Approximate Bayesian Computing for Spatial Extremes

Robert J. Erhardt

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2012

Approved by:

Richard L. Smith, Advisor

Lawrence Band, Reader

Joseph Ibrahim, Reader

Chuanshu Ji, Reader

Haipeng Shen, Reader

ABSTRACT

ROBERT J. ERHARDT: Approximate Bayesian Computing for Spatial Extremes.
(Under the direction of Richard L. Smith.)

Statistical analysis of max-stable processes used to model spatial extremes has been limited by the difficulty in calculating the joint likelihood function. This precludes all standard likelihood-based approaches, including Bayesian approaches. Here we present a Bayesian approach through the use of approximate Bayesian computing. This circumvents the need for a joint likelihood function and instead relies on simulations from the (unavailable) likelihood. This method is compared with an alternative approach based on the composite likelihood. When estimating the spatial dependence of extremes, we demonstrate that approximate Bayesian computing can provide estimates with a lower mean square error than the composite likelihood approach, though at an appreciably higher computational cost.

As this approach very naturally incorporates parameter uncertainty into predictions, it is well suited for use in pricing weather derivatives to manage environmental risks. We discuss the construction and pricing of such weather derivatives. The method described utilizes results from spatial statistics and extreme value theory to first model extremes in the weather as a max-stable process, and then use these models to simulate payments for a general collection of weather derivatives. These simulations capture the spatial dependence of payments. Incorporating results from catastrophe ratemaking, we show how this method can be used to compute risk loads and premiums for weather derivatives which are renewal-additive.

We illustrate the performance of the approximate Bayesian computing method and weather derivative pricing with applications to United States temperature data. The first application considers pricing weather derivatives for temperature extremes in the Midwestern United States. The second application demonstrates the use of the approximate Bayesian computing method in estimating the risk of crop loss due to an unlikely freeze event in northern Texas.

Acknowledgments

To Mar and Nolan
with thanks to Richard
along with Larry, Joseph, Chuanshu, and Haipeng.

Contents

List of Figures

List of Tables

1

Introduction

Modeling of spatial extremes is motivated by the need to model and predict environmental extreme events such as hurricanes, floods, droughts, heat waves, and other high impact events. Though the data have a natural spatial domain, standard spatial statistics methods may fail to accurately model extremes. Models specifically designed for extremes are better suited. The urgency of focusing on extremes is increased when one considers the potential influence of climate change on the probability of such high impact events. We consider point referenced data, usually taken as daily or hourly measurements $y_{t,d}$ at locations $d = 1, ..., D$ for time points $t = 1, ..., T$. When modeling extremes, as a first step one takes block maxima over some temporal block (usually one year) and obtains block maxima data $y_{i,d}$ where i is the block. In an environmental setting, for example, the data might be annual maxima at each of D locations.

For a single location, univariate extreme value theory provides a full range of tools to analyze the data. This theory is well developed and documented (Coles, 2001; de Haan and Ferreira, 2006; Embrechts, et. al., 1999; Resnick, 1987, 2007). When one considers several locations at once, multivariate extreme value theory is a natural extension. Multivariate models often work well for lower dimensions, but if the data have a natural spatial domain and the dimension grows rapidly, spatial extreme value theory becomes useful. Spatial extremes are the infinite dimensional generalization of multivariate extremes. The goal then

is to fit these block maxima data to a spatial process model so that the spatial dependence may be estimated. One promising class of models are max-stable processes. These arise as the limiting distribution of the maxima of independent and identically distributed random fields. A number of max-stable process models have been described (Schlather, 2002; Kabluchko et. al., 2009) and one unpublished model was described by Smith in 1990. The statistical analysis of these models is limited by the unavailability of the joint likelihood function. However, the bivariate distributions are available in closed-form. This allows one to write down the pairwise log-likelihood, which is the sum (taken over all unique pairs of locations) of all bivariate log-likelihoods, and is thus also a composite log-likelihood. Numerical maximization of the composite likelihood yields estimates of the parameters which are consistent and asymptotically normal (Padoan et. al., 2010; Lindsay, 1988). Maximum composite likelihood estimation has been the only method so far for analyzing max-stable processes which is widely applicable, implemented computationally (`R` package `SpatialExtremes`), and for which a viable asymptotic theory exists.

In this dissertation we develop a Bayesian alternative for analyzing the dependence of spatial extremes. It circumvents the need for the joint likelihood, and instead relies only on simulations. This approach, termed approximate Bayesian computing, has been successfully applied in many areas, including extreme values (Bortot et. al., 2007). We show three implementations of the approximate Bayesian computing approach for analyzing spatial extremes. The first two rely on the bivariate distribution function, and like the composite likelihood approach they consider the spatial dependence through all unique pairs of locations. The third and most successful approach extends beyond pairs, and is able to consider higher order k-tuples for $k \geqslant 3$. This feature is an important benefit of the approximate Bayesian computing approach over all pairwise approaches. We show that the approximate Bayesian computing method can result in a lower mean square error compared to the competing composite likelihood approach when estimating the spatial dependence.

We also discuss how this Bayesian approach naturally incorporates parameter uncertainty into predictions, which is a central task in the field of extremes.

The method is computationally intensive, but open to a number of computational enhancements as well. The simplest implementation based on rejection sampling can be done in parallel, and we demonstrate this. A more efficient implementation takes advantage of adaptive computing, in which the sampler targets regions of the parameter space which have shown greater promise. Not only does this approach reduce the computational cost, but it also makes the choice of prior distribution an easier one. We advocate choosing minimally informative, independent uniform priors on the natural parameters space Φ, and show how adaptive approximate Bayesian computing can smoothly move from these minimally informative priors to the target posterior distribution at a reasonable computational cost. Furthermore, computational considerations of various enhancements to the algorithm are discussed.

We connect the statistical methodology with an application of rising interest in the insurance industry - how to price weather derivatives for use as a risk management tool. Weather derivatives are contingent contracts whose payments are determined by the difference between some underlying weather measurement and a pre-specified strike value. They provide a useful risk management tool for any party facing weather risk. They also provide investments which are often uncorrelated with more traditional financial instruments, allowing investors to diversify. The first weather derivative was developed in 1996, and by 1999 derivatives and their options were being traded on the Chicago Mercantile Exchange (Kunreuther and Michel-Kerjan, 2009).

Finally, we demonstrate the use of max-stable processes for pricing weather derivatives for extremes, both from a frequentist and Bayesian perspective. The Bayesian approach has the added advantage of naturally incorporating parameter uncertainty into estimated risk measures and premiums, resulting in larger but more accurate estimates of risk. In a second

application, we demonstrate how the approach can be used to extrapolate models from regions of data to regions of interest, and serve as a rare event simulator. These simulators often serve as an entry point to catastrophe ratemaking in the insurance industry.

2

Extremes

2.1 Univariate and Multivariate Extremes

Let $Y_1, ..., Y_n$ be univariate i.i.d. replicates from some distribution function F, and define $M_n = \max(Y_1, ..., Y_n)$ as the maximum of the n random variables. The distribution of M_n can be obtained exactly assuming F is known. In practice, then one could estimate F from all of the data $Y_1, ..., Y_n$ and estimate the distribution of M_n as $P(M_n \leqslant z) = F^n(z)$, but this approach has two drawbacks. The first is that even minor discrepancies in estimating F result in large discrepancies in F^n, particularly in the tails of F. Put another way, why should we expect a model which fits the bulk of data to also be a good fit in the tails? A second drawback is that in the limit as $n \to \infty$, F^n does not converge to a non-degenerate distribution.

Instead, we model renormalized maxima $\frac{M_n - b_n}{a_n}$ for sequences $a_n > 0$ and b_n. If there exist sequences $a_n > 0$ and b_n such that

$$\lim_{n \to \infty} P\left(\frac{M_n - b_n}{a_n} \leqslant z\right) \to G(z)$$

for some non-degenerate distribution function G, then G is a member of one of the three

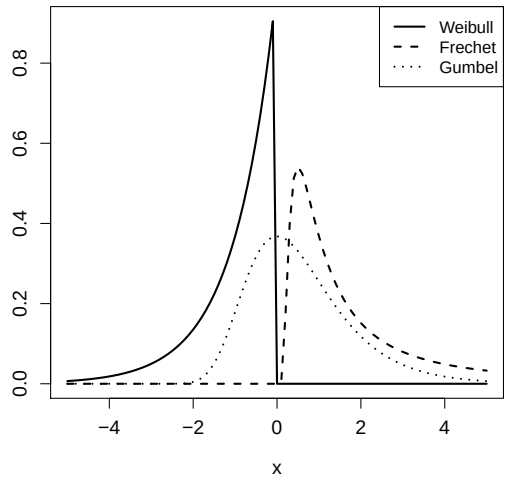

Figure 2.1: Standard unit-Fréchet, Weibull, and Gumbel density functions. These are obtained by setting $a = 1, b = 0$, and $\alpha = 1$ in equations (2.1), (2.2), and (2.3).

following families:

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty \tag{2.1}$$

$$II : G(z) = \begin{cases} 0, & z \leqslant b \\ \exp\{-\left(\frac{z-b}{a}\right)^{-\alpha}\}, & z > b \end{cases} \tag{2.2}$$

$$III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b \\ 1, & z \geqslant b, \end{cases} \tag{2.3}$$

with parameters $a > 0$, b, and $\alpha > 0$ for types II and III. This result is known as the *Three-Types Theorem* (Fisher and Tippett, 1928), and the three families are Gumbel, Fréchet, and Weibull, respectively. The Fréchet case corresponds to a heavy tailed distribution, Gumbel is intermediate, and Weibull has a bounded upper limit. Standard density functions for the three families are shown in Figure 2.1.

A challenge to working with the three-types representation is that in practice, first one

must choose one of the families, and then all subsequent inference assumes the choice to be correct. To avoid this, the three families are often written as a single *Generalized Extreme Value* (GEV) family, with distribution function

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

Here $a_+ = \max(a, 0)$, and μ, σ, and ξ are the location, scale, and shape parameters, respectively (Coles, 2001; Gnedenko, 1943). The sign of the shape parameter ξ corresponds to the three classical extreme values distributions: $\xi > 0$ is Fréchet with support $z \in [\mu - \sigma/\xi, +\infty)$, $\xi < 0$ is Weibull with support $z \in (-\infty, \mu - \sigma/\xi]$, and $\xi \to 0$ is Gumbel with support $z \in (-\infty, +\infty)$. When using this model, one allows the estimate of ξ to guide which of the three types is selected.

The Generalized Extreme Value distribution G has the property of max-stability, understood as follows: if $Y_1, ..., Y_n$ are i.i.d. from G, then $\max(Y_1, ..., Y_n)$ also has distribution G, meaning

$$G^n(A_n z + B_n) = G(z)$$

for appropriate sequences $A_n > 0$ and B_n. In fact, a distribution is max-stable if and only if it is a member of the GEV family (Leadbetter et. al., 1983). If block maxima are taken over a block size large enough to allow the GEV to be a valid approximation, then if one further increased the block size (from monthly to annual maxima, for example) the GEV model would still hold, with only a change in the three parameters. While these results for the GEV family assume i.i.d. data, this assumption can be relaxed and the limiting distribution still holds so long as certain mixing conditions are satisfied (Leadbetter et. al., 1983).

A useful member of the GEV family is the unit-Fréchet distribution, with distribution

function

$$P(Z \leqslant z) = \exp\left(-\frac{1}{z}\right).$$

The simplicity of the distribution function is helpful when one considers multivariate and ultimately spatial extremes. Any member of the GEV family may be transformed to have unit-Fréchet margins as follows: if Z has a GEV distribution, and a new variable U is defined as

$$U = \left(1 + \xi\frac{Z-\mu}{\sigma}\right)^{1/\xi}, \tag{2.4}$$

then U has unit-Fréchet margins. This transformation assumes that the parameters are known. If the parameters are unknown, they may first be estimated and then the transformation to U is taken. Ultimately, when we have extreme values data in a spatial setting, the first step will be to transform data at each location to unit-Fréchet by fitting all marginal distributions. Then we will proceed to analyze the spatial dependence among sites once every location has been transformed. Thus, for the remainder of this dissertation there is no loss of generality when one assumes unit-Fréchet margins.

The GEV model can be fit to observed data using maximum likelihood estimation. Call the parameter vector ϕ. This parameter can be as simple as three fixed parameters, as $\phi = (\mu, \sigma, \xi)$. Alternatively, one can model the GEV parameters using temporal or spatial covariates. A few examples include $\mu = \mu_1 + \mu_2 \cdot t$, where t is time, or $\sigma = \sigma_1 + \sigma_2 \cdot lat + \sigma_3 \cdot lon + \sigma_4 \cdot elev$, which considers effects of latitude, longitude, and elevation on the scale parameter. No matter the structure of the parameter ϕ, define the density function $g(z;\phi) = \frac{d}{dz}G(z;\phi)$. Then, the maximum likelihood estimate of ϕ is

$$\hat{\phi}_{MLE} = \text{argmax}_{\phi} \prod_i g(z \mid \phi) \tag{2.5}$$

This maximization is often done numerically, and has been implemented in a number of software programs including `R` (R Development Core Team, 2010) using the function `fgev`

in the package `evd`. The density function for the fitted model is obtained by plugging in the maximum likelihood estimate as $g(z; \hat{\phi}_{MLE})$.

We may extend this approach to handle multivariate extremes. Let $(X_{i1}, ..., X_{iD})$, $i = 1, ..., n$ be a $D-$dimensional random vector and let $M_n = (M_{n1}, ..., M_{nD})$ be the vector of componentwise maxima, where $M_{nd} = \max(X_{1d}, ..., X_{nd})$ for $d = 1, ..., D$. It is worth noting that M_n will not appear in the data record unless the occurrence times of each element's block maximum happen to coincide. In a spatial context, this vector M_n might refer to the annual maxima of some variable at D locations. A non-degenerate limit for M_n exists if there exist sequences $a_{nd} > 0$ and b_{nd}, $d = 1, ..., D$ such that

$$\lim_{n\to\infty} P\left(\frac{M_{n1} - b_{n1}}{a_{n1}} \leqslant z_1, ..., \frac{M_{nD} - b_{nD}}{a_{nD}} \leqslant z_D\right) = G(z_1, ..., z_D).$$

Then G is a multivariate extreme value distribution (MEVD), and is max-stable if there exist sequences $A_{nd} > 0$, B_{nd}, $d = 1, ..., D$ such that, for any $n > 1$

$$G^n(z_1, ..., z_D) = G(A_{n1}z_1 + B_{n1}, ..., A_{nD}z_D + B_{nD}).$$

The marginal distributions of a multivariate extreme value distribution are all necessarily GEV distributions. Thus, for each margin one can define a transformation like the one shown in equation (2.4) with parameter (μ_d, σ_d, ξ_d) and transform to unit-Fréchet. Since all GEV distributions can be transformed into unit-Fréchet, all MEVD can be transformed into multivariate unit-Fréchet, and thus we may assume, without loss of generality, that all MEVD have unit Fréchet margins. This is because the domain of attraction condition is preserved under monotone transformations of the marginal distributions (Resnick, 1987). Thus for D fixed locations, the joint distribution function can be written as

$$P(Z(x_1) \leqslant z_1, ..., Z(x_D) \leqslant z_D) = \exp\left(-V(z_1, ..., z_D)\right) \tag{2.6}$$

where $V(z_1, ..., z_D)$ is the exponent measure first described by Pickands (1981). This function takes the form

$$V(z) = D \cdot \int_{\Delta_D} \max_{d=1,...,D} \frac{w_d}{z_d} H(\, dw) \tag{2.7}$$

where $\Delta_D = w \in \mathbb{R}_+^D \mid w_1 + ... + w_D = 1$ is the $D-1$ dimensional simplex, and the angular (or spectral) measure H is a probability measure on Δ_D which determines the dependence structure of the random vector. Due to the common marginal distributions, H has moment conditions $\int_{\Delta_D} w_d H(\, w) = 1/D$ for $d = 1, ..., D$. Max-stability implies that for all N,

$$P(Z_1 \leqslant z_1, ..., Z_D \leqslant z_D)^N = \exp(-N \cdot V(z_1, ..., z_D)) = \exp(-V(z_1/N, ..., z_D/N))$$

with the final equality following from the homogeneity property of the exponent measure. The measure also satisfies two bounds: if all locations are independent, $V(z_1, ..., z_D) = 1/z_1 + ... + 1/z_D$; if all locations are totally dependent, $V(z_1, ..., z_D) = \max(1/z_1, ..., 1/z_D)$. Thus, we always have $\max(1/z_1, ..., 1/z_D) \leqslant V(z_1, ..., z_D) \leqslant 1/z_1 + ... + 1/z_D$.

There are two challenges to working with the spectral representation of the joint distribution function shown in equation (2.6). First, even if we assume that a closed form for the exponent measure can be found by solving equation (2.7), the joint density function undergoes a combinatorial explosion as the dimension D increases. Differentiating $\exp(-V)$ with respect to the values $z_1, ..., z_D$ leads to a rapid growth in terms:

- $-V_1 \exp(-V)$ (first partial derivative)
- $(V_1 V_2 - V_{12}) \exp(-V)$ (second partial derivative)
- $(-V_1 V_2 V_3 + V_{12} V_3 + V_{13} V_2 + V_{23} V_1 - V_{123}) \exp(-V)$ (third partial derivative)
- ...

where V_i is the partial derivative of V with respect to z_i. Thus even if a reasonable choice for V can be found, as the dimension D increases one is left with an unwieldy likelihood

function, which may be difficult to maximize. More common, though, is the situation where closed-form expressions for the exponent measure cannot be obtained by solving equation (2.7). This holds for all of the widely used max-stable process models, resulting in an unavailable joint likelihood function.

2.2 Background on Spatial Statistics

The basic object in spatial statistics is a stochastic process $Y(x), x \in X$ where X is a subset of $\mathbb{R}^p$, usually with $p = 2$. Let

$$\delta(x) = \mathbb{E}(Y(x)), \quad x \in X$$

be the mean of the process defined for all of X, and assume that the variance of $Y(x)$ exists everywhere in X. Then the process can be rewritten as

$$Y(x) = \delta(x) + e(x)$$

where $\delta(x)$ is the non-random mean function and $e(x)$ is a zero-mean stochastic process. One often models the mean of the process with covariates, i.e. $\delta(x) = W(x)^T\beta$, where $W(x)$ are covariates and β is a vector of regression covariates.

The process is said to be *Gaussian* if for any $D \geqslant 1$ and locations $x_1, ..., x_D$, the vector $(e(x_1), ..., e(x_D))$ has a mean-zero multivariate normal distribution, which in turn implies that the vector $(Y(x_1), ..., Y(x_D))$ has a multivariate normal distribution with mean $(\delta(x_1), ..., \delta(x_D))$. The process is *strictly stationary* if the joint distribution of $(Y(x_1), ..., Y(x_D))$ is the same as $(Y(x_1 + h), ..., Y(x_D + h))$ for any $h \in X$ and for any D

points $x_1, ..., x_D$. For a Gaussian process, strict stationarity implies

$$\text{Cov}(Y(x_1), Y(x_2)) = C(x_1 - x_2) \quad \text{for all} \quad x_1, x_2 \in X.$$

That is, the covariance of the process at any two locations is some function C which depends only on the separation vector between points, and not the particular locations. This is also called second-order stationarity. Next, we define the *variogram* through the relation

$$\text{Var}(Y(x_1) - Y(x_2)) = 2\gamma(x_1 - x_2)$$

where the quantity 2γ is the variogram, and γ is the *semi-variogram*. Under the assumption of strict (or second-order) stationarity,

$$\gamma(h) = C(0) - C(h) = C(0)(1 - \rho(h))$$

where $\rho(h)$ is the correlation between two locations separated by vector h. Further, if we have $\gamma(h) = \gamma(||h||)$ for all $h \in X$, meaning if the semi-variogram only depends on h through its length $||h||$, then the process is *isotropic*. The correlation function $\rho(h)$ is then usually chosen from one of the valid families of correlation functions for Gaussian processes. A few common choices of isotropic, stationary correlation functions are the Whittle-Matérn,

$$\rho(h) = c_1 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{h}{c_2}\right)^{\nu} K_\nu \left(\frac{h}{c_2}\right), \quad 0 \leqslant c_1 \leqslant 1, c_2 > 0, \nu > 0, \tag{2.8}$$

Cauchy,

$$\rho(h) = c_1 \left\{1 + \left(\frac{h}{c_2}\right)^2\right\}^{-\nu}, \quad 0 \leqslant c_1 \leqslant 1, c_2 > 0, \nu > 0, \tag{2.9}$$

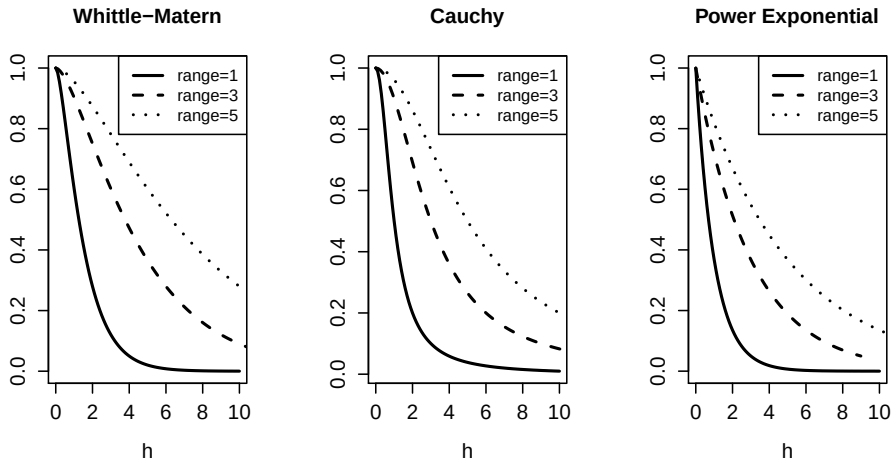

Figure 2.2: A few correlation functions for Gaussian processes. Left panel: Whittle-Matérn. Center panel: Cauchy. Right panel: Power Exponential. Each function is shown with nugget $c_1 = 1$, smooth $\nu = 1$, and range $c_2 = 1, 3$, and 5, using equations (2.8), (2.9), and (2.10).

and powered exponential

$$\rho(h) = c_1 \exp\left\{-\left(\frac{h}{c_2}\right)^{\nu}\right\} \quad 0 \leqslant c_1 \leqslant 1, c_2 > 0, 0 < \nu \leqslant 2, \tag{2.10}$$

where c_1, c_2 and ν are the nugget, range, and smooth parameters, Γ is the gamma function and K_ν is the modified Bessel function of the third kind with order ν. A few sample correlation functions are shown in Figure 2.2.

It is common to fix the nugget as $c_1 = 1$, which forces $\rho(h) \to c_1 = 1$ as $h \to 0$. This is a reasonable assumption for many environmental processes, and we make this assumption throughout this dissertation and do not attempt to model the nugget. We should clarify, though, that the theory and methods described in this dissertation would apply even if $c_1 \neq 1$, and so this restriction is not required. Throughout the remainder of this dissertation, the unknown spatial dependence parameter is called $\phi = (c_2, \nu)$, and unless stated otherwise we will assume spatial extremes models which are both stationary and isotropic. Methods for handling non-isotropic models are discussed in the next section.

2.3 Spatial Extremes and Max-stable Processes

Max-stable processes arise as the infinite dimensional generalization of multivariate extreme value theory. Let $Z(x), x \in X \subseteq \mathbb{R}^p$ be a spatial process. If for all $n \geqslant 1$, there exists sequences $a_n(x), b_n(x)$, $x \in X$ such that for any $x_1, ..., x_D \in X$,

$$P^n \left(\frac{Z(x_d) - b_n(x_d)}{a_n(x_d)} \leqslant z(x_d), d = 1, ..., D \right) \to G_{x_1,...,x_D}(z(x_1), ..., z(x_D))$$

then $G_{x_1,...,x_D}$ is a multivariate extreme value distribution. If the above holds for all possible subsets $x_1, ..., x_D \in X$ for any $D \geqslant 1$, then the process is max-stable.

The definition of a max-stable process as the infinite dimensional generalization of the multivariate extreme value distribution gives a well-defined model, but not an obvious way of constructing such a process. A conceptual construction with spectral representation was given by de Haan (de Haan, 1984; de Haan and Ferreira, 2006). Let $Y(x)$ be a non-negative stationary process on $\mathbb{R}^p$ such that $\mathbb{E}(Y(x)) = 1$ at each x. Let Π be a Poisson process on $\mathbb{R}_+$ with intensity $s^{-2}ds$. If $Y_i(x)$ are independent replicates of $Y(x)$, then

$$Z(x) = \max s_i \cdot Y_i(x), \quad x \in X$$

is a stationary max-stable process with unit Fréchet margins . From this, the joint distribution may be represented as

$$P(Z(x) \leqslant z(x), x \in X) = \exp\left(-\mathbb{E}\left[\sup_{x \in X} \frac{Y(x)}{z(x)} \right] \right),$$

where $\mathbb{E}\left[\sup_{x \in X} \frac{Y(x)}{z(x)}\right]$ is the exponent measure $V(z)$ shown in equation (2.6). Varying the choice of the process $Y(x)$ gives different max-stable processes. Smith (unpublished manuscript, 1990) constructed a process known as the Gaussian extreme value process. Let $(s_i, x_i), i \geqslant 1$ denote the points of a Poisson process on $(0, \infty) \times \mathbb{R}^p$ with intensity measure

$s^{-2}dsdx$. Take $f(x, x_i)$ to be the multivariate Gaussian density function,

$$f(x, x_i) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - x_i)^T\Sigma^{-1}(x - x_i)\right).$$

Then $Z(x) = \max_i s_i f(x, x_i)$ is a max-stable process with unit-Fréchet margins. Smith also introduced the "rainfall-storms" interpretation: think of $\mathbb{R}^p$ as the space of storm centers, s_i as the magnitude of the i^{th} storm, and $f(x, x_i)$ as the shape of the storm centered at position x_i. The maximum of independent storms at each location x is taken to be the max-stable process. With this framework, the bivariate distribution function of the max-stable process Z can be written as

$$P(Z_1 \leqslant z_1, Z_2 \leqslant z_2) = \exp\left[-\frac{1}{z_1}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_1}{z_2}\right)\right] \qquad (2.11)$$

where Φ is the standard normal distribution function, $a^2 = (x_1 - x_2)^T\Sigma^{-1}(x_1 - x_2)$, and Σ is the covariance matrix of f with covariance σ_{12} and standard deviations σ_1 and σ_2. The dependence parameter a represents a transformed distance between the two sites, and the limits $a \to 0$ and $a \to \infty$ correspond to perfect dependence and independence, respectively. Figure 2.3 shows one realization of this process.

Schlather (2002) introduced a more flexible set of models for max-stable processes by taking $Y(x)$ to be any stationary Gaussian process (and not just a multivariate normal density) with finite expectation. He considered a stationary Gaussian process Y on $\mathbb{R}^p$ with correlation function $\rho(\cdot)$ and finite mean $\mu = \mathbb{E}\max(0, Y(x)) \in (0, \infty)$. Let s_i be a Poisson process on $(0, \infty)$ with intensity measure $\mu^{-1}s^{-2}ds$. Then

$$Z(x) = \max_i s_i \max(0, Y_i(x))$$

is a stationary max-stable process with unit-Fréchet margins. The bivariate distribution

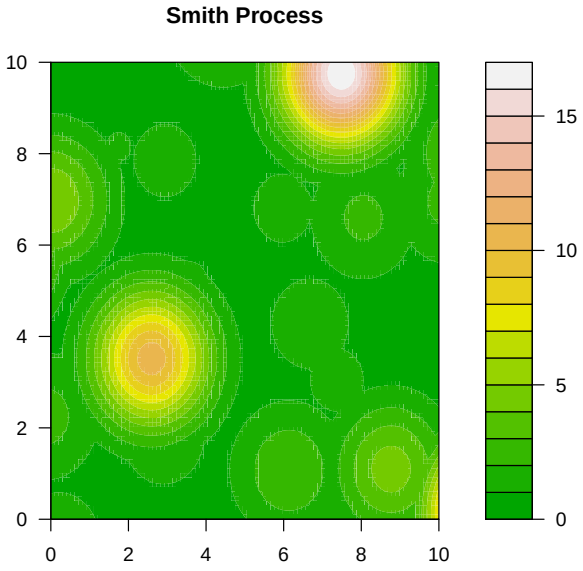

Figure 2.3: Gaussian extreme process with parameters $(\sigma_1 = \sigma_2 = 9/8, \sigma_{12} = 0)$

.

function is

$$P(Z_1 \leqslant z_1, Z_2 \leqslant z_2) = \exp\left[-\frac{1}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}\right)\left(1+\sqrt{1-2(\rho(h)+1)\frac{z_1 z_2}{(z_1+z_2)^2}}\right)\right] \quad (2.12)$$

where $\rho(h)$ is the correlation of the underlying Gaussian process Y and $h = ||x_1 - x_2||$. The correlation is chosen from one of the valid families of correlations for Gaussian processes, such as those shown in equations (2.8), (2.9) and (2.10). Following the rainfall-storms interpretation from Smith, the Schlather model takes maxima over a series of storms with the same dependence structure, but their realizations vary stochastically. This allows storms to have random shapes, unlike the deterministic multivariate normal shapes of the Smith model. Figure 2.5 shows one realization of a process with the Whittle-Matérn correlation function. This is generally considered a more realistic representation of an environmental process than the Gaussian extreme value process. In this dissertation we

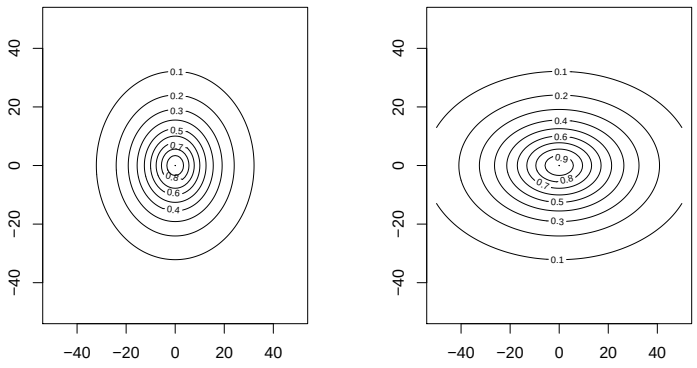

Figure 2.4: Whittle-Matérn correlation function with nugget=1, range=10, and smooth=1 for the isotropic case (left panel) and non-isotropic case (right panel).

focus on the Schlather model exclusively, but the methodology can be applied to any parametrically specified max-stable process from which realizations can be simulated.

Although the correlation functions $\rho(\cdot)$ shown above all assume isotropy, the Schlather model does not explicitly require this assumption. Ribatet (2011) gives a convenient method for extending the approach to non-isotropic data through a space warping argument. Given a valid, isotropic correlation function $\rho(\cdot)$, one may define an elliptical correlation function $\rho_e(\Delta x) = \rho(\sqrt{\Delta x^T A \Delta x})$ where Δx is the vector between two locations, and the matrix A handles the space-warping into an elliptical measure of distance (and would contain additional dependence parameters). An example of an elliptical correlation function in $\mathbb{R}^2$ is shown in Figure 2.4.

One drawback to the Schlather model is that it cannot attain the case of independence for extremes as distance $h \to \infty$. To overcome this problem, the process $Y(x)$ can be restricted to a random set $\mathcal{B}$, i.e.,

$$Z(x) = \max_i s_i \max(0, Y_i(x)) I_{\mathcal{B}_i}(x - x_i)$$

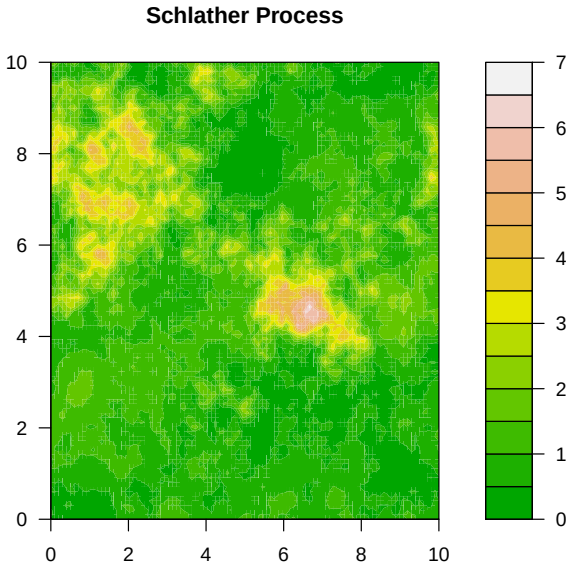

Figure 2.5: Extremal Gaussian process with Whittle-Matérn correlation with nugget $c_1 = 1$, range $c_2 = 3$, and smooth $\nu = 1$.

where $I_{\mathcal{B}}$ is the indicator function for a random set $\mathcal{B} \subset X$ and x_i. If Y_i is again a Gaussian process, then the bivariate distribution function can be written as

$$\exp\left\{-\left(\frac{1}{z_1}+\frac{1}{z_2}\right)\left[1-\frac{\alpha(h)}{2}\left(1-\sqrt{1-2(\rho(h)+1)\frac{z_1 z_2}{(z_1+z_2)^2}}\right)\right]\right\}$$

where $\alpha(h) = \mathbb{E}|\mathcal{B} \cap (h + \mathcal{B})|/\mathbb{E}(|\mathcal{B}|) \in [0, 1]$. This modification permits independent extremes in the limit as $h \to \infty$. One possible choice for $\mathcal{B}$ is a disc of radius r, which implies $\alpha(h) = \{1 - |h|/(2r)\}_+$, which equals 0 when $|h| > 2r$. Choices for $\mathcal{B}$ were explored by Davison and Gholamrezaee (2010).

Kabluchko et. al. (2009) proposed an alternative specification for the $Y(\cdot)$ processes, one with a weaker assumption than second-order stationarity. Let $Y(x) = \exp\{\epsilon_s(x) - \frac{1}{2}\sigma^2(x)\}$ where $\epsilon_s(x)$ is a Gaussian process with stationary increments and $\sigma^2(x) = Var\{\epsilon(x)\}$. Then the process defined is called the *Brown-Resnick process*. The bivariate CDF transformed

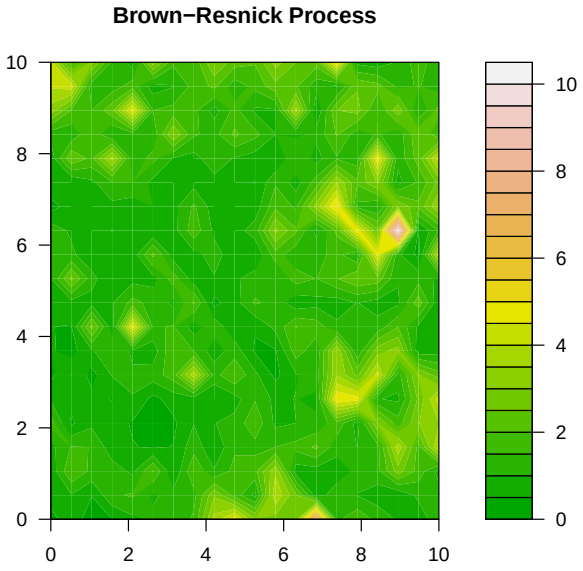

Figure 2.6: Brown-Resnick process with Whittle-Matérn correlation with nugget $c_1 = 1$, range $c_2 = 3$, and smooth $\nu = 0.5$.

to unit Fréchet margins is the same as the Smith model where the dependence parameter $a^2 = \gamma(h)$, and $\gamma(\cdot)$ is the variogram of $\epsilon(\cdot)$. The closed form of the bivariate distributions for the Brown-Resnick process associated to the variogram γ are given by

$$\begin{aligned} P & \{Z(x_1) \leqslant z_1, Z(x_2) \leqslant z_2\} \\ = & \exp\left\{ -\frac{1}{z_1}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_1}{z_2}\right)\right\} \end{aligned} \quad (2.13)$$

where Φ is the standard normal distribution function and h is the Euclidean distance between location x_1 and x_2. A realization is shown in Figure 2.6.

Each of the max-stable processes introduced above share some common features. First, they are all well defined, and can all be simulated using the point process approach outlined by de Haan (1984) with different choices of stochastic process $Y(x)$. Additional results on the simulation of max-stable processes are also found in Schlather (2002), Oesting et. al.

(2012), and Dombry et. al. (2012). Second, all finite dimensional distributions follow a MEVD, and we know that all marginal distributions at all locations $x_1, ..., x_D$ follow a GEV distribution, all of which can thus be transformed to unit-Fréchet by the one-to-one transformations

$$Y(x_d) = \left(1 + \xi(x_d)\frac{Z(x_d) - \mu(x_d)}{\sigma(x_d)}\right)^{1/\xi(x_d)}.$$

Without loss of generality then, one may assume that a max-stable process has unit-Fréchet margins. Finally, bivariate distribution functions can be written out as shown in equations (2.11), (2.12), and (2.13), and in each case the spatial dependence parameters $\Sigma, \rho(\cdot)$, or $\gamma(\cdot)$ appear in these bivariate functions. To generalize and simplify notation for the remainder of this dissertation, we will call the generic dependence parameter ϕ.

2.4 The Extremal Coefficient

Let $Z(x)$ be a stationary, isotropic max-stable random field with unit-Fréchet margins. As shown in equation (2.6), for D fixed locations the joint distribution function for a max-stable process can be written as

$$P(Z(x_1) \leqslant z_1, ..., Z(x_D) \leqslant z_D) = \exp\{-V(z_1, ..., z_D)\}$$

where $V(z_1, ..., z_D)$ is the exponent measure first described by Pickands (1981). Max-stability implies that for all N,

$$P(Z_1 \leqslant z_1, ..., Z_D \leqslant z_D)^N = \exp\{-NV(z_1, ..., z_D)\} = \exp\{-V(z_1/N, ..., z_D/N)\}.$$

The final equality arises as a consequence of the homogeneity property of the exponent measure. For each of the studied classes of max-stable processes, for $D \leqslant 2$ locations the function $V(\cdot)$ can be written out explicitly, but for $D \geqslant 3$ it cannot (an exception

to this is for the Smith model, where the trivariate distribution can be written in closed form if $X \subset \mathbb{R}^2$, explored in Genton et. al. (2011). However, for the other max-stable processes only univariate and bivariate distributions are available in closed-form). If we further consider the joint distribution of D locations evaluated at the same value z, we get

$$P(Z(x_1) \leqslant z, ..., Z(x_D) \leqslant z) = \exp\left\{-\frac{\theta(x_1, ..., x_D)}{z}\right\}$$

where $\theta(x_1, ..., x_D) = V(1, ..., 1)$ is the extremal coefficient for the D locations. Since the bounds on the extremal coefficient $V(z_1, ..., z_D)$ are $1/z_1 + ... + 1/z_D$ and $\max(1/z_1, ..., 1/z_D)$, bounds on the extremal coefficient are D and 1, respectively, with a value of D corresponding to complete independence and a value of 1 corresponding to complete dependence. The value can be thought of as the number of effectively independent locations among the D under consideration.

Many results are available for the pairwise extremal coefficient, which arises when considering any pair of locations,

$$P(Z(x_1) \leqslant z, Z(x_2) \leqslant z) = \exp\left(-\frac{\theta(x_1, x_2)}{z}\right)$$

Since the bivariate distribution functions are available in closed form equation (2.12) for the Schlather process, one may write out the pairwise extremal coefficients explicitly as

$$\theta(h) \quad = \quad 1 + \left\{\frac{1 - \rho(h; \phi)}{2}\right\}^{1/2} \tag{2.14}$$

where $h = ||x_1 - x_2||$. One may estimate the pairwise extremal coefficients directly from the data, and then through those estimates obtain an estimate of $\rho(\cdot)$. Smith (Smith 1990, unpublished manuscript) and Coles and Dixon (1999) proposed an estimate of the pairwise extremal coefficients as follows. First, assume that the field $Z(\cdot)$ has been transformed to

unit-Fréchet. This means that $1/Z(\cdot)$ is unit exponential, and $1/\max(Z(x_1), Z(x_2))$ is exponential with mean $1/\theta(x_1, x_2)$. A simple estimator then is

$$\hat{\theta}(x_1, x_2) = \frac{n}{\sum_{i=1}^{n} 1/\max(z_i(x_1), z_i(x_2))} \tag{2.15}$$

where i is the index for the block. In this dissertation, we move beyond the pairwise extremal coefficient and also focus on the tripletwise extremal coefficient, which is defined for any triplet of locations in the relation

$$P(Z(x_j) \leqslant z, Z(x_k) \leqslant z, Z(x_l) \leqslant z) = \exp\left\{-\frac{\theta(x_j, x_k, x_l)}{z}\right\}.$$

Since the trivariate distribution function for $Z(\cdot)$ is unavailable, so too is any closed-form expression for $\theta(x_j, x_k, x_l)$. However, following the same argument as in the pairwise case, we may estimate the coefficients using the estimator

$$\hat{\theta}(x_j, x_k, x_l) = \frac{n}{\sum_{i=1}^{n} 1/\max(z_i(x_j), z_i(x_k), z_i(x_l))} \tag{2.16}$$

where i is the index for the block. These estimated triplets will serve a key function in the approximate Bayesian computing algorithm. This argument may be extended to estimate all k-point extremal coefficients for any collection of k locations with $k \geqslant 3$.

2.5 Maximum Composite Likelihood Estimation

A barrier to fitting max-stable processes to data is that closed-form expressions for the joint likelihood can only be written out in low dimensional settings. The likelihood for the Smith model in $\mathbb{R}^2$ can be written out for dimension $D \leqslant 3$ (Genton et. al., 2011), but the likelihood for all other max-stable processes can only be written for dimension $D \leqslant 2$ (and we write this dissertation for this more general case). This means if the data are

observed at $D > 2$ locations in space, the joint likelihood cannot be written in closed form. Padoan et. al. (2010) proceeded with a likelihood-based approach to fitting max-stable processes by substituting a composite likelihood for the unavailable joint likelihood. We first introduce composite likelihoods, then show the connection to max-stable processes.

If $f(z;\phi)$ is a statistical model for data z and we have a set of measurable events $\{\mathcal{A}_i : i = 1, ..., m\}$, then a composite log-likelihood is a weighted sum of log-likelihoods for each event (Lindsay, 1988; Varin, 2008)

$$\ell_C(\phi; Z) = \sum_i w_i \cdot \log f(z \in \mathcal{A}_i; \phi).$$

One example of a composite log-likelihood is the pairwise log-likelihood, defined as

$$\ell_C(\phi; z) = \sum_{i=1}^{n} \sum_{d=1}^{D-1} \sum_{d'=d+1}^{D} \log f(z_{i,d}, z_{i,d'}; \phi),$$

where each term $f(z_{i,d}, z_{i,d'}; \phi)$ is a bivariate marginal density function based on locations d and d'. The two inner summations sum over all unique pairs, while the outer sums over the n i.i.d. replicates. Similar to the full likelihood function, the parameter which maximizes a composite log likelihood can be found, and is termed a *maximum composite likelihood estimate*, or MCLE. Under suitable regularity conditions (Lindsay, 1988) (Cox and Reid, 2004), the maximum composite likelihood estimator is consistent and asymptotically normal as

$$\hat{\phi}_{MCLE} \sim \mathcal{N}(\phi, \tilde{I}) \quad \text{with} \quad \tilde{I} = H(\phi)J^{-1}(\phi)H(\phi),$$

where $H(\phi) = \mathbb{E}(-H_\phi \ell_C(\phi; Z))$ is the expected information matrix, $J(\phi) = V(D_\phi \ell_C(\phi; Z))$ is the covariance of the score, H_ϕ is the Hessian matrix, D_ϕ is the gradient vector, and V is the covariance matrix. When one has the full likelihood, $H(\phi) = J(\phi)$, but in the composite likelihood setting these matrices are not equal.

Padoan et. al. (2010) used the composite likelihood to model the joint spatial dependence of extremes, and implemented their work in the `R` package `SpatialExtremes` (R Development Core Team, 2010). The maximum composite likelihood estimator $\hat{\phi}_{MCLE}$ is found numerically. The variance of the estimate is found through

$$\hat{H}(\hat{\phi}_{MCLE}) = -\sum_{i=1}^{n}\sum_{d=1}^{D-1}\sum_{d'=d+1}^{D} H_{\phi}\log f(z_{i,d}, z_{i,d'}; \hat{\phi}_{MCLE})$$

$$\hat{J}(\hat{\phi}_{MCLE}) = -\sum_{i=1}^{n}\sum_{d=1}^{D-1}\sum_{d'=d+1}^{D} D_{\phi}\log f(z_{i,d}, z_{i,d'}; \hat{\phi}_{MCLE}) D_{\phi}\log f(z_{i,d}, z_{i,d'}; \hat{\phi}_{MCLE})^{T}.$$

In general we call the dependence parameter ϕ, but for the Smith model the spatial dependence parameter is Σ, for the Schlather model it is the parameter embedded within the Gaussian correlation function $\rho(h;\phi)$, and for the Brown-Resnick model it is the parameter embedded within the variogram $\gamma(h;\phi)$. Notice for each of these models the target parameter shows up in the corresponding bivariate density functions, and thus also in the pairwise log-likelihood.

Model selection is based on minimizing the *composite likelihood information criteria* (CLIC) (Varin and Vidoni, 2005), equal to

$$-2\ell_C(\widehat{\phi}_{MCLE}; Z) - \text{tr}\left(\hat{J}(\hat{\phi}_{MCLE})\hat{H}(\hat{\phi}_{MCLE})^{-1}\right),$$

where the second term is the pairwise log-likelihood penalty term.

Thus fitting a max-stable process proceeds in two stages. We begin with an observed set of spatial extremes data, for locations $x_1, ..., x_D$. For each location, we transform the GEV data to unit-Fréchet margins by first estimating all GEV parameters $\hat{\mu}(x_d), \hat{\sigma}(x_d), \hat{\xi}(x_d), d = 1, ..., D$, then use these to transform all margins to unit-Fréchet using equation (2.4). Next, the maximum composite likelihood estimate for the dependence parameter $\widehat{\phi}_{MCLE}$ of the max-stable process is obtained using the composite likelihood approach on the transformed

data. The result is a fitted model for the extremes at the D specific locations, with spatial GEV parameter $(\hat{\mu}(x_d), \hat{\sigma}(x_d), \hat{\xi}(x_d), d = 1, ..., D)$ and spatial dependence parameter $\hat{\phi}$.

3

Approximate Bayesian Computing

Approximate Bayesian Computing is at its heart a *Bayesian* method, meaning it seeks to move from a prior distribution to a posterior distribution following Bayes's Rule:

$$\pi(\phi \mid Z) = \frac{f(Z \mid \phi)\pi(\phi)}{\int f(Z \mid \phi)\pi(\phi)\, d\phi} \tag{3.1}$$

Here, Z is taken to be the observed data, ϕ is the unknown parameter of interest, $\pi(\phi)$ is the prior distribution and $\pi(\phi \mid Z)$ is the posterior distribution. Typically, the likelihood $f(Z \mid \phi)$ is available in closed form, which allows for the possibility of an exact calculation using equation (3.1). As the dimension of the parameter space Φ increases, computing the integral in the denominator becomes increasingly complicated, to the point where most contemporary Bayesian applications do not even attempt to analytically evaluate it. Instead, the integral may be circumvented using a Monte-Carlo approach. If it is possible to generate random variables $\phi_1, ..., \phi_m$ from $\pi(\phi)$, then by the Law of Large Numbers we may approximate any function of the posterior $g(\phi \mid Z)$ by (Robert, 2007)

$$\frac{1}{m}\sum_{i=1}^{m} g(\phi_i) f(Z \mid \phi_i) \to \int g(\phi) f(Z \mid \phi)\pi(\phi)\, d\phi \quad \text{(almost surely)}$$

and similarly if an i.i.d. sample $\phi_i, ..., \phi_m$ can be produced from $\pi(\phi \mid Z)$, then the average

$$\frac{1}{m}\sum_{i=1}^{m} g(\phi_i) \to \frac{\int g(\phi) f(Z \mid \phi)\pi(\phi)\, d\phi}{\int f(Z \mid \phi)\pi(\phi)\, d\phi} \quad \text{(almost surely).} \tag{3.2}$$

Further, when $\text{var}(g(\phi) \mid Z)$ is finite, the Central Limit Theorem holds and the error remains of order $1/\sqrt{m}$ regardless of the dimension of Φ.

As computational complexities of the problem grow, a powerful technique for approximating the posterior is to use *Markov Chain Monte Carlo* (MCMC) methods. Rather than attempt to construct an independent sample $\phi_1, ..., \phi_m$ from the posterior, we instead construct a Markov chain ϕ^m which has stationary distribution equal to $\pi(\phi \mid Z)$. We accept the fact that our chain will be a series of dependent draws, but we gain an overwhelming amount of computational power that dramatically increases the reach of Bayesian methods.

One of the most popular MCMC approaches is the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et. al., 1953), which proceeds as follows:

1. Start with arbitrary initial value ϕ^0 and set $m = 0$

2. Generate ϕ' from some proposal distribution $q(\phi' \mid \phi^m)$

3. Define
$$\alpha = \min\left(1, \frac{f(Z \mid \phi')\pi(\phi')q(\phi^m \mid \phi')}{f(Z \mid \phi^m)\pi(\phi^m)q(\phi' \mid \phi^m)}\right)$$

4. Take $\phi^{m+1} = \phi'$ with probability α, and stay at ϕ^m otherwise

The algorithm defines a Markov chain whose stationary distribution is the target posterior $\pi(\phi \mid Z)$. After a sufficient burn-in period, values from this chain may be used as a collection of particles from $\pi(\phi \mid Z)$, and quantities of interest can be empirically estimated (equation (3.2)).

3.1 Approximate Bayesian Computing

To compute the exact posterior, or to implement the MCMC algorithm, one needs the closed-form expression for the likelihood $f(Z \mid \phi)$, but there are many cases where the likelihood function is either analytically intractable or computationally prohibitive. These settings were common in evolutionary genetics literature in the 1990s and 2000s, and led to a series of approximations based on simulations used to circumvent the need for the likelihood function. We review some of the major papers in this section.

Tavaré et. al. (1997) investigated the coalescence time (time to most recent common ancestor) for a random sample of n sequences of DNA. The target of their study was the simple posterior distribution $\pi(\phi \mid Z)$, where Z is the full data available. Standard Bayesian techniques failed since an explicit expression for $P(Z \mid \phi)$ was unavailable for all but the most trivial cases. Instead, they drew $\phi' \sim \pi(\phi)$ and accepted ϕ' if and only if

$$P(S = s \mid \phi') > cU$$

where S is a low, fixed dimension summary statistic, s is the value of S for observed data Z, U is a random uniform variable on (0,1), and c is a constant satisfying $c \geqslant \max_\phi P(S = s \mid \phi)$. Thus, in lieu of the unavailable likelihood, draws from the prior were accepted with probability proportional to $P(S = s \mid \phi)$. This is, in some sense, the core of ABC methods. The authors also began the discussion of how to rely on a summary statistic S when it is not a sufficient statistic, a theme that will occur again and again in the ABC literature.

Fu and Li (1997) extended the idea by adding a second simulation step for greater generality. They first drew $\phi' \sim \pi(\phi)$, but next simulated a data set $Z' \mid \phi'$, and accepted the draw if the observed and simulated summary statistics s and s' matched. The introduction of a simulated data set Z' needed at each iteration of the ABC algorithm began to dramatically increase the computation time of ABC methods.

Weiss and von Haeseler (1998) further extended the idea by replacing the single summary statistic with vectors of statistics s and s', accepting ϕ' whenever $||s - s'|| \leqslant \epsilon$ for an appropriate metric $||\cdot||$ and threshold ϵ. They simulated from a grid of values for ϕ and not from a true prior, but Pritchard et. al. (1999) reintroduced a proper prior distribution to the method.

The first reference often cited by statisticians developing the theory of ABC methods is Beaumont et. al. (2002). This was the first paper to try to obtain a smooth, functional form for the posterior density rather than simply a posterior sample. Beaumont proposed the use of kernel smoothing through equations

$$\hat{\pi}(\phi_0 \mid s) = \frac{\sum_i K_\Delta(\phi_i' - \phi_0) K_\epsilon(||s_i' - s||)}{\sum_i K_\epsilon(||s_i' - s||)} \tag{3.3}$$

where K_Δ and K_ϵ are kernels with bandwidths Δ and ϵ respectively. This led to an estimate of the posterior mean

$$\hat{\beta} = \frac{\sum_i \phi_i' \cdot K_\epsilon(||s_i' - s||)}{\sum_i K_\epsilon(||s_i' - s||)}. \tag{3.4}$$

The kernel K_Δ is a smooth, symmetric function centered at each accepted draw ϕ_i'. The kernel $K_\epsilon(||s_i' - s||)$ weights the posterior in preference of particles ϕ_i' with smaller values of $||s_i' - s||$. Typically, the kernel K_ϵ is taken as the indicator function $I_\epsilon(t) = 1 \iff t \leqslant \epsilon$, in which case equations (3.3) and (3.4) reduce to

$$\hat{\pi}(\phi_0 \mid s) = \frac{\sum_i K_\Delta(\phi_i' - \phi_0) I_\epsilon(||s_i' - s||)}{\sum_i I_\epsilon(||s_i' - s||)} \tag{3.5}$$

$$\hat{\beta} = \frac{\sum_i \phi_i' \cdot I_\epsilon(||s_i' - s||)}{\sum_i I_\epsilon(||s_i' - s||)} \tag{3.6}$$

From equation (3.6), it is easiest to see that as the bandwidth ϵ increases, the posterior

mean converges to the prior mean. As the bandwidth ϵ increases, more and more of the $\phi_i' \sim \pi(\phi)$ are accepted. Ultimately, all draws are accepted, so the posterior mean equals the prior mean (the same holds for medians, or any other function of the posterior).

This demonstrates that for any nonzero bandwidth ϵ, the resulting posterior mean will be biased back towards the prior mean. The same holds for all functions of the posterior distribution, meaning that one must always be aware that ABC methods are biased "back towards the prior", but this bias disappears as $\epsilon \to 0$.

Marjoram et. al. (2003) gave a nice (5 page!) statistical summary of ABC methods, and explicitly stated what can be considered the basic ABC-Rejection (ABC-REJ) algorithm with its two most common concessions:

ABC-REJ Algorithm

1. Draw $\phi' \sim \pi(\phi)$
2. Simulate data Z' from $f(Z \mid \phi')$, and compute summary $S' = s'$
3. Accept ϕ' if $d(s, s') \leqslant \epsilon$, and return to step 1. (This is equivalent to $I_\epsilon(||s' - s||)$ in the notation of equations (3.5) and (3.6).)

The use of summary statistic S, distance function $d(\cdot)$, and threshold ϵ ensures that the acceptance probability is workably high. Choosing these quantities is necessarily a trade-off between accuracy of the approximation and computational efficiency. The following two limits hold:

- If $\epsilon \to 0$, then $f(\phi \mid d(s, s') \leqslant \epsilon) \to \pi(\phi \mid s)$
- If $\epsilon \to \infty$, then $f(\phi \mid d(s, s') \leqslant \epsilon) \to \pi(\phi)$

As stated earlier, in practice ϵ will be some positive number larger than 0, so in practice the approximate posterior will retain some degree of bias back towards the prior. Further,

if the statistic S is sufficient for parameter ϕ, then the first limiting distribution is equal to $\pi(\phi \mid Z)$, the exact posterior distribution.

Marjoram et. al. (2003) showed how ABC methods may be integrated into ABC-MCMC with Metropolis-Hastings as follows:

1. Use transition kernel $q(\phi \to \phi')$

2. Generate data $Z' \sim f(Z \mid \phi')$

3. If $d(S, S') \leqslant \epsilon$, go to step 4. Otherwise, stay at ϕ and return to step 1.

4. Calculate
$$\alpha = \min\left(1, \frac{\pi(\phi')q(\phi' \to \phi)}{\pi(\phi)q(\phi \to \phi')}\right)$$

5. Accept ϕ' with probability α, otherwise stay at ϕ and return to step 1.

The stationary distribution of this chain is $f(\phi \mid d(S, S') \leqslant \epsilon)$. The key difference between ABC-MCMC and ordinary MCMC is that the likelihood $f(Z \mid \phi)$ is not available in the computation of α. In this particular implementation, the ABC-MCMC algorithm has a non-zero probability of moving from ϕ to ϕ' only when the distance between s and s' is below the threshold ϵ, which is equivalent to an accept step in the ABC-REJ algorithm. If flat priors are chosen with $\pi(\phi')/\pi(\phi) = 1$ and a symmetric transition kernel (such as a random walk) is selected then $q(\phi' \to \phi) = q(\phi \to \phi')$, and the ABC-MCMC is equivalent to moving only when an ABC-REJ acceptance occurs. As our preference is for flat, minimally informative priors and a random walk is the most natural choice for a transition kernel, there is no added value in choosing this implementation of ABC-MCMC over ABC-REJ. One could modify the transition probability α to involve a transition kernel other than the one shown in equation (3.7), but we found far greater success with adaptive computing, and discuss this in Section 4.2.

More recent literature within the statistics community focuses on the underlying theory of ABC methods within the familiar Bayesian framework. From this base, it becomes easier to envision improvements to the algorithm by drawing from computational results in traditional Bayesian statistics. Improvements to the efficiency allow the threshold ϵ to be set to a lower value and/or more informative but computationally summaries S to be utilized, both of which improve the approximation.

Sisson and Fan (2010) drew the connection between ABC methods and augmented Bayesian statistics. The target is the posterior distribution $\pi(\phi \mid Z) \propto f(Z \mid \phi)\pi(\phi)$, where Z is the observed data. ABC methods facilitate the computation by introducing an auxiliary parameter Z' (a simulated dataset) on the same space as observed data Z. Thus the ABC method actually computes

$$\pi_{ABC}(\phi, Z' \mid Z) \propto \pi(Z \mid Z', \phi)\pi(Z' \mid \phi)\pi(\phi).$$

Integrating out the simulated dataset yields the target posterior of interest

$$\pi_{ABC}(\phi \mid Z) \propto \pi(\phi) \int \pi(Z \mid Z', \phi)\pi(Z' \mid \phi)\, dZ'.$$

When $\pi(Z \mid Z', \phi)$ is exactly a point mass at the point $Z' = Z$ and zero everywhere else, the posterior is recovered exactly. This is likely to occur with probability 0 (for continuous data), or probability close to zero (for discrete but high dimensional data), so in practice the form is usually taken to be

$$\pi(Z \mid Z', \phi) = \frac{1}{\epsilon} K\left(\frac{|S(Z') - S(Z)|}{\epsilon}\right)$$

Under this form, the intractable likelihood is weighted in regions where $S(Z') \approx S(Z)$. When S is a sufficient statistic and in the limit as $\epsilon \to 0$, we have $\lim_{\epsilon \to 0} \pi_{ABC}(\phi \mid Z) = \pi(\phi \mid Z)$. Under the most familiar kernel,

$$\pi_\epsilon(Z \mid Z', \phi) \propto 1 \text{ if } d(S(Z'), S(Z)) \leqslant \epsilon \tag{3.7}$$

then K becomes a uniform density kernel.

3.2 Approximate Bayesian Computing for Spatial Extremes

Here we utilize the theory of max-stable processes to construct appropriate summary statistics to implement the approximate Bayesian computing algorithm for fitting max-stable processes to spatial extremes data. The challenge is to find a statistic which is highly informative (ideally sufficient) for ϕ, but also of low dimension and quickly computable, otherwise the cost of the ABC algorithm might be unreasonably high. In the following few subsections we discuss the construction of three summary statistics. The first two are based on pairs of data, but the third and most successful extends to triplets (and in principle all k-tuples for any $k \geqslant 3$).

3.2.1 The Madogram Method

Let $Z(x)$ be a stationary, isotropic max-stable random field with Generalized Extreme Value margins with $\xi < 1$. The madogram is defined as:

$$m(h) = \frac{1}{2}\mathbb{E}|Z(x+h) - Z(x)|,$$

and its natural estimator is defined as

$$\hat{m}(h) = \frac{1}{2n}\sum_{i=1}^{n} |z_i(x) - z_i(x+h)|, \tag{3.8}$$

where $z_i(x)$ is the realization of the i^{th} observed process at position x. This estimator is unbiased. Cooley et. al. (2006) showed the relationship between the madogram and the extremal coefficient $\theta(h)$. If the Generalized Extreme Value shape parameter $\xi < 1$, then the madogram $m(h)$ and extremal coefficient $\theta(h)$ verify

$$\theta(h) = \begin{cases} u_\beta + \frac{m(h)}{\Gamma(1-\xi)} & \text{if } \xi < 1 \text{ and } \xi \neq 0 \\ \exp\left(\frac{m(h)}{\sigma}\right) & \text{if } \xi = 0, \end{cases}$$

where $u_\beta = \left(1 + \xi\frac{u-\mu}{\sigma}\right)_+^{1/\xi}$ and $\Gamma(\cdot)$ is the Gamma function. Note in particular that for unit-Gumbel margins (with $\xi = 0$ and $\sigma = 1$), we have the simple relationship $m(h) = \log\theta(h)$. We will exploit this simple relationship by first transforming all margins of a max-stable process to unit-Gumbel (and not the usual unit-Fréchet). This is easily done by taking the log of data with unit-Fréchet margins.

Thus assuming that the marginal parameters of the process are known, the estimator of the madogram is unbiased, and we have a closed-form expression for the madogram as a function of the underlying correlation $\rho(h;\phi)$, which is the target of our method. We can naturally define a residual as $e(h) = \hat{m}(h) - \log\theta(h)$. Thus, for the Schlather model, plugging in equations (2.14) and (3.8) we obtain residuals

$$e(h) = \frac{1}{2n}\sum_{i=1}^{n}|z_i(x) - z_i(x+h)| - \log\left\{1 + \left(\frac{1-\rho(h;\phi)}{2}\right)^{1/2}\right\}.$$

The parameter value which minimizes the sum of squared residuals is the ordinary least squares estimator, equal to

$$\hat{\phi}_{OLS} = \text{argmin}_\phi \sum_h e(h)^2. \tag{3.9}$$

The summary statistic S is chosen to be the ordinary least squares fit to the madogram,

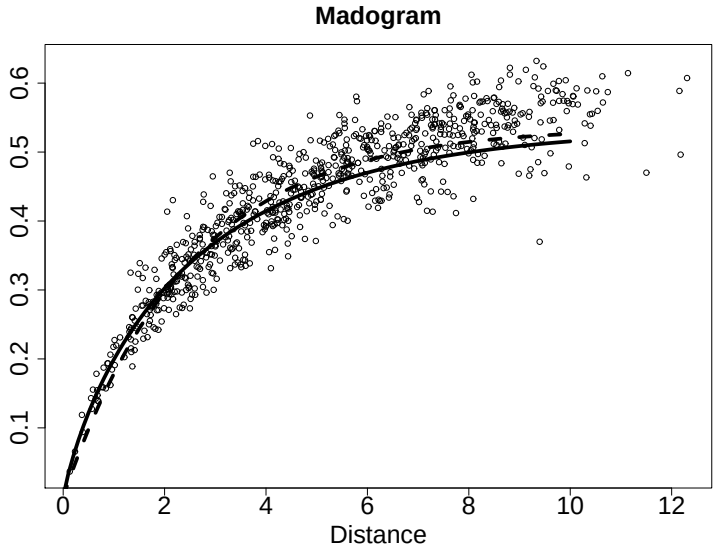

Figure 3.1: Example of a madogram (solid line), estimate (points), and ordinary least squares fit (dashed line). The entire dashed line is the summary statistic, defined by parameter $\hat{\phi}_{OLS}$.

subject to the constraint that it be a valid madogram. Mathematically, this is

$$S = \log\{\theta(h; \hat{\phi}_{OLS})\}. \tag{3.10}$$

An example of a madogram, estimate, and summary statistic is shown in Figure 3.1.

The procedure for utilizing the summary statistic is as follows. For observed data Z, the madogram is estimated and the OLS fit is obtained using equation (3.9). Then the summary statistic $S = s$ is computed using equation (3.10). For each successive iteration of the approximate Bayesian computing algorithm, a simulated data set Z' is obtained from parameter $\phi' \sim \pi(\phi)$. The madogram is estimated and an OLS fit to the madogram $S' = s'$ is obtained. What remains is some means of computing the distance between s and s'. We have chosen this as

$$d(s, s') = \int |s(h) - s'(h)|\, dh.$$

The integral of the absolute differences between the two curves s and s' is computed numerically, and taken as a measure of the distance between s and s'. The final step in approximate Bayesian computing is to accept ϕ' for the posterior if $d(s, s') \leqslant \epsilon$ for some suitably chosen ϵ.

The output of this is a collection of M particles $\phi'_1, ..., \phi'_M$ which is taken to be a sample from the approximate posterior. From this, we computed the correlation function $\rho(h; \phi)$ evaluated for each ϕ'_m. The analog of a posterior mean in this setting is the pointwise mean of all accepted functions,

$$\hat{\rho}(h) = \frac{1}{M} \sum_{m=1}^{M} \rho(h; \phi'_m). \tag{3.11}$$

We use the pointwise mean when evaluating the performance in a simulation study.

3.2.2 The Pairwise Extremal Coefficient Method

This approach is very similar to the preceding madogram approach, but instead of fitting a smooth curve to the madogram we fit the curve directly to the pairwise extremal coefficients. We define the residual as $e(h) = \hat{\theta}(h) - \theta(h)$. Plugging in equations (2.14) and (2.15), the parameter value which minimizes the sum of squared residuals is equal to

$$\hat{\phi}_{OLS} = \text{argmin}_{\phi} \sum_{h} e(h)^2.$$

The summary statistic S is chosen to be the ordinary least squares fit to the extremal coefficient, subject to the constraint that it be a valid extremal coefficient. Mathematically, this is

$$S = \theta(h; \hat{\phi}_{OLS}). \tag{3.12}$$

The remainder proceeds exactly as in the madogram method, using the summary shown in equation (3.12).

3.2.3 The Tripletwise Extremal Coefficient Method

Both the madogram approach and the pairwise extremal coefficient approach rely on pairs of locations. This is also true for the composite likelihood approach (Padoan et. al., 2010). A natural improvement is an approximate Bayesian computing method which moves beyond pairs and considers higher order k-tuples. The use of triplets was explored by Genton et. al. (2011), but only for the Smith model (Smith, 1990), a small subset of max-stable processes that does not include the Schlather model. In this section we use the estimated triplet extremal coefficients from equation (2.16) as the basis for the summary statistic $S(\cdot)$, and thus utilize information from triplets in the estimation of Schlather max-stable processes.

The number of unique sets of triplets in a set of data with D locations is $\binom{D}{3} = \frac{D(D-1)(D-2)}{6}$, which grows quite rapidly as D increases. For example, with only $D = 20$ locations we have 1140 unique triplets. This combinatorial explosion as D increases poses a problem for an approximate Bayesian computing approach. Higher dimensional summaries can only decrease the probability of acceptances, which may quickly leave an approach uncomputable in any practical sense. On the other hand, the uncertainty in estimating a single triplet extremal coefficient using equation (2.16) can be quite large (as compared with the known bounds [1,D]), so there is a natural desire to group estimates into homogeneous groups and take averages to reduce the uncertainty in estimation. The idea then is to group the $\binom{D}{3}$ triplets into K groups, which are ideally homogeneous within groups, heterogeneous across groups, and all such that $K \ll \binom{D}{3}$.

To reduce the dimension of the summary, we group these $\binom{D}{3}$ triplets into K groups using Ward's method (Ward, 1963). This method only requires a measure of distance between items, and the number of groupings. A triplet of locations is a triangle between 3 points, which produces 3 Euclidean distances $A = (a_1, a_2, a_3)$. To measure the distance

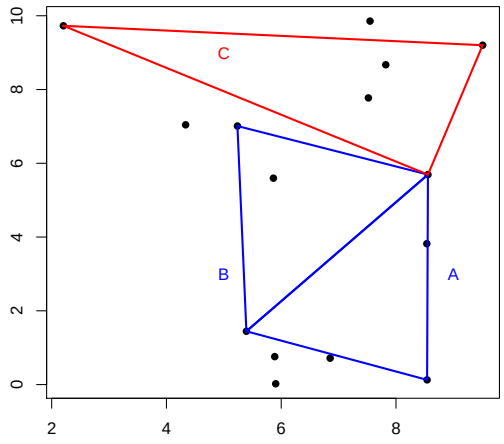

Figure 3.2: A simple example of 3 triplets labeled A, B, and C. Using the distance measure in equation 3.13, we see $dist(A, B) = 0.16$ while $dist(A, C) = 4.21$

between two triplets A and B, we take

$$dist(A, B) = \min_{\pi} \sum_{i} |a_i - b_{\pi(i)}| \tag{3.13}$$

where $\pi(i)$ is a permutation of $\{1, 2, 3\}$. Two sets of triplets A and B with identical lengths but different rotations, translation, and reflection would give a distance measure of zero. Such two sets should also have the same theoretical tripletwise extremal coefficients since the underlying field is isotropic and stationary. On the other hand, as two triangles become more dissimilar in their respective lengths, the distance measure will increase. Thus the clustering is based entirely on the geometry of the locations, and not on the actual estimates of the tripletwise extremal coefficients. An example is shown in Figure 3.2

For a data set with D locations, the first step is to compute an upper triangular dissimilarity matrix of size $\binom{D}{3}$ by $\binom{D}{3}$ which contains all distances computed using equation (3.13). In our simulations based on $D = 20$ locations, we chose to group the triplets into $K = 100$ clusters. This was selected to achieve a balance between maintaining within-group

homogeneity and ensuring that enough triplets fall into each group to reduce variability in group averages. Ward's method is a hierarchical algorithm which first assigns each item to its own cluster, and then merges two clusters chosen to minimize the overall increase in the sum of squares (which is the sum of squared distances from each item to its cluster center). Thus, the sum of squares begins at zero, and Ward's method proceeds by merging items which would result in the smallest increase. In our setting this clustering only needs to be done once, since all of the simulated draws will be at the same locations as the observed data. The requirement to enumerate all $\binom{D}{3}$ triplets is a practical limitation to how large D may be. For values of D where the dissimilarity matrix is computable, it may be time consuming to run the clustering algorithm.

The $\binom{D}{3}$ triplet extremal coefficients are estimated for the observed data using equation (2.16), and then values are averaged within the K clusters. The result is the summary of the observed data, $s = (\bar{\theta}_1, ..., \bar{\theta}_K)$. Next, we begin the approximate Bayesian computing procedure. Independent draws from the prior $\phi' \sim \pi(\phi)$ are taken. The parameter space is $\Phi = (0, \infty) \times (0, \infty)$, except when a powered exponential is used in which case it is $\Phi = (0, \infty) \times (0, 2]$. For each draw from the prior, a max-stable process with unit-Fréchet margins is simulated on the same locations and for the same number of years as the observed data. We estimate all triplet extremal coefficients for this simulated data, compute $s' = (\bar{\theta}'_1, ..., \bar{\theta}'_K)$, and use the sum of the absolute deviations as the distance metric d:

$$d(s, s') = \sum_{k=1}^{K} |s_k - s'_k|. \tag{3.14}$$

This entire process is repeated I times. The result is a collection of candidate parameter values $(\phi'_i, d_i), i = 1, ..., I$, which are then filtered as $(\phi'_i : d_i \leqslant \epsilon)$. This final filtration is an independent and identically distributed collection of M particles drawn from $\pi(\phi \mid d(s, s') \leqslant \epsilon)$, which for very small ϵ may be taken as an approximation to the true posterior. For each particle one can compute the spatial correlation function $\rho(h; \phi'_m)$. Pos-

terior standard deviations are obtained by simply regarding $\rho(\phi'_m)$ as an independent and identically distributed collection of draws from the posterior, and empirical 95% credible intervals can be constructed.

3.2.4 Simulation Study

We study the performance of the approximate Bayesian computing algorithm for estimating the spatial dependence of a Schlather process with Whittle-Matérn correlation $\rho(c_1 = 1, c_2, \nu)$. Simulations were conducted in `R`. We specified uniform, independent priors on [0,10] for the range c_2 and smooth ν parameters. This nicely spans the range of possible dependence functions on the space X (see Figure 3.3), and is consistent with the preference for minimally informative priors. While this prior may not be the most efficient choice, it does suffice to show the advantages of approximate Bayesian computing over the composite likelihood approach. We make the comparison using mean square error as our measure of performance. The simulations were all carried out for $n = 100$ years of data at $D = 20$ locations drawn from a uniform distribution on a 10 by 10 grid.

For each dataset we estimated the spatial dependence using both the composite likelihood approach and the approximate Bayesian computing approach shown in equation (3.11). Figure 3.3 shows an example. Approximate Bayesian computing was done with $I = 1,000,000$ draws. Due to the substantial computing time needed, we ran the simulations in parallel on 50 nodes on a research computing cluster, with each node only responsible for simulating 20,000 datasets. In parallel, total computing time for one dataset in one model was around 8 hours for the madogram method (which contains a numeric optimization step for each iteration), but often faster for the ABC pairwise and ABC tripletwise approaches. Given this constraint, we chose to limit the number of repetitions to only 5 replications for each model. In all there were 6 models, therefore 30 simulation runs (in the next chapter we discuss a faster implementation with more simulations).

The output from each simulation was filtered as $(\phi_i' : d_i \leqslant \epsilon_P)$, where ϵ_P is the 0.02% percentile of d_i. This ensures exactly $M = 200$ particles are accepted for the approximate posterior distribution for each simulation. We found that the spacings of the $(D = 20)$ locations can shift the overall distribution of d_i, so for identical model specifications one may need different thresholds of ϵ to ensure enough particles are accepted. Thus, it is better not to specify a fixed threshold ϵ but instead set as a very low percentile.

We judged relative performance of the methods based on estimating the true correlation $\rho(h; \phi_{TRUE})$, not in estimating the true parameter ϕ_{TRUE}. Two very different parameters ϕ_1 and ϕ_2 can produce similar correlations $\rho(h; \phi_1) \approx \rho(h; \phi_2)$ (by moving the range and smooth parameters in opposite directions, for example). A diffuse posterior distribution of ϕ can actually produce a tight posterior distribution of $\rho(h; \phi)$; we have observed that the ABC method shows this behavior. Thus, the comparison is made between the true correlation function $\rho(h; \phi_{TRUE})$ and estimated correlation function under the various approaches: $\hat{\rho}(h) = \rho(h; \hat{\phi}_{MCLE})$ for the composite likelihood method, and for the ABC approaches the pointwise posterior mean in equation (3.11).

Mean square error was computed as a numeric approximation to

$$\text{MSE} = \int_{\{h>0:\ \rho(h;\phi_{TRUE})\geqslant 0.1\}} \left(\rho(h; \phi_{TRUE}) - \hat{\rho}(h)\right)^2 \, dh \tag{3.15}$$

Taking the interval over the range $\{h > 0 :\ \rho(h; \phi_{TRUE}) \geqslant 0.1\}$ focuses the comparison on the regions of higher spatial correlation, which are of greater interest. If we computed mean square error over the entire range of $\rho(\cdot)$, results of this dissertation would not differ in any meaningful way. We stress that the pointwise mean in equation (3.11) is used only to compare the ABC methods with the composite likelihood approach in the simulation study. When the ABC approach is used alone in practice, one would use the full approximate posterior to handle prediction, credible intervals, and assess uncertainty. The applications in Chapter 6 show examples of this.

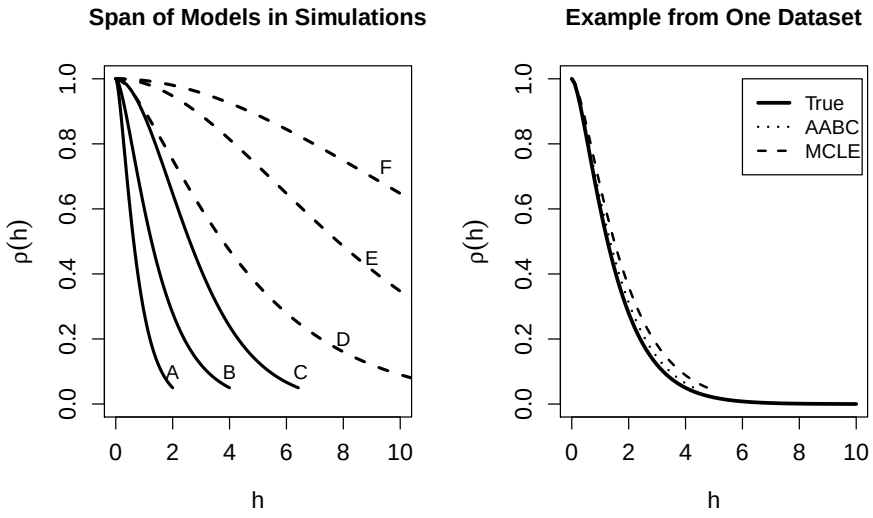

Figure 3.3: Left: Span of models included in simulation study. Models range from short-range dependence (A) to long-range dependence (F) on the scale of simulated data. In the three models labeled A, B, and C (solid lines) the adaptive ABC tripletwise approach outperformed MCLE. Right: Example of approximate Bayesian computing (dashed line) and maximum composite likelihood (dotted line) estimates from one model run, as compared to the true model (solid line).

Table 3.1: Mean Square Error using equation (3.15) for both the composite likelihood and three approximate Bayesian computing methods. Reported values are averages from 5 simulations of each of the six models. Standard error estimates are shown in brackets. The values reported have order of magnitude is 10^{-4} (multiply each entry by 10^{-4} to show the true value). The final column shows the relative reduction in MSE when using the approximate Bayesian computing tripletwise method, as compared to the composite likelihood method.

Model	ABC Madogram	ABC Pairwise	ABC Tripletwise	Composite Likelihood	Reduction: ABC Tripletwise vs. Composite Likelihood
A ($c_2 = 0.5, \nu = 1$)	99 [68]	45 [27]	15 [9]	26 [13]	40.1%
B ($c_2 = 1, \nu = 1$)	207 [81]	262 [102]	125 [77]	146 [121]	14.5%
C ($c_2 = 1, \nu = 3$)	314 [104]	272 [37]	103 [46]	140 [43]	26.6%
D ($c_2 = 3, \nu = 1$)	98 [38]	284 [155]	99 [69]	163 [74]	39.4%
E ($c_2 = 3, \nu = 3$)	385 [208]	555 [241]	287 [100]	283 [105]	−1.3%
F ($c_2 = 5, \nu = 3$)	269 [92]	645 [174]	70 [41]	485 [374]	85.6%

Results are shown in Table 3.1. The first two columns show the performance of the two pairwise ABC methods, the madogram and extremal pairwise coefficient methods, followed by the ABC tripletwise and MCLE approaches. The average MSE over 5 runs is lowest for the ABC tripletwise in 5 of 6 models, and essentially tied with composite likelihood in the sixth. Within each model specification, having only five repetitions means standard error estimates remain large, and it is difficult to make firm conclusions. However, when viewed as a whole, these 30 runs do show evidence in favor of the ABC tripletwise approach. The ABC tripletwise method outperformed the ABC pairwise method in 25 out of 30 of the model runs, and it outperformed the composite likelihood approach in 19 out of 30 runs. The ABC tripletwise method also gave the best estimate out of the four methods in 16 out of 30 runs, whereas the composite likelihood was best in only 8 of 30 runs. We found the greatest correlation in performance between the ABC madogram and ABC pairwise methods (0.613), as expected, since these two methods are the most similar and utilize essentially the same information in the summary statistics. These findings motivated the larger simulation study discussed in the next chapter.

4

Computational Enhancements

Having established the basic ABC algorithm in Chapter 4, here we discuss modifications to improve estimation and decrease the computational cost. We state right at the start that we do not resort to Markov Chain Monte Carlo, which is often the solution to many Bayesian computational challenges. Although there are implementations of ABC which utilize MCMC (Bortot et. al., 2007), we found much greater benefit by relying on adaptive (or sequential) computing, which moves from a diffuse prior to an approximate posterior into multiple stages. The method begins with a diffuse prior, and ABC-REJ is first used to produce a first approximation π_1 at threshold ϵ_1. This then serves as the prior for a second ABC step, which produces a second approximation π_2 with a lower threshold $\epsilon_2 < \epsilon_1$, and so on. This is the major improvement discussed in detail in this chapter. We also document and discuss other computational issues associated with the algorithm.

4.1 Weighting the Summary Statistic

Recall that our summary statistic is $s = (\bar{\theta}_1, ..., \bar{\theta}_K)$, where $\bar{\theta}_k$ is the mean of the N_k estimated extremal tripletwise coefficients which fall into group k. We have shown that estimating an extremal tripletwise coefficient is equivalent to estimating the rate of an exponential distribution, and we used the maximum likelihood estimator for the latter.

Thus, we know our estimates of $\theta_{i,j,k}$ are consistent and asymptotically normal, and further by the Central Limit Theorem we know $\text{var}(\bar{\theta}) = \text{var}(\hat{\theta})/N_k$, where N_k is the number of triplets we average over in group k.

To capture the overall discrepancy between observed and simulated data Z and Z', we chose $d(s, s') = \sum_{k=1}^{K} |s_k - s'_k|$, so the k^{th} element of the sum captures the absolute difference for cluster group k, in which there are N_k items averaged for both s and s'. Consider two groups k_1 and k_2 with $N_{k_1} < N_{k_2}$ items, respectively, and suppose the discrepancies between observed and simulated summaries are roughly equal, i.e. $|s_{k_1} - s'_{k_1}| \approx |s_{k_2} - s'_{k_2}|$. Intuitively, we want the ABC algorithm to down-weight this discrepancy for group k_1 since there are fewer items N_{k_1}, our estimates of the tripletwise extremal coefficients are more variable, and thus there is less evidence the observed and simulated summaries truly differ. The discrepancy in group k_2 is more informative. A natural improvement, then, is to weight the summary statistic to reflect the unequal numbers of triplets in each group. Thus, a standardized distance measure is

$$d_2(s, s') = \sum_{k=1}^{K} \sqrt{N_k} |s_k - s'_k| \tag{4.1}$$

where N_k is the number of items in group k. This measure appropriately weights discrepancies more for groups with more members, and down-weights discrepancies for groups with only a few members.

4.2 Adaptive Approximate Bayesian Computing

Motivated by the promising results for the ABC tripletwise approach shown in Table 3.1, we carried out a second simulation study using a more computationally efficient adaptive approximate Bayesian computing (AABC) algorithm to more closely compare the performance of the ABC tripletwise and MCLE approaches. The aim of this simulation study

was to increase computational efficiency and the number of simulations per model, but avoid the use of parallel computing. Beaumont et. al. (2009) describe the adaptive algorithm in detail. It is a sequential algorithm which uses ABC rejection sampling to produce a first approximation, and then re-samples from this first approximation for second round of ABC rejection sampling to produce a subsequent approximation. This allows the second stage of ABC to sample more efficiently, thus increasing the efficiency of the algorithm. We implemented the AABC algorithm exactly as described by Beaumont et. al. (2009). Specifically, the steps were to:

1. Run the ABC rejection algorithm exactly as described in subsection 3.2.3, but with $I = 100,000$ simulations to produce a first approximation $\phi_1^{(1)}, ..., \phi_J^{(1)}$ (the $J = 500$ particles filtered as $(\phi_i' : d_i \leqslant \epsilon_P)$, where ϵ_P is the 0.5% percentile of d_i)

2. Compute Ω as twice the empirical variance of $\phi_1^{(1)}, ..., \phi_J^{(1)}$.

   (a) Resample a particle ϕ^* from $\phi_1^{(1)}, ..., \phi_J^{(1)}$

   (b) Mutate using kernel $K(\phi' \mid \phi^*) = \mathcal{N}(\phi^*, \Omega)$, where $\mathcal{N}(\cdot)$ is a multivariate normal density with mean ϕ^* and covariance matrix Ω.

   (c) Simulate $Z' \mid \phi'$, compute summary s' and weighted distance $d_2(s, s')$ from equation (4.1)

   (d) (Repeat 100,000 times)

3. Filter the 100,000 particles as $(\phi_i' : d_i \leqslant \epsilon_P)$, where ϵ_P is the 0.5% percentile, ensuring exactly 500 particles are accepted. Call these $\phi_m^{(2)}, m = 1, ..., 500$.

4. For accepted particle $\phi_m^{(2)}$ compute rescaled weight

$$w_m \propto \frac{1}{\sum_{j=1}^{J} \frac{1}{J} \cdot \mathcal{N}(\phi_m^{(2)} \mid \phi_j^{(1)}, \Omega)}$$

where $\mathcal{N}(\phi_m^{(2)} \mid \phi_j^{(1)}, \Omega)$ is the density of a multivariate normal with mean $\phi_j^{(1)}$ and variance Ω evaluated at the point $\phi_m^{(2)}$. Note that the rescaled weights w_m sum to one.

The only additional consequence of this adaptive algorithm is that we have produced a weighted sample from the approximate posterior, and thus have to modify our estimate of the correlation function from equation (3.11) to now be

$$\hat{\rho}(h) = \frac{1}{M} \sum_{m=1}^{M} w_m \cdot \rho(h; \phi_m^{(2)}), \tag{4.2}$$

where w_m is the rescaled weight from step 4 above.

4.2.1 Simulations

The more efficient AABC approach could be run without the use of any parallel computing, freeing up nodes, which allowed for 30 runs in each model (thus $6 \cdot 30 = 180$ simulations in total). We chose an initial sampling of 100,000 and a re-sampling of 100,000 to keep the computational cost to around 8 hours per run. This means the performance of AABC as shown in Table 4.1 is roughly what a user might expect when analyzing a dataset on a single computer in a single day. One example run is shown in Figure 4.1. The AABC method resulted in a lower MSE for the three short range processes (A, B, and C) but a larger MSE for the three longer range processes (D, E, and F). Clearly, the adaptive ABC approach was not shown to outperform MCLE for all of the models, but there is a clear statistical benefit for the short-range processes. We discuss this more in chapter 7.

4.3 Clustering and the Choice of K

In the implementation discussed thus far, Ward's method (Ward, 1963) was used to cluster the $\binom{D}{3}$ extremal coefficient estimates into K groups. To cluster T objects using Ward's

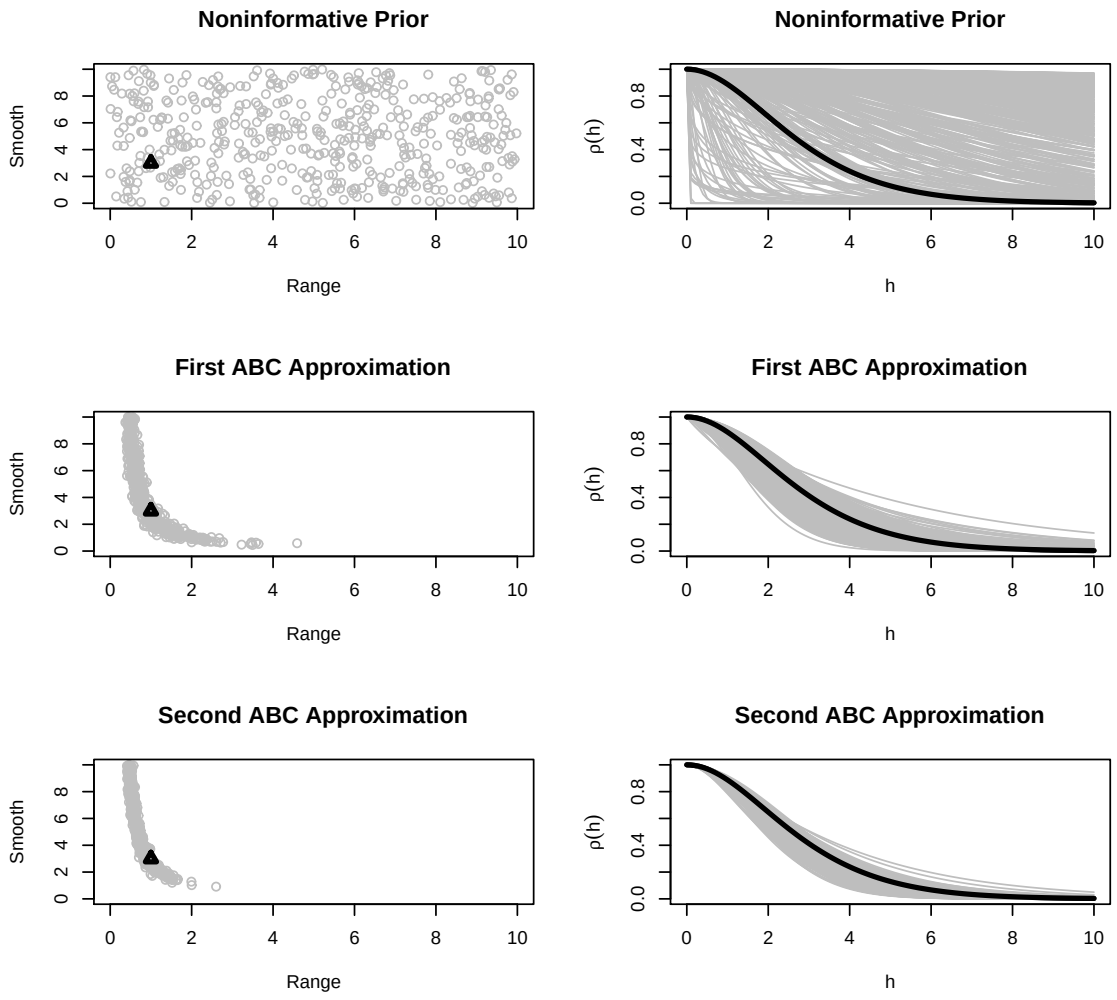

Figure 4.1: An example of the AABC method. The top row shows the noninformative prior, the middle shows the first ABC approximation, and the third row shows the final adaptive ABC approximation to the posterior. The left column shows 500 draws from each distribution on the parameter space with true parameter value (shown as the triangle), while the right column shows the same draws on the functional space along with the true correlation function (shown as the heavy black line).

.

Table 4.1: Mean square error of the pointwise mean correlation function estimates (taken with respect to the true correlation function) for the MCLE and AABC methods. Reported values are averages from 30 simulations of each of the six models. Standard error estimates are shown in brackets. The values reported have order of magnitude is 10^{-4} (multiply each entry by 10^{-4} to show the true value).

Model	MCLE	AABC
A ($c_2 = 0.5, \nu = 1$)	265 [134]	217 [23]
B ($c_2 = 1, \nu = 1$)	330 [36]	115 [33]
C ($c_2 = 1, \nu = 3$)	162 [33]	76 [17]
D ($c_2 = 3, \nu = 1$)	225 [44]	395 [26]
E ($c_2 = 3, \nu = 3$)	158 [7]	238 [11]
F ($c_2 = 5, \nu = 3$)	47 [8]	79 [6]

method, the computation is $O(T^2)$, which is a result of the necessary upper triangular matrix of dimension $T \times T$. Computing this matrix becomes costly as T increases. Recall that for a data set with D fixed locations, there are $\binom{D}{3} = O(D^3)$ extremal tripletwise coefficients to be clustered, so under Ward's method the overall cost if $O(D^6)$. We found in practice that this limited the ABC method with Ward's method to small spatial datasets.

An alternative is to use the k-means++ algorithm (Arthur and Vassilvitskii, 2006; Ostrovsky et. al., 2006). Call the location vectors of the J items to be clustered $x_j, j = 1, ..., J$.

1. Initialize K cluster means $m_1, ..., m_K$ (more on this later).

2. For each of the j items, compute the distance to each current cluster centroid m_i as

$$d_i(x_j)^{(t)} = ||x_j - m_i^{(t)}||$$

and assign the item to the cluster with the smallest distance. Repeat for all j.

3. Update the cluster means as

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

where $|S_i^{(t)}|$ is the number of items in the i^{th} group, and repeat steps 2-3 until all j items remain in their clusters.

The main advantage is that for $\binom{D}{3}$ items, one only needs to compute $\binom{D}{3} \cdot K$ terms, which is $O(D^3)$. This substantially reduces the computational complexity as D is increased. To initialize the K cluster means, one can simply randomly choose K of the locations. There are more elegant ways of ensuring a reasonable initial set, however. A simple one based on biased sampling as follows:

1. Randomly choose the first centroid as m_1.

2. For each of the j points, compute the distance to m_1 as

$$d(x_j)^{(1)} = ||x_j - m_1^{(1)}||$$

and sample an additional point with probability $\propto d(x_j)^{(1)}$. Set m_2 equal to this point.

3. Given centroids $m_i, i = 1, ..., k$, for each of the remaining $J - k$ points compute

$$d(x_j)^{(k)} = \min_i ||x_j - m_i^{(k)}||$$

and sample an additional point with probability $\propto d(x_j)^{(k)}$. Set m_{k+1} equal to this point.

4. Repeat until K means are selected.

This approach encourages initial cluster means to span the full space of items. Figures 4.2 and 4.3 show examples of clustering for $D = 20$ locations and $K = 24$ groups using both Ward's method and the k-means++ algorithm. There is some evidence of sub-optimal clustering, but that is a feature of all clustering routines which do not enumerate every single cluster combination.

There are a few other challenges: if a cluster becomes empty during the algorithm, the centroid no longer exists, and the approach may fail. To correct for this, we simply re-sampled a point at random from the triplet space to serve as the new centroid, and then continued. Perhaps more significantly, the number of clusters must be specified in advance, and it is not obvious how to handle this. To address this, we repeated simulations for a sequence of K values to determine an optimal choice. These are shown in Figures 4.5 and 4.6, and discussed later.

To further study the benefits of k-means++ over Ward's method, we ran a simulation study and recorded the run times for various clustering algorithms. For each run, we randomly placed D data points on a 10 by 10 grid, and used Ward's method along with two versions of the k-means++ clustering algorithm. Both Ward's and the first k-means++ algorithm clustered the items into $K = \lceil \sqrt{\frac{D(D-1)(D-2)}{12}} \rceil$ clusters, where $\lceil \cdot \rceil$ is the ceiling function which rounds the argument up to the next nearest integer. An alternate k-means++ algorithm fixes the total number of clusters at $K = \min(\lceil \sqrt{\frac{D(D-1)(D-2)}{12}} \rceil, 100)$ which for $D \geqslant 50$ results in a smaller number of clusters, and thus presumably a lower mean run time. Results are shown in Figure 4.4.

The ABC algorithm we have outlined requires a choice of K, the number of clusters of tripletwise extremal coefficients and dimension of the summary statistic. This should be regarded as a tuning parameter. Increasing the number of clusters K helps to increase the within-group homogeneity but simultaneously results in fewer triplets per group. Conversely, reducing the number of clusters K reduces within-group homogeneity but increases

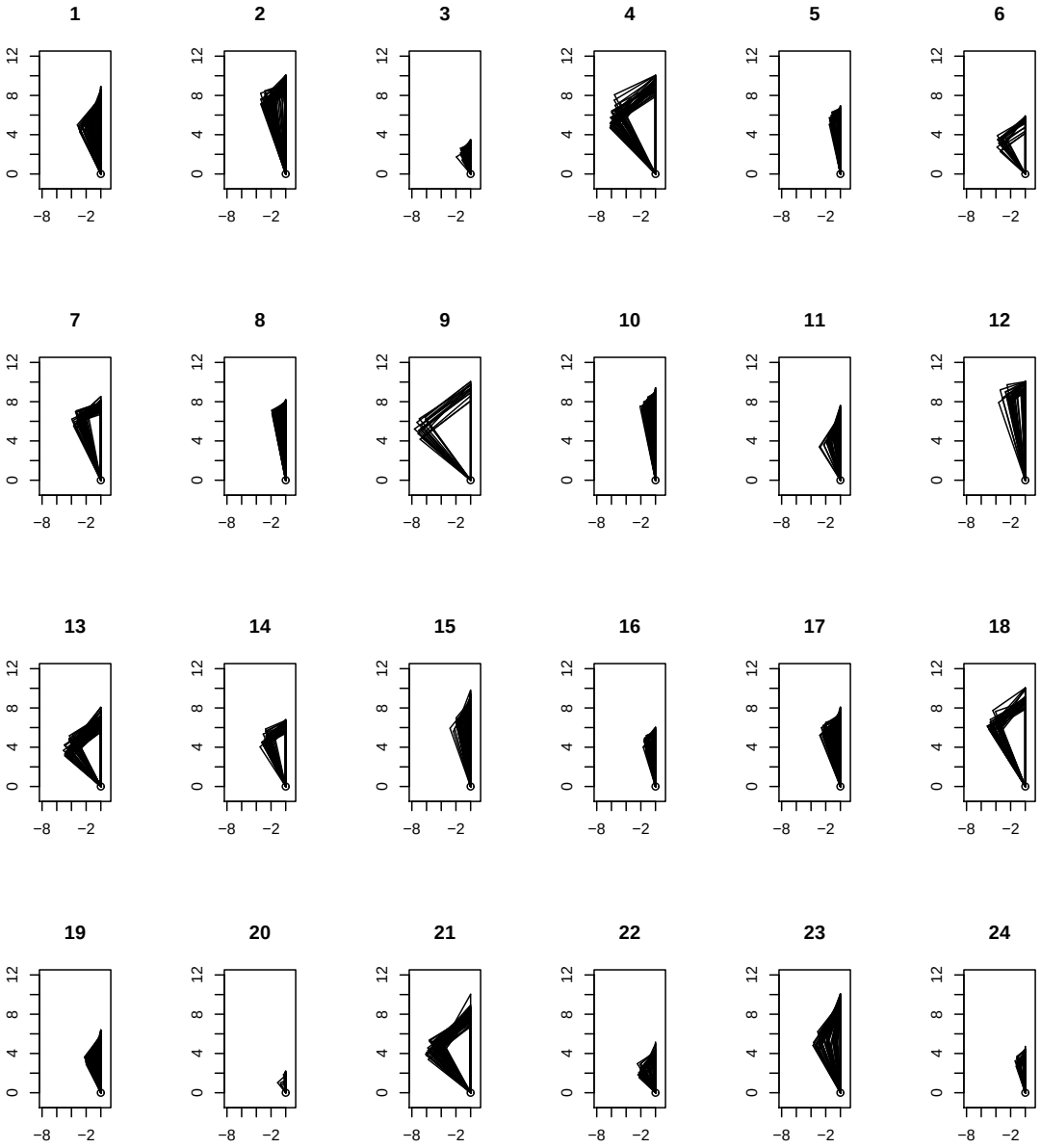

Figure 4.2: Results from clustering with the k-means++ method. Data from $D = 20$ locations (thus 1140 triplets) are clustered into $K = 24$ clusters using the k-means++ method. Each triangle is drawn as follows: first, the longest length is drawn vertically from the fixed origin. The second longest length is drawn to the left, and the shortest length connects. Thus, all triangles share a single point (the origin), and are oriented in the same manner. This helps to see the degree of homogeneity in each cluster.

.

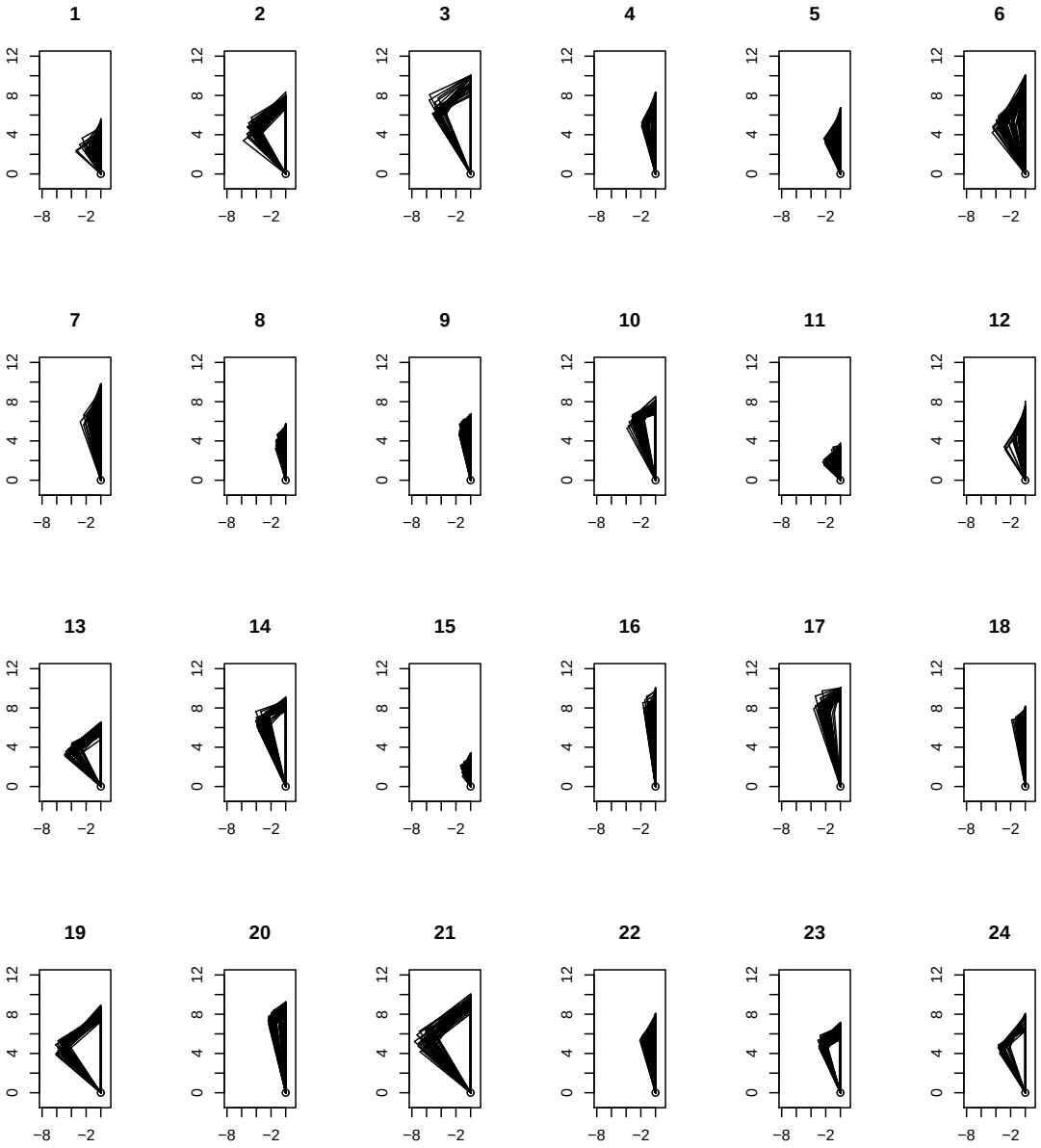

Figure 4.3: Results from clustering with Ward's method. Data from $D = 20$ locations (thus 1140 triplets) are clustered into $K = 24$ clusters using the k-means++ method. Each triangle is drawn as follows: first, the longest length is drawn vertically from the fixed origin. The second longest length is drawn to the left, and the shortest length connects. Thus, all triangles share a single point (the origin), and are oriented in the same manner. This helps to see the degree of homogeneity in each cluster.

.

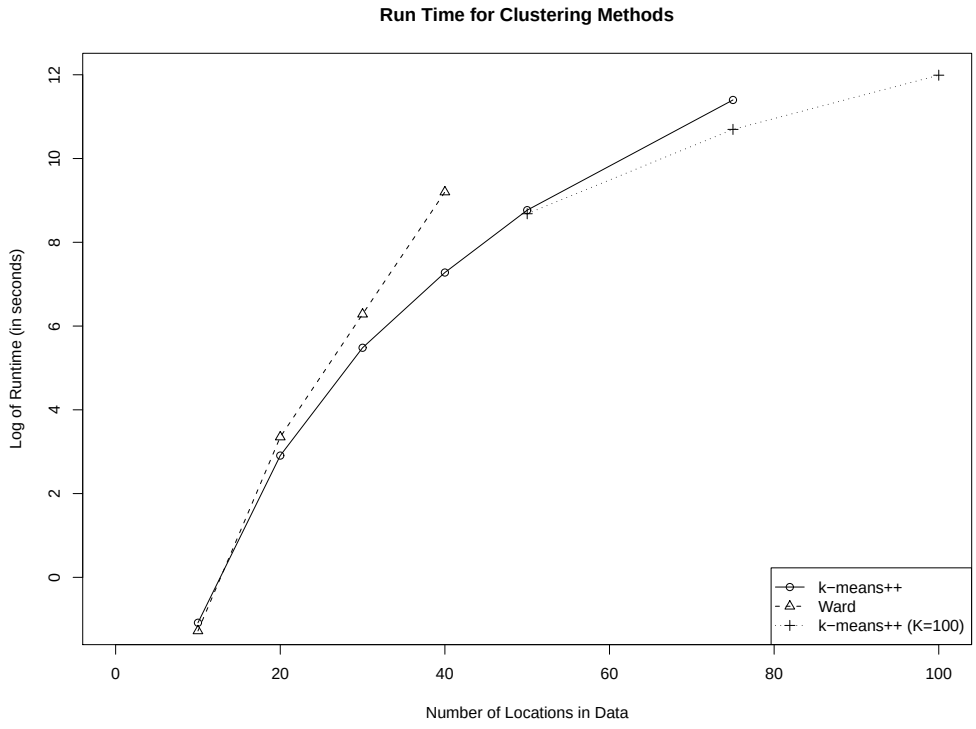

Figure 4.4: Mean run times for three clustering methods, taken over 10 runs for each point. The solid line shows the log of mean run times for the k-means++ algorithm with $K = \lceil\sqrt{\frac{D(D-1)(D-2)}{12}}\rceil$, which is Mardia's convention (Mardia et.al., 1980). The dashed line uses Ward's method, also with $K = \lceil\sqrt{\frac{D(D-1)(D-2)}{12}}\rceil$. The dotted line uses the k-means++ algorithm, but fixes $K = 100$. For $D \geqslant 50$, $K = 100$ is a smaller number of clusters than the number obtained from Mardia's convention.

.

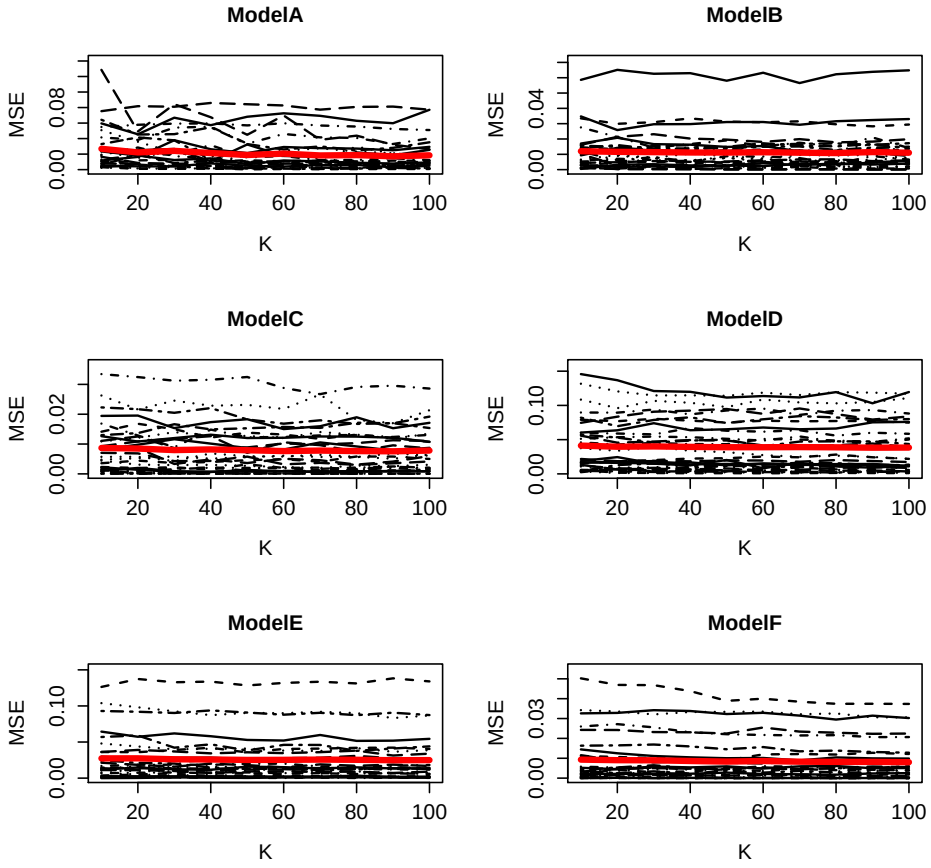

Figure 4.5: Mean square error as a function of K (the number of clusters) for $K = 10, ..., 100$ for the six model choices using ABC-REJ. Averages over the 30 runs are shown by the heavy red line.

the number of triplets per group. The aim then is to balance these two competing benefits.

We ran the ABC-REJ algorithm for $D = 20$ locations, $Y = 100$ years and $K = 10, 20, ..., 100$ clusters to investigate the impact K has on the performance. Results are shown in Figures 4.5 and 4.6. Somewhat surprisingly, we see very little impact in performance. There is a very slight decrease in average MSE as K increases to 100, but for most runs in most models the performance of the ABC algorithm is not affected greatly on the range $K = 10, ..., 100$. We conclude by suggesting that the user simply keep in mind the trade-off involved.

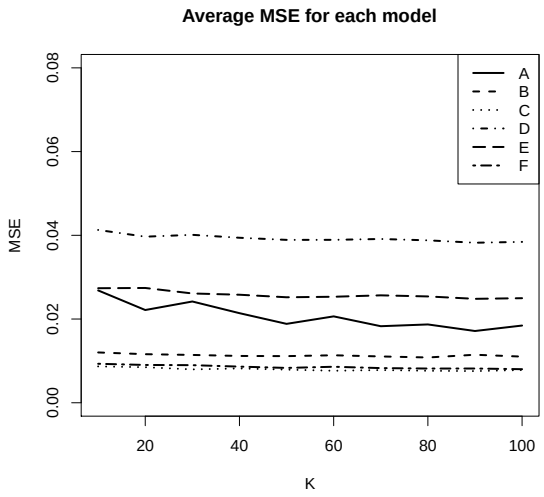

Figure 4.6: Average mean square error (taken over 30 runs) for the six models using ABC-REJ. These are equivalent to the heavy red lines shown in Figure 4.5.

4.4 Extending Beyond Triplets

Extremal coefficients cannot be written out explicitly for $D \geqslant 3$, due to the unavailable joint likelihood function. Erhardt and Smith (2012) discuss extremal coefficients based on k-tuples, and the focus was on $k = 3$. In principle, the approach could be extended to any k-tuples for $k > 3$ as well. Here we show the dramatic rise on computational cost of considering higher order k-tuples. A max-stable process defined at D fixed point-referenced locations allows for $\binom{D}{k}$ unique k-tuples, each of which can be used to define an extremal k-wise coefficient. The number of unique k-tuples is thus $O(D^k)$. Figure 4.7 shows this growth (on the log scale) for $k = 3, 4$, and 5.

The ABC approach reduces the dimension of the summary statistic by grouping these k-tuples into K groups according to some clustering algorithm. All clustering algorithms require a measure of dissimilarity, and we have used the following approach to clustering triplets: a triplet of locations is a triangle between 3 points, which produces 3 Euclidean

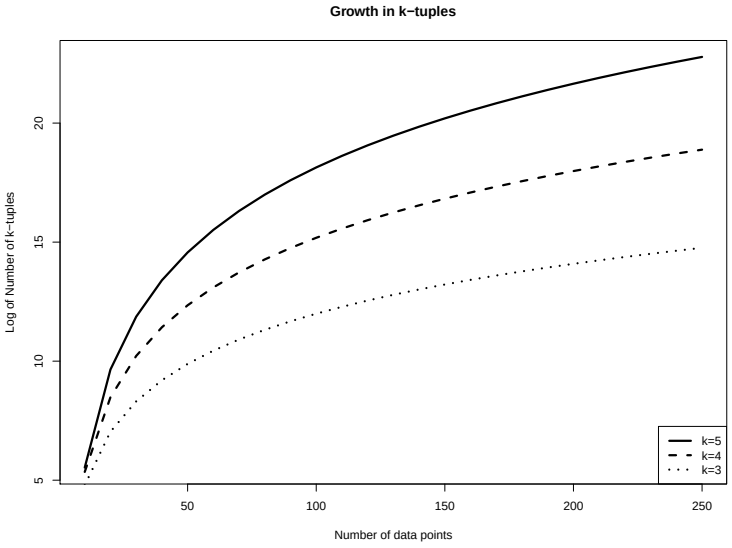

Figure 4.7: The number of unique k-tuples in a data set with D locations.

.

distances $A = (a_1, a_2, a_3)$. To measure the distance between two triplets A and B, we take

$$dist(A, B) = \min_{\pi} \sum_{i}^{3} |a_i - b_{\pi(i)}| \tag{4.3}$$

where $\pi(i)$ is a permutation of $\{1, 2, 3\}$. That is, we compare only the three lengths of the two triplets, and find the permutation which results in the smallest sum of absolute deviations. There are 6 such permutations, so a single distance computation between two triplets requires taking a minimum over 6 permutations.

To generalize, let us now consider a k-tuple of points on the plane. These points produce $\sum_{i=1}^{k} i = k(k-1)/2$ unique line segments. So, if $k = 3$ there are 3 unique line segments; for $k = 4$, there are 6, and so on. Following the same idea as above, let us call two k-tuples A and B, with respective line segment distances $(a_1, ..., a_{k(k-1)/2})$ and $(b_1, ..., b_{k(k-1)/2})$. Generalizing equation (4.3), the distance between them would be

Table 4.2: Number of permutations when computing the distance between two k-tuples using equation (4.4).

k	$[k(k-1)/2]!$
3	6
4	720
5	3,628,800
...	...

$$dist(A, B) = \min_{\pi} \sum_{i=1}^{k(k-1)/2} |a_i - b_{\pi(i)}| \tag{4.4}$$

where $\pi(i)$ is a permutation of $\{1, 2, ..., k(k-1)/2\}$. The number of permutations in π is $[k(k-1)/2]!$. Thus, a single distance computation must take the minimum over a vastly increasing number of items. Table 4.2 shows how quickly these add up.

Using this definition of distance, the computation of distance between two k-tuples is $O([k(k-1)/2]!)$. We stress this is only to compute the distance between two singular objects we have denoted as A and B; there are $\binom{D}{k}$ total objects in a spatial data set of D locations. To run the overall clustering, Ward's method requires $O(D^{2k})$ total distance calculations, and k-means requires $O(D^k)$. It should be clear that either approach quickly leads to overwhelming computational cost. Only $k = 3$ or $k = 4$ could seriously be implemented using the distance function we have described in equation (4.4). Otherwise, some alternative definition of distance would be needed.

4.5 Choosing a Threshold ϵ

The choice for selecting a threshold ϵ is driven almost entirely by practical considerations. One would always prefer a smaller threshold ϵ to ensure a high quality approximation to the posterior. The trade-off is fewer accepted particles. Thus, the threshold should be selected to be as small as possible so long as enough particles are accepted. We have

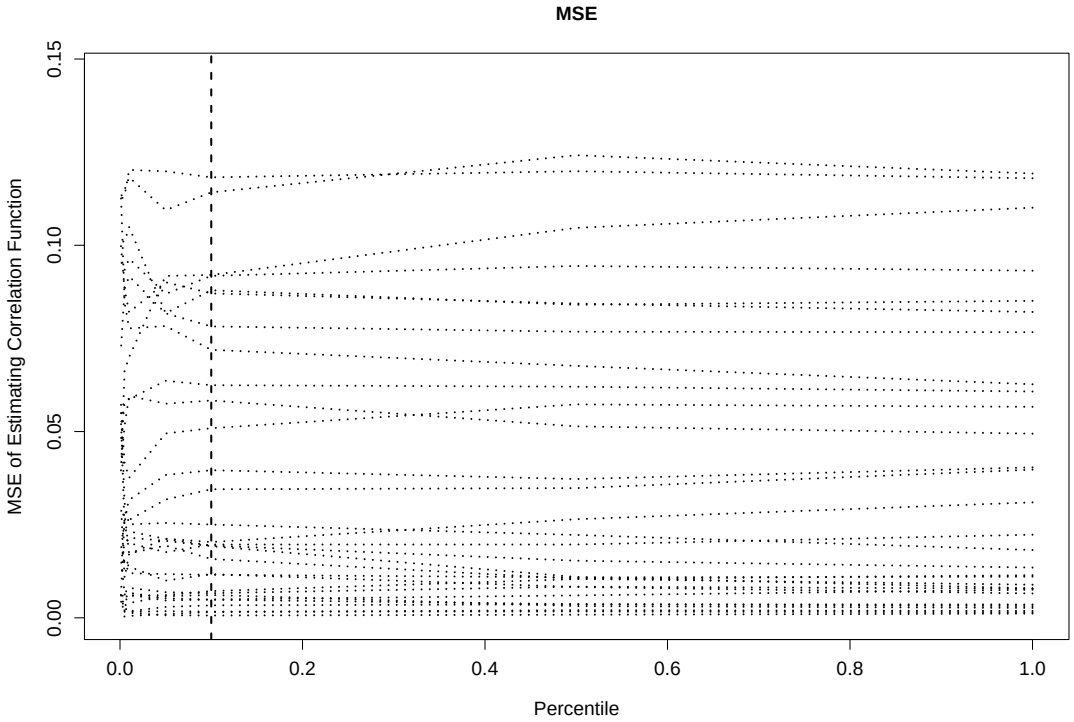

Figure 4.8: This shows the relative stability in estimating MSE using various percentiles for 5,000,000 draws from Adaptive Approximate Bayesian Computing. These lines correspond to the 30 individual model runs from a Whittle-Matérn with range $c_2 = 3$ and smooth $\nu = 1$. Here we have advocated selecting the 0.1%, which ensures exactly 500 particles are accepted for the approximate posterior.

handled this by setting ϵ to be a very low percentile of the distances, chosen to ensure a reasonable fixed number of particles are accepted. An example of this trade-off is shown in Figure 4.8.

Verifying that the resulting threshold is in fact low enough to obtain a quality posterior approximation is handled through a robust simulation study, such as the one shown in Table 4.1. Tables 4.4 and 4.5 also contain some results on how the selection of the threshold affects the performance of the method.

4.6 Selecting a Prior Distribution

We first point out a critical difference between approximate Bayesian computing and other forms of Bayesian computing when it comes to the prior distribution. In ordinary Bayesian computing, one seeks a collection of draws from the posterior distribution $\pi(\theta \mid Z) \propto \pi(\theta) \cdot f(Z \mid \theta)$. Closed-form expressions are usually available, though of course it may be difficult to compute the normalizing constant. To gain efficiency, one can sample from a different density function $g(\theta)$ chosen to closely resemble the posterior. With an appropriate accept/reject step, the result is a collection of particles from $\pi(\theta \mid Z)$. The key point here is that one can define one prior $\pi(\theta)$ but sample from a distinct distribution $g(\theta)$ for efficiency.

In approximate Bayesian computing, without a closed-form expression for the likelihood one cannot write out $\pi(\theta \mid Z) \propto \pi(\theta) \cdot f(Z \mid \theta)$, and so there is no method for identifying an alternative distribution $g(\theta)$ to increase sampling efficiency. Whichever distribution candidate parameter values are sampled from, *that is the prior* $\pi(\theta)$. Thus, if one wishes to define non- (or minimally-) informative priors, necessarily this is what the ABC algorithm must draw candidate parameter values from. The result can be a severe loss of computational efficiency.

Fortunately, adaptive computing can move from a non-informative prior to a high quality approximate posterior by splitting the transition up into a smoother sequence. If the mixture of Bayesian and Frequentist viewpoints can be forgiven here, so long as one chooses a prior containing an open neighborhood around the true parameter value, any diffuse prior can be used. Adaptive computing combined with diffuse priors is thus a very good general answer to the question "how does one choose the prior for ABC?"

There is one minor addition when it comes to defining what a "diffuse" prior is. In the simulation study we advocated the use of independent uniform priors on the parameter space $\Theta = c_2 \times \nu$, with upper limits sufficiently high to effectively cover the full range of

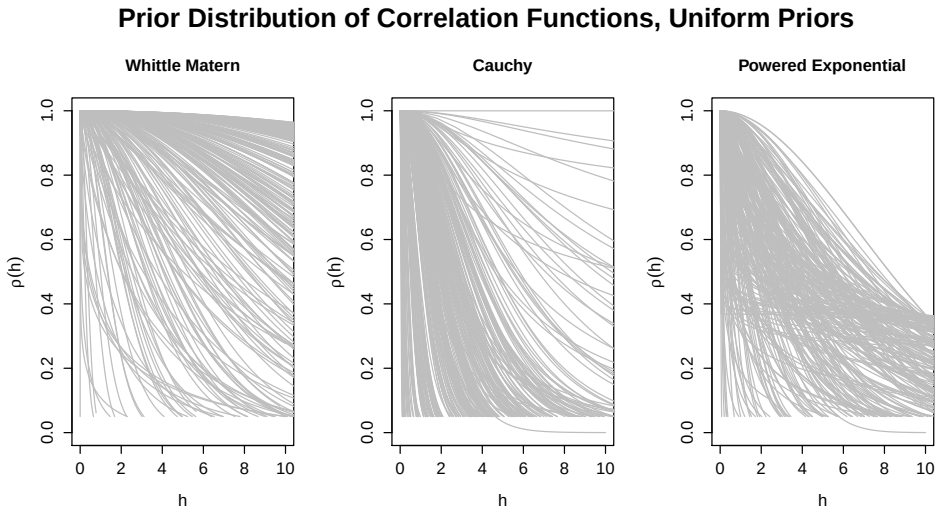

Figure 4.9: This shows the non-even distribution of correlation functions using independent, uniform priors for the range c_2 and smooth ν. For the Whittle-Matérn and Cauchy, priors are independent exponentials $c_2 \sim U(0,10), \nu \sim U(0,10)$, and for the powered exponential priors are $c_2 \sim U(0,10), \nu \sim U(0,2)$. Each figure shows 250 random draws.

spatial dependence ranges on the space of observed data. While these priors appear to show no preference on the space Θ, in fact they do tend to favor different spatial processes, as shown in Figure 4.9. For the Whittle-Matérn, longer range processes are over-sampled, and for the Cauchy and powered exponential shorter-range processes are over-sampled.

Alternatively, one can select a non-uniform prior on the space Θ which results in a more homogeneous sampling of spatial processes. By selecting independent exponential priors for the range and smooth, for example, one obtains a more even sampling of dependence ranges for spatial processes, as shown in Figure 4.10.

4.7 Simulations and Asymptotics

Here we demonstrate the benefit of increasing the number of i.i.d. replicates (Y in the notation throughout this dissertation) of spatial extremes data along with the benefit of increasing the number of observation locations (D). The latter is sometimes referred to as

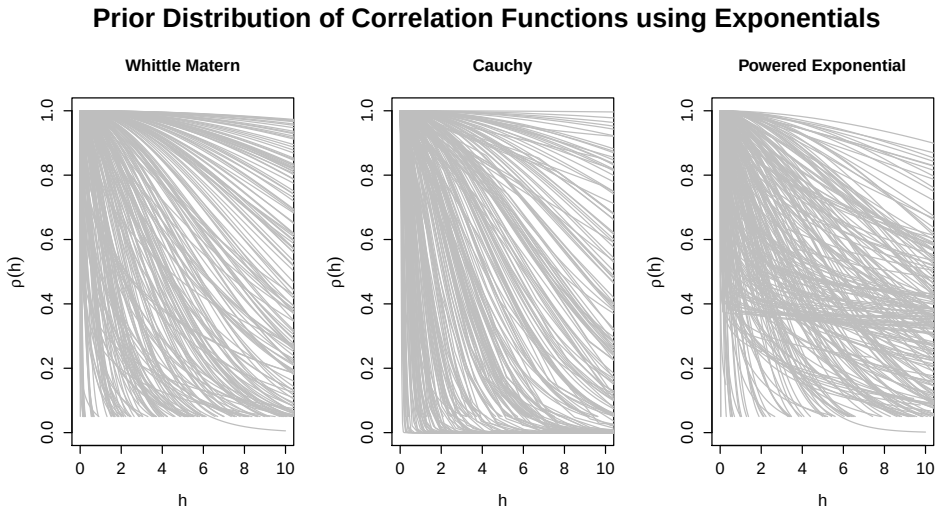

Figure 4.10: This shows the roughly even distribution of correlation functions using priors for the range c_2 and smooth ν. For the Whittle-Matérn, priors are independent exponentials $c_2 \sim Exp(\lambda = 3/h), \nu \sim Exp(\lambda = 3/h)$, for the Cauchy priors are independent exponentials $c_2 \sim Exp(\lambda = 1/2h), \nu \sim Exp(\lambda = 1/2h)$, and for the powered exponential priors are $c_2 \sim Exp(\lambda = 1/h), \nu \sim U(0, 2)$. Each figure shows 250 random draws.

"infill asymptotics", meaning the number of observed locations D increases while the spatial domain X remains fixed (this is distinguished from "increasing domain asymptotics", in which additional locations are added in a way which also expands the spatial domain X). We simulated 30 max-stable processes from a Schlather process with Whittle-Matérn ($c_1 = 1, c_2 = 1, \nu = 1$) correlation for $D = 20$ and 40 years, for $Y = 50, 100$, and 250 years. This as the observed data Z. We then ran the AABC algorithm precisely as outlined in Section 4.2. Table 4.3, 4.4 and 4.5 show results from these simulations. As expected, the mean square error falls as Y increases for fixed D, and also as D increases for fixed Y.

Table 4.3: Mean square error (taken over 30 runs) for the AABC method analyzing data from a Schlather model with Whittle-Matérn ($c_1 = 1, c_2 = 1, \nu = 1$) process for various numbers of data points D and years Y. Standard error estimates are shown in brackets. All values have order 10^{-4} (multiply each entry by 10^{-4} to show the true value).

D \ Y	50	100	250
20	172 [28]	143 [36]	97 [29]
40	160 [62]	119 [37]	31 [9]

Table 4.4: Mean square error (taken over 30 runs) for the AABC method analyzing data from a Schlather model with Whittle-Matérn ($c_1 = 1, c_2 = 1, \nu = 1$) process for $D = 20$ data points. Standard error estimates are shown in brackets. All values have order 10^{-4} (multiply each entry by 10^{-4} to show the true value). Simulations are broken out by number of years Y and threshold ϵ, shown as a percentile.

ϵ (percentile)	Particles	50	100	250
.10	5000	197 [32]	151 [40]	131 [37]
.05	2500	182 [28]	149 [39]	118 [33]
.01	500	172 [28]	143 [36]	97 [29]
.05	250	181 [30]	140 [34]	91 [28]
.001	50	181 [31]	130 [30]	81 [25]

Table 4.5: Mean square error (taken over 30 runs) for the AABC method analyzing data from a Schlather model with Whittle-Matérn ($c_1 = 1, c_2 = 1, \nu = 1$) process for $D = 40$ data points. Standard error estimates are shown in brackets. All values have order 10^{-4} (multiply each entry by 10^{-4} to show the true value). Simulations are broken out by number of years Y and threshold ϵ, shown as a percentile.

ϵ (percentile)	Particles	50	100	250
.10	5000	166 [60]	125 [39]	55 [14]
.05	2500	160 [59]	120 [39]	43 [12]
.01	500	160 [62]	119 [37]	31 [9]
.05	250	156 [62]	120 [37]	30 [9]
.001	50	157 [58]	103 [27]	26 [9]

5

Weather Derivatives

5.1 Introduction to Weather Derivatives

Richards et. al. (2004) give a list of 5 elements common to all weather derivatives. These include (a) an underlying weather index, (b) a well-defined time period, (c) the weather station used for reporting, (d) the payment attached to the index value, and (e) the strike value which first triggers payment. The intention is for the buyer of the derivative to be compensated by the seller for amounts which roughly correspond to actual business losses. Ideally, these losses are perfectly correlated with the payments of the weather derivative, though in practice this is rarely achieved. Tailoring the contract to the specific needs of one buyer reduces its general appeal in a secondary market, and thus lowers the value of the contract.

Weather derivatives offer benefits to the buyer and seller not found in traditional insurance. The buyer does not need to have an insurable interest, and they do not need to demonstrate an actual loss to receive payment. The loss payment itself is generally proportional to the difference between the weather index and strike value. Furthermore, weather derivatives offer the buyer the opportunity to sell (or even buy back) the contract in a secondary market such as the Chicago Mercantile Exchange, or over the counter. There are no such markets for traditional insurance products. Derivatives are also in general

more lightly regulated, and payments are often considered taxable.

From the seller's point of view, weather derivatives offer several advantages over traditional insurance. Weather derivatives avoid the higher administrative and loss adjustment expenses of insurance contracts. They also eliminate concern for moral hazard, morale hazard, and fraud, as the event triggering the payment is easily verified and completely beyond the buyer's control. When used to insure crops, weather derivatives help reduce the perceived information asymmetry associated with crop insurance, wherein farmers often have more information about their individual risk than insurers (Goodwin and Smith, 1995).

We consider the use of weather derivatives to provide protection against high impact, low probability business losses caused by extremes in the weather. Though not technically insurance, we model losses and price derivatives using actuarial techniques originally developed to price insurance. The weather derivatives we consider define some payment $L = L(M; s, t)$, where M is the unknown weather random variable, and s and t are the pre-specified strike and limit values (occasionally we only write L or $L(m)$ to refer to the loss to simplify notation). Examples of three types of derivatives with payments based on high exceedances include

1. $L = \alpha$ if $\{M \geqslant s\}$ and 0 otherwise

2. $L = \beta \cdot (m - s)$ if $\{M \geqslant s\}$ and 0 otherwise

3. $L = \beta \cdot (m - s)$ when $\{s \leqslant M \leqslant t\}$ and $L = \beta \cdot (t - s)$ when $\{M \geqslant t\}$, and 0 otherwise,

where α and β are dollar values, and m is the realization of random variable M. The first provides a flat payment whenever the event $\{M \geqslant s\}$ occurs, the second provides a proportional payment based on the difference $(m - s)$, while the third limits the total payment.

Unlike most derivatives, weather derivatives do not have an underlying tradable asset, and thus many pricing approaches based on financial theory are inappropriate. Jewson and Brix (2005) provide an excellent reference and discussion of pricing techniques for weather derivatives. The pricing approach we take is based on computing expected losses and expected loss variability. We also show the estimation of several common risk measures in actuarial science. The first step is to compute the expected payout $\mathbb{E}(L) = \int L(m)g(m)\,dm$, where $g(m)$ is the density function of weather variable M. For the three derivatives, shown, expected payments are

1. $\mathbb{E}(L) = \int_s^\infty \alpha \cdot g(m)\,dm = \alpha \cdot P(M \geqslant s)$

2. $\mathbb{E}(L) = \int_s^\infty \beta \cdot (m-s)g(m)\,dm$

3. $\mathbb{E}(L) = \int_s^t \beta \cdot (m-s)g(m)\,dm + \beta \cdot (t-s) \cdot P(M \geqslant t)$.

When viewed from the point of view of insurance, the quantities $\mathbb{E}(L)$ are the *pure premiums* of the contracts. In the absence of any expenses, profit, risk loadings and time-value financial considerations, this is the price of the contract for the buyer. Such contracts are fairly straightforward to price once one has an accurate estimate of the density function $g(m)$ in the region where $m \geqslant s$. Further, one can compute all second moments as

1. $\mathbb{E}(L^2) = \int_s^\infty \alpha^2 \cdot g(m)\,dm = \alpha^2 \cdot P(M \geqslant s)$

2. $\mathbb{E}(L^2) = \int_s^\infty \{\beta \cdot (m-s)\}^2 g(m)\,dm$

3. $\mathbb{E}(L^2) = \int_s^t \{\beta \cdot (m-s)\}^2 g(m)\,dm + \{\beta \cdot (t-s)\}^2 \cdot P(M \geqslant t)$,

and from these compute the variance $\mathrm{var}(L) = \mathbb{E}(L^2) - (\mathbb{E}(L))^2$. This information is often incorporated into the premium by adding a risk load $R(L)$, which is added to the pure premium as $P = \mathbb{E}(L) + R(L)$. Risk loads are meant to account for the additional risk

taken on by writing derivatives with larger variability of losses. Commonly, risk loads are a function of the variance (or standard deviation) of loss (Feldblum, 1990). Mango (1998) describes several common risk loadings such as $R(L) = \lambda \cdot \sqrt{\text{var}(L)}$, or $R(L) = \lambda \cdot \text{var}(L)$, where λ is a dollar amount chosen to satisfy some risk tolerance criteria.

Actuaries further consider a set of alternative risk measures designed to measure the risks associated with an insurance policy (Dhaene et. al., 2006; Denuit et. al., 2005). Five common measures are:

1. The Value-at-Risk, or VaR, defined as $\text{VaR}(L; p) = F_L^{-1}(p)$, where L is a loss random variable, F is the cumulative distribution function, and $0 < p < 1$. The Value-at-Risk is simply a quantile of loss, typically taken with p close to one which gives a high loss quantile.

2. The Tail Value-at-Risk, or TVaR, defined as $\text{TVaR}(L; p) = \frac{1}{1-p} \int_p^1 \text{VaR}(L; t)\, dt$. TVaR is the arithmetic average of quantiles of L, defined for all p on. TVaR is designed to convey some information of the thickness of the tail of L in regions above a high quantile $\text{VaR}(L; p)$.

3. The Conditional Tail Expectation, defined as $\text{CTE}(L; p) = \mathbb{E}(L \mid L > \text{VaR}(L; p))$. This is the expected loss given a high threshold is exceeded.

4. The Conditional Value-at-Risk, or $\text{CVaR}(L; p) = \text{CTE}(L; p) - \text{VaR}(L; p)$, which shows the gap between CVaR and CTE

5. The Expected Shortfall, or $\text{ES}(L; p) = \mathbb{E}\left[(L - \text{VaR}(L; p))_+\right]$. If one sold an insurance policy for premium P and held a reserve equal to $P + \text{VaR}$, the expected shortfall describes the expected additional loss above the reserve.

These five risk measures are closely related, particularly when the loss random variable

L is continuous. A few common relations are:

$$\text{TVaR}(L;p) = \text{VaR}(L;p) + \frac{1}{1-p}\text{ES}(L;p),$$

$$\text{CTE}(L;p) = \text{VaR}(L;p) + \frac{1}{1-F_L(\text{VaR}(L;p))}\text{ES}(L;p),$$

and, if L is continuous,

$$\text{TVaR}(L;p) = \text{CTE}(L;p).$$

Next, consider a portfolio of K weather derivatives, with aggregate payment $L = L_1 + ... + L_K$. The expected aggregate payment is simply the sum of the individual expected payments,

$$\mathbb{E}(L) = \sum_{k=1}^{K} \mathbb{E}(L_k).$$

However, when we allow for possible dependence among contracts, the variance of the aggregate payment is

$$\text{var}(L) = \sum_{k=1}^{K} \text{var}(L_k) + \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} 2 \cdot \text{cov}(L_k, L_{k'}) \tag{5.1}$$

The first issue when pricing weather derivatives (extreme or otherwise) is to properly address the positive correlation among contracts, with its resulting impact on aggregate loss variability as shown in equation (5.1). It is easy to envision positively correlated payments in a portfolio of weather derivatives, since weather variables are often positively correlated in space. This correlation implies the variance of the aggregate payment exceeds the sum of individual payment variances, and thus risk loadings priced individually would be insufficient for the portfolio as a whole.

Next, consider the challenges when focusing on extreme weather events. Examples of

such events may include the maximum daily temperature exceeding some high threshold, the minimum daily temperature falling below some low threshold, or the minimum monthly rainfall falling below some low threshold. These events can often be written as $\{\max Y \geqslant s\}$ or $\{\min Y \leqslant s\}$ where Y is some weather-related random variable and s is a pre-specified strike value. In all cases we are defining a contract not based on "typical" weather patterns of temperature or precipitation, but on extremes. What is needed to accurately price these contracts is the distribution function of extreme events. Furthermore, when one considers a collection of K derivatives defined at different locations, one must carefully consider if the dependence of extremes is different from the dependence exhibited by non-extreme events. Putting the two issues together, what is needed to price weather derivatives for extreme events is a model that (1) directly targets extremes, and (2) properly incorporates the spatial correlation of weather extremes. One could further extend this to models which incorporate time dependence of extremes as well. However, here we focus on the spatial dependence of weather derivatives for extremes, as that is the largest omission of current methodology.

We begin with the problem of pricing a single weather derivative in the Section 5.2. This is handled through the Generalized Extreme Value distribution (Embrechts, et. al., 1999), which is the only permissible limit of the maxima of independent, identically distributed univariate random variables. We demonstrate the approach by pricing a weather derivative based on extreme summer temperatures in Phoenix, Arizona. In Sections 5.3 and 5.4, we extend our weather derivative pricing model to multiple locations through the use of max-stable processes. These processes capture dependence in spatial extremes. Through large numbers of simulations, we can estimate all marginal variances, covariances, and risk measures when pricing a portfolio of weather derivatives. This information is ultimately incorporated into risk loads added to the pure premiums. From the method presented, pure premiums, risk loads and risk measures for a collection of K spatially dependent weather

derivatives can be obtained.

5.2 Pricing a Weather Derivative at a Single Location

5.2.1 Pricing a Contract Through Simulations

Once we have a fitted model, we can use this to estimate the necessary pure premium and risk loadings through a large collection of simulations. The first two moments of the unknown payment variable L, are

$$\mathbb{E}(L^d) = \int L(m; s, t)^d g(m)\, dm$$

where $L(m; s, t)^d$ is the loss payment for realization $M = m$ raised to the d^{th} power ($d = 1$ or 2), and $g(m)$ is the density function of the maxima. The first type of weather derivative discussed has $L(m; s, t)^d = \alpha^d$ for $m \geqslant s$, which means the integral can be evaluated exactly as

$$\alpha^d \cdot P(M \geqslant s) = \alpha^d \cdot \left(1 - G(s; \hat{\phi})\right) \tag{5.2}$$

Our ultimate goal is to describe a general approach to estimating risk for a collection of weather derivatives. Even if we chose derivatives with a sufficiently simple mathematical payment structure such that closed-form expressions for the moments and risk measures could be obtained, once we extended to the spatial case with $D > 2$ locations we would not have the joint likelihood function for aggregate loss $L = L_1 + ..., +L_D$. Since the primary motivation for using max-stable processes to price weather derivatives is to properly incorporate spatial dependence, we will be forced to drop the search for closed-form expressions for risk measures well before we meet our stated goal. Therefore we focus on numeric approximations to all moments and risk measures, and demonstrate how the approach extends to cover the spatial case. All numeric estimates will rely on a large i.i.d.

sample $M_i \sim G(m)$ for $i = 1, ..., I$, and for each draw we compute the payment $L(m_i)$ for a chosen type of weather derivative.

To estimate the first and second moments, assuming that $\mathbb{E}(|L(M)|^2)$ is finite, by the Strong Law of Large Numbers sample means converge to the first and second moments as (Robert, 2007)

$$\frac{1}{I}\sum_{i=1}^{I} L(m_i) \to \mathbb{E}(L(M)) = \int L(m)g(m)\,dm \quad \text{(almost surely)} \tag{5.3}$$

and

$$\frac{1}{I}\sum_{i=1}^{I} L(m_i)^2 \to \mathbb{E}(L(M)^2) = \int L(m)^2 g(m)\,dm \quad \text{(almost surely).} \tag{5.4}$$

Furthermore, if the fourth moment is finite, as $\int L(m)^4 g(m)\,dm < \infty$, then by the Central Limit Theorem we know that the sample average in equation (5.3) is asymptotically normal with variance $\text{var}(L(M))/I$, and the sample average in equation (5.4) is asymptotically normal with variance $\text{var}(L(M)^2)/I$.

Similarly, all five of the risk measures described (VaR, CTE, TVar, CVar, ES) can be numerically estimated as a function of p once an approximate cumulative distribution function for L is available. Value-at-Risk is estimated as a simple quantile of the loss cumulative distribution function:

$$\widehat{\text{VaR}}(L; p) = w_1 \cdot L_{[j]} + w_2 \cdot L_{[j+1]} \tag{5.5}$$

where $j/n \leqslant p < (j+1)/n$, $L_{[j]}$ is the j^{th} order statistic, and $w_1 + w_2 = 1$ are the weights on the two order statistics. There are various methods for choosing the weights (Hyndman and Fan, 1996), but as the number of simulated points n grows the discrepancies between the various methods for quantile estimation vanish, so we will not focus on them here. For

Expected Shortfall we use

$$\widehat{\mathrm{ES}}(L;p) = \frac{1}{I}\sum_{i=1}^{I} \max\left(L(m_i) - \widehat{\mathrm{VaR}}(L;p), 0\right). \tag{5.6}$$

Conditional Tail Expectation can then be estimated as

$$\widehat{\mathrm{CTE}}(L;p) = \widehat{\mathrm{VaR}}(L;p) + \frac{1}{1-p}\widehat{\mathrm{ES}}(L;p). \tag{5.7}$$

Finally, Conditional Variance is the difference of CTE and VaR as

$$\widehat{\mathrm{CVaR}}(L;p) = \widehat{\mathrm{CTE}}(L;p) - \widehat{\mathrm{VaR}}(L;p). \tag{5.8}$$

One example of a risk-loaded premium based on marginal variance is

$$\hat{P} = \hat{E}(L) + \hat{R}(L) = \frac{1}{I}\sum_{i=1}^{I} L(m_i) + \lambda \cdot \left[\frac{1}{I}\sum_{i=1}^{I} L(m_i)^2 - \left(\frac{1}{I}\sum_{i=1}^{I} L(m_i)\right)^2\right] \tag{5.9}$$

for some dollar amount λ, chosen to satisfy some risk tolerance criteria.

5.2.2 Example: Extreme Temperature in Phoenix, Arizona

As an example of how this model may be used, consider pricing a weather derivative with payments whenever the maximum daily summer temperature in the city of Phoenix, AZ exceeds some high threshold s. On June 26, 1990, Phoenix airport was forced to close because the temperature exceeded 122 degrees Fahrenheit. Aircraft operating manuals did not provide information for takeoff and landing procedures in temperatures above 120 degrees Fahrenheit. The closure caused the predictable sort of economic disruption which accompanies airport closures. We envision a weather derivative as a useful tool in this situation.

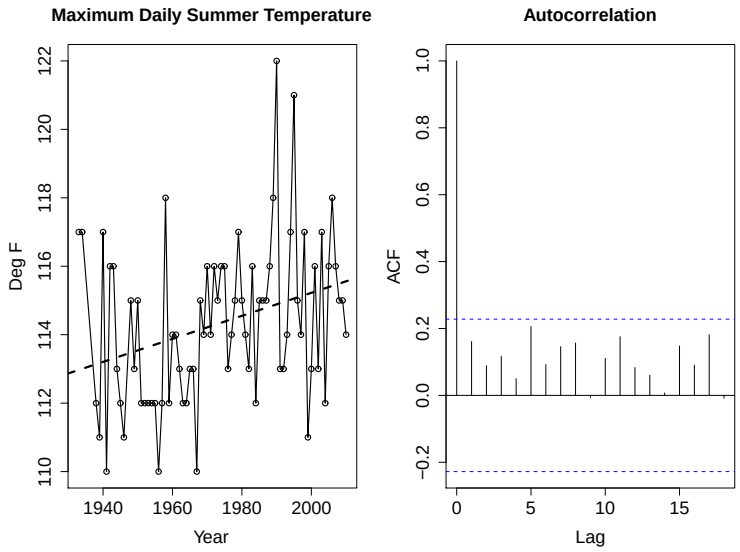

Figure 5.1: Left: Maximum annual summer temperature at Phoenix International Airport from 1933-2010 (with some missing values). The dashed line shows the annual trend, with statistically significant positive slope (p-val = 0.007). Right: Empirical autocorrelation of maximum daily summer temperatures. There is no evidence annual maximum daily temperatures are autocorrelated, as the value at all lags greater than 1 falls below the 95% confidence interval line obtained from white noise sequences.

To price the derivative, we collect maximum daily summer temperatures at the Phoenix airport, $y_{i,j}$ for year i and day j, where $j = 1, ..., 92$ (the 92 days in June, July, and August) for years 1933 to 2010. This data comes from the National Climatic Data Center. For each year i, we take the block maximum $m_i = \max(y_{i,1}, ..., y_{i,92})$, and model these annual maxima m_i as a Generalized Extreme Value distribution. Plotting these data, we observe evidence of a slight positive trend over time (Figure 5.1). A simple linear model of maximum temperature versus year shows a statistically significant positive slope of 0.03363, with p-value 0.007. We also find no evidence that annual maximum temperatures are autocorrelated.

We estimate the GEV parameter ϕ using maximum likelihood estimation, as shown in equation (2.5), but with the possibility of a trend on the location parameter, as $\mu = \mu_1 + \mu_2 \cdot t$ where t is year. Thus, the GEV parameter here is actually $\phi = (\mu_1, \mu_2, \sigma, \xi)$. The maximum

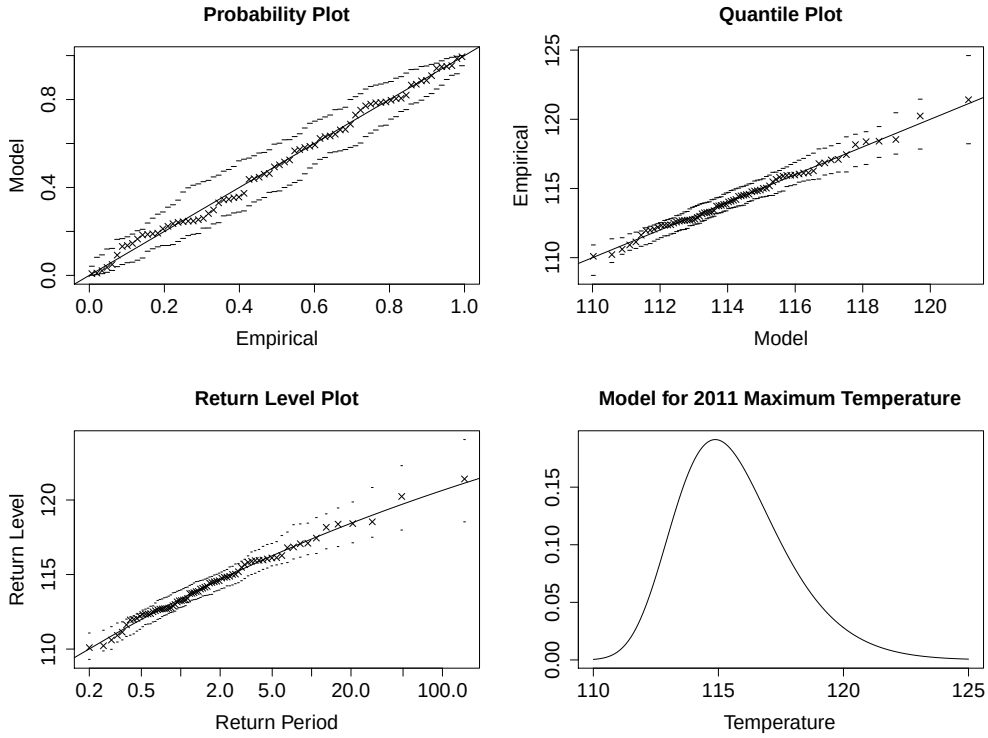

Figure 5.2: Diagnostics from the maximum likelihood fit to the Phoenix summer temperature data. Top left: comparison of empirical and model probabilities. Top right: comparison of empirical and model quantiles. Bottom left: The return period is the expected number of years required for the process to exceed the corresponding return level. Bottom right: Model density function for 2011 maximum summer temperature in Phoenix AZ.

likelihood estimates (with standard errors shown in brackets) are $\hat{\mu}_1 = 113.367$ [0.250], $\hat{\mu}_2 = 0.035$ [0.011], $\hat{\sigma} = 1.931$ [0.176], and $\hat{\xi} = -0.090$ [0.078]. Figure 5.2 shows some common diagnostics and the return level plot. The return level plot shows the expected number of years before an exceedance of a certain level is reached. This is the same as the reciprocal of the probability of a specified exceedance, and forms the basis for statements such as describing an event as a"once every 50 years" event.

With the fitted model for maximum summer temperature, we can estimate the first and second moments of various weather derivative payments in the year 2011. Estimated moments for the three types of derivatives from Section 5.1 are shown in Tables 5.1, 5.2,

Table 5.1: Phoenix, AZ example: Model first and second moments for the type 1 weather derivative with flat payment $L = 1000$ paid whenever the maximum daily temperature $M \geqslant s$ in the year 2011, using equation (5.2).

Threshold s	114	116	118	120	122	124
$\hat{E}(L)$	759.11	391.84	144.11	41.61	9.72	1.79
$\hat{E}(L^2) \cdot 10^{-3}$	759.11	391.84	144.11	41.61	9.72	1.79

Table 5.2: Phoenix, AZ example: Model first and second moments for the type 2 weather derivative with proportional payment $L = 1000 \cdot (m - s)$ in the year 2011, for varying thresholds of s. Estimates are based on $I = 1,000,000$ Monte Carlo draws used with using equations (5.3) and (5.4).

Threshold s	114	116	118	120	122	124
$\hat{E}(L)$	1,882.13	732.20	224.57	56.39	11.59	1.87
$\hat{E}(L^2) \cdot 10^{-3}$	7,336.56	2,369.34	627.82	137.98	24.45	3.26

5.3, and 5.4. As the limit $t \to \infty$, payments under the second and third types are equal.

We also use the simulated losses to provide estimates of each of the risk measures using equations (5.5), (5.6), (5.7), and (5.8). These are shown in Figure 5.3.

Tables like these can be used to price a wide range of weather derivatives. Consider a weather derivative with payment $1000 \cdot (M - 118)$ for $M \leqslant 125$ and 7000 for $M \geqslant 125$, where M is the maximum summer temperature in Phoenix. Using equation (5.9) with $\lambda = 0.0001$, the tables show the pure premium should be $223.89 + 0.0001 \cdot (618.16 \cdot 10^3 - 223.89^2) = 280.69$.

Table 5.3: Phoenix, AZ example: Model first moments for the type 3 weather derivative with proportional payment $L = 1000 \cdot (m - s)$ up to limit $1000 \cdot (t - s)$ in the year 2011, for varying thresholds of s and t. Estimates are based on $I = 1,000,000$ Monte Carlo draws used with using equations (5.3) and (5.4).

t \ s	114	116	118	120	122	124
119	1,766.86	616.93	109.30			
121	1,855.93	705.99	198.37	30.18		
123	1,877.34	727.41	219.78	51.59	6.80	
125	1,881.44	731.51	223.89	55.70	10.90	1.18
∞	1,882.13	732.20	224.57	56.39	11.59	1.87

Table 5.4: Phoenix, AZ example: Model second moments for the type 3 weather derivative with proportional payment $L = 1000 \cdot (m - s)$ up to limit $1000 \cdot (t - s)$ in the year 2011, for varying thresholds of s and t. Estimates are based on $I = 1,000,000$ Monte Carlo draws used with using equations (5.3) and (5.4). Values shown have order 10^3 (multiply each entry by 10^3 to show the true value).

t \ s	114	116	118	120	122	124
119	5,894.14	1,383.33	98.21			
121	6,914.34	2,050.63	412.61	26.27		
123	7,243.24	2,294.71	571.86	100.69	5.84	
125	7,321.98	2,357.23	618.16	130.77	19.70	0.97
∞	7,336.56	2,369.34	627.82	137.98	24.45	3.26

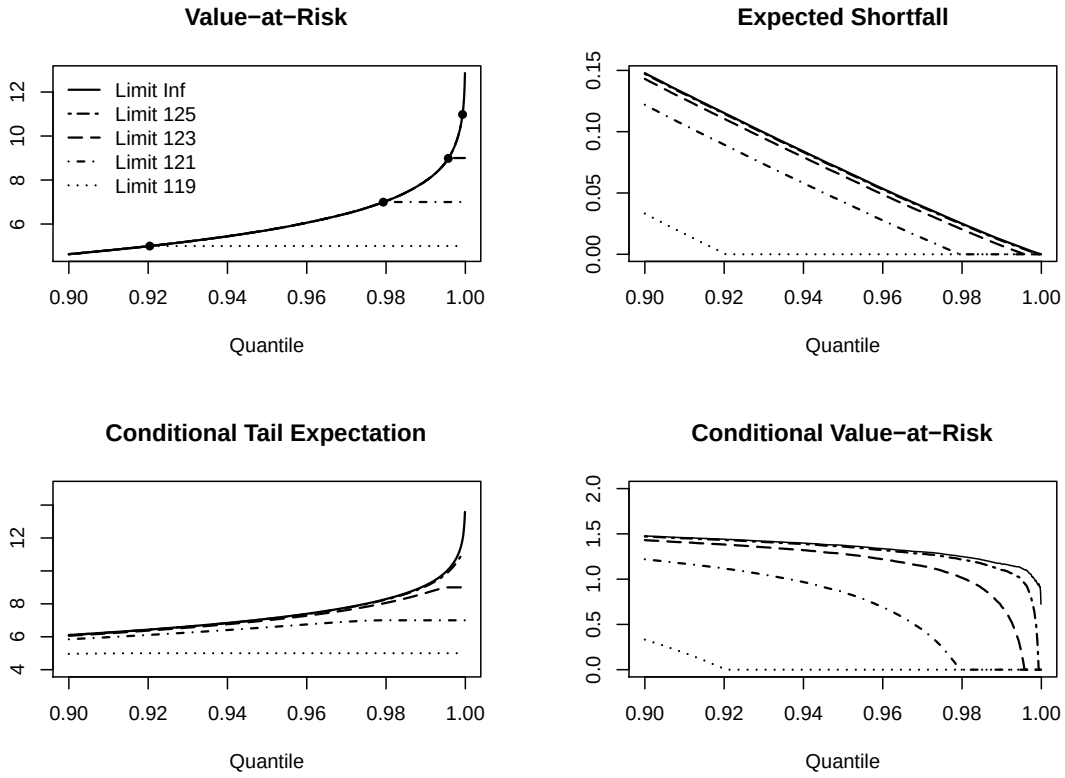

Figure 5.3: Empirical Estimates of the four risk measures for the Phoenix, AZ temperature example. Shown are losses with a strike temperature of 114 degrees and limits of 119, 121, 123, 125 and ∞. Estimates are based on $I = 1,000,000$ simulated losses from the fitted model.

.

Premiums for other limits, strike values, and payment structures can be estimated from the same general approach once a fitted model has been obtained.

5.3 Simulating Losses at Multiple Locations

From the fitted model, we can simulate a collection of max-stable processes $Z_i(x_k), i = 1, ..., I$ for the same $k = 1, ..., K$ locations as the observed data, and then transform back to the original scale of extremes data by transforming margins using the estimated GEV parameter $\hat{\phi}$. From this we can compute the payments from weather derivatives at each location as $L(m_{i,k}; s, t)$.

To price a collection of K weather derivatives, with jointly dependent losses, there are several quantities of interest. First, the total variability of loss payments is

$$\text{var}\left(\sum_{k=1}^{K} L_k\right) = \sum_{k=1}^{K} \text{var}(L_k) + \sum_{k=k'+1}^{K} \sum_{k'=1}^{K} 2 \cdot \text{cov}(L_k, L_{k'}). \tag{5.10}$$

Now consider a portfolio of $K - 1$ derivatives, with the seller deciding whether or not to write a K^{th} derivative. The additional derivative will increase the total portfolio variance by

$$MV_K = \text{var}\left(\sum_{k=1}^{K} L_k\right) - \text{var}\left(\sum_{k=1}^{K-1} L_k\right). \tag{5.11}$$

The covariance for any two derivatives at locations x_k and $x_{k'}$ is

$$\text{cov}(L_k, L_{k'}) = E(L_k \cdot L_{k'}) - E(L_k)E(L_{k'}) \tag{5.12}$$

The quantities in equations (5.10), (5.11) and (5.12) can each be estimated from a large

collection of I simulations. The total variance of the portfolio is

$$\widehat{\text{var}}\left(\sum_{k=1}^{K} L_k\right) = \left[\frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} L(m_{i,k})^2 - \left(\frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} L(m_{i,k})\right)^2\right], \tag{5.13}$$

the marginal variance for adding a K^{th} derivative is

$$\widehat{MV_K} = \widehat{\text{var}}\left(\sum_{k=1}^{K} L(m_{i,k})\right) - \widehat{\text{var}}\left(\sum_{k=1}^{K-1} L(m_{i,k})\right), \tag{5.14}$$

and the covariance between any two derivatives is

$$\widehat{\text{cov}}(L_k, L_{k'}) = \left(\frac{1}{I}\sum_{i=1}^{I} L(m_{i,k}) \cdot L(m_{i,k'})\right) - \left(\frac{1}{I}\sum_{i=1}^{I} L(m_{i,k})\right)\left(\frac{1}{I}\sum_{i=1}^{I} L(m_{i,k'})\right) \tag{5.15}$$

As before, all estimates for the actuarial risk measures will be obtained from empirical estimation on the I simulated aggregate losses.

5.3.1 Simulated Example

We simulated a max-stable process with parameters chosen to mimic annual temperature maxima in North America. The process had unit-Fréchet margins and Whittle-Matérn covariance with dependence parameter $\phi_2 = (c_1, c_2, \nu) = (1, 3, 1)$ for 75 years at 20 locations randomly placed on a 10 by 10 grid. Call the vertical dimension latitude (lat) and the horizontal longitude (lon). To make this data consistent with annual temperature maxima, at each location we transformed to the GEV scale by specifying parameters $\mu(x) = 110 - lat/2$, $\sigma(x) = 1.5 + lat/5$, and $\xi(x) = -0.1$. The basic idea was to imagine higher latitude locations having overall lower extreme temperatures, but higher variability of extremes. We used these transformations for each of the 20 locations to produce a max-stable process with GEV$(\mu(x), \sigma(x), \xi(x))$ margins. We fix this as the "observed" data.

Next, we analyzed these data using composite likelihood estimation. For each location

Table 5.5: Simulated payments for weather derivatives paying 1 when $T \geqslant 112$. This information shows the marginal variance for adding the 20^{th} policy is 20.874 - 19.753 = 1.121.

Event	L_1	L_2	...	L_{19}	$\sum_{k=1}^{19} L_k$	L_{20}	$\sum_{k=1}^{20} L_k$
1	0	0	...	0	0	0	0
2	1	0	...	1	8	1	9
3	1	1	...	0	4	0	4
...	...	...	...	...	...	...	...
1,000,000	0	0	...	0	1	0	1
Mean	0.209	0.162	...	0.123	3.028	0.069	3.097
Variance	0.165	0.136	...	0.108	19.753	0.064	20.874

x_k, we obtained a maximum likelihood estimate $\hat{\phi}_1(x_k) = \hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k), k = 1, ..., K$ using equation (2.5) and used these to transform each margin to unit-Fréchet using equation (2.4). We fit a max-stable process with Whittle-Matérn correlation with nugget parameter 1, and obtained maximum composite likelihood estimate $\hat{\phi}_{2MCLE} = (\hat{c}_2, \hat{\nu})$. Using this fitted model, we simulated a large number of processes and transformed them to the temperature scale using $\hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k)$ at each location x_k.

Thus we have described a means of simulating $i = 1, ..., I$ extreme temperature events $m_{i,k}$ for locations $x_1, ..., x_K$ from our fitted model. This information was used to estimate payments for weather derivatives by computing $L(m_{i,k}; s, t)$. Table 5.5 shows results from one simulation. We simulated 1,000,000 extreme temperature events at the same 20 locations, and used these to compute payments $L_{i,k}$ for $i = 1, ..., 1,000,000$ and $k = 1, ..., K$. Shown are payments for a weather derivative paying 1 when $T \geqslant 112$, the aggregate payment $\sum_{k=1}^{19} L_k$, and the payments for a possible weather derivative L_{20}. The marginal variance of adding the 20^{th} derivative was estimated as $20.874 - 19.753 = 1.121$, which is clearly much larger than 0.064, the variance of the derivative ignoring dependence terms. We also estimated the actuarial risk measures, shown in Figure 5.4.

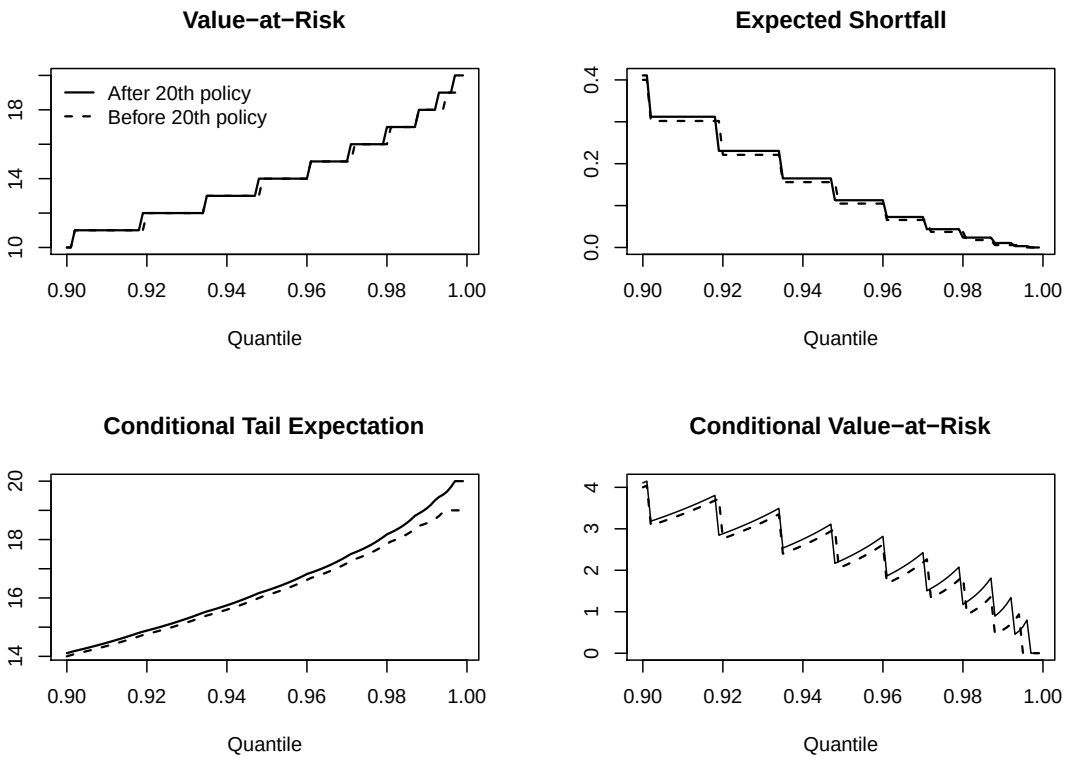

Figure 5.4: Empirical Estimates of the four risk measures for the simulation discussed in Section 5.3.1. The saw-tooth pattern arises because the aggregate loss distribution of L is discrete on support 0,1,...,20, and not because of any poor quality estimation

.

5.3.2 Simulation Study of Performance

We evaluated the performance of this method in estimating the marginal variance of adding a 4^{th} weather derivative to an existing portfolio composed of L_1, L_2, and L_3. This quantity is key to pricing a risk load for L_4. We randomly placed $K = 25$ locations on a 10 by 10 grid, and randomly selected 4 of these to represent locations of weather derivatives. The target quantity was MV_4, the marginal variance of adding a fourth derivative. We estimated this quantity using two methods:

1. Estimate MV_4 using equation (5.14), which accounts for spatial dependence by fitting a max-stable process and uses simulations from the model, with fitted parameter is $\widehat{\phi} = (\hat{\mu}(x_1), \hat{\sigma}(x_1), \hat{\xi}(x_1), ..., \hat{\mu}(x_4), \hat{\sigma}(x_4), \hat{\xi}(x_4), \hat{c_2}, \hat{\nu})$.

2. Estimate MV_4 using equation (5.13), which fits a GEV to the data at location $k = 4$ but does not account for spatial dependence among the derivatives, with fitted parameter is $\widehat{\phi} = (\hat{\mu}_4, \hat{\sigma}_4, \hat{\xi})$.

Call the true full parameter $\phi = (\mu(x_1), \sigma(x_1), \xi(x_1), ..., \mu(x_K), \sigma(x_K), \xi(x_K), c_2, \nu)$, and the estimated parameter $\hat{\phi}$. Call the true marginal variance $MV(\phi)$, and an estimate $\widehat{MV}(\hat{\phi})$. The true marginal variance was found by simulating $I = 1,000,000$ realizations of a max-stable process under the true parameter ϕ, and using equation (5.14). Method 1 uses the same approach, but with estimated parameter $\hat{\phi}$. The first measure of error we use is mean relative error,

$$MRE = \frac{1}{J} \sum_{j=1}^{J} \frac{\left(\widehat{MV}_j(\hat{\phi}) - MV_j(\phi) \right)}{MV_j(\phi)} \tag{5.16}$$

where $j = 1, ..., 500$ refers to a simulation run. This choice preserves the sign of estimation error. Results are shown in Figure 5.5. Here, we see the peril of ignoring spatial dependence of losses in a collection of weather derivatives. The right column shows that as the range of

Table 5.6: Mean absolute error (MAE) in estimating MV_4, the marginal variance of adding a fourth weather derivative to a portfolio. Estimates are based on 25 locations and are computed from equation (5.17) for 50, 100, 250, and 500 years of data. For each spatial dependence range shown, error falls as the number of years of data increases. Values shown have order 10^{-3} (multiply each entry by 10^{-3} to show the true value).

Range c_2	Y=50	Y=100	Y=250	Y=500
Short 0.5	160	121	71	51
Medium 3	233	139	108	64
Long 8	231	151	87	64

the spatial dependence increases, the underestimation bias of estimating marginal variance MV_4 increases. The left column shows the unbiased results obtained from incorporating dependence using the method of this dissertation.

We also show the asymptotic results in Table 5.6. Here, we use a slight variant of estimation error called mean absolute error,

$$MAE = \frac{1}{J}\sum_{j=1}^{J}\frac{|\widehat{MV}_j(\hat{\phi}) - MV_j(\phi)|}{MV_j(\phi)}. \tag{5.17}$$

This choice does not preserve the sign of error, but is more suited to showing asymptotic results. Results from 150 simulations are shown in Table 5.6, using a variety of number of years and dependence ranges. For all dependence ranges shown, the error in estimation falls as more data is available.

5.4 Pricing a Collection of Weather Derivatives

Insurers price a new policy based on how much *marginal risk* is added to their entire portfolio. Under many common forms of insurance (automobile, homeowners, etc.) losses are independent, and the increase in total portfolio variance exactly equals the variance of loss of the new policy (all covariances are zero). It thus does not matter in what order policies are added to a portfolio. However, when losses are correlated the order policies

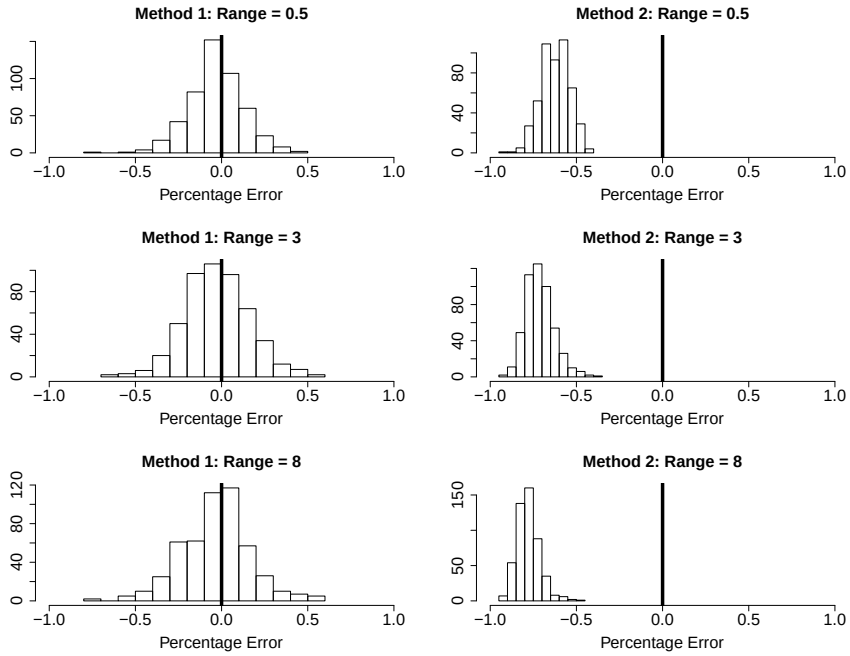

Figure 5.5: The mean relative error in estimating the marginal variance MV_4 under three spatial dependence scenarios, with 500 simulations in each. The top row corresponds to a short-range dependence process, the middle row is medium-range, and the third row shows long-range spatial dependence. The left column uses the approach outlined in this dissertation, which incorporates spatial dependence by first fitting a max-stable process to the temperature, and thus recognizes the correlation of losses. The right column ignores spatial dependence. Results are plotted as mean relative error from the true marginal variance, using equation (5.16).

.

are priced does matter. This has been recognized and discussed in actuarial journals (Feldblum, 1990; Kreps, 1990; Philbrick, 1991; Gogol, 1992), and is best illustrated by a toy example.

Consider two correlated policies, L_1 and L_2, and assume the risk load will be based on marginal variance. Let us first view the portfolio as a whole, and determine a total risk load. When both policies are active in a portfolio, it should be clear that the total portfolio risk load is $\lambda\left(\text{var}(L_1) + \text{var}(L_2) + 2 \cdot \text{cov}(L_1, L_2)\right)$. The additional risk arising from the correlated losses is $2\lambda \cdot \text{cov}(L_1, L_2)$, and this would have to be reflected in the premiums for L_1 and L_2 somehow. But next, consider building up the portfolio one policy at a time:

1. First, policy L_1 is priced and the risk load is $\lambda \cdot \text{var}(L_1)$ (there are no other policies yet in the portfolio, and thus no covariance terms).

2. Next, L_2 is added, and the marginal risk load is $\lambda(\text{var}(L_2) + 2 \cdot \text{cov}(L_1, L_2))$. Notice the entire portfolio risk load would be added to the premium for L_2 because it was priced after L_1.

3. When L_1 renews in the following year, it is being added to a portfolio consisting of L_2 and thus receives risk load $\lambda(\text{var}(L_1) + 2 \cdot \text{cov}(L_1, L_2))$. The renewal premium for L_1 is larger than the premium paid in step 1.

4. When L_2 later renews, the marginal risk load is again $\lambda \cdot (\text{var}(L_2) + 2 \cdot \text{cov}(L_1, L_2))$. Thus the premium for L_2 remains unchanged.

The sum of these renewal risk loads is $\lambda\left(\text{var}(L_1) + \text{var}(L_2) + 4 \cdot \text{cov}(L_1, L_2)\right)$, which does not equal the total portfolio risk load $\lambda\left(\text{var}(L_1) + \text{var}(L_2) + 2 \cdot \text{cov}(L_1, L_2)\right)$. The covariance has been double counted. The point here is that if one computes risk loads for correlated policies based only on marginal variance, then the sum of individual risk loads will not equal the required total portfolio risk load.

The approach we take to pricing a portfolio of dependent weather derivatives follows the work of Mango (1998). In Mango's terminology, we use the covariance-share method, which apportions the total covariance between policies L_j and L_K and computes risk loads as

$$R(L_K) = \lambda \left(\text{var}(L_K) + 2 \sum_{j=1}^{K-1} a_{j,K} \cdot \text{cov}(L_K, L_j) \right)$$

for any $0 \leqslant a_{j,K} \leqslant 1$. This quantities $a_{j,K}$ are chosen to split the respective covariance terms and ensure the sum of individual renewal risk loads matches the total portfolio risk load. One reasonable choice splits the total covariance in proportion to the expected losses of policies j and K, as

$$a_{j,K} = \frac{\mathbb{E}(L_K)}{\mathbb{E}(L_j) + \mathbb{E}(L_K)}.$$

Under this choice, we always have $a_{j,K} + a_{K,j} = 1$, so the risk loads will always be renewal-additive. Relevant quantities are estimated from large numbers of event simulations, and the risk load is

$$\hat{R}(L_K) = \lambda \left(\widehat{\text{var}}(L_K) + 2 \sum_{j=1}^{K-1} \hat{a}_{j,K} \cdot \widehat{\text{cov}}(L_K, L_j) \right) \tag{5.18}$$

using equations (5.13) and (5.15), where

$$\hat{a}_{j,K} = \frac{\frac{1}{I} \sum_{i=1}^{I} L(m_{i,K})}{\frac{1}{I} \sum_{i=1}^{I} L(m_{i,j}) + \frac{1}{I} \sum_{i=1}^{I} L(m_{i,K})}. \tag{5.19}$$

This procedure ensures that the sum of individual risk loads will equal to required total portfolio risk load, and is demonstrated in the application in Section 6.1.

6

Applications

In this chapter we illustrate how the ABC tripletwise method can be used to fit max-stable processes to spatial extremes data, and how these models can then be used in actuarial or risk management applications. In the first example, extreme summer temperatures in the Midwest are modeled as a max-stable process with the intention of pricing a collection of weather derivatives at a set of locations. First we use the composite likelihood approach to build a model and estimate all quantities associated with actuarial risk, and then we use the ABC approach to do the same. Risk estimates based on the ABC model are larger, a direct consequence of incorporating parameter uncertainty. In the second example, we model temperature extremes (this time October minima) and demonstrate how this fitted model can be used as a rare-event generator, providing a large set of simulated rare events which goes beyond the data record and serves as a starting point for catastrophe ratemaking in insurance.

6.1 Pricing Derivatives for Midwestern Temperature

6.1.1 Composite Likelihood Approach

We illustrate the methodology on US temperature data. The data, freely available from the National Climatic Data Center (`http://cdiac.ornl.gov/ftp/ushcn_daily/`),come from 39 locations in the midwestern United States with complete summer (June 1 - August 31) temperature records from 1895 to 2009. All sites are located between longitudes 93 and 103 degrees west, and latitudes 37 to 45 degrees north, shown in Figure 6.1. We use all 39 locations to estimate the max-stable process, but we only consider weather derivatives at 4 of these locations, labeled 1-4 and drawn with triangles on the figure. Call the maximum summer temperature at these $K = 4$ locations $M_{i,k}$, with payments $L_{i,k}$ defined as

1. $L_{i,1} = 1000$ if $\{M_{i,1} \geqslant 107\}$ and 0 otherwise
2. $L_{i,2} = 300 \cdot (M_{i,2} - 105)$ when $\{105 \leqslant M_2 \leqslant 110\}$, 1500 when $\{M_{i,2} \geqslant 110\}$, and 0 otherwise
3. $L_{i,3} = 200 \cdot (M_{i,3} - 105)$ if $\{M_{i,3} \geqslant 105\}$ and 0 otherwise
4. $L_{i,4} = 200 \cdot (M_{i,3} - 102)$ if $102 \leqslant \{M_{i,4} \leqslant 112\}$, 2000 when $\{M_{i,4} \geqslant 112\}$ and 0 otherwise

These four particular choices for payments are arbitrary, and thresholds were selected simply to target extremes in the temperature data. The application proceeds in two steps. The first is to use data from all 39 locations to fit a max-stable process in the study region, and the second is to then simulate temperature events from the fitted model only at locations 1-4 to estimate the renewal-additive risk load and premium for adding a weather derivative at location 4.

To investigate the possibility of a trend in maximum daily temperatures over time, we fit simple linear models to maximum daily temperature versus year, but found only 4 out of

39 locations showed statistically significant slopes at the $p = 0.01$ level (this lower level was selected to reduce the false-positive rate which occurs with multiple tests). Furthermore, all four slopes were negative. We also fit GEV models to data from each station allowing for a time-varying location parameter as $\mu_k = \mu_{k,0} + \mu_{k,1} \cdot t$, where t is year, but found only 7 differed significantly from 0 (again at the p=0.01 level), and again, all were negative. These 7 locations included the four identified from the simple linear regression models, along with three additional locations. These locations were spread throughout the study region, and showed no discernible spatial pattern or clustering. We concluded that there was no evidence of a widespread shift in maximum temperatures over time throughout the entire region, and dropped the time-varying GEV location parameter. However, just as a precaution we also conducted a separate analysis of the data including these 7 negative trends, but found it had little impact on the results.

We fit ordinary GEV models to each station, and obtained maximum likelihood estimate $\hat{\phi} = (\hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k))$ for $k = 1, ..., 39$. Diagnostics like those shown in Figure 5.2 gave no indication the GEV was inappropriate for any of these locations. These fitted models were used to transform data at each location to unit-Fréchet. Next we assessed the appropriateness of using a max-stable process for the dependence. For the Schlather model, one method of detecting anisotropy would be to fit a model to $Z(Ax)$, where z is the 2 dimensional spatial coordinate and A is a 2 by 2 matrix which warps space (see 2.4). The four parameters in A could be estimated along with the dependence parameter ϕ, and one could check if $\hat{A}$ suggested anisotropy. However, we chose the more direct method of fitting a Smith process to the data, which automatically incorporates such a matrix Σ, and did not see strong evidence of anisotropy. The parameter estimate of covariance $\widehat{\Sigma}$ were $\hat{\Sigma}_{11} = 2.064$ [0.020] $\approx \hat{\Sigma}_{22} = 1.897$ [0.020], and $\hat{\Sigma}_{12} = -0.085$ [0.009] ≈ 0, where the number in brackets is the standard error of the estimate (when $\Sigma_{11} = \Sigma_{22}$ and $\Sigma_{12} = 0$, we have perfect isotropy). Having determined the data showed no significant evidence

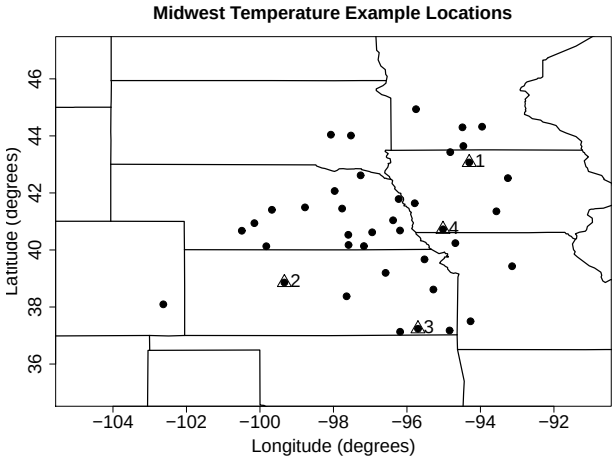

Figure 6.1: Locations of the 39 stations in the Midwest temperature example used to fit the max-stable process to maximum summer temperature. The locations in triangles labeled 1-4 are the places where weather derivatives are priced.

of anisotropy, we then turned back to the more flexible Schlather model with Whittle-Matérn, Cauchy, and powered exponential correlation functions, and found the Whittle-Matérn to be the best with the lowest CLIC score. Using the Whittle-Matérn correlation model, we obtained maximum composite likelihood estimates of the range and smooth as $\hat{c}_2 = 4.6819$ [1.2975] and $\hat{\nu} = 0.3155$ [0.04625], where the number in brackets is the standard error of the estimate.

Next, we simulated $I = 1,000,000$ max-stable processes from our fitted model at the four locations with weather derivatives. Using the GEV estimates $(\hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k))$ for $k = 1, 2, 3, 4$ we transformed the unit-Fréchet margins to GEV at each location. Thus, we had simulations of maximum summer temperatures $M_{i,k}$ for $i = 1, ..., 1,000,000$ at the $k = 4$ locations. From these, we computed the payments $L_{i,k}$ under the four contracts considered. Table 6.1 shows a few of these simulations. Figure 6.2 shows the estimated risk measures.

Table 6.1: Payments for weather derivatives in the Midwestern temperature example. $I = 1,000,000$ simulated extreme temperature events are simulated at locations 1-4 in Figure 6.1. The covariance share quantities $a_{j,K}$ are estimated using equation (5.19).

Event	L_1	L_2	L_3	$\sum_{k=1}^{3} L_k$	L_4	$\sum_{k=1}^{4} L_k$
1	0	0	0	0	0	0
2	0	1500	0	1500	241.1445	1741.145
3	0	0	0	0	0	0
4	0	1500	543.84	2043.84	450.36	2494.20
...	...	...	...	...	...	...
1,000,000	0	0	0	0	0	0
Mean	14.804	619.938	211.762	846.504	195.932	1042.436
Variance ($\cdot 10^{-3}$)	14.585	375.737	178.093	796.398	211.635	1426.356
$\text{Cov}(L_k, L_4)(\cdot 10^{-3})$	21.13	87.674	100.358	209.161		
$\hat{a}_{k,4}$	0.101	0.4192	0.4798			

From equation (5.18) and the information in the table, we compute the risk load for contract L_4 as

$$\widehat{R}(L_4) = \{211.635 + 2\,(0.4798 \cdot 100.358 + 0.4192 \cdot 87.674 + 0.1010 \cdot 21.13)\} \cdot 10^3 \cdot \lambda = 385,712 \cdot \lambda.$$

This is 61.2% of the total increase of $(1426.356 - 796.398) \cdot 10^3 \cdot \lambda = 629,967 \cdot \lambda$. The remainder would be apportioned to the risk loads of first three derivatives as they renew.

When we included the 7 time-varying location parameters, we computed a risk load of $375,298 \cdot \lambda$, a reduction of only 2.7%. This alleviates concerns that we might have wrongly ignored trends in the GEV location parameter μ. If the inclusion of trends on the GEV location parameters $\mu_k, k = 1, ..., 39$ had resulted in a substantially larger risk estimate, it might warrant the inclusion of trends as the more conservative choice. However, the limited statistical evidence of trends combined with such a small reduction in the risk estimate supports dropping them altogether.

We have described a means of pricing a collection of extreme weather derivatives based on simulations from max-stable processes. Naturally, there will be some error between the

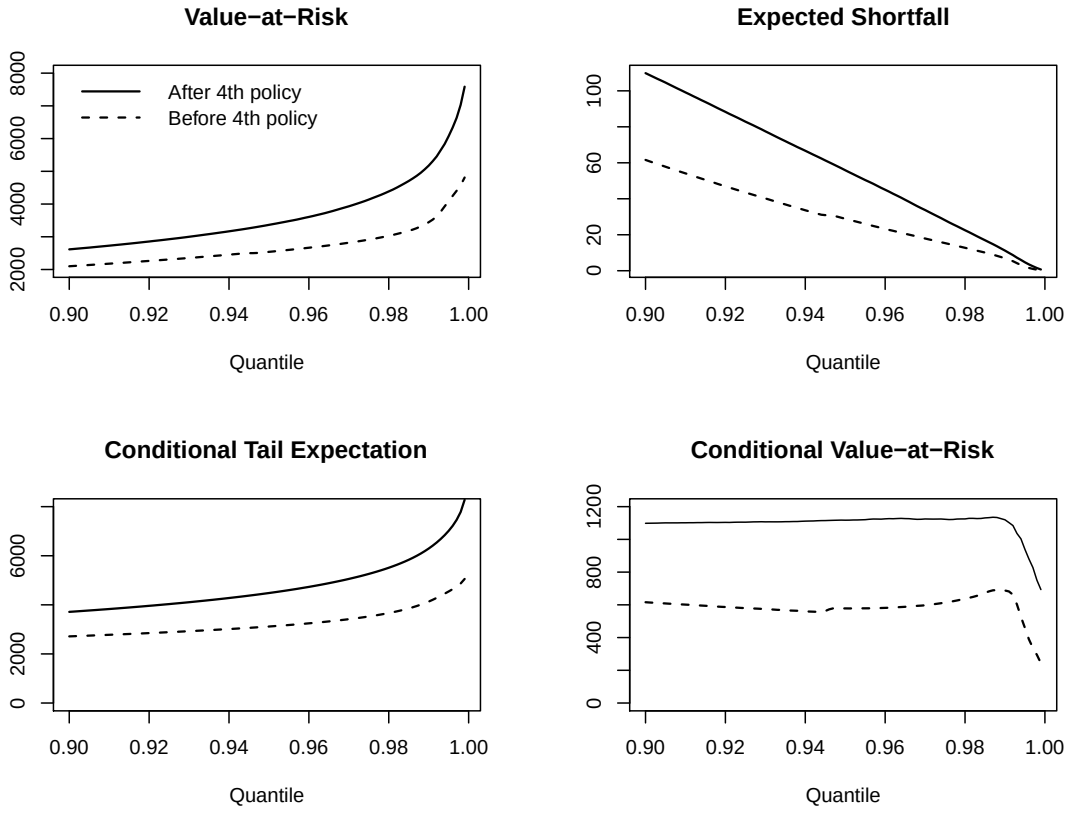

Figure 6.2: Empirical Estimates of the four risk measures for the Midwest temperature data. Measures are estimated before and after the fourth policy is added, this the increase shown is the marginal risk added by the fourth policy.

collection of simulated payments and actual payments. We discuss the errors introduced from model selection, simulations, and parameter estimation.

We have taken the approach to modeling spatial extremes as max-stable processes with Generalized Extreme Value margins, and naturally this model may not be appropriate for some spatial extremes data. For weather derivatives with payments based on maxima (or minima) of some weather variable, models based on block maxima (or minima) of the data make the most sense, and certainly the GEV distribution has appealing asymptotic properties for these data. Coles (2001) discusses diagnostics to check the validity of the GEV for the marginal data. Max-stable processes very naturally extend the GEV to the spatial domain, and are thus the logical choice for spatial block maxima data. While a goal is to extend the approach presented in this dissertation to include non-stationary fields, at present this approach can only handle stationary fields. Within the class of stationary max-stable processes, there are some choices of models. One can model the GEV parameters $\mu, \sigma,$ and ξ with spatial, temporal, or other covariates, and one can consider different correlation functions $\rho(h)$ for the spatial dependence of the max-stable process. In this dissertation we did not show much detail on model selection, however the paper by Padoan et. al. (2010) shows the use of composite likelihood information criteria to handle model selection questions like these.

The computational cost of simulations from a max-stable process is minimal, and thus one can simulate hundreds of thousands or millions of events with relative ease. Errors arising from numerical approximation in estimating the moments of payments assuming some fitted model are thus likely to be quite small, and can be made arbitrarily smaller with greater numbers of simulated events.

The largest source of error in this approach is likely to come from parameter risk - that is, the error in estimating the GEV and max-stable process parameters ϕ_1 and ϕ_2. In the next subsection we discuss how approximate Bayesian computing can incorporate

parameter risk. Within the composite likelihood framework, reducing parameter risk is best handled through fitting the process to more weather data: more years of data, more locations of data, or ideally both. A point worth stressing is that the data used to fit the max-stable process can (and probably should) contain far more locations than the portfolio of weather derivatives. By adding additional points of data to fit the process, one reduces the parameter risk associated with estimating ϕ_2, the spatial dependence parameter.

Our analysis is really a two-step procedure: the first transforms GEV margins to unit-Fréchet by obtaining parameter estimate $\hat{\phi_1} = (\hat{\mu}(x_1), \hat{\sigma}(x_1), \hat{\xi}(x_1), ..., \hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k))$, and then in a second step we fit a max-stable process to the transformed data to obtain dependence parameter estimates $\hat{\phi_2} = (\hat{c_2}, \hat{\nu})$. We should point out that a single step procedure is possible, and is implemented in the package `SpatialExtremes`, but this has two drawbacks. The first is that the numeric optimization of the likelihood needs to maximize a high dimension parameter. In our example on Midwestern temperature data, the dimension would be $39 \cdot 3 + 2 = 119$ (and even larger if we kept time-varying GEV location parameters). The dimension raises concerns that the numeric optimizer may converge to a local maxima, not the global one. A second drawback is that single-step maximization of a max-stable process can be a painfully slow process, requiring orders of magnitude more time than a two-step procedure. With these drawbacks in mind, the two-step procedure was selected.

One final comment is the potential mismatch between past and future weather extremes, particularly in the context of climate change. One can model the location μ and scale σ parameters of the GEV with time covariates to allow for the possibility of non-stationary maxima in time. It is much less common to model the shape parameter ξ as anything other than a fixed number. We have illustrated the use of time covariates for modeling the location parameter as $\mu = \mu_1 + \mu_2 \cdot t$ in the Phoenix airport temperature example. We caution readers not to extrapolate models such as these too far into the future.

6.1.2 Adaptive Approximate Bayesian Computing Approach

Here we repeat the example from the previous section using the AABC method instead of the MCLE approach. Our aim is to incorporate parameter uncertainty into estimates of the risk measures and risk premium.

As before, we fit ordinary GEV models to each station, and obtained maximum likelihood estimate $\hat{\phi}_1 = (\hat{\mu}(x_k), \hat{\sigma}(x_k), \hat{\xi}(x_k))$ for $k = 1, ..., 39$. These fitted models were used to transform data at each location to unit-Fréchet. We next ran the AABC procedure to obtain an approximate posterior distribution for the range and smooth parameters. Priors were independent, uniform(0,10).

1. Run the ABC rejection algorithm exactly as described in the section 4.1 but with $I = 100,000$ simulations to produce a first approximation $\phi_1^{(1)}, ..., \phi_J^{(1)}$ (the $J = 1000$ particles filtered as $(\phi_i' : d_i \leqslant \epsilon_P)$, where ϵ_P is the 0.1% percentile of d_i)

2. Compute Ω as twice the empirical variance of $\phi_1^{(1)}, ..., \phi_J^{(1)}$.

   (a) Resample a particle ϕ^* from $\phi_1^{(1)}, ..., \phi_J^{(1)}$

   (b) Mutate using kernel $K(\phi' \mid \phi^*) = \mathcal{N}(\phi^*, \Omega)$

   (c) Simulate $Z' \mid \phi'$, compute summary s' and weighted distance $d_2(s, s')$ from equation (4.1)

   (d) (Repeat 50,000 times)

3. Filter the 50,000 particles as $(\phi_i' : d_i \leqslant \epsilon_P)$, where ϵ_P is the 1.0 % percentile and d_i is the measure d_2 applied to the i^{th} particle. This ensures exactly 500 particles are accepted. Call these $\phi_m^{(2)}, m = 1, ..., 500$.

4. For accepted particle $\phi_m^{(2)}$ compute weight

$$w_m \propto \frac{1}{\sum_{j=1}^{J} \frac{1}{J} \cdot \mathcal{N}(\phi_m^{(2)} \mid \phi_j^{(1)}, \Omega)}$$

where $\mathcal{N}(\phi_m^{(2)} \mid \phi_j^{(1)}, \Omega)$ is the density of a multivariate normal with mean $\phi_j^{(1)}$ and variance Ω evaluated at the point $\phi_m^{(2)}$.

The result of this algorithm is a weighted sample of particles $\phi_1, ..., \phi_{500}$ with weights $w_1, ..., w_{500}$. These are shown in Figure 6.3. Our aim is to use these particles to simulate a collection of $I = 1,000,000$ extreme temperature events to price the collection of weather derivatives. To incorporate parameter uncertainty, we sampled one of the accepted ϕ_m with sampling probability w_m, and simulated a max-stable process conditional on this draw ϕ_m at the four locations $k = 1, 2, 3, 4$. We used the estimated parameter $\hat{\phi}_1$ to transform this unit-Fréchet process back to the temperature scale, and then computed the four losses L_1, L_2, L_3, and L_4 accordingly. This was repeated $I = 1,000,000$ times, and results are shown in Table 6.2. Again using equation (5.18) and the information in Table 6.2, we computed the risk load for contract L_4 as

$$\widehat{R}(L_4) = \{212.842 + 2\,(0.479 \cdot 103.224 + 0.414 \cdot 89.153 + 0.107 \cdot 23.122)\} \cdot 10^3 \cdot \lambda = 390,497 \cdot \lambda.$$

Notice this is slightly larger (1.24%) than $385,712 \cdot \lambda$, the estimated risk load when ignoring parameter uncertainty. In this particular case the overall increase in the risk load when incorporating the parameter uncertainty of spatial dependence parameter ϕ is somewhat small, though this does not mean it would always be small, or that such a small increase would always result in a negligible increase in a risk-based premium. We also obtain empirical estimates of the four risk measures, before and after the fourth policy is added. These are shown in Figure 6.4.

As expected, empirical estimates of all four risk measures are greater when the AABC approach is used instead of the MCLE approach. This is shown in Table 6.3. For example, the estimated marginal increase in VaR at the 95^{th} percentile was 828.8 when using the AABC approach, but 824.4 when using the MCLE approach. In fact, marginal increases

Table 6.2: Payments for weather derivatives in the Midwestern temperature example using the AABC approach to fitting the max-stable process. $I = 1,000,000$ simulated extreme temperature events are simulated at locations 1-4 in Figure 6.1. The covariance share quantities $a_{j,K}$ are estimated using equation (5.19).

Event	L_1	L_2	L_3	$\sum_{k=1}^{3} L_k$	L_4	$\sum_{k=1}^{4} L_k$
1	0	181.87	0	181.87	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	1434.04	781.45	2215.49	2804.67	
...	...	...	...	...	...	...
1,000,000	0	0	0	0	0	0
Mean	15.06	625.984	209.213	850.257	196.721	1046.978
Variance ($\cdot 10^{-3}$)	14.833	376.614	176.989	798.956	212.842	1442.798
$\text{Cov}(L_k, L_4)(\cdot 10^{-3})$	23.122	89.153	103.224	215.5		
$\hat{a}_{k,4}$	0.107	0.414	0.479			

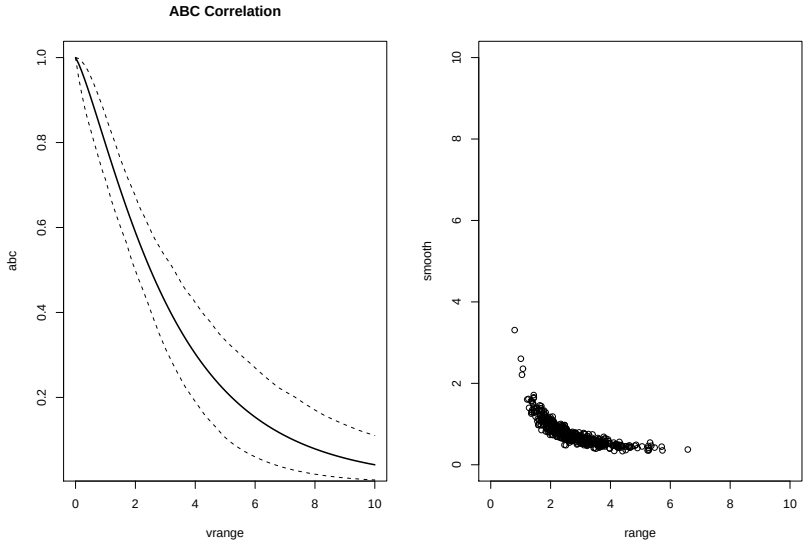

Figure 6.3: Left Panel: AABC posterior shown on the space of correlation functions. The solid line is the pointwise posterior mean (equation (4.2)) and the dashed lines form an empirical estimate of a pointwise 95% credible interval. Right Panel: The 500 accepted particles from the AABC posterior shown on the parameter space $\nu \times c_2$.

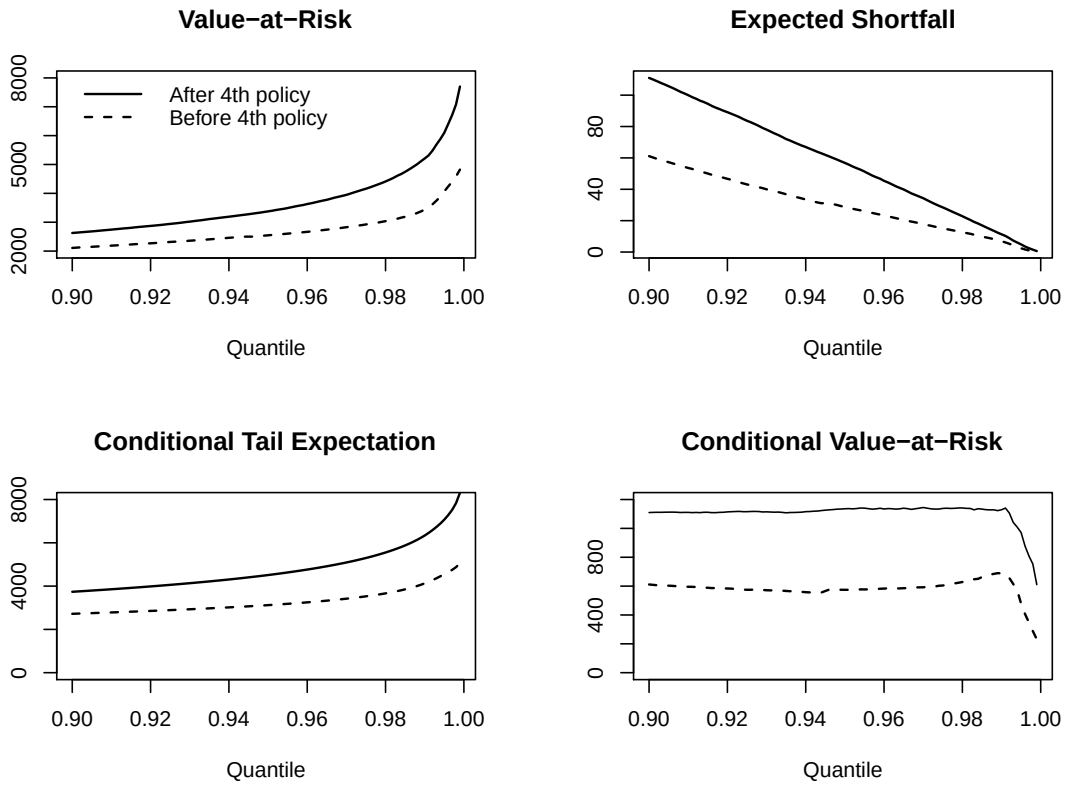

Figure 6.4: Empirical estimates of the four risk measures for the Midwest temperature data, using the AABC approach to fit the max-stable process. Measures are estimated before and after the fourth policy is added, this the increase shown is the marginal risk added by the fourth policy.

Table 6.3: Estimated marginal increases in the four risk measures at the 90^{th} and 95^{th} percentiles, using both the AABC and MCLE approach. In all cases, the AABC approach estimates a slightly higher marginal increase in risk load, a reflection of the incorporated parameter uncertainty.

	90^{th} percentile		95^{th} percentile	
Risk Measure	AABC	MCLE	AABC	MCLE
VaR	518.6	516.7	828.8	824.4
ES	49.9	48.6	28.1	27.34
CTE	1017.1	1002.7	1390.4	1371.3
CVaR	498.5	485.9	561.7	546.9

for all four risk measures were higher when the AABC method was used instead of the MCLE. This is a direct consequence of incorporating parameter into predictions.

6.2 Frost Risk for the Texas Cotton Industry

We further illustrate the methodology of the approximate Bayesian computing tripletwise extremal coefficient approach on US temperature data in northern Texas, with the aim of modeling the acreage of cotton at risk of an October freeze. Data on crop losses taken from the United States Department of Agriculture Risk Management Agency shows that between 1989 and 2008, Texas cotton losses caused by freezing totaled $108,478,787. Of these losses, fully 67.8% ($73,642,461) occurred in the month of October.

The data are daily minimum temperature data taken from 30 gauged sites centered around northern Texas in the United States, freely obtained from the Global Historical Climatology Network. All sites are between 104 and 98 degrees west longitude and 31 to 37 degrees north latitude. We required stations have at least 90% of daily October values for at least 90% of the years between 1935 and 2009. This region is shown in Figure 6.5. Also shown are the 58 counties which jointly comprise the four Texas agricultural districts responsible for 82.3% of all Texas upland cotton acreage (in 2009). For each year and location, we took the minimum daily temperature in the month of October. The aim of

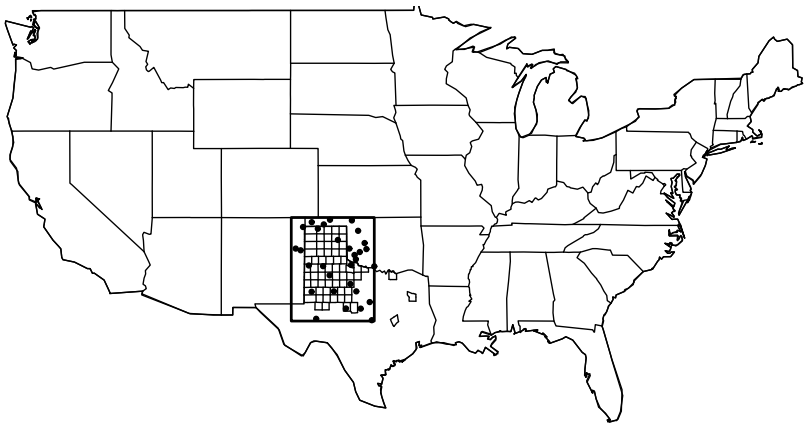

Figure 6.5: Locations of the 30 gauged sites and 58 counties in the primary cotton growing region of Texas.

the analysis was to estimate the spatial dependence of the process through $\rho()$, and use this information to estimate the number of acres of cotton at risk of an early freeze through simulations.

First we transformed data at each location to unit-Fréchet by fitting the marginal univariate data to the Generalized Extreme Value distribution and obtained maximum likelihood estimates of the location-specific Generalized Extreme Value parameters $\mu(x), \sigma(x), \xi(x)$ for $x_1, ..., x_{30} \in X$. These estimates are shown in Figure 6.6. The location parameter and scale parameter were both influenced heavily by spatial location, whereas the shape parameter showed no discernible relationship to spatial placement. We viewed this transformed data as a realization from a max-stable process with unit-Fréchet margins, and proceeded with the ABC approach to fitting max-stable processes.

Next we estimated the tripletwise extremal coefficients for the 4060 unique triplets using equation (2.16). We clustered these into $K = 100$ groups using Ward's method. Our summary statistic for the data was the average within K groups, $s = (\bar{\theta}_1, ..., \bar{\theta}_{100})$ for $K = 100$ groups. We considered the Powered Exponential correlation function, with uniform priors for the range c_2 as U[0, 10] and for the smooth ν as U[0, 2] (since this is the

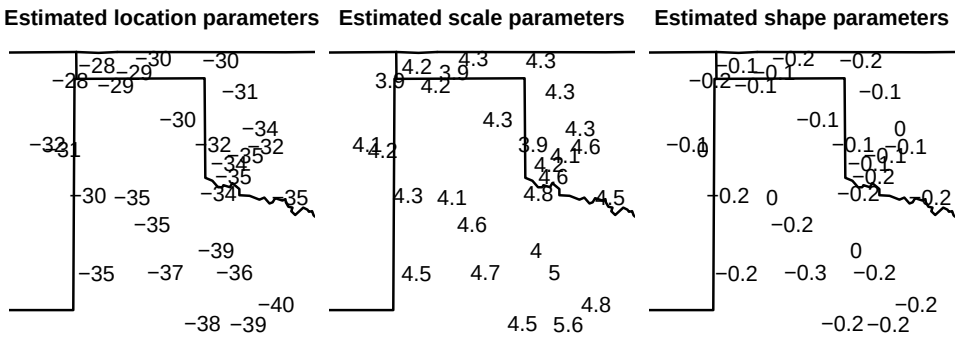

Figure 6.6: Estimates of the Generalized Extreme Value parameters at each of the 30 gauged locations. The location $\mu(x)$ and scale $\sigma(x)$ show spatial dependence, whereas the shape parameter $\xi(x)$ does not. Heavy lines are US state borders.

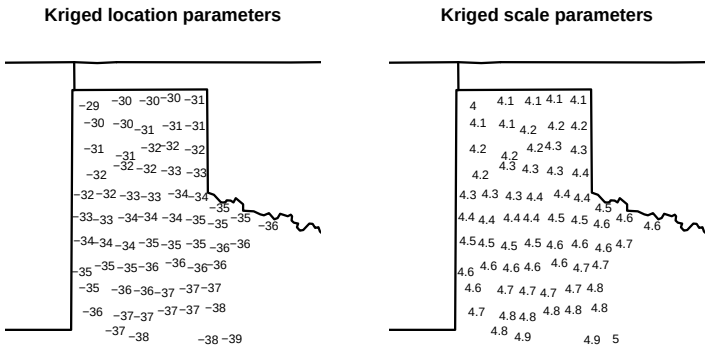

Figure 6.7: Kriged Generalized Extreme Value parameters at each of the 58 ungauged locations. Heavy lines are US state borders.

full permissible range). For this model, this prior allows for a full range of spatial processes on the scale of the observed data in X. We drew 1,000,000 draws from the prior, and ran the approximate Bayesian computing algorithm.

The threshold ϵ was set as the 0.02% percentile of d_i. This ensured exactly 200 particles were accepted for the approximate posterior. The pointwise mean of these 200 accepted functions $\rho(h;\phi')$ was taken as the estimate of the correlation function. Approximate pointwise 95% credible intervals were estimated as the pointwise 2.5% and 97.5% quantiles, taken at each h. This is shown in Figure 6.8.

The primary aim of this application was to estimate the number of acres of cotton at risk of an October freeze. To extrapolate the max-stable process to the 58 ungauged county centroids, we obtained the county centroids from the US Census Bureau, and extrapolated location-specific Generalized Extreme Value parameters for each of these by using the standard spatial Kriging in the `R` package `fields`. We found estimates of the location $\mu(x)$ scale $\sigma(x)$ parameters varied with location, whereas the shape parameter $\xi(x)$ showed no discernible relationship to spatial location (Figure 6.6). Thus, for an ungauged location x' we used Kriged values of $\mu(x')$ and $\sigma(x')$, but took $\xi(x') = \frac{1}{30}\sum_{i=1}^{30}\hat{\xi}(x_i)$, the average of the 30 estimates $\hat{\xi}(x)$ from the gauged locations. Kriged values for the location and scale are shown in Figure 6.7.

We generated simulations from the fitted model not with the intent of matching them to observed data, but rather to calculate a distribution of cotton losses in hypothetical future years under the same climate. We sampled ϕ (with replacement) from the approximate posterior and simulated a max-stable process with unit-Fréchet margins and Powered Exponential correlation at the 58 ungauged county centroids. We used the Kriged location-specific Generalized Extreme Value parameters to transform this back to a temperature scale at each of the 58 county centroids. If the minimum October temperature of the county centroid fell below the chosen threshold, we assumed all acres within the county

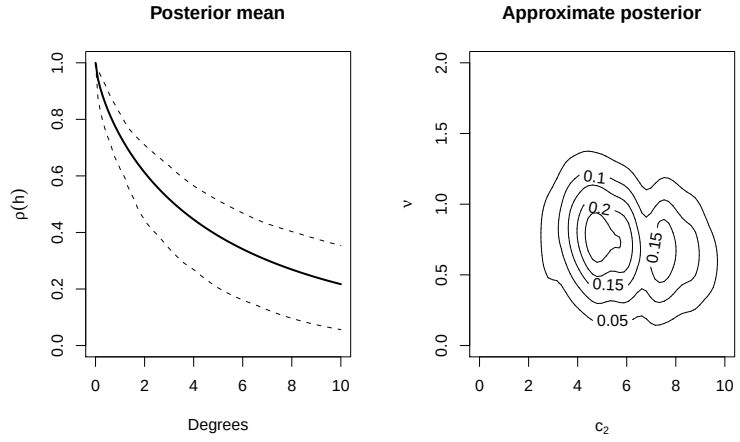

Figure 6.8: Left: approximate Bayesian computing estimate of spatial correlation of the max-stable process, with pointwise 95% credible interval estimates shown as dotted lines. Right: Approximate posterior of the range and smooth parameters.

were exposed to the temperature event. Figure 6.9 shows the log-density of exposed counties for 10,000 simulations at various temperature thresholds. This method of simulation based on the approximate posterior very naturally incorporates parameter uncertainty into the quantity of interest.

Of particular interest to the insurance and agricultural communities are the number of simulations resulting in intermediate exposure, say between 0.5 and 3.5 million acres (contrasted with the all-or-nothing scenarios of 0 or the full 4.1 million acres exposed). This provides evidence that these counties are not completely dependent with respect to an October freeze, and lends additional support to the idea that it may be possible to offer financial or insurance products to protect against crop losses caused by this freeze peril. The method shown here allows for realistic estimation of a distribution of insurance losses, going beyond the empirical distribution calculated from past data. This information could be useful to an actuary interested in calculating the expected payout associated with an insurance policy.

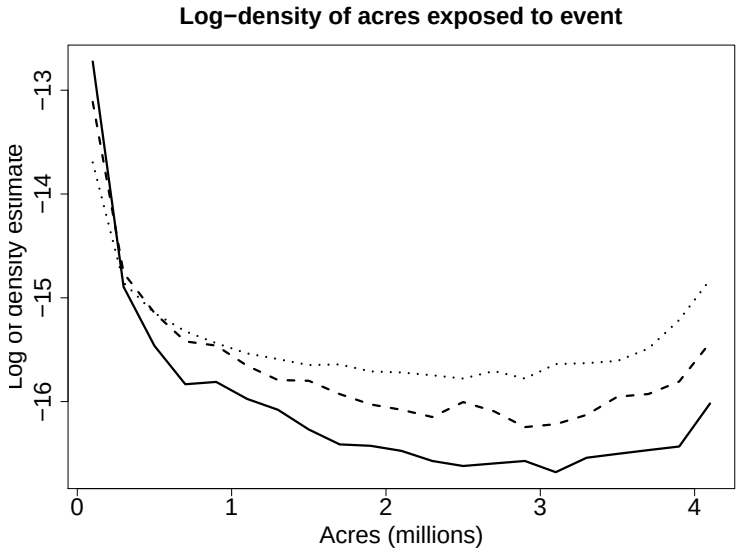

Figure 6.9: A graph showing the number of acres of cotton exposed to minimum temperature event, taken from 10,000 simulations. The temperature events are minimum October temperature falling below 32 degrees F (dotted line), 30 degrees F (dashed line), and 28 degrees F (solid line).

7

Discussion

In this dissertation we have demonstrated an implementation of the approximate Bayesian computing algorithm which can be used for fitting max-stable processes to spatial extremes data. We have shown this approach can result in a lower mean square error than the competing composite likelihood approach for certain classes of processes. We have shown how the ABC approach can consider the spatial dependence through triplets and higher order k-tuples, a methodological improvement over all other inferential methods for max-stable processes. We have outlined the computational costs and challenges associated with the ABC approach, and suggested several strategies for reducing the computational cost. Finally, we have demonstrated how such fitted models can be used to estimate risk of weather extremes and price weather derivatives to serve as insurance against such risks, all in such a way that parameter uncertainty is incorporated into these estimates. These are the advantages of the ABC approach.

Without the joint likelihood function, it is difficult to obtain theoretical results for the ABC method outlined in this dissertation. Identifying optimal choices for tuning parameters is, at present, impossible. This is true not only for this dissertation but all ABC methods which rely on insufficient statistics and finite computing power. Our aim was to demonstrate that particular choices of tuning parameters were acceptable through wide simulation studies. The simulations in this dissertation in no way exhaust the full range of

ABC implementations possible. There are open questions as to how quartets, quintets, or higher order k-tuples may be incorporated into an improved summary statistic, and also there are open questions as to how more efficient ABC samplers could allow the threshold ϵ to be reduced and thus improve the ABC posterior approximation. We are continuing to work on these questions, but the implementation in this dissertation should serve as a solid foundation on which improved ABC implementations can be built.

One area which we have not discussed is model selection. When fitting max-stable processes, there are choices of both the structure of the max-stable model (Schlather vs. alternatives) and of the spatial correlation function within the model. Ideally, we would like a method of deciding amongst those models. Varin and Vidoni (2005) introduced the composite likelihood information criteria (CLIC), which works well with a composite likelihood framework, but this does not have a logical Bayesian interpretation. The typical approach used in approximate Bayesian computing has been to obtain approximate Bayes factors, which is the more logical choice within the Bayesian context. Both Pritchard et. al. (1999) and Blum (2010) followed this approach. However, recent literature has cast a shadow over the quality of approximating Bayes factors using ABC methods (Robert et. al., 2011; Sisson and Fan, 2010). Robert et. al. (2011) in particular found that even with sufficient statistics for two models under consideration, the Bayes factor obtained from ABC cannot always be trusted. In the more realistic setting with in-sufficient statistics and thus wider loss of information, there is even less reason to trust an ABC Bayes factor. We are left without a clear ABC model selection procedure at this time, and in the applications above, we used the standard CLIC as a first step to determine which model might be best, and then proceeded with ABC.

A natural question to ask at this point is: do the results in this dissertation suggest any improvements which could be made to the composite likelihood approach? While we have treated the MCLE approach as a fixed benchmark, there are several avenues for improve-

ment. It may be, for example, that the ABC approach outperformed MCLE for short-range processes in part because it is better at down-weighting the influence of distant groupings of points than the composite likelihood approach. If the distance between locations is much greater than the range of correlation $\rho(h;\phi)$, then there is very little information about the parameter ϕ to be found in such data. For a short-range process, most pairs (or triplets) which can be formed by D locations would fall into this category. The clustering step of the ABC algorithm may help minimize the influence of such non-informative groupings. If this is true, then an obvious improvement to the composite likelihood approach would be to transition to a weighted composite likelihood, or perhaps a subset of the full pairwise likelihood in which not all pairs are considered (equivalent to a weighted composite likelihood with 0-1 weights). How such weights or subsets could be chosen is an interesting but unanswered question. There is a danger of circular reasoning (the range of dependence would be used to determine weights or subsets, but the dependence is the target of the inferential method), but perhaps empirical or non-parametric measures of dependence could first be used to guide the formation of a composite likelihood.

However, the same can then be said of the ABC approach. There are certainly other ways of clustering of locations, or assigning weights to clusters. Potential improvements to the MCLE approach suggest similar improvements to the ABC approach, and vice versa. It is our hope that both methods will advance, and new ones will be identified. In the meantime, the ABC approach is the only inferential method which can handle higher order k-tuples of locations for $k \geqslant 3$. As for computational cost, over time this tends to fall partly as a result of faster computers but also partly as a result of improved algorithms. Since the full dataset Z is by definition a sufficient statistic for ϕ, if computational cost weren't an issue then one could always recover the exact posterior $\pi(\phi \mid d(Z, Z') \leqslant \epsilon)$ for an arbitrarily small $\epsilon \to 0$. As computational cost falls (and it surely will), the performance of the ABC approach will correspondingly rise.

Bibliography

Arthur, D., Vassilvitskii, S. (2006). How Slow is the k-means method? *Proceedings of the 22nd Annual Symposium of Computational Geometry*, pp. 144-153.

Beaumont, M.A., Zhang, W., Balding, D.J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162:2002, 2025-2035.

Beaumont M., Cornuet J., Marin J., Robert C. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96, 4, 983-990.

Blanchet, J. and Davison, A. (2011). Spatial Modeling of Extreme Snow Depth. *Annals of Applied Statistics*, Vol 5, No 3, 1699-1725.

Blum, M. (2010). Approximate Bayesian Computation: a non-parametric perspective. *Journal of the American Statistical Association*, 105: 1178-1187.

Bortot, P., Coles, S. G., and Sisson, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association*, 102:84-92.

Brown, B., Resnick, S. (1977). Extreme Values of Independent Stochastic Processes. *Journal of Applied Probability*, 14: 732-739.

Coles, S., Dixon, M. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2:5-23.

Coles, S., Heffernan, J., Tawn, J. (1999). Dependence Measures for Extreme Value Analyses. *Extremes*, Volume 2, Number 4, 339-365.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. London, Springer.

Cooley D., Naveau P., Poncet P. (2006). Variograms for spatial max-stable random fields. *Dependence in Probability and Statistics*, edited by Bertail P., Doukhan P., Soulier P.; Springer Lecture Notes in Statistics 187.

Cox, D., Reid, N. (2004). a note on pseudo-likelihood constructed from marginal densities. *Biometrika*, 91:729-737.

Cressie, N. (1993). *Statistics for Spatial Data*. New York, Wiley Interscience.

Davison, A., Gholamrezaee, M. Geostatistics of Extremes, Preprint, 2010. URL `http://stat.epfl.ch`

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

de Haan, L. and Ferreira, A. *Extreme Value Theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

Denuit, M., Dhaene, J., Goovaerts, M., Kaas, R. (2005). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley.

Dhaene, J., Vanduffel, S., Goovaerts, M., Kaas, R., Tang, Q., Vyncke, D. (2006). Risk Measures and Comonotonicity: A Review. *Stochastic Models*. 22:573-606.

Dombry, C., Éyi-Minko, F., and Ribatet, M. Conditional Simulations of Brown-Resnick Processes. http://arxiv.org/pdf/1112.3891.pdf.

Embrechts, P., Resnick, S., Samorodnitsky, G. (1999). Extreme Value Theory as a Risk Management Tool. *North American Actuarial Journal* 26, pp. 30-41.

Erhardt, R., Smith, R. (2012). Approximate Bayesian Computing for Spatial Extremes. *Computational Statistics and Data Analysis* 56:1468-1481.

Feldblum, S. (1990). Risk Loads for insurers. *Proceedings of the Casualty Actuarial Society* LXXVII, 160-195.

Fisher, R., Tippett, L. (1928). On the estimation of the frequency distributions of the largest or smallest in a sample. *Proceedings of the Cambridge Philosophical Society* 24, 180-190.

Fu, Y., Li, W. (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol*, 14:195-199.

Genton, M., Ma, Y., Sang, H. (2011). On the likelihood function of Gaussian max-stable processes. *Biometrika* 98: 481-488.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44, 423-453.

Gogol, D. (1992). Discussion of Kreps: Reinsurer Risk Loads from Marginal Surplus Requirements. *Proceedings of the Casualty Actuarial Society* LXXIX, 362-366.

Goodwin, B., Smith, V. (1995). *The Economics of Crop Insurance and Disaster Aid* The AEI Press, 1995.

Hastings W.K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.

Hyndman, R. and Fan, Y. (1996) Sample quantiles in statistical packages. *American Statistician*, 50:361-365

Jewson, S. and Brix, A. (2005). *Weather Derivative Valuation*. Cambridge University Press, 2005.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology,* 7(1):no. 23.

Kabluchko Z., Schlather M., de Haan L. (2009). Stationary Max-stable fields associated to negative definite functions. *The Annals of Probability*, 37:5 2042-2065.

Kreps, R. (1990). Reinsurer Risk Loads from Marginal Surplus Requirements. *Proceedings of the Casualty Actuarial Society* LXXVII, 196-203.

Kunreuther, H., Michel-Kerjan, E. (2009). *At War with the Weather: Managing Large-Scale Risks in a New Era of Catastrophes* The MIT Press, 2009.

Leadbetter, M., Lindgren, G., Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*. New York, Springer Verlag.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221-239.

Mango, D. (1998). An Application of Game Theory: Property Catastrophe Risk Load *Proceedings of the Casualty Actuarial Society* LXXXV, pp. 157-186.

Mardia, K., Kent, J., Bibby, J. (1980) *Multivariate Analysis*, Academic Press.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100:15324 - 15328.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21, 1087-1092.

Mikosch, T. (2006). Copulas: Tales and Facts. *Extremes*, 9:3-20.

Oesting, M., Kabluchko, Z., Schlather, M. (2012). Simulation of Brown-Resnick Processes. *Extremes*, 15:89-107.

Ostrovsky, R., Rabani, Y., Schulman, L., Swamy, C. (2006). The Effectiveness of Lloyd-Type Methods for k-means Problem", *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 165-174.

Padoan S., Ribatet M., Sisson S. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263-277

Peterson, T. and Vose, R. (1997). An overview of the Global Historical Climatology Network temperature data base. *Bulletin of the American Meteorological Society* 78 (12): 2837-2849.

Pickands, J. (1981). Multivariate extreme value distributions. *Proceedings of the 43rd Session of the International Statistics Institute*, 859-878.

Philbrick, S. (1991). Discussion of Feldblum: Risk Loads for Insurers. *Proceedings of the Casualty Actuarial Society* LXXVIII, 56-63.

Pritchard J., Seielstad M., Perez-Lezaun A., Feldman M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16:1791-1798.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Resnick, S. *Extreme values, regular variation, and point processes*, Volume 4 of *Applied Probability. A series of the Applied Probability Trust.* Springer-Verlag, New York, 1987.

Resnick, S. *Heavy-tail phenomena.* Springer Series in Operations Research and Financial Engineering. Springer, New York, 2007. Probabilistic and statistical modeling.

Ribatet, M., *Spatial Extremes: Modelling Spatial Extremes*, 2011. URL `http://CRAN.R-project.org/package=SpatialExtremes`. R package version 1.8-1.

Ribatet, M., Cooley, D., Davison, A. (2011). Bayesian Inference from Composite Likelihoods, with an Application to Spatial Extremes. *Statistica Sinica* (accepted 2011).

Richards, T., Manfredo, M., Sanders, D. (2004). Pricing Weather Derivatives. *American Journal of Agricultural Economics*, 86(4), 1005-1017.

Robert C. (2007). The Bayesian Choice: From Decision Theoretic Foundations to Computational implementation, second ed. Springer.

Robert, C., Cornuet, J.-M., Marin, J.-M. and Pillai, N. (2011). Lack of confidence in approximate Bayesian computational (ABC) model choice. PNAS 108(37), 15112-15117

Schlather M. (2002). Models for stationary max-stable random fields. *Extremes*, 5.1, 33-44.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 104:1760-1765. Errata (2009), 106:168-89.

Sisson S. A., and Fan, Y. (2010). *Handbook of Markov Chain Monte Carlo*, Eds. S. P. Brooks, A. Gelman, G. Jones and X.-L. Meng. Chapman and Hall/CRC Press.

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

Tavaré, S., Balding, D., Griffiths, R., Donnelly, P. (1997). Inferring Coalescence Times from DNA Sequence Data. *Genetics*, 145:505-518.

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

Varin, C. (2008). On Composite Marginal Likelihoods. *Advances in Statistical Science*, 92:1-28.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236.

Weiss G., A. von Haeseler (1998). Inference of population history using a likelihood approach. *Genetics* 149: 1539-1546.