} \begin{bmatrix} \sigma^2 & -\gamma(h) \\ -\gamma(h) & \sigma^2 \end{bmatrix}$$

where the determinant of Σ is $|\Sigma| = \sigma^4 - \gamma(h)^2$ and h corresponds to the euclidean distance between two positions where the Gaussian process is observed. This allows one to write the pairwise log likelihood for y_i, y_j

$$\begin{aligned} \log f(y_i, y_j) = \quad & -\log 2\pi - \frac{1}{2}\log|\Sigma| \\ & - \frac{\sigma^2(y_i-\mu)^2 - 2\gamma(h)(y_i-\mu)(y_j-\mu) + \sigma^2(y_j-\mu)^2}{2\{\sigma^4 - \gamma(h)^2\}}. \end{aligned}$$

In order to get an estimate for the covariance matrix of the maximum composite likelihood estimate $\hat{\theta}_C$, the matrix J needs to be estimated as described in section 1.1. Therefore the gradient of the pairwise log likelihood needs to be calculated.

The gradients with respect to μ and σ^2 can be obtained as functions of the covariance function $\gamma(h)$.

$$\begin{aligned} \nabla_\mu \log f(y_i, y_j) &= \frac{(y_i-\mu) + (y_j-\mu)}{\sigma^2 + \gamma(h)} \\ \nabla_{\sigma^2} \log f(y_i, y_j) &= -\frac{1}{\sigma^2} + \frac{\sigma^2(y_i-\mu)^2 - 2(y_i-\mu)(y_j-\mu)\gamma(h) + \sigma^2(y_j-\mu)^2}{2\sigma^2\{\sigma^4 - \gamma(h)^2\}} \end{aligned}$$

For the powered exponential covariance function the following results are obtained for

the gradients with respect to β and ν:

$$\begin{aligned}
\nabla_\beta \log f(y_i, y_j) &= \frac{\gamma(h)\partial\gamma(h)/\partial\beta}{\sigma^4 - \gamma(h)^2} - \sigma^2\{(y_i-\mu)^2 + (y_j-\mu)^2\}\frac{\gamma(h)\partial\gamma(h)/\partial\beta}{\{\sigma^4-\gamma(h)^2\}^2} \\
&\quad + (y_i-\mu)(y_j-\mu)\frac{\partial\gamma(h)/\partial\beta\{\sigma^4-\gamma(h)^2\} + 2\gamma(h)^2\partial\gamma(h)/\partial\beta}{\{\sigma^4-\gamma(h)^2\}^2} \\
\nabla_\nu \log f(y_i, y_j) &= \frac{\gamma(h)\partial\gamma(h)/\partial\nu}{\sigma^4 - \gamma(h)^2} - \sigma^2\{(y_i-\mu)^2 + (y_j-\mu)^2\}\frac{\gamma(h)\partial\gamma(h)/\partial\nu}{\{\sigma^4-\gamma(h)^2\}^2} \\
&\quad + (y_i-\mu)(y_j-\mu)\frac{\partial\gamma(h)/\partial\nu\{\sigma^4-\gamma(h)^2\} + 2\gamma(h)^2\partial\gamma(h)/\partial\nu}{\{\sigma^4-\gamma(h)^2\}^2},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial\gamma(h)}{\partial\beta} &= \nu\gamma(h)\frac{h^\nu}{\beta^{\nu+1}} \\
\frac{\partial\gamma(h)}{\partial\nu} &= -\gamma(h)\log\left(\frac{h}{\beta}\right)\left(\frac{h}{\beta}\right)^\nu.
\end{aligned}$$

Concerning the Cauchy covariance function, the following identities hold:

$$\begin{aligned}
\nabla_\beta \log f(y_i, y_j) &= \frac{2\nu h^2\gamma(h)^2}{\beta^3\left\{1+(\frac{h}{\beta})^2\right\}\{\sigma^4-\gamma(h)^2\}}\left[1 + \frac{(y_i-\mu)(y_j-\mu)}{\gamma(h)}\right. \\
&\quad \left. - \frac{\sigma^2(y_i-\mu)^2 - 2(y_i-\mu)(y_j-\mu)\gamma(h) + \sigma^2(y_j-\mu)^2}{\sigma^4-\gamma(h)^2}\right] \\
\nabla_\nu \log f(y_i, y_j) &= -\frac{\gamma(h)^2\log\left\{1+(\frac{h}{\beta})^2\right\}}{\{\sigma^4-\gamma(h)^2\}}\left[1 + \frac{(y_i-\mu)(y_j-\mu)}{\gamma(h)}\right. \\
&\quad \left. - \frac{\{\sigma^2(y_i-\mu)^2 - 2(y_i-\mu)(y_j-\mu)\gamma(h) + \sigma^2(y_j-\mu)^2\}}{\sigma^4-\gamma(h)^2}\right]
\end{aligned}$$

Finally, for the Whittle-Matérn covariance function the gradient with respect to β can be obtained as follows

$$\begin{aligned}
\nabla_\beta \log f(y_i, y_j) &= \frac{\gamma(h)\partial\gamma(h)/\partial\beta}{\sigma^4 - \gamma(h)^2} - \sigma^2\{(y_i-\mu)^2 + (y_j-\mu)^2\}\frac{\gamma(h)\partial\gamma(h)/\partial\beta}{\{\sigma^4-\gamma(h)^2\}^2} \\
&\quad + (y_i-\mu)(y_j-\mu)\frac{\partial\gamma(h)/\partial\beta\{\sigma^4-\gamma(h)^2\} + 2\gamma(h)^2\partial\gamma(h)/\partial\beta}{\{\sigma^4-\gamma(h)^2\}^2},
\end{aligned}$$

where

$$\frac{\partial\gamma(h)}{\partial\beta} = -\frac{\gamma(h)\nu}{\beta} - \frac{\sigma^2 h^\nu}{\beta^{\nu+2}}\frac{2^{1-\nu}}{\Gamma(\nu)}\left[\frac{h}{2\sin(\pi\nu)}\left\{\sin(\pi(1+\nu))K_{1+\nu}\left(\frac{h}{\beta}\right) + \sin(\pi(\nu-1))K_{\nu-1}\left(\frac{h}{\beta}\right)\right\}\right],$$

where $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is the modified Bessel function of third kind of order $\nu > 0$.

As the modified Bessel function of third kind of order $\nu > 0$ is not differentiable with respect to the parameter ν, it is not possible to obtain the gradient of the pairwise log likelihood with respect to ν.