UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES
Mathematical Sciences

Bayesian Optimal Designs for the Gaussian Process Model

by

Maria Adamou

Thesis submitted for the degree of Doctor of Philosophy
September 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Mathematical Sciences

Doctor of Philosophy

Bayesian Optimal Designs for the Gaussian Process Model

by Maria Adamou

This thesis is concerned with methodology for finding Bayesian optimal designs for the Gaussian process model when the aim is precise prediction at unobserved points. The fundamental problem addressed is that the design selection criterion obtained from the Bayesian decision theoretic approach is often, in practice, computationally infeasible to apply.

We propose an approximation to the objective function in the criterion and develop this approximation for spatial and spatio-temporal studies, and for computer experiments. We provide empirical evidence and theoretical insights to support the approximation.

For spatial studies, we use the approximation to find optimal designs for the general sensor placement problem, and also to find the best sensors to remove from an existing monitoring network. We assess the performance of the criterion using a prospective study and also from a retrospective study based on an air pollution dataset. We investigate the robustness of designs to misspecification of the mean function and correlation function in the model through a factorial sensitivity study that compares the performance of optimal designs for the sensor placement problem under different assumptions.

In computer experiments, using a Gaussian process model as a surrogate for the output from a computer model, we find optimal designs for prediction using the proposed approximation. A comparison is made of optimal designs obtained from commonly used model-free methods such as the maximin criterion and Latin hypercube sampling via both the space-filling and prediction properties of the designs.

For spatio-temporal studies, we extend our proposed approximation to include both space and time dependency and investigate the approximation for a particular choice of separable spatio-temporal correlation function. Two cases are considered: (i) the temporal design is fixed and an optimal spatial design is found; (ii) both optimal temporal and spatial designs are found.

For all three of the application areas, we found that the choice of optimal design depends on the degree and the range of the correlation in the Gaussian process model.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Maria Adamou, declare that the thesis entitled "Bayesian Optimal Designs for the Gaussian Process Model" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date:

Acknowledgements

First and foremost I would like to thank my supervisors, Professor David Woods, Professor Susan Lewis and Professor Sujit Sahu, for all their assistance, feedback and advice. This PhD would not have been achievable without their invaluable support and guidance.

I gratefully acknowledge the funding sources that made my PhD work possible, a studentship from the Engineering and Physical Sciences Research Council and a Mathematics Research Fellowship from the Mathematical Sciences Academic Unit.

Last but not least, I owe my loving thanks to my family. My parents, Adamos and Efsevia, for always believing in me, supporting and encouraging me. I never would have made it here without you. My beloved grandparents, Antonis and Zabelou, and my siblings, Kyriakos and Izambela, for their constant love and encouragement throughout all these years. And most of all my partner Thodoris who has been by my side throughout this PhD, living every single minute of it, for his unwavering love and support.

List of Notation

Symbol	Description
Θ	Parameter space, $\tilde{\boldsymbol{\theta}} \in \Theta$
$\tilde{\boldsymbol{\theta}}$	Vector of all unknown parameters
$\mathbf{y}$	Data vector
$\mathcal{Y}$	Data space
$f(\mathbf{y}\|\tilde{\boldsymbol{\theta}})$	Likelihood function given $\tilde{\boldsymbol{\theta}}$
$\pi(\tilde{\boldsymbol{\theta}}\|\mathbf{y})$	Posterior density of $\tilde{\boldsymbol{\theta}}$
$\pi(\tilde{\boldsymbol{\theta}})$	Prior density of $\tilde{\boldsymbol{\theta}}$
n	Number of design points
d	Number of controllable variables
$\mathcal{X}_{\mathcal{P}}$	Continuous region of interest for prediction
$\mathcal{X}$	Design space or study region
$\mathbf{x}_i$	$d \times 1$ design point
ξ	Design, a collection of n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$
ξ^*	Optimal design under a statistical criterion
Ξ	Set of all possible designs
$\Psi(\xi)$	Objective function for design selection
$\text{Eff}(\xi)$	Relative efficiency of design ξ
$Z(\mathbf{x})$	Gaussian process
$\mathbf{h}$	Separation vector, $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$
$K(\mathbf{h})$	Covariance function for two design points separated by $\mathbf{h}$
$\rho(\mathbf{x}_i, \mathbf{x}_j)$	Correlation function between points $\mathbf{x}_i$ and $\mathbf{x}_j$

d_{ij}	Distance metric
$\mathbf{C}(\boldsymbol{\theta})$	Correlation matrix, with ijth entry $\rho(\mathbf{x}_i, \mathbf{x}_j)$
$\boldsymbol{\theta}$	Vector of correlation parameters
ϕ	Decay parameter in the correlation function $\rho(\cdot)$
ν	Smoothness parameter in the correlation function $\rho(\cdot)$
σ^2	Variance of the Gaussian process
τ^2	Nugget
δ^2	Noise-to-signal ratio, $\delta^2 = \tau^2/\sigma^2$
$\boldsymbol{\Sigma}$	The covariance matrix for $\mathbf{y}$, $\boldsymbol{\Sigma} = \mathbf{C}(\boldsymbol{\theta}) + \delta^2\mathbf{I}$
$\mathbf{f}(\mathbf{x}_i)$	$k \times 1$ vector of known fixed regression functions
$\mathbf{F}$	$n \times k$ model matrix
k	Number of regression functions
$\boldsymbol{\beta}$	$k \times 1$ vector of trend parameters
$\epsilon(\mathbf{x}_i)$	Measurement error at $\mathbf{x}_i$
$\mathbf{Y}$	$n \times 1$ vector of data
$\mathbf{x}_{n+1}$ or $\mathbf{x}_p$	$d \times 1$ prediction point
y_{n+1} or $y(\mathbf{x}_p)$	Future observation at the prediction $\mathbf{x}_p$
$\mathbf{f}_{n+1}$ or $\mathbf{f}_p$	$k \times 1$ vector of known fixed regression functions for $\mathbf{x}_p$
$\boldsymbol{\omega}$	$n \times 1$ vector of covariances between the response at each of the existing n inputs and the response at the new point
$\boldsymbol{\beta}_0$	Known prior mean for trend parameters $\boldsymbol{\beta}$
$\mathbf{R}$	Known $k \times k$ covariance matrix for trend parameters $\boldsymbol{\beta}$
a	Known shape parameter for the prior distribution of variance σ^2
b	Known scale parameter for the prior distribution of variance σ^2
$\boldsymbol{\beta}^*$	Mean of the conditional posterior distribution for $\boldsymbol{\beta}$. $\boldsymbol{\beta}^* = (\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} + \mathbf{R}\boldsymbol{\beta}_0)$
$\mathbf{V}^*$	Covariance matrix of the conditional posterior distribution for $\boldsymbol{\beta}$. $\mathbf{V}^* = (\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}$
a^*	Shape parameter for the conditional posterior distribution for $\sigma^2\|\phi, \delta^2$. $a^* = a + n/2$

b^*	Scale parameter for the conditional posterior distribution for $\sigma^2\|\phi,\delta^2$. $b^* = b + \frac{1}{2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-\mathbf{1}}\mathbf{F}^\top]^{-\mathbf{1}}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_\mathbf{0})\right]$
μ^*	Mean parameter for the conditional posterior predictive distribution for $y_{n+1}\|\mathbf{y},\phi,\delta^2$. $\mu^* = (\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0+[\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}+(\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}]\mathbf{y}$
Σ^*	Variance parameter for the conditional posterior predictive distribution for $y_{n+1}\|\mathbf{y},\phi,\delta^2$. $\Sigma^* = (1+\delta^2) - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} + (\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}(\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})^\top$
α	State of nature
$\mathcal{A}$	Set of all states of nature
$\gamma(\mathbf{y})$	Decision as a function of the data, chosen according to the aim of the experiment
$\hat{\gamma}(\mathbf{y})$	Optimal decision
$\mathcal{G}$	Set of all possible decisions
$L(\gamma(\mathbf{y}),\boldsymbol{\alpha};\xi)$	Loss function defined on the space $\mathcal{G}\times\mathcal{A}$
a_i	ith abscissa for general quadrature method
w_i	ith weight assigned to a_i for general quadrature method
m	Number of quadrature points
$L^I(\phi,\delta^2)$	The integrated likelihood

Chapter 1

Introduction

Data collected from correlated processes arise in many diverse application areas, including studies in environmental and ecological science where the response or characteristics of interest may vary across space and/or time. The data are often used to build models for predicting the process at unobserved points in some continuous region of interest, $\mathcal{X}_{\mathcal{P}} \subseteq \mathbb{R}^d$ for $d \geq 1$ and integer. Popular models are derived from the Gaussian process, under the belief that observations made at points close in space or time tend to have similar values.

Gaussian process models are also widely used in analysing data from computer experiments as they provide a very flexible class of models for approximating complex surfaces.

In this thesis, we address the problem of how to choose a set of $n > 1$ design points $\mathbf{x}_i = (x_{1i}, \ldots, x_{di})^\top$, $i = 1, \ldots, n$, constituting a design $\xi = \{\mathbf{x}_1 \ldots, \mathbf{x}_n\}$, from a design space or study region $\mathcal{X} \subseteq \mathcal{X}_{\mathcal{P}}$, to obtain precise prediction from a Gaussian process model. This problem is of great importance in many applications. The structure and strength of the correlations in the Gaussian process model affect various properties, such as the prediction variance or the inter-point distance of an optimal design, and hence we have to take these correlations into account in finding the "best" designs.

In this thesis, we take a Bayesian approach to finding optimal and efficient designs using selection criteria formulated from the objectives of the experiment. In addition to the applications described below, the methods in this thesis apply to nonparametric regression generally and to similar machine-learning problems.

1.1 Motivating Examples

1.1.1 Environmental application

In this section, an example of spatio-temporal data is presented which is used later in the thesis to demonstrate new methods for finding an optimal design for the collection

Figure 1.1: Network of 122 monitoring stations in the eastern USA.

of spatial data. Modelling such data has its origins in mining applications to predict ore grade and in the use of a type of Gaussian process modelling known as kriging (Krige, 1951). Matheron (1963) developed essential elements of spatial statistics such as the concepts of stationarity, isotopy and variograms, and nowadays we use these aspects of the "Matheron school" as explanatory tools in the statistical data analysis.

In the early days of design of experiments, researchers tried to find ways to take account of spatial correlation in the collection and analysis of data which led to principles of experimental design, such as randomisation and blocking, being applied in agricultural experiments; see, for example, Yates (1970) for a review.

A further spatial problem of wide application, including in environmental science, is the sensor placement problem, i.e. how to make an optimal choice of sensor locations within the geographical region of interest in order to obtain, from the available resources, the most precise predictions of the response at unobserved locations. See, for example, Diggle and Lophaven (2006); Zimmerman (2006); Uciński and Maciej (2010).

There are many examples affecting our everyday life that indicate the importance of spatial statistics. Concerns about climate change have led to the measurement of, for example, sea levels. Air pollution levels are regularly monitored because chemicals such as sulphur dioxide and nitrogen dioxide may cause disease or damage the environment.

One particular example concerns the use of monitoring networks for collecting data to measure the level of pollutants in water or air. The map in Figure 1.1 shows a network of 122 monitoring stations in the eastern USA which measure the amount of a chemical

deposited at each site. This chemical deposition is responsible for damage to lakes, forests and streams. Observations from two stations that are geographically close tend to be very similar. The 122 monitoring stations collected weekly data for 52 weeks in 2001.

Usually the cost of maintaining such a large monitoring network is high. In these circumstances, a common problem is how to reduce the number of stations with minimum loss of precision in the predictions of the response at unobserved locations.

1.1.2 Computer experiments

Computer experiments are increasingly used in many fields. Engineers and scientists routinely use deterministic computer models to study actual or theoretical physical and social systems. There are many examples of scientific and technological developments that use computer models, or simulators, to greatly reduce costly physical experimentation, or where physical experiments are infeasible:

- In the design of an aircraft wing, computational fluid dynamics models are used to calculate the air flow over a wing.
- Finite element models are used for pre-clinical testing of hip replacement implants to understand the scenarios in which an implant fails.
- In drug development, molecular modelling is an important part of investigating, explaining and predicting the properties of potential drug candidates.

Although obtaining data from computer models has several advantages over experiments on real processes, they can be expensive and slow to run. For this reason, Sacks et al. (1989) proposed constructing a surrogate model, specifically a Gaussian process model, which is simpler and much faster to run. This approach is now often used to approximate expensive computer models.

In the computer experiments setting, the design problem is how to choose a set, ξ, of points where the computer model will be run to obtain simulated data that allow precise predictions from the surrogate model.

1.2 Aim and Objectives

The aim of this thesis is to develop methodology for Bayesian design of experiments to enable precise predictions to be obtained from the fitted Gaussian process model.

The work differs from previous research in the area by incorporating uncertainty in all the parameters of the model used to describe the data. The main methodological advance is a new approach to design selection using a proposed approximation to the integrated variance of the posterior predictive distribution. This approximation

makes feasible the computation of Bayesian optimal designs. This approximation is investigated and found to be supported by theoretical and numerical studies.

Specific objectives of the thesis are to:

1. review the area of decision-theoretic design for correlated data, especially related to spatial data and computer experiments
2. develop and validate a decision-theoretic design criterion and associated novel approximations for its efficient implementation
3. develop, from these approximations, an efficient method of conducting sensitivity studies on how the choice and efficiency of an optimal design is affected by varying the values of hyperparameters in the prior distributions
4. apply the new methodology to find optimal and near-optimal designs for spatial processes (i.e. find optimal sets of locations)
5. demonstrate the methodology on the general sensor placement problem and the example in Section 1.1.1
6. apply the methodology to computer experiments for deterministic simulators, and compare the designs obtained to standard designs
7. extend the methods to find designs for spatio-temporal models (i.e. find sets of locations and/or times).

1.3 Thesis Organisation

The remainder of the thesis is organised as follows.

In Chapter 2, we describe Gaussian process models and their key elements. The Gaussian process is introduced and the Bayesian approach to Gaussian process modelling is reviewed. We also provide the posterior predictive distributions for Gaussian process models, and give some derivations and details for those distributions used in the remainder of the thesis.

Chapter 3 introduces optimal design theory and describes the decision theoretic framework used to obtain Bayesian optimal designs. We apply this approach to the Gaussian process model and define the design selection criterion that we use to obtain optimal designs. We provide methodology for finding optimal designs for prediction when the parameters of the model are unknown. We propose and investigate a mathematical approximation to the expected loss using the squared error loss function which reduces dependence on Monte Carlo integration and quadrature methods. This new proposed approximation is key to overcoming the computational challenges of the Bayesian method so that optimal designs can be obtained. We also describe the particular coordinate exchange algorithm that we use to find designs, together with a brief review

of existing algorithms. We apply these methods to the numerical approximation of Bayesian design objective functions in later chapters.

Chapter 4 gives efficient methods of investigating (i) the robustness of the choice of an optimal design to varying the hyperparameter values of the prior distributions, and (ii) the sensitivity of the efficiencies of a given optimal design when the hyperparameter values are changed. The methods are demonstrated for $d = 2$ dimensions.

In Chapter 5, we apply the methodology from Chapters 3 and 4 to find optimal designs for the general sensor placement problem. The majority of the literature on design for spatial data focuses on the frequentist approach or considers the correlation parameters to be known. For our Bayesian method, we apply the closed-form approximation developed in Chapter 3. The accuracy of the approximation is supported by a numerical study.

In Chapter 6, we apply our methodology for spatial design to the problem from Section 1.1.1 of deciding which stations should be dropped from the monitoring network.

Computer experiments are addressed in Chapter 7 where we apply our methodology to higher dimensions ($d = 3$). The closed-form approximation to the objective function is investigated, as in Chapter 5, for $d = 3$. We find Bayesian optimal designs for $d = 2$, $n = 3, 5, 7$ points and compare them with designs in the literature. Further designs for higher dimensions, larger numbers of points and two different correlation structures are investigated.

In Chapter 8, we investigate an extension of our general methods to find Bayesian optimal designs for Gaussian process models that include both space and time dependency. For example, observations made hourly or daily at a sampling location may be correlated over time. We extend our closed-form approximation of the objective function to account for both spatial and temporal correlation. For a particular form of correlation structure, numerical studies are provided which support the approximation. Bayesian optimal designs are found using this approximation and their properties are discussed.

Finally, in Chapter 9, we discuss the research contributions in this thesis, their implications and future research directions.

Chapter 2

Gaussian Process Models

In this chapter, Gaussian process models are described in detail and the main concepts and methods used in this thesis are introduced. We begin by defining a Gaussian process and discussing its properties, particularly those arising from characteristics of the correlation function. After a brief introduction to Bayesian inference, we describe the Bayesian approach to Gaussian process modelling. Based on the literature, we give formulations and derivations of the prior, posterior and predictive distributions which are used in the following chapters.

2.1 Introduction

For data of the form $(\mathbf{x}, y(\mathbf{x}))$, where $y(\mathbf{x})$ denotes the response measured at a specific point $\mathbf{x}$, we assume that there is function $g(\mathbf{x})$ which approximates the mean relationship between the point and the response. If we are able to make assumptions about the form of this function, for example that it is a low-order polynomial, then well-known parametric methods, such as linear regression, can be applied to estimate it. However when the response is highly complex, the explicit form of this function is often unknown. We then seek to infer the function from the given data using nonparametric methods. Basically, the use of nonparametric methods allows the data to speak for themselves.

We start with the assumption that

$$y(\mathbf{x}) = g(\mathbf{x}) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon}$ represents the noise or measurement error and the function $g(\mathbf{x})$ is unspecified. The main objective of nonparametric regression is to estimate the unknown function $g(\mathbf{x})$. Some common approaches are local polynomial regression, spline methods and Gaussian process modelling.

In this thesis, we use Gaussian process modelling because it has the following advantages (Rasmussen and Williams, 2006; Kaufman et al., 2008; Gramacy and Lee, 2008):

(i) it provides a stochastic interpretation of the data without requiring additional assumptions on the errors in the response

(ii) it provides flexibility through the choice of specifications of the correlation function

(iii) conditionally conjugate prior distributions are available to simplify the calculations required for obtaining predictions

(iv) spline methods and local polynomial regression can be derived as special cases via particular specifications of the correlation function.

Gaussian process models have a long history dating back to the 1940s, when they were used for time series by Kolmogorov (1941) and Wiener (1949). They are now well-established in time series, spatial and spatio-temporal statistics, computer experiments and machine learning. In the second of these areas, also known as geostastistics, making predictions from a Gaussian process model is known as kriging, after Krige (1951). Theory and methodology were developed by Matheron (1963); comprehensive reviews of Gaussian process modelling for prediction in geostatistics can be found in Cressie (1993).

O'Hagan (1978) used Gaussian processes to describe the behaviour of an unknown mathematical function. More than a decade later, Sacks et al. (1989) proposed the use of Gaussian process models in the design and analysis of deterministic computer experiments (i.e. $\boldsymbol{\epsilon} = 0$). A more recent area where the Gaussian process model has been applied is machine learning, see Rasmussen and Williams (2006).

2.2 Gaussian Process

A stochastic process indexed by point $\mathbf{x} \in \mathcal{X}_{\mathcal{P}} \subseteq \mathbb{R}^d$ is a set of real random variables $\{Z(\mathbf{x}); \mathbf{x} \in \mathcal{X}_{\mathcal{P}}\}$. Examples of $\mathbf{x}$ and $\mathcal{X}_{\mathcal{P}}$ are:

(i) $\mathbf{x}$ = time and $\mathcal{X}_{\mathcal{P}} = [0, \infty)$

(ii) $\mathbf{x} = (l_1, l_2)^\top$ where l_1, l_2 are longitude and latitude, respectively, and $\mathcal{X}_{\mathcal{P}} = [-180^o, 180^o] \times [-90^o, 90^o]$.

We say that a stochastic process is a Gaussian Process when, for any finite integer $n \geq 1$ and any choice of points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}_{\mathcal{P}}$, the joint distribution of the $n \times 1$ vector $\mathbf{Z} = (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^\top$ has a multivariate normal distribution

$$\mathbf{Z} \sim N(\mathbf{m}^*, \boldsymbol{\Sigma}^*),$$

where $\mathbf{m}^*$ is the $n \times 1$ mean vector and $\mathbf{\Sigma}^*$ is the $n \times n$ covariance matrix with ijth element $\text{Cov}[Z(\mathbf{x}_i), Z(\mathbf{x}_j)]$.

We now summarise two important properties which a Gaussian process may possess.

2.2.1 Stationarity

There are several different forms of stationarity for a Gaussian process.

A Gaussian process is said to be *strictly stationary* if, for any given $n \geq 1$, any set of points $\mathbf{x_1}, \ldots, \mathbf{x}_n \in \mathcal{X}_\mathcal{P}$ and any *separation* vector $\mathbf{h} \in \mathbb{R}^d$, defined such that $\mathbf{x}_i + \mathbf{h} \in \mathcal{X}_\mathcal{P}$ for all $i = 1, \ldots, n$, the joint distributions of $Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)$ and $Z(\mathbf{x}_1 + \mathbf{h}), \ldots, Z(\mathbf{x}_n + \mathbf{h})$ are the same.

A Gaussian process has *weak* or *second-order stationarity* if it has

- constant mean i.e. $\mathbb{E}\left[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})\right] = 0 \quad \forall\, \mathbf{x} \in \mathcal{X}_\mathcal{P}$, and
- $\text{Cov}[Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})] = K(\mathbf{h}), \quad \forall\, \mathbf{x} \in \mathcal{X}_\mathcal{P}$,

where $K(\mathbf{h})$ is called the *covariance function* and depends only on $\mathbf{h}$. This latter property means that the covariance between any two of the random variables depends only on the separation vector $\mathbf{h}$. It follows that a Gaussian process that is second-order stationary is also strictly stationary.

If either of the above types of stationarity do not hold then the process is called *non-stationary.*

2.2.2 Isotropy

A Gaussian process is *isotropic* if it is second order stationary and has covariance function $K(\mathbf{h})$ which depends upon the separation vector $\mathbf{h}$ only through the distance $\|\mathbf{h}\|$ between two points $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$, where $\|\cdot\|$ denotes a distance metric. A process which is not *isotropic* is called *anisotropic.* An anisotropic process may be defined as, $\text{Cov}(Z(\mathbf{x}),\ Z(\mathbf{x} + \mathbf{h}_1)) \neq \text{Cov}(Z(\mathbf{x}),\ Z(\mathbf{x} + \mathbf{h}_2))$ for $(\mathbf{h}_1, \mathbf{h}_2)$ such that $\|\mathbf{h}_1\| = \|\mathbf{h}_2\|$. Hence changes in different variables can influence the response differently.

2.3 Isotropic Correlation Functions

Often, the parametrisation of an isotropic covariance function has the following form for the ijth entry:

$$\mathbf{Cov}_{ij} = \sigma^2 \rho(d_{ij}; \boldsymbol{\theta}) \quad \text{if } i \neq j \quad \text{for} \quad d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|, \quad \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_\mathcal{P} \tag{2.2}$$

where $0 \leq \rho(d_{ij}; \boldsymbol{\theta}) \leq 1$, σ^2 is the constant global variance parameter, and $\boldsymbol{\theta}$ is a vector of correlation parameters.

The covariance matrix corresponding to (2.2) is

$$\mathbf{Cov} = \sigma^2 \mathbf{C}(\boldsymbol{\theta}), \tag{2.3}$$

where $\mathbf{C}(\boldsymbol{\theta})$ is an $n \times n$ matrix of correlations with ijth entry $\rho(d_{ij}; \boldsymbol{\theta})$.

A correlation matrix $\mathbf{C}(\boldsymbol{\theta})$ must also have the properties that

- $\mathbf{C}(\boldsymbol{\theta})$ is positive definite
- $\rho(d_{ii}; \boldsymbol{\theta}) = 1$ for all $\mathbf{x}_i \in \mathcal{X}_{\mathcal{P}}$
- $\mathbf{C}(\boldsymbol{\theta})$ is symmetric i.e. $\rho(d_{ij}; \boldsymbol{\theta}) = \rho(d_{ji}; \boldsymbol{\theta})$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_{\mathcal{P}}$.

Therefore choices of functional form for the correlation function are very restricted.

Many isotropic correlation functions have been proposed which are feasible and fairly simple to apply. In the next subsection, we introduce several such functions that have applications in describing spatial and temporal correlation and, more recently, are used in computer experiments and machine learning.

2.3.1 Examples of families of parametric, isotropic correlation functions

In many applications, it is assumed that responses measured at points close together should have similar values for $Z(\mathbf{x})$. For this reason, several correlation functions have been formulated to include a decay parameter which controls the rate at which the correlation decays with distance. As this parameter increases, the correlation between the observations at each fixed pair of points decreases and vice versa. Hence the decay parameter controls the distance at which two observations become almost independent. This distance is known as the *effective range*. A further parameter is used to control the smoothness of the functions drawn from the Gaussian process. These two parameters control the shape of realisations of the Gaussian process.

(a) The most widely used family of correlation functions is the *Matérn* class and was introduced by Matérn (1960). The function is given by

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \frac{1}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu}\|\mathbf{x}_i - \mathbf{x}_j\|\phi)^{\nu} K_{\nu}(2\sqrt{\nu}\|\mathbf{x}_i - \mathbf{x}_j\|\phi), \tag{2.4}$$

where $\boldsymbol{\theta} = (\phi, \nu)^{\top}, \phi > 0, \nu > 0$ are the decay and smoothness parameters, respectively, and K_{ν} is the modified Bessel function of order ν and $\Gamma(\nu)$ is the gamma function.

Four special cases of (2.4) are given by

1. when $\nu = 0.5$, $\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \exp(-\phi d_{ij})$ and is known as the exponential correlation function

2. when $\nu = 1.5$, $\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = (1 + \sqrt{3}\phi d_{ij}) \exp(-\sqrt{3}\phi d_{ij})$

3. when $\nu = 2.5$, $\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = (1 + \sqrt{5}\phi d_{ij} + \frac{5}{3}\phi^2 d_{ij}^2) \exp(-\sqrt{5}\phi d_{ij})$

4. when $\nu \to \infty$, $\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \exp(-\phi^2 d_{ij}^2)$, the Gaussian correlation function.

Figure 2.1 shows the correlation functions for cases 1-3 with Euclidean distance, $d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$ and $\phi = 1$. We see that the exponential correlation function decreases most rapidly. This means that, for example, for points at distance 2 units apart, the corresponding random variables will have less correlation and hence realisations are likely to be less similar than when the correlation is described by the other two functions.

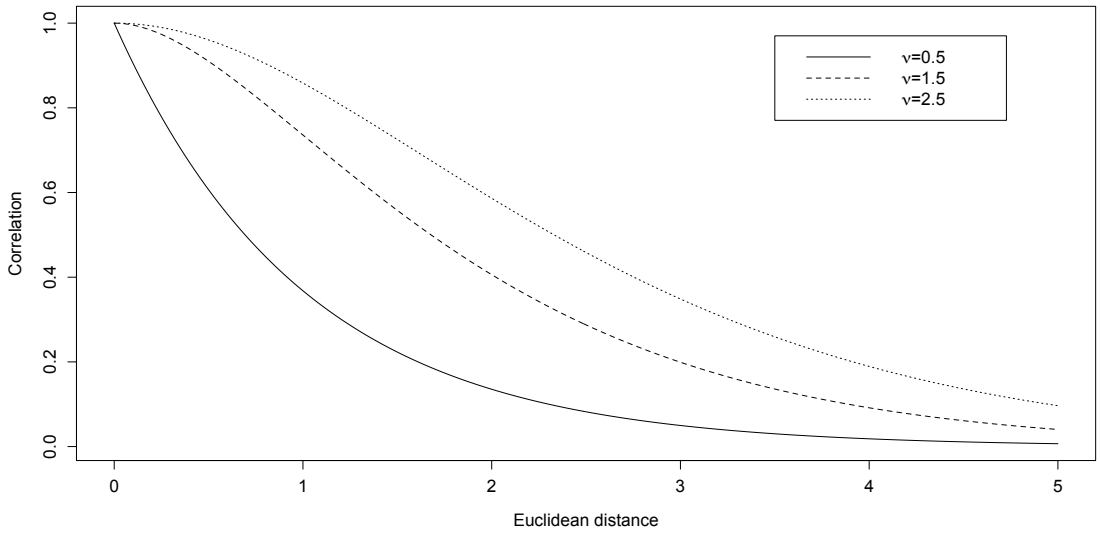


Figure 2.1: The Matérn correlation function with decay parameter $\phi = 1$ for $\nu = 0.5$, 1.5 and 2.5.

(b) Another important family of correlation functions is the *powered exponential*, see, for example, Diggle et al. (1998), which has the form

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \exp\left(-(\phi\|\mathbf{x}_i - \mathbf{x}_j\|)^{\nu}\right), \tag{2.5}$$

where $\boldsymbol{\theta} = (\phi, \nu)^{\top}, \phi > 0, 0 < \nu < 2$ are the decay and smoothness parameters, respectively. For $0 < \nu < 2$, the process is continuous at the origin but not differentiable, except for $\nu = 2$ when the Gaussian correlation function is obtained. When $\nu = 1$, the exponential correlation function again results. The powered exponential family is less flexible than the *Matérn* due to the differentiability properties of the process $Z(\mathbf{x})$. The Matérn correlation function results in a process that is $\lceil \nu \rceil - 1$ times differentiable; in contrast, the process with the *powered exponential* is either nowhere differentiable for

$0 < \nu < 2$, or infinitely differentiable for $\nu = 2$. This advantage of the Matérn explains why this family is more widely used than the powered exponential (Diggle and Ribeiro, 2007).

In Chapter 7 we are going to investigate the separate non-isotropic Matérn and power exponential correlation functions which are extensions of these isotropic functions.

In the next section we present the essential elements of the statistical model which uses a Gaussian process to model observations.

2.4 Gaussian Process Model and Prediction

2.4.1 Statistical model

We adopt the following general model and notation. There are n data points of the form $(\mathbf{x}_i, y(\mathbf{x}_i)), i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes the ith design point within the study region $\mathcal{X}$ and $y(\mathbf{x}_i)$ denotes an observation taken at $\mathbf{x}_i$ on a single realisation of the Gaussian process. In general, we assume $y(\mathbf{x}_i)$ are observed with noise and hence describe them by the following statistical model, referred to as the Gaussian process model:

$$y(\mathbf{x}_i) = \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta} + Z(\mathbf{x}_i) + \epsilon(\mathbf{x}_i), \qquad i = 1, \ldots, n, \tag{2.6}$$

where

- $\mathbf{f}(\mathbf{x}_i) = (f_1(\mathbf{x}_i), \ldots, f_k(\mathbf{x}_i))^\top$ is a $k \times 1$ vector of known fixed regression functions.
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{k-1})^T$ is a $k \times 1$ vector which contains unknown model parameters, often called trend parameters or regression coefficients.
- $Z(\mathbf{x}_i)$ is a Gaussian process with mean zero and covariance matrix which models the dependency between $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$ through specification of the covariance

  $$\mathrm{Cov}(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \sigma^2 \rho(d_{ij}; \boldsymbol{\theta})$$

  for some known correlation function $\rho(d_{ij}; \boldsymbol{\theta})$ from a specified parametric family, where d_{ij} is a measure of the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$, $\boldsymbol{\theta}$ is the vector of correlation parameters and σ^2 is the constant variance.
- $\epsilon(\mathbf{x}_i)$ represents the measurement error or noise associated with repeat observation at $\mathbf{x}_i$. We assume that $\epsilon(\mathbf{x}_i)$ and $\epsilon(\mathbf{x}_j)$ $(i, j = 1, \ldots, n, i \neq j)$ are independent and identically normally distributed with zero mean and variance τ^2 (the nugget). Also $\epsilon(\mathbf{x}_i)$, $Z(\mathbf{x}_j)$ are assumed independent.

That is, we have

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2 \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I}),$$

where $\mathbf{Y}$ is the $n \times 1$ vector of data, $\mathbf{F}$ is the $n \times k$ model matrix and the variance covariance matrix $\sigma^2\mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I}$ has ijth entry $\sigma^2\rho(d_{ij};\boldsymbol{\theta}) + \tau^2\mathbf{1}\{\mathbf{x}_i = \mathbf{x}_j\}$, with $\mathbf{1}$ an indicator function taking value 1 if $\mathbf{x}_i = \mathbf{x}_j$, and 0 otherwise.

In geostatistics, the Gaussian process Z in (2.6) is used to model the spatial correlation between random variables at two points or locations $\mathbf{x}_i \in \mathbb{R}^d$ with $d = 2$. The Gaussian process models the response that might be measured at any point in some geographical region, $\mathcal{X}_\mathcal{P}$, of interest.

In spatio-temporal applications, the Gaussian process Z is used to model the spatio-temporal correlation between two points that are (location, time) vectors $\mathbf{x}_i \in \mathbb{R}^2 \times [0, \infty)$. The model (2.6) is used for a phenomenon evolving through both space and time.

The Gaussian process model (2.6) is also very widely used to describe output from a computer experiment since the form of the function that maps a point $\mathbf{x} \in \mathcal{X}$ into an output $y(\mathbf{x})$ is unknown. The Gaussian process is used to represent the available prior information about the unknown function.

2.4.2 Predictions

In order to make inferences about an observation at a new point $\mathbf{x}_{n+1}$ using model (2.6), we need to define the predictive distribution for the random variable $y_{n+1} = y(\mathbf{x}_{n+1})$.

In general, following Banerjee et al. (2004, Ch. 2), for two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ described by (2.6), the joint distribution conditional on all the unknown parameter $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}$ and τ^2 is given in matrix form by

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}}\\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}}\end{pmatrix}\right),$$

where $\boldsymbol{\Sigma_{21}} = \boldsymbol{\Sigma_{12}}^\top$. The conditional distribution of $\mathbf{y}_1|\mathbf{y}_2$ is normal with mean and variance:

$$\begin{aligned}\mathbb{E}(\mathbf{y}_1|\mathbf{y}_2) &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\mathbf{y}_2 - \boldsymbol{\mu}_2),\\ \mathrm{Var}(\mathbf{y}_1|\mathbf{y}_2) &= \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}}.\end{aligned}$$

In our context, we let $\mathbf{y}_1 = y_{n+1}$ and $\mathbf{y}_2 = \mathbf{y}$ then it follows that

$$\boldsymbol{\mu}_1 = \mathbf{f}_{n+1}^\top\boldsymbol{\beta} \qquad \text{and} \qquad \boldsymbol{\mu}_2 = \mathbf{F}\boldsymbol{\beta},$$

where $\mathbf{f}_{n+1} = \mathbf{f}(\mathbf{x}_{n+1})$ is the $k \times 1$ vector of regression functions for $\mathbf{x}_{n+1}$, $\mathbf{F}$ is the $n \times k$

matrix of regression functions with ijth element $f_j(\mathbf{x}_i)$ and

$$\boldsymbol{\Sigma_{11}} = \sigma^2 + \tau^2, \qquad \boldsymbol{\Sigma_{12}} = \sigma^2\boldsymbol{\omega}^\top = \tilde{\boldsymbol{\omega}}^\top, \quad \text{and} \qquad \boldsymbol{\Sigma_{22}} = \sigma^2\mathbf{C} + \tau^2\mathbf{I}.$$

Here $\boldsymbol{\omega} = [\rho(d_{n+1,1};\boldsymbol{\theta}), \ldots, \rho(d_{n+1,n};\boldsymbol{\theta})]^\top$ is the $n \times 1$ vector of correlations between the response at each of the existing inputs and the response at the new point where we want to predict. Substituting these values into the above mean and variance formulae, we obtain that $y_{n+1}|\mathbf{y}$, conditional on $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2$, is normal with:

$$\mathbb{E}(y_{n+1}|\mathbf{y}) = \mathbf{f}_{n+1}^\top\boldsymbol{\beta} + \tilde{\boldsymbol{\omega}}^\top[\sigma^2\mathbf{C} + \tau^2\mathbf{I}]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}), \tag{2.7}$$

$$\mathrm{Var}(y_{n+1}|\mathbf{y}) = \sigma^2 + \tau^2 - \tilde{\boldsymbol{\omega}}^\top[\sigma^2\mathbf{C} + \tau^2\mathbf{I}]^{-1}\tilde{\boldsymbol{\omega}}. \tag{2.8}$$

If we make the assumption that all the parameters are known, then the conditional expectation (2.7) and variance (2.8) are the simple kriging predictor and simple kriging variance respectively. The simple kriging predictor minimises the mean square error (MSE) in geostatistical methods.

Predictions based on the Gaussian process model (2.6) are popular for many reasons: the predictor is semi-parametric, i.e. the large scale variation is modelled by the mean function and is specified by a regression, and the small scale deviations from the mean are described by a stationary Gaussian process. Also, when the parameters are known then the conditional expectation (2.7) and variance (2.8) used to predict at a new point are simple to obtain.

2.4.3 Bayesian Gaussian process model

The Gaussian process model may be used with both frequentist and Bayesian approaches to designing a study. In this thesis we follow a fully Bayesian approach. We prefer to analyse the experimental data using this method since it seems a natural way to describe our prior information and beliefs and to update these using information from the data. Further, a Bayesian approach allows us to take uncertainty into account in a more coherent way than the frequentist approach.

A key difference between the frequentist and Bayesian modelling approaches is that the latter incorporates any prior knowledge of the parameters by treating the parameters as random variables and assigning a prior distribution to them. By doing this, we take into account previous knowledge and uncertainty about the parameters. The full specification is typically called the Bayesian Gaussian process model. It is a hierarchical model as it is based on the probability theory that the joint distribution of random variables can be decomposed into a series of conditional distributions and a marginal distribution.

The literature on the Bayesian Gaussian process model was developed initially by Ki-

tanidis (1986) and there is now a substantial literature on this field including Le and Zidek (1992), Handcock and Stein (1993), Banerjee et al. (2004) and the references therein. For a complicated response surface, Bayesian hierarchical modelling provides a flexible framework for both estimation and prediction problems. We follow the three stages of the model, described by Banerjee et al. (2004, Ch. 5) and Wikle (2010):

$$\begin{aligned} &\textbf{Stage 1:} \text{ Data model: } \textit{data} \mid \textit{process}, \textit{parameters} \\ &\textbf{Stage 2:} \text{ Process model: } \textit{process} \mid \textit{parameters} \\ &\textbf{Stage 3:} \text{ Parameter model: } \textit{parameters} \end{aligned} \tag{2.9}$$

The first stage specifies the distribution of the data conditional on the process and the parameters that describe the model. The second stage describes the distribution of the process given the parameters, and the third stage specifies the distribution of all the unknown parameters, denoted by $\tilde{\boldsymbol{\theta}}$, and takes into account the uncertainty in the model due to the unknown parameters.

Before collecting the data, information is often available about $\tilde{\boldsymbol{\theta}}$, typically obtained from subject experts or from previous data sets. This information is used to provide a specification of the prior density $\pi(\tilde{\boldsymbol{\theta}})$ for the values of the model parameters. After the data $\mathbf{y}$ are gathered, they are used to update the prior distribution and calculate the posterior density $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$ using Bayes theorem.

Bayes Theorem. Suppose that there are two random variables $\mathbf{y}$ and $\tilde{\boldsymbol{\theta}}$ with joint probability density functions (pdf) $f(\mathbf{y}|\tilde{\boldsymbol{\theta}})$ and $\pi(\tilde{\boldsymbol{\theta}})$, respectively. Then the posterior density of $\tilde{\boldsymbol{\theta}}$ given $\mathbf{y}$ is:

$$\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y}) = \frac{f(\mathbf{y}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})}{\int_{-\infty}^{\infty} f(\mathbf{y}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}}} \propto f(\mathbf{y}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}}), \tag{2.10}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ are the data, $\tilde{\boldsymbol{\theta}}$ is the vector of unknown parameters, $\pi(\tilde{\boldsymbol{\theta}})$ is the prior density and $f(\mathbf{y}|\tilde{\boldsymbol{\theta}})$ the likelihood of the data given the unknown parameters. This likelihood function describes our belief that $\mathbf{y}$ would be the outcome if we knew $\tilde{\boldsymbol{\theta}}$ to be true. Inference about the parameters proceeds from the posterior distribution. Basically, $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$ describes our beliefs that $\tilde{\boldsymbol{\theta}}$ is the true value, having observed the data $\mathbf{y}$. The denominator $\int_{-\infty}^{\infty} f(\mathbf{y}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}}$ is called the normalising constant; usually we find the posterior distribution up to a normalising constant.

An important objective of Bayesian inference is prediction and this is addressed through the predictive distribution. Suppose that we have a sample of observations $y_1, \ldots, y_n$ and we want to predict y_{n+1}. Then, we need to find the predictive distribution that represents the uncertainty in a future observation given the previous observations. That

is, we require the posterior predictive density given by

$$\pi(y_{n+1}|\mathbf{y}) = \int_{-\infty}^{\infty} f(y_{n+1}|\tilde{\boldsymbol{\theta}}, \mathbf{y})\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})d\tilde{\boldsymbol{\theta}},$$

where $f(y_{n+1}|\tilde{\boldsymbol{\theta}}, \mathbf{y})$ is the conditional distribution of y_{n+1} given $\boldsymbol{\theta}$ and data $\mathbf{y}$. The prediction density is obtained as an average over the posterior density $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$, which contains all the information that we know about the parameter $\tilde{\boldsymbol{\theta}}$. See, for example, Gelman et al. (2003, Ch. 1).

We can express the hierarchical structure (2.9) for the Bayesian Gaussian process model as

$$\begin{aligned}
\text{Data model:} \quad & \mathbf{Y}|\boldsymbol{\beta}, \mathbf{Z}, \tau^2 \sim N(\mathbf{F}\boldsymbol{\beta} + \mathbf{Z}, \tau^2\mathbf{I}) \\
\text{Process model:} \quad & \mathbf{Z}|\boldsymbol{\theta}, \sigma^2 \sim N(0, \sigma^2\mathbf{C}(\boldsymbol{\theta})) \\
\text{Parameter model:} \quad & \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2 \text{ and joint prior distribution } \pi(\cdot)
\end{aligned}$$

An alternative representation of the hierarchical structure is obtained by combining the data and process models. This *marginal formulation of the model* is obtained by integrating out the process model so that the data model depends only on the parameters. We obtain

$$\begin{aligned}
\text{Data model:} \quad & \mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2 \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I}) \\
\text{Parameter model:} \quad & \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2 \text{ and } \pi(\cdot)
\end{aligned}$$

The model specification requires assignment of a prior distribution to the unknown parameters $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2)^\top$. A common approach is to separate this prior distribution into sections and there are two cases:

(a) assume the parameters are independent, then $\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\boldsymbol{\theta})\pi(\tau^2)$

(b) assume the trend parameters and the variance are independent of the correlation parameters and the nugget effect, and $\boldsymbol{\theta}$ and τ^2 are independent, then $\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\theta})\pi(\tau^2)$.

In general, specific choices of prior distributions are often made to facilitate computations. It is common to verify that the posterior analysis is not very sensitive to the choice of the prior distribution.

2.4.4 Prior specification of the parameter model

Bayesian inference depends on the prior distribution that we choose to represent our beliefs. Generally there are two types of priors: informative and non-informative priors. The former expresses specific and definite information about the unknown parameter

while the latter expresses vague and general information. Jeffrey's priors are a common example of non-informative priors. More details about choosing a prior distribution can be found in Berger (1985, Ch. 3). As discussed in Banerjee et al. (2004, Ch. 4), it is better to choose informative priors for the unknown parameters in order to avoid the problem of improper posteriors.

For the trend parameters and the variance $(\boldsymbol{\beta}, \sigma^2)$, we can consider conjugate priors, i.e. the resulting posterior distribution belongs to the same family as the prior distribution, or improper priors. A common choice of conjugate prior for $\boldsymbol{\beta}$ and σ^2, is the multivariate normal and inverse gamma, respectively. For an improper prior, $\pi(\boldsymbol{\beta}, \sigma^2) = 1/\sigma^2$ is often used.

The covariance matrix depends on the unknown vector of parameters $\boldsymbol{\theta}$ of the correlation function. Usually $\boldsymbol{\theta}$ contains the decay and smoothness parameters, i.e. $\boldsymbol{\theta} = (\phi, \nu)^\top$. In this thesis, in line with common practice, we assume that the smoothness parameter ν is known and fixed, and assign a prior distribution to the decay parameter ϕ. To avoid singularity of $\mathbf{C}(\boldsymbol{\theta})$, we specify $\phi > 0$. Hence a uniform prior on $(0, b_1)$, $b_1 \in \mathbb{R}^+$ might be considered appropriate or, alternatively, a prior with flexibility in the shape and scale such as a log-normal or inverse gamma distribution.

When the nugget is included in the model, it is useful to reparameterise the two types of variance σ^2 and τ^2, to facilitate computations. Two model parameterisations for the variances have been proposed by Yan et al. (2007) and Diggle and Ribeiro (2007). The former authors proposed

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2_\kappa, \phi, \kappa \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2_\kappa[(1-\kappa)\mathbf{C}(\phi) + \kappa\mathbf{I}]), \tag{2.11}$$

where $\sigma^2_\kappa = \sigma^2 + \tau^2$ and $\kappa = \tau^2/\sigma^2_\kappa$. It has the advantage that κ has bounded support and this makes easier the use of some types of Monte Carlo Markov Chain (MCMC) sampling techniques.

The parameterisation of the latter authors is

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2 \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2[\mathbf{C}(\phi) + \delta^2\mathbf{I}]), \tag{2.12}$$

where $\delta^2 = \tau^2/\sigma^2$ is the ratio of the nugget to the process variation. We choose to follow the second parameterisation since it has the advantage of being scale-free and is also more commonly used in the literature and in practice.

In subsequent chapters, we parameterise the variance components (σ^2, τ^2) by (σ^2, δ^2) where $\delta^2 = \tau^2/\sigma^2$. From now on, we denote the reparametrised covariance matrix $\sigma^2[\mathbf{C}(\phi) + \delta^2\mathbf{I}]$ by $\sigma^2\boldsymbol{\Sigma}$.

We consider two cases

- δ^2 known and assigned fixed values

- δ^2 unknown and assigned a discrete uniform prior or a continuous uniform prior distribution.

Throughout this thesis, we choose conjugate priors for $(\boldsymbol{\beta}, \sigma^2)$ for algebraic convenience when deriving an approximation for the design criterion in Chapter 3. We investigate the impact of different choices of $\boldsymbol{\beta}, \phi$ and δ^2 on design selection in later chapters.

In order to make predictions from a Gaussian process model, we require the posterior distributions of the parameters and the predictive distribution. These are derived in the following two sections.

2.4.5 Predictive distribution when covariance parameters are known

Throughout this section, we derive Bayesian inference results for the Gaussian model when the covariance parameters ϕ and δ^2 are assumed fixed and known. We allow for uncertainty only in the trend $\boldsymbol{\beta}$ and variance σ^2. In this case, the posterior distribution for $\boldsymbol{\beta}$ and the posterior predictive distribution can be derived analytically taking into account uncertainty in both $\boldsymbol{\beta}$ and σ^2. Full derivations of conditional and marginalised posterior distributions of the parameters can be found in Gelman et al. (2003, Ch. 15).

For fixed ϕ and δ^2, the conjugate joint prior distribution for $\boldsymbol{\beta}$ and σ^2 is the Normal-Inverse Gamma distribution. Hence the prior densities for $\boldsymbol{\beta}$ and σ^2 are:

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2),$$

where $\pi(\boldsymbol{\beta}|\sigma^2) \sim N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R}^{-1})$ and $\pi(\sigma^2) \sim IG(a, b)$, and the inverse gamma distribution, $IG(a, b)$ has density proportion to

$$\pi(\sigma^2|a, b) \propto (\sigma^2)^{-(a+1)} \exp\{-b\sigma^{-2}\},$$

$\boldsymbol{\beta}_0$ is the known prior mean, $\mathbf{R}^{-1}$ is a known symmetric, $k \times k$ matrix and a, b are known hyperparameters. Therefore,

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2) &= \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{(2a+k)}{2}+1} \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{R}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}\right]. \qquad (2.13)\end{aligned}$$

Equation (2.13) defines the Normal-Inverse-Gamma prior density and we write

$$\pi(\boldsymbol{\beta}, \sigma^2) \sim NIG(\boldsymbol{\beta}_0, \mathbf{R}^{-1}, a, b).$$

The likelihood function for (2.6) is given by:

$$f(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\phi,\delta^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})\right\}. \tag{2.14}$$

Using Bayes theorem (2.10), the prior density (2.13) is combined with the likelihood (2.14), and the resulting posterior density is given by:

$$\begin{aligned}\pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\phi,\delta^2) &\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\phi,\delta^2)\pi(\boldsymbol{\beta},\sigma^2)\\ &\propto (\sigma^2)^{-(\frac{k+2a^*}{2}+1)} \times \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \mathbf{V}^{*-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+b^*\right]\right\},\end{aligned} \tag{2.15}$$

where

$$\begin{aligned}\boldsymbol{\beta}^* &= (\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y}+\mathbf{R}\boldsymbol{\beta}_0),\\ \mathbf{V}^* &= (\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1},\\ a^* &= a+n/2,\\ b^* &= b+\frac{1}{2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right].\end{aligned} \tag{2.16}$$

The posterior distribution given by (2.15) can be identified as $NIG(\boldsymbol{\beta}^*,\mathbf{V}^*,a^*,b^*)$. Hence it belongs to the conjugate family for the Gaussian process model.

The marginal posterior distribution for the unknown parameter $\boldsymbol{\beta}$ is given by:

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y},\phi,\delta^2) &= \int \pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\phi,\delta^2)d\sigma^2\\ &\propto \left[1+\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\mathbf{V}^{*-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{2b^*}\right]^{-(\frac{2a^*+k}{2})}.\end{aligned} \tag{2.17}$$

Hence, $\boldsymbol{\beta}$ follows a multivariate t-distribution

$$\boldsymbol{\beta}|\mathbf{y},\phi,\delta^2 \sim t_{2a^*}\left[k,\boldsymbol{\beta}^*,\frac{b^*}{a^*}\mathbf{V}^*\right]. \tag{2.18}$$

The marginal posterior for σ^2 is obtained by integrating the joint posterior distribution (2.15) over the trend parameter as follows:

$$\begin{aligned}\pi(\sigma^2|\mathbf{y},\phi,\delta^2) &= \int \pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\phi,\delta^2)d\boldsymbol{\beta}\\ &\propto (\sigma^2)^{-(\frac{2a+n}{2}+1)} \exp\left\{\frac{1}{\sigma^2}(b+\frac{1}{2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^T[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0))\right\}\\ &\sim IG(a^*,b^*).\end{aligned} \tag{2.19}$$

The next step is to derive the marginal distribution of the data. Since $\mathbf{y}|\boldsymbol{\beta},\sigma^2,\phi,\delta^2 \sim$

$N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma})$ and $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R}^{-1})$, we conclude that the density is proportional to the exponential of a quadratic form. Thus, the distribution of the data marginal to $\boldsymbol{\beta}$ but conditional on σ^2, ϕ and δ^2 is normal with mean and variance obtained from the laws of total expectation and total variance.

The law of total expectation for three random variables X, Y, Z states that

$$\mathbb{E}(Y|Z) = \mathbb{E}_{X|Z}(\mathbb{E}(Y|X, Z)). \tag{2.20}$$

The law of total variance for three random variables X, Y, Z states that

$$\text{var}(Y|Z) = \mathbb{E}_{X|Z}(\text{var}(Y|X, Z)) + \text{var}_{X|Z}(\mathbb{E}(Y|X, Z)). \tag{2.21}$$

Based on (2.20) and (2.21) we have

$$\begin{aligned} \mathbb{E}(\mathbf{y}|\sigma^2, \phi, \delta^2) &= \mathbb{E}[\mathbb{E}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] = \mathbf{F}\boldsymbol{\beta}_0, & (2.22)\\ \text{var}(\mathbf{y}|\sigma^2, \phi, \delta^2) &= \mathbb{E}[\text{var}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] + \text{var}[\mathbb{E}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] \\ &= \sigma^2[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top], & (2.23) \end{aligned}$$

and

$$\mathbf{y}|\sigma^2, \phi, \delta^2 \sim N(\mathbf{F}\boldsymbol{\beta}_0, \sigma^2[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]).$$

Then the density marginal to σ^2, obtained by integrating out σ^2, is:

$$\begin{aligned} \pi(\mathbf{y}|\phi, \delta^2) &= \int \pi(\mathbf{y}|\sigma^2, \phi, \delta^2)\pi(\sigma^2)d\sigma^2 \\ &\propto \int \left(\frac{1}{\sigma^2}\right)^{\frac{n+2a}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0) + b\right]\right\} d\sigma^2 \\ &\propto \left[1 + \frac{(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)}{2b}\right]^{-(\frac{2a+n}{2})}. \end{aligned} \tag{2.24}$$

Equation (2.24) indicates that the marginal distribution of $\mathbf{y}$ is a multivariate t-distribution:

$$\mathbf{y}|\phi, \delta^2 \sim t_{2a}\left[n, \mathbf{F}\boldsymbol{\beta}_0, \frac{b}{a}\left[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top\right]\right]. \tag{2.25}$$

To obtain the predictive distribution, we know that the marginal posterior $\pi(\sigma^2|\mathbf{y}, \phi, \delta^2)$ is given by (2.19) and $\pi(y_{n+1}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi, \delta^2)$ is a normal distribution with mean and variance given by (2.7) and (2.8), respectively. It follows that the mean and variance of $\pi(y_{n+1}|\mathbf{y}, \sigma^2, \phi, \delta^2)$, again a normal distribution, can be found using the total laws of expectation and variance, given in (2.20) and (2.21) respectively. Hence, we have that:

$$y_{n+1}|\mathbf{y}, \sigma^2, \phi, \delta^2 \sim N(\mu^*, (\sigma^*)^2), \tag{2.26}$$

where μ^* and $(\sigma^*)^2$ given by:

$$\begin{aligned}
\mu^* &= \mathbb{E}(y_{n+1}|\mathbf{y}, \sigma^2, \phi, \delta^2) \\
&= \mathbb{E}[\mathbb{E}(y_{n+1}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] \\
&= \mathbb{E}(\mathbf{f}_{n+1}^\top \boldsymbol{\beta} + \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})) \\
&= (\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 \\
&\quad + [\boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1} + (\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}]\mathbf{y} \qquad (2.27) \\
(\sigma^*)^2 &= \mathrm{var}(y_{n+1}|\mathbf{y}, \sigma^2, \phi, \delta^2) \\
&= \mathbb{E}[\mathrm{var}(y_{n+1}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] + \mathrm{var}(\mathbb{E}[y_{n+1}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi, \delta^2)] \\
&= \sigma^2(1 + \delta^2) - \sigma^2 \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} \\
&\quad + \sigma^2(\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}(\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})^\top. \qquad (2.28)
\end{aligned}$$

The interpretation of the three components in the expression for $(\sigma^*)^2$ is (i) the variability without taking account of the information provided by the data, (ii) the reduction in variability due to conditioning on the data, and (iii) the increase in variability due to the uncertainty in the estimate of regression coefficients $\boldsymbol{\beta}$.

There is a relationship between the predictions obtained from the Bayesian and frequentist approaches when a flat prior is assumed for $\boldsymbol{\beta}$, i.e. the prior variance for the trend parameters is large. In particular, for flat prior the $k \times k$ matrix $\mathbf{R}$ does not exist, ($\mathbf{R} = 0$). The frequentist approach can be interpreted as prediction which takes into account the uncertainty in the trend parameters.

- when $\beta_i = 0$ $(i = 1, \ldots, k-1)$ and σ^2, ϕ and δ^2 are fixed, the prediction is known as "ordinary kriging", common in geostatistics, with μ^* and $(\sigma^*)^2$ in (2.28) known as the ordinary kriging mean and the ordinary kriging variance

- when there exist at least one $\beta_i \neq 0$ $(i = 1, \ldots, k-1)$ and σ^2, ϕ and δ^2 are fixed, then the prediction method is known as "universal kriging" with μ^* and $(\sigma^*)^2$ in (2.28) known as the universal kriging mean and the universal kriging variance.

In general, for any other choice of prior distribution, e.g. the conjugate prior, the predictions from frequentist and Bayesian approaches do not coincide.

The posterior predictive density can be found by integrating out the unknown σ^2:

$$\begin{aligned}
\pi(y_{n+1}|\mathbf{y}, \phi, \delta^2) &= \int \pi(y_{n+1}|\mathbf{y}, \sigma^2, \phi, \delta^2)\pi(\sigma^2|\mathbf{y}, \phi, \delta^2)d\sigma^2 \\
&\propto \left[1 + \frac{(y_{n+1} - \mu^*)^2}{2b^*\Sigma^*}\right]^{-\frac{(2a^*+1)}{2}}. \qquad (2.29)
\end{aligned}$$

Equation (2.29) indicates that the posterior predictive distribution for y_{n+1} at a new

point $\mathbf{x}_{n+1}$ is a univariate t-distribution:

$$y_{n+1}|\mathbf{y},\phi,\delta^2 \sim t_{2a^*}\left[1,\mu^*,\frac{b^*\Sigma^*}{a^*}\right], \tag{2.30}$$

where the mean and the variance are μ^* and $b^*\Sigma^*/(a^*-1)$ respectively, with $\Sigma^* = (1+\delta^2) - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega} + (\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}(\mathbf{f}_{n+1}^\top - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})^\top$, and b^*, a^* and μ^* given by (2.16) and (2.27).

2.4.6 Predictive distribution when covariance parameters are unknown

In practice, we will usually not know the values of the decay parameter ϕ and noise-to-signal ratio δ^2. Hence, realistically, we need to allow uncertainty in all of the model parameters. We distinguish between two cases:

- ϕ unknown; δ^2 known and fixed (Case (i))
- both ϕ and δ^2 unknown (Case (ii))

In both cases we cannot derive analytical forms for the posterior distribution for parameters and or the posterior predictive distributions.

We assign a normal distribution for the trend parameter $\boldsymbol{\beta}$, i.e. $N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R}^{-1})$, inverse gamma for the variance σ^2, i.e. $IG(a,b)$. We denote by

$$\begin{aligned} L^I(\phi,\delta^2) &= \int f(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\phi,\delta^2)\pi(\boldsymbol{\beta},\sigma^2)d\boldsymbol{\beta}d\sigma^2 \\ &= \frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{\left[b+\frac{1}{2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}, \end{aligned} \tag{2.31}$$

the integrated likelihood with respect the unknown β and σ^2.

Case (i). We consider a proper prior density for ϕ and the joint prior distribution

$$\pi(\boldsymbol{\beta},\sigma^2,\phi) = \pi(\boldsymbol{\beta},\sigma^2)\pi(\phi),$$

with $\pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\phi,\delta^2)$ given by (2.15). The marginal posterior density for ϕ is

$$\begin{aligned} \pi(\phi|\mathbf{y},\delta^2) &= \frac{\pi(\boldsymbol{\beta},\sigma^2,\phi|\mathbf{y},\delta^2)}{\pi(\boldsymbol{\beta},\sigma^2|\mathbf{y},\phi,\delta^2)} \\ &\propto \frac{f(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\phi,\delta^2)\pi(\boldsymbol{\beta},\sigma^2|\phi,\delta^2)\pi(\phi)}{\pi(\boldsymbol{\beta}|\mathbf{y},\sigma^2,\phi,\delta^2)\pi(\sigma^2|\mathbf{y},\phi,\delta^2)} \\ &\propto \frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{\left[b+\frac{1}{2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}\pi(\phi), \end{aligned}$$

i.e.

$$\pi(\phi|\mathbf{y}, \delta^2) \propto L^I(\phi, \delta^2)\pi(\phi). \tag{2.32}$$

Case (ii). We consider proper densities for each of ϕ and δ^2. The prior distributions for ϕ and δ^2 can be continuous or discrete. We adopt the following joint prior distribution:

$$\pi(\boldsymbol{\beta}, \sigma^2, \phi, \delta^2) = \pi(\boldsymbol{\beta}, \sigma^2)\pi(\phi)\pi(\delta^2).$$

The posterior distribution for the parameters is then given by:

$$\pi(\boldsymbol{\beta}, \sigma^2, \phi, \delta^2|\mathbf{y}) = \pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \phi, \delta^2)\pi(\phi, \delta^2|\mathbf{y}), \tag{2.33}$$

with $\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \phi, \delta^2)$ given by (2.15). It follows that the marginal posterior distribution $\pi(\phi, \delta^2|\mathbf{y})$ can be derived as follows:

$$\begin{aligned}
\pi(\phi, \delta^2|\mathbf{y}) &= \frac{\pi(\boldsymbol{\beta}, \sigma^2, \phi, \delta^2|\mathbf{y})}{\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \phi, \delta^2)} \\
&\propto \frac{f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)\pi(\boldsymbol{\beta}, \sigma^2|\phi, \delta^2)\pi(\phi, \delta^2)}{\pi(\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \phi, \delta^2)\pi(\sigma^2|\mathbf{y}, \phi, \delta^2)} \\
&\propto \frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{\left[b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}\pi(\phi)\pi(\delta^2),
\end{aligned}$$

i.e.

$$\pi(\phi, \delta^2|\mathbf{y}) \propto L^I(\phi, \delta^2)\pi(\phi)\pi(\delta^2). \tag{2.34}$$

The problem for both cases is that the marginal posterior distributions (2.32) and (2.34) are not standard distributions and hence the predictive distribution $\pi(y_{n+1}|\mathbf{y})$ cannot be expressed analytically. The Bayesian prediction is based on the predictive distribution which is given by:

$$\begin{aligned}
\text{Case (i)} \quad \pi(y_{n+1}|\mathbf{y}) &= \int \pi(y_{n+1}|\mathbf{y}, \phi, \delta^2)\pi(\phi|\mathbf{y}, \delta^2)d\phi \\
\text{Case (ii)} \quad \pi(y_{n+1}|\mathbf{y}) &= \iint \pi(y_{n+1}|\mathbf{y}, \phi, \delta^2)\pi(\phi, \delta^2|\mathbf{y})d\phi d\delta^2.
\end{aligned} \tag{2.35}$$

In the majority of the literature, conditional predictions are made from the Gaussian process model, with estimates of the unknown decay and noise-to-signal ratio parameters "plugged-in" to equations such as (2.30) (e.g. Zhu and Stein (2006), Zimmerman (2006)). If we compare this with the Bayesian counterpart (2.35), we can see that is the weighted average of the plug-in approach with weights corresponding to the posterior density $\pi(\phi, \delta^2|\mathbf{y})$, which incorporates information from the available data.

A similar problem exists for the posterior density $\pi(\boldsymbol{\beta}|\mathbf{y})$, which cannot be expressed in closed form:

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y}) &= \iiint \pi(\boldsymbol{\beta}, \sigma^2, \phi, \delta^2|\mathbf{y}) d\sigma^2 d\phi d\delta^2 \\ &= \iint \pi(\boldsymbol{\beta}|\mathbf{y}, \phi, \delta^2)\pi(\phi, \delta^2|\mathbf{y}) d\phi d\delta^2.\end{aligned}$$

If δ^2 is known, then the posterior distribution reduces to:

$$\pi(\boldsymbol{\beta}|\mathbf{y}) = \int \pi(\boldsymbol{\beta}|\mathbf{y}, \phi, \delta^2)\pi(\phi|\mathbf{y}, \delta^2) d\phi.$$

These integrals do not have an analytical solution and, as a result, a numerical evaluation is required. We employ either Monte Carlo integration or quadrature methods, overviewed in Section 3.5.

Evaluation of these integrals using numerical methods is computationally expensive, especially in the context of optimal design where we have to optimise a function that is an integral of $\mathbf{y}$. For this reason, we propose in the next chapter an approximation to overcome this problem that avoids time consuming Monte Carlo integration.

2.5 Summary

This chapter has introduced a number of key concepts for Gaussian process models and reviewed the Bayesian approach which will be used throughout this thesis. For conjugate prior distributions for the trend and variance parameters, we define the posterior and predictive distributions, when the decay and noise-to-signal ratio parameters are either known or unknown.

Chapter 3

Optimal Design and the Proposed Design Selection Criterion

3.1 Introduction

This chapter begins with a brief overview and comparison of frequentist and Bayesian theory for optimal designs. An outline is then given of the decision theoretic framework that is used to derive an objective function, the expected loss, to be minimised in a Bayesian design selection criterion when the aim of an experiment is precise prediction at unobserved points. This aim is common in spatial, spatio-temporal and computer experiments. We then propose a new approximation to the expected loss that avoids the computational burden usually associated with finding Bayesian optimal designs. Algorithms for finding optimal designs are briefly reviewed and the coordinate exchange algorithm, used to find designs in later chapters, is described.

3.2 Brief Overview of Approaches to Design Selection

In this thesis, we develop highly efficient and optimal designs using a Bayesian approach. As the majority of the literature develops designs from a frequentist perspective, in this section we briefly overview the frequentist and Bayesian methods. Much of the theory of frequentist optimal design goes back to the work of Kiefer and Wolfowitz (Kiefer and Wolfowitz, 1959), including the introduction of the well known "alphabetic optimality" criteria. Atkinson et al. (2007) is a useful source of information on the theory of optimal design including optimality criteria. The theory for Bayesian optimal design is more recent and was introduced by Lindley (1956) and Chaloner (1984).

A design ξ^* of size n is defined as optimal, with respect to a specific criterion, by

comparison with the set Ξ of all possible designs of the same size. A criterion for a design to be optimal in Ξ is defined through an objective function, Ψ, reflecting the aim of the experiment, which is to be minimised or maximised.

In the frequentist approach to design, many criteria have Ψ formulated as a function of the Fisher information matrix $M(\xi), \xi \in \Xi$, as the maximum likelihood estimator of $\hat{\tilde{\boldsymbol{\theta}}}$ has asymptotic distribution $N(\tilde{\boldsymbol{\theta}}, (nM(\xi))^{-1})$; see Atkinson et al. (2007) for details. The most popular frequentist design criteria are A-optimality and D-optimality, defined as follows:

Definition 3.1. A design ξ^* is A-optimal if

$$\Psi_A(\xi^*) = \min_{\xi \in \Xi} \text{tr}(M(\xi)^{-1}).$$

This criterion seeks to minimise the average of the variances of the parameter estimators.

Definition 3.2. A design ξ^* is D-optimal if

$$\Psi_D(\xi^*) = \max_{\xi \in \Xi} [\det(M(\xi))]^{1/p}.$$

This criterion maximises the generalised variance of the parameter estimators, i.e. it minimises the volume of the ellipsoidal confidence region for $\tilde{\boldsymbol{\theta}}$.

The Bayesian approach to design of experiments uses prior information about the parameters (see Section 3.3). Chaloner and Verdinelli (1995) gave Bayesian selection criteria that correspond to the above frequentist D- and A-criteria for a normal linear model with conjugate prior distributions. A Bayesian design is D-optimal if it maximises the expected gain in Shannon Information, i.e. the gain in moving from a prior distribution to a posterior distribution. A Bayesian design is A-optimal if it minimises the expected squared error loss which is defined and used in the next section. Note that not every frequentist optimality criterion has a corresponding utility-based Bayesian criterion.

The main difference between the frequentist and Bayesian approaches to design is that the former may require assumptions about the values of the unknown parameters, while the latter accounts for the uncertainty in any prior knowledge of the parameters by treating the parameters as random variables and assigning a prior distribution to them. By doing this, we take into account previous knowledge and uncertainty about the parameters. A disadvantage of the frequentist approach is that a design that is optimal for one set of specified parameter values may not be optimal for a different set. The Bayesian approach suffers from this problem to a lesser degree because it

assigns a prior distribution to the parameters and integrates out the uncertainty in parameters by working with the posterior distributions (as in Section 2.4.3). Of course, the optimal designs may still be sensitive to the choice of prior distributions. The main drawback of Bayesian optimal design is the large computational burden usually incurred by optimising the objective function Ψ. This is because Ψ is often not analytically tractable, and its optimisation requires repeated approximation of an integral.

In this thesis, we often compare the performance of two designs using their relative efficiency, defined as follows.

Definition 3.3. The *relative efficiency* with respect to an objective function $\boldsymbol{\Psi}$ of design ξ_1 compared with a design ξ_2 is given by

$$\text{Eff}(\xi_1, \xi_2) = \frac{\boldsymbol{\Psi}(\xi_1)}{\boldsymbol{\Psi}(\xi_2)}.$$

When a design ξ is compared with an optimal design, ξ^*, we write

$$\text{Eff}(\xi) = \text{Eff}(\xi^*, \xi).$$

and call this measure the *efficiency* of design ξ.

In the following section, we give further details on the Bayesian approach, including the Bayesian design selection criterion for prediction when a quadratic loss function is assumed. In Section 3.4 we then propose a closed-form approximation for the objective function from that criterion under a Gaussian process model when the decay parameter ϕ and the noise-to-signal ratio δ^2 are unknown.

3.2.1 Space-filling designs

Designs based on geometric criteria, developed by Johnson et al. (1990) are in two categories: maximin distance designs and minimax distance designs. These criteria can be defined:

Maximin criterion: a maximin optimal design ξ^* maximises

$$\psi_{Mm}(\xi) = \min_{i \neq j} d(\mathbf{x}_i, \mathbf{x}_j), \quad \text{where} \quad \mathbf{x}_i, \mathbf{x}_j \in \xi,$$

Minimax criterion: a minimax optimal design ξ^* minimises

$$\psi_{mM}(\xi) = \max_{\mathbf{x}' \in \mathcal{X}} d(\mathbf{x}', \mathbf{x}), \quad \text{where} \quad \mathbf{x} \in \xi \quad \text{and} \quad d(\mathbf{x}', \mathbf{x}) = \min_{i,\ldots,n} d(\mathbf{x}', \mathbf{x}_i).$$

The maximin designs maximise the smallest distance between pairs of points in the design; in this way, no two points in the design are "too close" and the points are spread

throughout the region. On the other hand, minimax designs minimise the maximum distance from all the points in the region to their closest point in the design; here the points cover the design region.

The coverage and spread measures, related to the minimax and maximin designs respectively, are given by

$$\int_{\mathcal{X}} \left[\min_{\mathbf{x} \in \xi^*} d(\mathbf{x}, \mathbf{x}') \right] d\mathbf{x}', \tag{3.1}$$

and

$$\sum_{i=1}^{n} \left[\min_{\mathbf{x} \in \xi^* \setminus \{\mathbf{x_i}\}} d(\mathbf{x}, \mathbf{x_i}) \right], \tag{3.2}$$

see Bowman and Woods (2013), and references therein, for details.

Two classes of designs based on geometric criteria were developed by Johnson et al. (1990), maximin and minimax designs, and asymptotic optimality properties were presented for specific designs. In particular, for Gaussian process model (2.6) with constant mean function and all model parameters were known, they showed that maximin optimal designs are asymptotically D-optimal and the minimax optimal designs are asymptotically G-optimal for weak correlations (i.e. G-optimality seeks to minimise the maximum predictive variance). A drawback of maximin designs is that they place the majority of points on the boundary of the region and the interior region is not well explored. In general, minimax designs are computationally difficult to generate and for this reason are not widely used.

3.3 Bayesian Optimal Design via a Decision Theoretic Framework

Following Chaloner and Verdinelli (1995), the Bayesian design problem is formulated as a decision theoretic problem. "Decision theory" refers to decision making in the presence of statistical knowledge which can provide some information on the uncertainties involved in the problem.

The basic elements of a decision problem are:

- the "truth", or state of nature, denoted by $\boldsymbol{\alpha} \in \mathcal{A}$
- a decision denoted by $\gamma(\mathbf{y})$, a function of data, e.g. an estimator or prediction, and decision space $\mathcal{G}$
- a loss function $L(\gamma(\mathbf{y}), \boldsymbol{\alpha}; \xi)$ defined in the space $\mathcal{G} \times \mathcal{A}$. The loss function measures the consequence of choosing a particular decision $\gamma(\mathbf{y})$ when $\boldsymbol{\alpha}$ is the truth. The loss function is chosen according to the aim of the experiment; for example, the aim may be estimation of the unknown parameters or prediction at an unobserved point. Different choices of loss function lead to different optimality criteria.

An important use of experimental data is the prediction of responses at new points not included in the experiment. In this case the Bayesian decision theoretic framework uses the predictive distribution and a loss function involving the prediction of future observations is employed.

In what follows, we adopt the general notation for a prediction $y(\mathbf{x}_p) \in \mathcal{Y}$ at a new point $\mathbf{x}_p \in \mathcal{X}_\mathcal{P}$, instead of using y_{n+1} and $\mathbf{x}_{n+1}$, respectively as in Chapter 2. This notation facilitates the formulation of a design criterion to find optimal designs for making predictions at one or more points in $\mathcal{X}_\mathcal{P}$. In what follows, we concentrate on predictions over $\mathcal{X}_\mathcal{P}$.

3.3.1 Design for prediction

The Bayesian approach to design uses the predictive distribution of $y(\mathbf{x}_p)$ at a new point $\mathbf{x}_p$ to obtain an optimal design for a particular loss function as follows:

1. Formulate the expected loss with respect to the posterior predictive distribution, $\pi(y(\mathbf{x}_p)|\mathbf{y})$, for any decision $\gamma(\mathbf{y}) \in \mathcal{G}$, any design $\xi \in \Xi$ and the chosen loss function L as

$$\mathbb{E}[L(y(\mathbf{x}_p), \gamma(\mathbf{y}); \xi)|\mathbf{y}] = \int_\mathcal{Y} L(y(\mathbf{x}_p), \gamma(\mathbf{y}); \xi)\pi(y(\mathbf{x}_p)|\mathbf{y})dy(\mathbf{x}_p). \tag{3.3}$$

2. Derive the minimum of the expected loss with respect to the decision $\gamma(\mathbf{y})$.

3. Obtain the objective function by averaging the minimum expected loss over the marginal density of the data $\mathbf{y}$ and over the prediction region $\mathcal{X}_P$:

$$\Psi(\xi) = \int_{\mathcal{X}_\mathcal{P}} \int_\mathcal{Y} \min_{\gamma(\mathbf{y})\in\mathcal{G}} \mathbb{E}[L(y(\mathbf{x}_p), \gamma(\mathbf{y}); \xi)|\mathbf{y}]\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p. \tag{3.4}$$

4. Then an optimal design, ξ^*, for prediction is a design that minimises this objective function, i.e. $\xi^* = \arg\min_{\xi\in\Xi} \Psi(\xi)$.

We follow the above approach for the squared loss function given by

$$L(y(\mathbf{x}_p), \gamma(\mathbf{y}); \xi) = (y(\mathbf{x}_p) - \gamma(\mathbf{y}))^2. \tag{3.5}$$

The decision $\hat{\gamma}(\mathbf{y})$, that minimises the expected loss is found by substituting (3.5) into (3.3), expanding the squared loss function and setting the first derivative to zero. We

obtain:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\gamma(\mathbf{y})}\left[\int_{\mathcal{Y}} y(\mathbf{x}_p)^2\pi(y(\mathbf{x}_p)|\mathbf{y})dy(\mathbf{x}_p)\right] - \frac{\partial}{\partial\gamma(\mathbf{y})}\left[2\gamma(\mathbf{y})\int_{\mathcal{Y}} y(\mathbf{x}_p)\pi(y(\mathbf{x}_p)|\mathbf{y})dy(\mathbf{x}_p)\right] \\
&\quad + \frac{\partial}{\partial\gamma(\mathbf{y})}\left[\gamma(\mathbf{y})^2\int_{\mathcal{Y}} \pi(y(\mathbf{x}_p)|\mathbf{y})dy(\mathbf{x}_p)\right] \\
&= 2\gamma(\mathbf{y}) - 2\int_{\mathcal{Y}} y(\mathbf{x}_p)\pi(y(\mathbf{x}_p)|\mathbf{y})dy(\mathbf{x}_p), \qquad (3.6)
\end{aligned}
$$

from which it follows that the optimal decision with respect to the squared loss function is $\hat{\gamma}(\mathbf{y}) = \mathbb{E}[y(\mathbf{x}_p)|\mathbf{y}]$.

Under other loss functions, such as absolute or step loss, other predictions are optimal, median and mode respectively, and different objective functions result for design selection.

Substitution of $\gamma(\mathbf{y}) = \hat{\gamma}(\mathbf{y})$ into (3.4) gives

$$
\begin{aligned}
\Psi(\xi) &= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}} \mathbb{E}[L(y(\mathbf{x}_p), \mathbb{E}(y(\mathbf{x}_p)|\mathbf{y}); \xi)]\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p. \\
&= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}} \mathbb{E}\left[\{y(\mathbf{x}_p) - \mathbb{E}(y(\mathbf{x}_p)|\mathbf{y})\}^2\right]\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p. \\
&= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}} \mathrm{var}[y(\mathbf{x}_p)|\mathbf{y}]\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p. \qquad (3.7)
\end{aligned}
$$

A Ψ-optimal design minimises (3.7). We adopt the notation "Ψ-optimal" to denote a Bayesian optimal design for prediction under squared error loss.

3.4 Bayesian designs for prediction via the Gaussian process model

In this section we develop methodology for Bayesian design for prediction under squared error loss for the Gaussian process model (2.6), assuming all model parameters are uncertain.

In practice, the values of covariance parameters, ϕ and δ^2, are usually unknown and hence we need to take into account the uncertainty in their values. We saw in Section 2.4.5 that if ϕ and δ^2 are known, then the predictive distribution is a t-distribution. The inner integral in $\Psi(\xi)$ (3.7) is then analytically tractable. If ϕ is unknown, and either δ^2 is known or δ^2 is unknown (see Section 2.4.6 Case(i) and Case(ii)), then the posterior predictive distribution does not have a standard form, and we cannot find an analytical expression for the integral in (3.7). As a result, to evaluate $\Psi(\xi)$, we need to evaluate the variance of the posterior predictive distribution using Monte Carlo Markov Chain (MCMC) methods and also evaluate the integral with respect to the unknown data.

The repeated use of MCMC methods to obtain a Bayesian optimal design is time consuming and computationally prohibitive. For this reason, we propose a new-closed form approximation to $\Psi(\xi)$ which allows us to find Bayesian optimal designs when both ϕ and δ^2 are unknown, the most important general case, as well as ϕ unknown and δ^2 known.

Using the fact that we have an analytical expression for the predictive distribution when the parameters ϕ and δ^2 are known, see (2.30), we employ the law of total of variance (2.21) to obtain the variance of the posterior predictive distribution $\text{var}(y(\mathbf{x}_p)|\mathbf{y})$ when ϕ and δ^2 are unknown:

$$\begin{aligned}\text{var}(y(\mathbf{x}_p)|\mathbf{y}) &= \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{\text{var}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\} + \text{var}_{\phi,\delta^2|\mathbf{y}}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\} \\ &= \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{\text{var}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\} + \\ &\quad \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left[\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2) - \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}\right\}^2\right]. \qquad (3.8)\end{aligned}$$

Here, the expectations are with respect the posterior distribution of $\pi(\phi,\delta^2|\mathbf{y})$, given by (2.34), which is not a standard distribution. This leads to the following simple theorem.

Theorem 3.1. *Consider the Gaussian process model (2.6) and using the squared error loss function, the objective function with unknown correlation parameter ϕ and noise-to-signal ratio δ^2 decomposes into two integrals via replacement of $var(y(\mathbf{x}_p)|\mathbf{y})$ with the expression (3.8):*

$$\begin{aligned}\Psi(\xi) &= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}} var(y(\mathbf{x}_p)|\mathbf{y})\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p \\ &= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}} \left(\mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{var(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\} + var_{\phi,\delta^2|\mathbf{y}}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}\right)\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p \\ &= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}}\int_{\delta^2}\int_{\phi} var(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\pi(\phi,\delta^2|\mathbf{y})\pi(\mathbf{y})d\phi d\delta^2 d\mathbf{y}d\mathbf{x}_p + \\ &\int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}}\int_{\delta^2}\int_{\phi}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2) - \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}\right\}^2\pi(\phi,\delta^2|\mathbf{y})\pi(\mathbf{y})d\phi d\delta^2 d\mathbf{y}d\mathbf{x}_p \\ &= \Psi_1(\xi) + \Psi_2(\xi). \qquad (3.9)\end{aligned}$$

When δ^2 is assumed known, then objective function (3.9) reduces to

$$\begin{aligned}\Psi(\xi) &= \int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}}\int_{\phi} \mathbb{E}_{\phi|\mathbf{y},\delta^2}\left\{var(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}\pi(\phi|\mathbf{y},\delta^2)\pi(\mathbf{y})d\phi d\mathbf{y}d\mathbf{x}_p + \\ &\int_{\mathcal{X}_\mathcal{P}}\int_{\mathcal{Y}}\int_{\phi} var_{\phi|\mathbf{y},\delta^2}\left\{\mathbb{E}(y(\mathbf{x}_p)\mathbf{y},\phi,\delta^2))\pi(\phi|\mathbf{y},\delta^2\right\}\pi(\mathbf{y})d\phi d\mathbf{y}d\mathbf{x}_p \\ &= \Psi_1(\xi) + \Psi_2(\xi). \qquad (3.10)\end{aligned}$$

Here the expectation and the variance are with respect to the conditional posterior dis-

tribution of ϕ given by (2.32).

Proof of Theorem 3.1 follows directly from the application of the law of total variance and using (3.8).

Assumption 3.1. From now on, we assume conjugate prior distributions for the trend parameters $\boldsymbol{\beta}$ and the Gaussian process variance σ^2, that is a normal inverse-gamma distribution.

This assumption allows analytical calculation of the integrals with respect to the data $\mathbf{y}$ in $\Psi_1(\xi)$, given by

$$\begin{aligned}
\Psi_1(\xi) &= \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \int_{\delta^2} \int_{\phi} \text{var}(y(\mathbf{x}_p)|\mathbf{y}, \phi, \delta^2)\pi(\phi, \delta^2|\mathbf{y})\pi(\mathbf{y}) d\phi d\delta^2 d\mathbf{y} d\mathbf{x}_p \\
&= \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \int_{\delta^2} \int_{\phi} \text{var}(y(\mathbf{x}_p)|\mathbf{y}, \phi, \delta^2)\pi(\mathbf{y}|\phi, \delta^2)\pi(\phi, \delta^2) d\phi d\delta^2 d\mathbf{y} d\mathbf{x}_p \\
&= \int_{\mathcal{X}_\mathcal{P}} \int_{\delta^2} \int_{\phi}^{\Sigma^*} \int_{\mathcal{Y}} \frac{b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top [\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)}{2a + n - 2} \pi(\mathbf{y}|\phi, \delta^2) d\mathbf{y} \pi(\phi, \delta^2) d\phi d\delta^2 d\mathbf{x}_p \\
&= \frac{b}{a-1} \int_{\mathcal{X}_\mathcal{P}} \int_{\delta^2} \int_{\phi} \Sigma^* \pi(\phi, \delta^2) d\phi d\delta^2 d\mathbf{x}_p,
\end{aligned} \tag{3.11}$$

$$\Psi_2(\xi) = \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \int_{\delta^2} \int_{\phi} [(\mu^* - \mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*))(\mu^* - \mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*))^\top] \pi(\phi, \delta^2|\mathbf{y}) \pi(\mathbf{y}) d\phi d\delta^2 d\mathbf{y} d\mathbf{x}_p,$$

where the posterior mean is given by

$$\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*) \quad = \quad \int_{\mathcal{X}_\mathcal{P}} \int_{\delta^2} \int_{\phi} \mu^* \pi(\phi, \delta^2|\mathbf{y}) d\phi d\delta^2 d\mathbf{x}_p. \tag{3.12}$$

Here, $\Sigma^*, a, b, F, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \mathbf{R}, \mu*$ are defined in Section 2.4.5. Equation (3.11) follows from $\pi(\mathbf{y}|\phi, \delta^2)$ having a t-distribution, and using the quadratic form:

$$\mathbb{E}[\boldsymbol{\varepsilon}^T \Lambda \boldsymbol{\varepsilon}] = \text{tr}(\Lambda K) + \boldsymbol{\mu}^T \Lambda \boldsymbol{\mu}, \tag{3.13}$$

where $\boldsymbol{\mu}$ and K are the mean vector and variance-covariance matrix of $\boldsymbol{\varepsilon}$ respectively. Here we apply (3.13) with $\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)$ and $\Lambda = [\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}$. The mean and the variance-covariance of $(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)$ are $\boldsymbol{\mu} = 0$ and $K = \frac{2b}{2a-2}[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]$, respectively.

When the noise-to-signal ratio δ^2 is known, the objective function $\Psi(\xi)$ is given by

(3.10) and $\Psi_1(\xi)$ and $\Psi_2(\xi)$ are given by

$$\begin{aligned}\Psi_1(\xi) &= \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \mathbb{E}_{\phi|\mathbf{y},\delta^2}(\text{var}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2))\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p \\ &= \frac{b}{a-1}\int_{\mathcal{X}_\mathcal{P}} \int_{\phi} \Sigma^*\pi(\phi)d\phi d\mathbf{x}_p, \end{aligned} \tag{3.14}$$

$$\Psi_2(\xi) = \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \int_{\phi} [(\mu^* - \mathbb{E}_{\phi|\mathbf{y},\delta^2}(\mu^*))(\mu^* - \mathbb{E}_{\phi|\mathbf{y},\delta^2}(\mu^*))^\top]\pi(\phi|\mathbf{y},\delta^2)\pi(\mathbf{y})d\phi d\mathbf{y}d\mathbf{x}_p,$$

and the posterior mean is given by

$$\mathbb{E}_{\phi|\mathbf{y},\delta^2}(\mu^*) = \int_{\mathcal{X}_P} \int_{\phi} \mu^*\pi(\phi|\mathbf{y},\delta^2)d\phi d\mathbf{x}_p. \tag{3.15}$$

The density $\pi(\phi|\mathbf{y},\delta^2)$ is given by (2.32).

Conjecture 3.1. Consider the Gaussian process model (2.6) and the objective function $\Psi(\xi)$ from Theorem 3.1. Based on the Assumption 3.1 we conjecture that $\Psi_2(\xi) << \Psi_1(\xi) \Rightarrow \Psi(\xi) \approx \Psi_1(\xi) \; \forall \; \pi(\phi,\delta^2)$.

Finding optimal designs using a fully Bayesian approach is computationally infeasible. We need to minimise objective function $\Psi(\xi)$, which involves an analytically intractable high dimensional integral with respect to the data. Only very small examples of designs can be found using this objective function. However, using our proposed approximation we are able to overcome the computational burden associated with Bayesian optimal designs and find designs for larger examples.

Assessment on small examples, with e.g. four design points, have shown that finding designs by minimising $\Psi_1(\xi)$ is at least two or three orders of magnitude faster than finding designs minimising $\Psi(\xi)$. For many problems, this is the difference between the design search being computationally feasible or not.

We have both numerical and theoretical evidence to support the conjecture:

1. In Section 3.4.1, we outline some supporting theory.

2. Numerical evidence is presented in future chapters. In Chapters 5, 7 and 8, we numerically study the objective function $\Psi(\xi)$, approximated via Monte Carlo integration and quadrature, and the relative sizes of the two components, $\Psi_1(\xi)$ and $\Psi_2(\xi)$. For a wide variety of Gaussian process models and parameter values, we found that $\Psi_2(\xi) << \Psi_1(\xi)$ and hence we conjecture that $\Psi(\xi) \approx \Psi_1(\xi)$.

3. Our conjecture is in line with findings from Wu and Kaufman (2014) on the variance $\text{var}(y(\mathbf{x}_p)|\mathbf{y})$ of the posterior predictive distribution. Wu and Kaufman

Correlation function	Smoothness parameter, ν	$v(\phi)$	$\omega(\phi)$
Power exponential	$\nu \in (0, 2]$	ϕ^{ν}	$\phi^{2\nu}$
Matérn	$\nu < 1$	$\phi^{2\nu}$	ϕ^2
	$\nu = 1$	$\phi^2 \log(1/\phi)$	ϕ^2
	$1 < \nu < 2$	ϕ^2	$\phi^{2\nu}$
	$\nu = 2$	ϕ^2	$\phi^4 \log(1/\phi)$
	$\nu > 2$	ϕ^2	ϕ^4

Table 3.1: Asymptotic expansions of the power exponential (2.5), and Mátern correlation (2.4) functions as $\phi \to 0^+$.

(2014) investigated how the choice of prior distribution for the unknown parameters in a Gaussian process model affects the predictive performance of posterior distributions in spatial modelling. Analogously to (3.9), they used the law of total variance (2.21) to decompose the posterior predictive variance in two parts, conditional on the unknown correlation parameters. They conducted simulation studies and concluded that the first component, $\mathbb{E}_{\phi,\delta^2|\mathbf{y}}\left\{\text{var}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}$, dominates the total variance with the second component, $\text{var}_{\phi,\delta^2|\mathbf{y}}\left\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y},\phi,\delta^2)\right\}$, having negligible magnitude in comparison. They found this to be the case regardless of the choice of prior distributions. Therefore, their finding supports our numerical evidence that $\Psi_2(\xi)$ has less magnitude than $\Psi_1(\xi)$.

3.4.1 Supporting theory

The supporting theoretical evidence for Conjecture 3.1 relies on the limiting behaviour of the integrated likelihood $L^I(\phi, \delta^2)$. In order to study the limiting behaviour of $L^I(\phi, \delta^2)$, we make use of properties of the correlation function. We follow a similar approach as described by Berger et al. (2001), Kazianka and Pilz (2012) and Ren et al. (2012). All authors considered the case of examining the limiting behaviour of the integrated likelihood when non-informative prior distributions are assigned to $\boldsymbol{\beta}$ and σ^2. Berger et al. (2001) considered a model with an isotropic correlation function and no nugget in the model while Kazianka and Pilz (2012) and Ren et al. (2012) considered the case of a nugget. Kazianka and Pilz (2012) and Ren et al. (2012) examined the same problem but they considered a different parametrisation for the model. Kazianka and Pilz (2012) considered model (2.11) and Ren et al. (2012) model (2.12); see Section 2.4.4.

Throughout this section we use the asymptotic expansion of a continuous correlation function in order to express the correlation matrix $\mathbf{C}(\phi)$ as a sum of matrices required in Assumption 3.3. Therefore, a continuous correlation function, as it is described in

Section 2.3, is often assumed to have a Taylor expansion of the form

$$\rho(\phi; d) = 1 + v(\phi)h_1(d) + \omega(\phi)h_2(d) + r(\phi) \quad \text{as} \quad \phi \to 0^+, \tag{3.16}$$

for $v(\phi)$ and $\omega(\phi)$ known functions, $h_1(d)$ and $h_2(d)$ are known function of the distance d, where d is the Euclidean distance between input points, and $r(\phi)$ is a remainder term. The asymptotic expansions of power exponential (2.5) and Mátern correlation (2.4) functions given by Kazianka and Pilz (2012) are presented in Table 3.1. For $\nu = 1$, $h_1(d) = -d^2/2$ and $h_2(d) = (d^2/2)(\log(d) - \log(2) + 1/2 + c$ where c is Euler's constant. For $\nu > 1$ $h_1(d) = d^2/(4(1-\nu))$ and $h_2(d) = (d^2/2)^{2\nu}\Gamma(1-\nu)/\Gamma(1+\nu)$, where $\Gamma(\cdot)$ is the Gamma function, and for $\nu < 1$ $h_1(d)$ and $h_2(d)$ are switched.

For our mathematical results, we require the following assumptions to hold for the correlation function $\rho(\phi; d)$, taken from Berger et al. (2001) and Kazianka and Pilz (2012). We extend the results of Ren et al. (2012) to normal-inverse gamma conjugate prior distribution.

Assumption 3.2. $\rho(\phi; d)$ is a continuous function of $\phi > 0$ such that, for any $d \geq 0$, $\rho(\phi; d) = \rho^0(d\phi)$ where $\rho^0(\cdot)$ is a correlation function satisfying $\lim_{h\to+\infty} \rho^0(h) = 0$.

Assumption 3.3. As $\phi \to 0^+$, $\mathbf{C}(\phi) = \mathbf{1_n}\mathbf{1_n}^\top + v(\phi)\mathbf{D} + \omega(\phi)\mathbf{D}^* + \mathbf{G}(\phi)$, where $\mathbf{D}$ is a non-singular, fixed matrix with ijth entry $h_1(d_{ij})$, $\mathbf{1_n}\mathbf{1_n}^\top + v(\phi)\mathbf{D}$ is a positive definite matrix and $\mathbf{D}^*$ is a fixed matrix with ijth entry $h_2(d_{ij})$. Also $v(\phi)$, $\omega(\phi)$ and $\mathbf{G}(\phi)$ are continuous with respect to ϕ, satisfying $\frac{\omega(\phi)}{v(\phi)} \to 0$ and $\frac{\|\mathbf{G}(\phi)\|_\infty}{\omega(\phi)} \to 0$ as $\phi \to 0^+$. Here $\|\cdot\|_\infty$ denotes the matrix max-norm, that is $\|\mathbf{A}\|_\infty = \max_{i,j}|a_{ij}|$.

Assumptions 3.2 and 3.3 are satisfied for power exponential and Matérn correlation functions, see Berger et al. (2001).

Lemma 3.1. Consider the Gaussian process model (2.6), conjugate prior distributions (Assumption 3.1) and a correlation function that satisfies Assumptions 3.2 and 3.3. Then the integrated likelihood $L^I(\phi, \delta^2)$ is a continuous function for $(\phi, \delta^2) \in (0, \infty) \times (0, \infty)$ and has the following limiting behaviour:

(a) When $(\phi, \delta^2) \to (\phi, 0^+)$ or $(\phi, \delta^2) \to (\infty, \delta^2)$ or $(\phi, \delta^2) \to (\infty, 0^+)$ the respective limits of $L^I(\phi, \delta^2)$ exist and are positive.

(b) When $(\phi, \delta^2) \to (0^+, 0^+)$ or $(\phi, \delta^2) \to (0^+, \delta^2)$, i.e. when $\phi \to 0^+$ and δ^2 is known and fixed or $\delta^2 \to 0^+$, then

$$L^I(\phi, \delta^2) = \begin{cases} \mathcal{O}\big((\delta^2 + v(\phi))^{a+k/2+1/2}\big) & \text{if } \mathbf{1} \notin \mathcal{C}(\mathbf{F}) \\ \mathcal{O}\big((\delta^2 + v(\phi))^{a+k/2}\big) & \text{if } \mathbf{1} \in \mathcal{C}(\mathbf{F}) \end{cases}$$

where $\mathcal{C}(\mathbf{F})$ is the set of columns of $\mathbf{F}$.

(c) When $(\phi, \delta^2) \to (\infty, \infty)$ or $(\phi, \delta^2) \to (\phi, \infty)$ or $(\phi, \delta^2) \to (0^+, \infty)$, i.e. when

$\delta^2 \to \infty$ and ϕ is known and fixed or $\phi \to 0^+$, then

$$L^I(\phi, \delta^2) = \mathcal{O}\big((\delta^2)^{a+k/2}\big),$$

where $a > 0$ is the prior hyperparameter which corresponds to the shape parameter of the inverse-gamma prior for σ^2, and k corresponds to the number of trend parameters. Here the notation $(\phi, \delta^2) \to (\phi, \cdot)$, or $(\phi, \delta^2) \to (\cdot, \delta^2)$, indicates that we fix ϕ while δ^2 tends to a limit, or fix δ^2 as ϕ tends to a limit, respectively. Also we define $g_1(x) = \mathcal{O}\big(g_2(x)\big)$ if $|g_1| \leq M|g_2|$ for all $|x - x_0| < c$, for some positive numbers M, x_0 and c. It is used to describe the behaviour of the function $g_1(x)$ near the limit $x_0 < \infty$. If we want to describe the behaviour of the function $g_1(x)$ as $x \to \infty$ we define $g_1(x) = \mathcal{O}\big(g_2(x)\big)$ if $|g_1| \leq M|g_2|$ for all $x \geq x_0$, for some positive numbers M, x_0.

The proof of Lemma 3.1 can be found in Appendix A.1.

We can now use our understanding of the limiting behaviour of the integrated likelihood to provide insights to support the conjecture.

Recall that $\Psi_2(\xi)$, (3.9), is the variance of the mean of the predictive distribution conditional on the correlation parameters and noise-to-signal ratio, averaged across the joint posterior distribution of these unknown parameters and the data, i.e.

$$\Psi_2(\xi) = \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \int_{\delta^2} \int_{\phi} \Big\{ \mathbb{E}(y(\mathbf{x}_p)|\mathbf{y}, \phi, \delta^2) - \mathbb{E}_{\phi,\delta^2|\mathbf{y}}\big\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y}, \phi, \delta^2)\big\}\Big\}^2 \pi(\phi, \delta^2|\mathbf{y})\pi(\mathbf{y}) d\phi d\delta^2 d\mathbf{y} d\mathbf{x}_p, \tag{3.17}$$

where $\pi(\phi, \delta^2|\mathbf{y})$ is given by (2.34), and depends on the integrated likelihood.

We now consider the form of Ψ and Ψ_2 under the three cases from Lemma 3.1.

(a): When $(\phi, \delta^2) \to (\phi, 0)$, $\boldsymbol{\Sigma} = \mathbf{C}(\phi)$, a fixed and known matrix. The objective function (3.9) reduces to:

$$\Psi(\xi; \phi, \delta^2) = \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \text{var}[y(\mathbf{x}_p)|\mathbf{y}, \phi, \delta^2] \pi(\mathbf{y}) d\mathbf{y} d\mathbf{x}_p,$$

when we make the assumption that ϕ and δ^2 are known. The inner integral with respect the unknown data $\mathbf{y}$ is tractable. For the case of known ϕ and δ^2, and non-informative prior distributions on the trend parameters and the variance, the Bayesian optimal designs coincide with the designs obtained from the frequentist approach.

When $(\phi, \delta^2) \to (\infty, \delta^2)$ or $(\infty, 0)$, $\boldsymbol{\Sigma} = (\delta^2 + 1)\mathbf{I}$ or $\boldsymbol{\Sigma} = \mathbf{I}$ (if $\delta^2 = 0$) which corresponds to the case of a linear model for uncorrelated data. The criterion reduces to that for prediction for a linear model.

(b): When $(\phi, \delta^2) \to (0^+, 0^+)$ or $(\phi, \delta^2) \to (0^+, \delta^2)$, the integrated likelihood is bounded by polynomials of ϕ and δ^2. Since $a + k/2 + 1/2 > 1$ and $a + k/2 > 1$, then as $\phi \to 0^+$, and either $\delta^2 \to 0^+$ or δ^2 is fixed, the integrated likelihood goes to zero faster than

$(\delta^2+v(\phi))^{a+k/2+1/2}$ or $(\delta^2+v(\phi))^{a+k/2}$. As a result $L^I(\phi,\delta^2)$ tends to zero faster than ϕ and δ^2, and the posterior distribution $\pi(\phi,\delta^2|\mathbf{y})$ (2.34) yields very small values. When averaged across all possible data, $\Psi_2(\xi)\approx 0$. Similar to case (a) the linear model is a good approximation for Gaussian process model.

(c): When $(\phi,\delta^2)\to(\infty,\infty)$,$(\phi,\delta^2)\to(0^+,\infty)$ or $(\phi,\delta^2)\to(\phi,\infty)$, the integrated likelihood is bounded by polynomials in δ^2 and, as $a+k/2>1$ as $\delta^2\to\infty$, the ratio $L^I(\phi,\delta^2)/\delta^2$ goes to zero faster than $\delta^2\to\infty$. However, the rate of convergence to 0 of the integrated likelihood in this case is slower compared to case (b). Therefore, we expect that larger values of δ^2 will provide larger $\Psi_2(\xi)$ in this case compared with case (b). This is in line with our numerical evidence.

For case (b), numerical studies indicate $\Psi_2(\xi)$ is always of order at most 10^{-3} and, for case (c) when we assume very large values of δ^2, it is of order at most10^{-2}; in both cases, $\Psi_1(\xi)\leq 1$. In both cases, the integrated likelihood gets very small, and $\Psi_2(\xi)<<\Psi_1(\xi)$. In both cases, $\Psi_1(\xi)\leq 1$.

3.5 Bayesian Computation

Objective functions (3.9) and (3.10) both require numerical approximation. We introduce the two main methods of evaluating an intractable integral using numerical methods: Monte Carlo integration and deterministic quadrature. We derive the approximations to integrals in (3.9) and (3.10) that are needed to find Bayesian optimal designs.

Monte Carlo integration: The basic idea here is that summary statistics from a large sample from the distribution of $\tilde{\boldsymbol{\theta}}$ can be used to approximate, for example, the moments of the distribution.

Suppose we are interested in a function $f(\tilde{\boldsymbol{\theta}})$ of the parameters and can simulate a sample $\tilde{\boldsymbol{\theta}}^{(1)},\ldots,\tilde{\boldsymbol{\theta}}^{(N)}$ from the distribution $\pi(\tilde{\boldsymbol{\theta}})$, for example, using MCMC methods. Then we can approximate $\mathbb{E}\{f(\tilde{\boldsymbol{\theta}})\}$ as

$$\int_\Theta f(\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}}\simeq\frac{1}{N}\sum_{i=1}^{N}f(\tilde{\boldsymbol{\theta}}^{(i)}). \tag{3.18}$$

Monte Carlo methods are straightforward to implement and, through increasing sample size N, arbitrary precision can be obtained. Clearly, if $\pi(\tilde{\boldsymbol{\theta}})$ is a difficult distribution to sample from or $f(\cdot)$ is an expensive function to evaluate, the Monte Carlo method may have substantial computational cost.

Gaussian quadrature methods: This is a class of numerical techniques for approximating the integral in (3.18) when the form of $\pi(\tilde{\boldsymbol{\theta}})$ is known. The best choice of quadrature method depends on the location and shape of this distribution. The inte-

gral is approximated by a weighted sum of the integrand at particular points within the domain of integration. An m-point Gaussian quadrature rule yields an exact result for $f(\cdot)$ being a polynomial of degree $2m-1$ or less by choosing unequally spaced grid points a_i, called abscissae, and weights w_i. That is, we use the general approximation

$$\int_{\Theta} f(\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}} \simeq \sum_{i=1}^{m} w_i f(a_i).$$

Often, transformations will need to be applied to $\tilde{\boldsymbol{\theta}}$ prior to applying the quadrature rule. Further details about Gaussian quadrature techniques may be found, for example, in Kythe and Schaferkotter (2005).

Below we describe two quadrature methods used in this thesis.

1. **Gauss-Hermite quadrature for a log-normal distribution**

   Gauss-Hermite quadrature is suitable for approximating an integral of the general form

   $$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx. \tag{3.19}$$

   Consider a log-normal prior distribution for an unknown parameter $\tilde{\theta}$ in a Bayesian model:

   $$\pi(\tilde{\theta}) = \frac{1}{\tilde{\theta}\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(\tilde{\theta})-\mu)^2}{2\sigma^2}\right\}. \tag{3.20}$$

   We can evaluate an integral of the form

   $$I_1 = \int_0^{\infty} f(\tilde{\theta})\frac{1}{\tilde{\theta}\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(\tilde{\theta})-\mu)^2}{2\sigma^2}\right\} d\tilde{\theta}, \tag{3.21}$$

   by applying a transformation $x = \log(\tilde{\theta})$ to obtain

   $$I_1 = \int_{-\infty}^{\infty} f(e^x)\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx,$$

   and the further substitution $x = \mu + z\sigma\sqrt{2}$ to give

   $$\begin{aligned} I_1 &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f\left(e^{\mu+z\sigma\sqrt{2}}\right) e^{-z^2} dz \\ &\simeq \frac{1}{\sqrt{\pi}} \sum_{i=1}^{m} w_i f\left(e^{\mu+a_i\sigma\sqrt{2}}\right), \end{aligned} \tag{3.22}$$

   where a_i and w_i $(i = 1, \ldots, m)$ denote the abscissae and weights obtained from the Hermite polynomial (Kythe and Schaferkotter (2005, p.118-119)).

2. **Gauss-Legendre quadrature for a uniform prior distribution**

   Gauss-Legendre quadrature is suitable for approximating integrals of the general

form

$$\int_{-1}^{1} f(x)dx. \tag{3.23}$$

Suppose that the prior distribution of an unknown parameter $\tilde{\theta}$ is the uniform distribution on an interval $[a_1, b_1]$, with density

$$\pi(\tilde{\theta}) = \frac{1}{b_1 - a_1}, \qquad a_1 \leq \tilde{\theta} \leq b_1. \tag{3.24}$$

Then we can evaluate an integral of the form

$$I_1 = \int_{a_1}^{b_1} f(\tilde{\theta}) \frac{1}{b_1 - a_1} d\tilde{\theta}, \tag{3.25}$$

by transformation and application of Gauss-Legendre quadrature.

Let $x = \frac{2\tilde{\theta} - (b_1 + a_1)}{(b_1 - a_1)}$. Then

$$\begin{aligned} I_1 &= \frac{1}{b_1 - a_1} \int_{a_1}^{b_1} f(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{2} \int_{-1}^{1} f\left(\frac{b_1 - a_1}{2} x + \frac{b_1 + a_1}{2} \right) dx \\ &\simeq \frac{1}{2} \sum_{i=1}^{m} w_i f\left(\frac{b_1 - a_1}{2} a_i + \frac{b_1 + a_1}{2} \right), \end{aligned} \tag{3.26}$$

where a_i and w_i are obtained from the Legendre polynomial (Kythe and Schaferkotter (2005, p.115-117)).

In the next subsections, we apply these approximations to objective function $\Psi(\xi)$. Depending to the choice of prior distribution for ϕ and δ^2, we use either Gauss-Hermite or Gauss-Legendre quadrature to approximate $\Psi_1(\xi)$. For $\Psi_2(\xi)$, Monte Carlo integration is required in addition to the application of quadrature methods.

3.5.1 Approximating the objective function with continuous prior distributions for ϕ and δ^2

As discussed in Section 2.4.4, two possible prior distributions for the decay parameter ϕ are a uniform prior or a log-normal prior. The latter distribution allows us to express subjective prior beliefs that some values of ϕ are more likely than other values. Both of these prior distributions are continuous, and the objective function $\Psi(\xi)$ (3.9) has components $\Psi_1(\xi)$ and $\Psi_2(\xi)$ given by (3.11) and (3.12) respectively.

There is no analytical solution for the integrals with respect to ϕ and δ^2 in (3.11), and we use Gaussian quadrature methods to approximate them numerically. Function $\Psi_1(\xi)$ is evaluated directly using quadrature methods. For $\Psi_2(\xi)$ the calculations are more

complicated as the integral is a function of the posterior density of $\pi(\phi, \delta^2|\mathbf{y})$, given by (2.34). We require the following normalising constant to obtain the probability density function:

$$\pi(\mathbf{y}) = \int_{\delta^2}\int_{\phi} \frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{b+\frac{1}{2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}\pi(\phi)\pi(\delta^2)d\phi d\delta^2. \tag{3.27}$$

This marginal distribution can be approximated using quadrature methods. The joint posterior density ϕ and δ^2 is then:

$$\pi(\phi, \delta^2|\mathbf{y}) = \frac{1}{\pi(\mathbf{y})}\left\{\frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{b+\frac{1}{2}\left[(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}\pi(\phi)\pi(\delta^2)\right\}. \tag{3.28}$$

The integrals with respect to the prediction region $\mathcal{X}_\mathcal{P}$, in (3.11) and (3.12), are approximated by a grid of prediction points, for example a regular grid or points chosen using Latin Hypercube sampling.

Uniform distribution for ϕ and δ^2 and Gauss-Legendre quadrature: Assuming both ϕ and δ^2 have uniform prior distributions (3.24):

$$\pi(\phi) = \begin{cases} \frac{1}{b_1-a_1}, & \text{for } a_1 \le \phi \le b_1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \pi(\delta^2) = \begin{cases} \frac{1}{b_2-a_2}, & \text{for } a_2 \le \delta^2 \le b_2 \\ 0, & \text{otherwise .} \end{cases}$$

We denote by $f_1(\phi, \delta^2, \mathbf{x}_p)$ the integrand in equation (3.11):

$$f_1(\phi, \delta^2, \mathbf{x}_p) = 1+\delta^2-\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}+(\mathbf{f}_p^\top-\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R})^{-1}(\mathbf{f}_p^\top-\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})^\top,$$

and with repeated application of formula (3.26) we have the approximation:

$$\Psi_1(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{2}\frac{1}{2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 f_1\left(\frac{b_1-a_1}{2}a_i^1+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_j^2+\frac{b_2+a_2}{2}, \mathbf{x}_p\right)d\mathbf{x}_p. \tag{3.29}$$

Here a_i^1 and w_i^1 are obtained from the Legendre polynomials for ϕ, and a_j^2 and w_j^2 are obtained from the Legendre polynomials for δ^2.

For Ψ_2 we first generate a random sample $\mathbf{y}_k$, $k = 1, \ldots, N$, from $\pi(\mathbf{y})$ via $\pi(\phi)$, $\pi(\delta^2)$ and $\pi(\mathbf{y}|\phi, \delta^2)$ and, for each $\mathbf{y}_k$, Gauss-Lagendre quadrature is applied to approximate the integrals over the prior distributions for ϕ and δ^2. Finally Monte Carlo integration is applied. Substituting (3.28) into (3.12), and denoting by $f_2(\phi, \delta^2, \mathbf{x}_p, \mathbf{y})$ the integrand in the equations (3.12):

$$f_2(\phi, \delta^2, \mathbf{x}_p, \mathbf{y}) = \frac{1}{\pi(\mathbf{y})}\frac{[(\mu^*-\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*))(\mu^*-\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*))^\top]|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}+\mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{[b+\frac{1}{2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)]^{a+\frac{n}{2}}}.$$

The posterior mean, $\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*)$, with respect to the posterior distribution of ϕ and δ^2 is approximated using quadrature

$$\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{4}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 \mu^*\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^2+\frac{b_2+a_2}{2}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p,$$

and with repeated application of formula (3.26) and Monte Carlo integration we have the approximation

$$\begin{aligned}\Psi_2(\xi) &\simeq \int_{\mathcal{X}_\mathcal{P}} \int \frac{1}{4}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_i^2 f_2\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^2+\frac{b_2+a_2}{2}, \mathbf{x}_p, \mathbf{y}\right)\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p \\ &\simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{4}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 f_2\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^2+\frac{b_2+a_2}{2}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p.\end{aligned} \tag{3.30}$$

Here a_i^1 and w_i^1 are again the nodes and weights obtained from the Legendre polynomials for ϕ, and a_j^2 and w_j^2 are obtained from the Legendre polynomials for δ^2.

Log-normal distribution for ϕ and Gauss-Hermite quadrature and uniform prior on δ^2 and Gauss-Legendre quadrature: Here, a log-normal prior distribution is assumed for ϕ, with

$$\pi(\phi) = \frac{1}{\phi\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\log(\phi)-\mu)^2}{2\sigma^2}\right\},$$

and a uniform prior distribution for δ^2

$$\pi(\delta^2) = \begin{cases}\frac{1}{b_2-a_2}, & \text{for } a_2 \le \delta^2 \le b_2 \\ 0, & \text{otherwise .}\end{cases}$$

We follow the same procedure as before but now apply both Gauss-Hermite quadrature (3.22) and Gauss-Legendre quadrature (3.26) to obtain

$$\Psi_1(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{\sqrt{\pi}}\frac{1}{2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 f_1\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \frac{b_2-a_2}{2}a_j^2+\frac{b_2+a_2}{2}, \mathbf{x}_p\right) d\mathbf{x}_p. \tag{3.31}$$

Now, we redefine a_i^1 and w_i^1 as the abscissae and weights, respectively, obtained from the Hermite polynomials for ϕ, and a_j^2 and w_j^2 to be obtained from the Legendre polynomials for δ^2.

For $\Psi_2(\xi)$, which is again more complicated to approximate, we apply Gauss-Hermite quadrature (3.22) for ϕ, Gauss-Legendre (3.26) for δ^2, and Monte Carlo integration

(3.18). We obtain

$$\Psi_2(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{\sqrt{\pi}}\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 f_2\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \frac{b_2-a_2}{2}a_j^2 + \frac{b_2+a_2}{2}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p,$$

and

$$\mathbb{E}_{\phi,\delta^2|\mathbf{y}}(\mu^*) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{\sqrt{\pi}}\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^2 \mu^*\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \frac{b_2-a_2}{2}a_j^2 + \frac{b_2+a_2}{2}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p, \tag{3.32}$$

where $\mu^*(\cdot)$ is given by (2.27) and is a function of ϕ and δ^2.

3.5.2 Continuous prior distribution for ϕ with fixed and known δ^2

When the noise-to-signal ratio δ^2 is known, the objective function $\Psi(\xi)$ is given by (3.10) and $\Psi_1(\xi)$ and $\Psi_2(\xi)$ are given by (3.14) and (3.15) respectively. The numerical evaluation of these integrals is again via quadrature, using

$$\Psi_1(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{2}\sum_{i=1}^{m_1} w_i^1 f_1\left(\frac{b_1-a_1}{2}a_i^1 + \frac{b_1+a_1}{2}, \mathbf{x}_p\right) d\mathbf{x}_p,$$

for $\phi \sim \text{Unif}(a_1, b_1)$, and

$$\Psi_1(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{\sqrt{\pi}}\frac{1}{2}\sum_{i=1}^{m_1} w_i^1 f_1\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \mathbf{x}_p\right) d\mathbf{x}_p, \tag{3.33}$$

for $\phi \sim \text{log-normal}(\mu, \sigma^2)$.

For $\Psi_2(\xi)$, the approximations again require Monte Carlo integration and quadrature

$$\Psi_2(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{m_1} w_i^1 f_2\left(\frac{b_1-a_1}{2}a_i + \frac{b_1+a_1}{2}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p,$$

for $\phi \sim \text{Unif}(a_1, b_1)$, and

$$\Psi_2(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{N}\frac{1}{\sqrt{\pi}}\sum_{k=1}^{N}\sum_{i=1}^{m_1} w_i^1 f_2\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \mathbf{x}_p, \mathbf{y}_k\right) d\mathbf{x}_p, \tag{3.34}$$

for $\phi \sim \text{log-normal}(\mu, \sigma^2)$.

Here a_i^1 , w_i^1 are the abscissae and weights from Legendre (uniform) or Hermite (log-normal) polynomials.

3.5.3 Choice of number of quadrature points

We choose the number of quadrature points, m_1 and m_2, by comparing Monte Carlo and quadrature methods as follows :

1. Approximate the prediction space $\mathcal{X}_{\mathcal{P}}$, by a 10×10 regular grid on $[-1, 1]^2$;
2. Generate 30 random designs, each with $n = 5$ points;
3. Generate a sample of size 10000 for ϕ from both the uniform and log-normal distributions;
4. Generate a sample of size 10000 for δ^2 from the uniform distribution (this step is omitted if δ^2 is know and fixed);
5. Evaluate the objective function $\Psi(\xi)$ using Monte Carlo integration (3.18), using samples from steps 3 and 4;
6. Evaluate $\Psi(\xi)$ with quadrature methods using a variety of different numbers of quadrature points.

We conclude that for $m_1 = 5$, the two methods of numerical evaluation (Monte Carlo and quadrature) give similar results with the difference between the two methods to be around 0.5%. In the rest of the thesis, numerical evaluation of the objective function is obtained using Gauss-Legendre and Gauss-Hermite methods with $m_1 = 5$ and $m_2 = 5$ points.

3.6 Algorithms for Finding Optimal Design

Minimising the objective functions (3.9) and (3.10) cannot be done algebraically and, as a result, we need to use fast algorithms to obtain optimal designs.

A fundamental problem is how to minimise the objective function using a computationally efficient algorithm. Generally, there are two kinds of algorithms: stochastic and greedy, and both seek the best solution through the following steps. We choose an initial solution for the optimisation problem, then the algorithm modifies this solution, the new solution is assessed, and these steps are repeated until an optimal or near-optimal solution is achieved.

An example of a stochastic algorithm is simulated annealing which was used by Zhu and Stein (2005), Zimmerman (2006) and Xia et al. (2006) to find optimal designs, and is also used in other areas of design of experiments, see Woods (2010).

An example of a greedy algorithm is an exchange algorithm. These algorithms add new designs points and remove existing points to improve the objective function. Exchange algorithms are classified, according to the way they add and delete points, into two categories: (i) those that choose points to add and delete sequentially, for example

Wynn's algorithm (Wynn, 1972), and (ii) those that choose points to add and delete simultaneously for example, Fedorov's algorithm (Fedorov, 1972), the modified Fedorov algorithm (Cook and Nachtsheim, 1980), and the k-exchange algorithm (Johnson and Nachtsheim, 1983). The coordinate exchange algorithm (Meyer and Nachtsheim, 1995) is a modification of the k-exchange algorithm and is effective for large designs, i.e. a large number of points and a large number of variables. The most commonly used algorithms in the design of experiments are the exchange algorithms because of their computational efficiency for large number of factors, their easy implementation for any design region and their adaptability for any design criterion.

An extensive review of exchange algorithms for the construction of exact designs was given by Meyer and Nachtsheim (1995). They indicated that Fedorov's algorithm is computationally expensive, whereas the k-exchange algorithm focuses only on k points and considers only single point exchanges. The approach that many small steps are better than large steps motivated the creation of the coordinate exchange algorithm. The idea behind the coordinate exchange algorithm is the "sub single point exchange", e.g. exchange of coordinates within each point. Meyer and Nachtsheim (1995) showed that this algorithm is faster than the k-exchange and still gives efficient designs. We employ coordinate exchange algorithms to find designs in this thesis.

An evaluation of the exchange algorithms used to construct spatial designs was given by Royle (2002). He modified the candidate set of points for possible exchanges so that the exchange algorithm is more efficient for large problems. He investigated two modifications: the "nearest neighbour" and "along coordinate axes". He compared these kinds of searches with the more traditional exchange algorithms, the original Fedorov and modified Fedorov, and found that the quality of the designs obtained is not affected by the search. He concluded that for "large" problems, such as the spatial design problem with many points, a combination of nearest neighbour and coordinate search may be preferable since it is less computationally expensive and has little impact on design quality.

In general exchange algorithms have gained more popularity than simulated annealing algorithms for finding an optimal design because of their simplicity of application. Moreover, the optimal designs obtained from exchange algorithms are as efficient as those from simulated annealing.

3.6.1 Coordinate exchange algorithm

The coordinate exchange algorithm proceeds element by element through the rows and columns of the design matrix. It is called coordinate exchange because, in each iteration, we consider possible changes for every element and each element is a coordinate of a point in the study region.

We modify the coordinate exchange algorithm of Meyer and Nachtsheim (1995) by

allowing a continuous, gradient based optimisation for each coordinate, rather than considering exchanges among a discrete set; see also Gotwalt et al. (2009).

At each step, we numerically optimise a single coordinate, keeping all other coordinates, both in that design point and in all other points, fixed. The algorithm can be described as follows:

1. Choose a random starting design, $\xi = (\mathbf{x}_1^0, \ldots, \mathbf{x}_n^0)$, where $\mathbf{x}_i^0 \in \mathcal{X} \subseteq \mathbb{R}^d$, i.e, $\mathbf{x}_i^0 = (x_{1i}^0, \ldots, x_{di}^0)$ with $x_{1i}^0, \ldots, x_{di}^0$ the coordinates of the ith point.

2. For each point, use a quasi-Newton algorithm to minimise the objective function with respect to each coordinate in turn, with all the other coordinates remaining fixed:

   Set $j = 1$:

   (a) Select the jth point, $\mathbf{x}_j = (x_{1j}, \ldots, x_{dj})$, and keep the remaining $(n-1)$ points fixed at their current values.

   (b) Set $i = 1$, find x_{ij} that minimises the objective function, keeping all other coordinates fixed.

   (c) Set $i = i + 1$, if $i \leq d$, repeat step (b). If $i = d + 1$, go to (d).

   (d) Set $j = j + 1$, if $j \leq n$, repeat (a) to (d).

3. When $j = n + 1$, set $j = 1$ and repeat steps (a) to (d). A new coordinate value replaces an existing value only if it decreases the value of the objective function. We repeat (a)-(d) until no decrease is obtained in the objective function for any new value of a coordinate.

3.7 Estimation

Bayesian optimal design for estimating trend parameters, $\boldsymbol{\beta}$, in a Gaussian process model can be found following a similar approach to that outlined in 3.3.1.

The main steps are as follows:

1. We find the expected loss with respect to the posterior distribution, $\pi(\boldsymbol{\beta}|\mathbf{y})$, for any decision (choice of estimator) $\boldsymbol{\gamma}(\mathbf{y}) \in \mathcal{G}$ and loss function $L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)$

$$\mathbb{E}[L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)|\mathbf{y}] = \int L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta}. \tag{3.35}$$

2. We minimise the expected loss with respect to the decision $\boldsymbol{\gamma}(\mathbf{y})$.

3. For any design $\xi \in \Xi$, where Ξ is the set of all possible designs, in order to obtain the objective function we average the minimum expected loss over the marginal

distribution of the data $\pi(\mathbf{y})$:

$$\Psi(\xi) = \int_{\mathcal{Y}} \min_{\boldsymbol{\gamma}(\mathbf{y})\in\mathcal{G}} \mathbb{E}[L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)|\mathbf{y}]\pi(\mathbf{y})d\mathbf{y}. \tag{3.36}$$

4. Then an optimal design, ξ^*, will be the one that minimises the objective function, i.e. $\xi^* = \arg\min_{\xi\in\Xi} \Psi(\xi)$.

Using the quadratic error loss function $L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi) = (\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{y}))^T(\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{y}))$, the expected loss is given by

$$\begin{aligned} \mathbb{E}[L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)|\mathbf{y}] &= \int L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} \\ &= \int (\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{y}))^T(\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{y}))\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta}. \end{aligned} \tag{3.37}$$

The decision $\boldsymbol{\gamma}(\mathbf{y}) \in \mathcal{G}$ which minimises the expected loss, i.e. $\min_{\boldsymbol{\gamma}(\mathbf{y})\in\mathcal{G}} \mathbb{E}[L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)|\mathbf{y}]$, is the posterior mean of $\boldsymbol{\beta}$:

$$\begin{aligned} 0 &= \frac{d}{d\boldsymbol{\gamma}(\mathbf{y})}\left[\int \boldsymbol{\beta}^T\boldsymbol{\beta}\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} - 2\boldsymbol{\gamma}(\mathbf{y})^T\int \boldsymbol{\beta}\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} + \boldsymbol{\gamma}(\mathbf{y})^T\boldsymbol{\gamma}(\mathbf{y})\int \pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta}\right] \\ &= \boldsymbol{\gamma}(\mathbf{y}) - \int \boldsymbol{\beta}\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} \\ \Rightarrow\ & \hat{\boldsymbol{\gamma}}(\mathbf{y}) = \int \boldsymbol{\beta}\pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} \\ \Rightarrow\ & \hat{\boldsymbol{\gamma}}(\mathbf{y}) = \mathbb{E}[\boldsymbol{\beta}|\mathbf{y}]. \end{aligned} \tag{3.38}$$

Therefore, the objective function is given by:

$$\begin{aligned} \Psi(\xi) &= \int_{\mathcal{Y}} \min_{\boldsymbol{\gamma}(\mathbf{y})\in\mathcal{G}} \mathbb{E}[L(\boldsymbol{\beta}, \boldsymbol{\gamma}(\mathbf{y}); \xi)|\mathbf{y}]\pi(\mathbf{y})d\mathbf{y} \\ &= \int_{\mathcal{Y}} \mathbb{E}[L(\boldsymbol{\beta}, \mathbb{E}(\boldsymbol{\beta}|\mathbf{y}); \xi)]\pi(\mathbf{y})d\mathbf{y} \\ &= \int_{\mathcal{Y}} \mathbb{E}\left[\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}^T\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}\right]\pi(\mathbf{y})d\boldsymbol{y} \\ &= \int_{\mathcal{Y}} \mathbb{E}\left[\mathrm{tr}\left[\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}^T\right]\right]\pi(\mathbf{y})d\mathbf{y} \\ &= \int_{\mathcal{Y}} \mathrm{tr}\left[\mathbb{E}\left[\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}\{\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta}|\mathbf{y})\}^T\right]\right]\pi(\mathbf{y})d\mathbf{y} \\ &= \int_{\mathcal{Y}} \mathrm{tr}\left[\mathrm{var}(\boldsymbol{\beta}|\mathbf{y})\right]\pi(\mathbf{y})d\mathbf{y}. \end{aligned} \tag{3.39}$$

Therefore, when the aim of the experiment is the estimation of the unknown trend parameters, the same decision theoretic approach can be applied. A design is Bayesian optimal when it minimises the trace of the variance covariance matrix of the posterior

distribution of the unknown regression coefficients averaged across the unknown data. We will not pursue designs for estimation further in this thesis.

3.8 Summary

Previous efforts in the literature to find designs for Gaussian process models for spatial data, spatio-temporal and computer experiments have generally assumed known covariance parameters values with the computational cost of a fully Bayesian approach proving prohibitive for design selection. Here we proposed a closed-form approximation to the objective function from a Bayesian decision theoretic approach. Our design criterion is derived as an approximation of the expected predictive variance and can be evaluated with reduced computational cost as it avoids making use of the Monte Carlo methods usually associated with Bayesian paradigms.

The aim of our designs is precise prediction of the response, and for this reason a quadratic loss function is chosen to represent the penalty for predicting a future observation. The objective function $\Psi(\xi)$ (3.7) is the average across the design region of the variance of the posterior distribution, for an individual prediction with the unknown data integrated out with respect to its marginal distribution. An alternative approach could be to consider the joint posterior distribution for the prediction of groups of points, taking into account possible correlation between the prediction points in a more coherent way.

In order to facilitate computations of the objective function, we made the assumption of conjugate prior distributions for the regression coefficients and the Gaussian process variance. However, in practice, other non-informative, prior distributions may be considered. In the following chapters, we will approximate non-informative prior distributions through change of the hyperparameters of the prior distributions for the regression coefficients. Also, in order to have a closed-form for the posterior densities conditional on the covariance parameters, we made a re-parametrisation using the noise-to-signal ratio as described in Chapter 2. Finally, the choice of the prior hyperparameters for the Gaussian process variance, σ^2, do not affect design selection as they affect the objective function through a multiplicative constant.

In Section 3.7, we briefly introduced the main steps for formulating the objective function when the aim of the experiment is the estimation of the trend parameters. Although in this thesis we do not consider the problem of finding optimal designs for estimation, optimal design for the regression coefficients may be applied in other contexts. For example, this approach can be applied in the area of experiments for estimating treatment effects in the presence of spatial trends in the units.

Chapter 4

Sensitivity Study

4.1 Introduction

This chapter explores the performance of Bayesian optimal designs for prediction obtained using a uniform or a log-normal prior distribution for the correlation parameter ϕ. Designs are found minimising (3.14), facilitated through incorporating the approximations described in Section 3.5.2. The main purpose of this chapter is to investigate the robustness of the choice of optimal design, and the sensitivity of the efficiency of an optimal design, to the values of hyperparameters of the prior distributions and also the function form of the mean and correlation and the size of the experiment.

In the study we assume the Gaussian process model (2.6) with unknown trend parameters, variance and decay parameter, and known noise-to-signal ratio. We find optimal designs when the aim is to predict over a 10×10 regular grid by minimising the closed form approximation $\Psi_1(\xi)$, (3.14), to the objective function $\Psi(\xi)$, (3.10). We perform the sensitivity study for designs in two dimensions, i.e. $d = 2$, and use Euclidean distance between two points in the study region $\mathcal{X} = [-1, 1]^2$, as would be suitable for spatial experiments (see Chapter 5).

The study uses a factorial design with five crossed factors and one nested factor, corresponding to features of the model, and experiments to assess simultaneously the effect of these factors on the performance of an optimal design. The crossed factors determine the number of runs, the mean function, the correlation function, the noise-to-signal ratio and the decay parameter. The nested factor is the hyperparameter, $\mathbf{R^{-1}}$, of the prior distribution of the regression coefficients; this is nested within the mean function. In total, 64 combinations of parameters are studied.

Factors	Levels	
	0	1
F_1	$n = 10$	$n = 30$
F_2	$M = \beta_0$	$M = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
F_3	$\nu = 0.5$	$\nu = 1.5$
F_4	$\delta^2 = 0$	$\delta^2 = 1$
F_5	$\phi \sim$ Unif$(0.1, 1)$	$\phi \sim$ log-normal$(-1.1, 1)$

Table 4.1: Five crossed factors together with their levels and coded values.

4.2 Factors and Study Design

In this section we give the general set up for our factorial study, and define and discuss the choice of the levels of each factor.

Five crossed factors are studied, each at two levels. The factors are listed in Table 4.1, together with their levels and coded values.

The first level of the mean function corresponds to a known regression functions including only the intercept term. The second level is a first order polynomial function of the variables. For the constant mean function ($F_2 = 0$), all the variation is assumed to be captured by the covariance structure of the Gaussian process model (2.6). Whereas, when we allow the mean function to be modelled as a linear function of the variables ($F_2 = 1$), variation is also described through the mean function.

Smoothness parameter $\nu = 0.5$ ($F_3 = 0$) corresponds to the exponential correlation function (Section 2.3), widely used in many applications in geostatistics. The second level ($F_3 = 1$) is chosen as $\nu = 1.5$, commonly used in both statistics and machine learning applications. Another common choice is $\nu = 2.5$; however this value is not considered here as this choice can result in very high correlation between two observations in the study region $\mathcal{X} = [-1, 1]^2$, and hence lead to problems inverting singular correlation matrices.

In this study, we consider the case of known and fixed noise-to-signal ratio δ^2. The first level ($F_4 = 0$) corresponds to a Gaussian process model (2.6) without a nugget effect, i.e. $\tau^2 = 0$. The second level ($F_4 = 1$) assumes the Gaussian process variance σ^2 is equal to the nugget τ^2.

Two prior distributions for ϕ ($F_5 = 0$ and $F_5 = 1$) are chosen to have the same prior mean. The chosen prior distributions for ϕ result in the correlation between observations at two points at the maximum Euclidean distance apart in this region, i.e. $d = \sqrt{8}$, to be between $[0.05 - 0.75]$ when $\nu = 0.5$. For the exponential correlation function, $\sqrt{8}$ is the effective range for the smaller value of ϕ; the distance beyond which the correlation between two observations is less than or equal to 0.05. For $\nu = 1.5$ the correlation for these prior values is between $[0.2 - 0.97]$. The effective range for $\nu = 1.5$ is $d = \sqrt{8}$ when $\phi = 1.7$. Note that for $\nu = 2.5$, the corresponding range of correlation

is $[0.4 - 0.99]$; this high correlation supports our decision not to choose this smoothness parameter.

Factor F_6 determines the prior variance of trend parameters, and is nested within factor F_2 (form of mean function).

$$F_6|(F_2 = 0) = \begin{cases} 0 & \Rightarrow R^{-1} = 0.25, \\ 1 & \Rightarrow R^{-1} = 4. \end{cases}$$

$$F_6|(F_2 = 1) = \begin{cases} 0 & \Rightarrow \mathbf{R^{-1}} = 0.25\mathbf{I}_3, \\ 1 & \Rightarrow \mathbf{R^{-1}} = 4\mathbf{I}_3. \end{cases}$$

The values of this factor are chosen to be either a scalar value 0.25 or a 3×3 matrix with diagonal elements 0.25. Otherwise, the scalar 4 or a 3×3 matrix with diagonal elements 4. The first level corresponds to a normal prior distribution for $\boldsymbol{\beta}$ with small prior variance, and hence more information about the trend parameters, and the second level indicates larger prior variance and hence a much less informative prior for the trend parameters.

The hyperparameters a and b for the inverse gamma prior distribution for σ^2 are kept constant for all the combinations of $F_1 - F_6$. These two parameters only affect the objective function through a multiplicative constant, see equation (3.14), and so do not affect the choice of a design or calculation of design efficiency. We set $a = 3$ and $b = 1$ to provide a prior distribution for σ^2 with finite variance.

4.3 Study Assessment

In this section, we assess the designs found for each combination of values for $F_1 - F_6$ in terms of quantitative changes in the location of design points, quantitative space-filling properties, and efficiencies under objective function $\Psi_1(\xi)$ (3.14).

We select Ψ-optimal designs for each of the 64 combinations of $F_1 - F_6$ as follows:

1. We generate 50 randomly selected starting designs from $\mathcal{X} = [-1, 1]^2$.

2. For each starting design, the coordinate exchange algorithm (Section 3.6.1) is used to find a design that minimises $\Psi_1(\xi)$.

3. From the 50 designs obtained by algorithmic search, we select the design that minimises $\Psi_1(\xi)$. (In the event of ties; a design is chosen at random from those with equal objective function values).

4.3.1 Robustness of design points and space-filling properties

We start by examining how the locations of the design points and the space-filling properties of the designs vary with the settings of $F_1 - F_6$.

We focus on the impact on the design of changing the values of F_5 and F_6, that is, the settings for ϕ and $\mathbf{R^{-1}}$.

For $n = 10$, the Ψ-optimal designs are presented in Figures 4.1-4.8 and for $n = 30$ in Figures A.1-A.8 in Appendix A.2. The figures display both the design points and contours of constant correlation between each point in the design region and the centre of the region, averaged across the prior values of ϕ. For all figures, the strength of the correlation is indicated by colour, where darker red colour indicates high correlation and lighter yellow indicates low correlation.

In addition to qualitative comparisons of the designs via plotting design points, we also assess the space-filling properties of the designs. The quantitative differences between designs are assessed in terms of inter-point distances. We choose to investigate the space-filling properties of our designs because space filling designs are a very popular alternative design choice for the Gaussian process models. Also the aim of our designs is prediction, and a space filling design covers the design region to ensure good predictions.

Table 4.2 shows the average inter-point distance between all points in each design. The average inter-point distance is defined as the $\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} d(\mathbf{x}_i, \mathbf{x}_j)$. Then we use analysis of variance (ANOVA) for the 64 combinations of the factor levels in Table 4.1 to decompose the variation in the inter-point distance. We do not perform a full

Combination							
F_3	F_4	F_5	F_6	$F_1 = 0$ $F_2 = 0$	$F_1 = 0$ $F_2 = 1$	$F_1 = 1$ $F_2 = 0$	$F_1 = 1$ $F_2 = 1$
0	0	0	0	1.288	1.321	1.224	1.201
0	0	1	0	1.276	1.282	1.238	1.186
0	0	0	1	1.266	1.355	1.224	1.186
0	0	1	1	1.276	1.357	1.220	1.235
1	0	0	0	1.317	1.322	1.246	1.250
1	0	1	0	1.279	1.294	1.229	1.218
1	0	0	1	1.318	1.336	1.245	1.251
1	0	1	1	1.282	1.311	1.233	1.231
0	1	0	0	1.255	1.442	1.281	1.326
0	1	1	0	1.240	1.493	1.267	1.332
0	1	0	1	1.295	1.594	1.270	1.361
0	1	1	1	1.284	1.635	1.288	1.361
1	1	0	0	1.371	1.674	1.389	1.491
1	1	1	0	1.317	1.633	1.334	1.495
1	1	0	1	1.442	1.764	1.407	1.538
1	1	1	1	1.377	1.758	1.347	1.510

Table 4.2: Average inter-point distances for 64 Ψ-optimal designs found for different combinations of settings of $F_1 - F_6$.

statistical analysis and we do not conduct any hypothesis testing. The ANOVA Table 4.3 the corresponding sum of squares.

From Table 4.3, factor F_1 explains 11% of the variability in the spread of design points; as the number of points increases, the average inter-point distance decreases, as would be expected. We focus the rest of our discussion on design with $n = 10$ points; our conclusions do not change greatly for $n = 30$.

We now discuss the impact of the different study factors on the designs based on Figures 4.1-4.8 and Tables 4.2 and 4.3. Each figure compares the effect of the values of $\mathbf{R}^{-1}$ and ϕ on the choice of the Ψ-optimal designs. To further demonstrate the impact of the range of the correlation on the choice of optimal design we provide contour plots. Contours display the average correlation between the centre point of the study region and each other point on a 100×100 grid, averaged across the prior distribution for ϕ.

Figure 4.1 corresponds to the case $F_1 = 0$ and $F_2 = 0$, and the first four rows of the Table 4.2. It allows us to assess the effect of the decay parameter and the prior hyper-parameter of the prior distribution for the trend parameters , F_5 and F_6, respectively, on the choice of Ψ-optimal design. For the four combinations $0000, 0010, 0001, 0011$, the design points are spread to cover the study region and also they have similar average inter-point distance, around 1.3, see Table 4.2.

When a nugget effect is included in the model, i.e. $F_4 = 1$, then the correlation decreases, indicated by the light yellow colour in Figures 4.3 and 4.4. The ranges of the average correlation for $\nu = 0.5$ are 0.2 and 0.18 for uniform and log-normal priors, respectively, and the corresponding values for $\nu = 1.5$ are 0.07 and 0.14. The plots of the eight combinations of factors in these figures indicates that the Ψ-optimal designs have similar space-filling designs to those obtained for $F_4 = 0$; the points tend

Factors	Sum of Squares	Percentage of variation
F_4	0.4021	35%
F_2	0.1823	16%
F_2F_4	0.1469	13%
F_1	0.1258	11%
F_3	0.0864	7%
F_1F_2	0.0601	5%
F_3F_4	0.0564	5%
F_6	0.0176	2%
F_4F_6	0.0084	1.5%
F_1F_6	0.0066	0.7%
F_3F_5	0.0056	0.6%
F_2F_6	0.0055	0.5%
$F_2F_3F_4$	0.0049	0.5%
F_2F_3	0.0034	0.4%

Table 4.3: Anova table: important factors and interactions with the corresponding sum of squares.

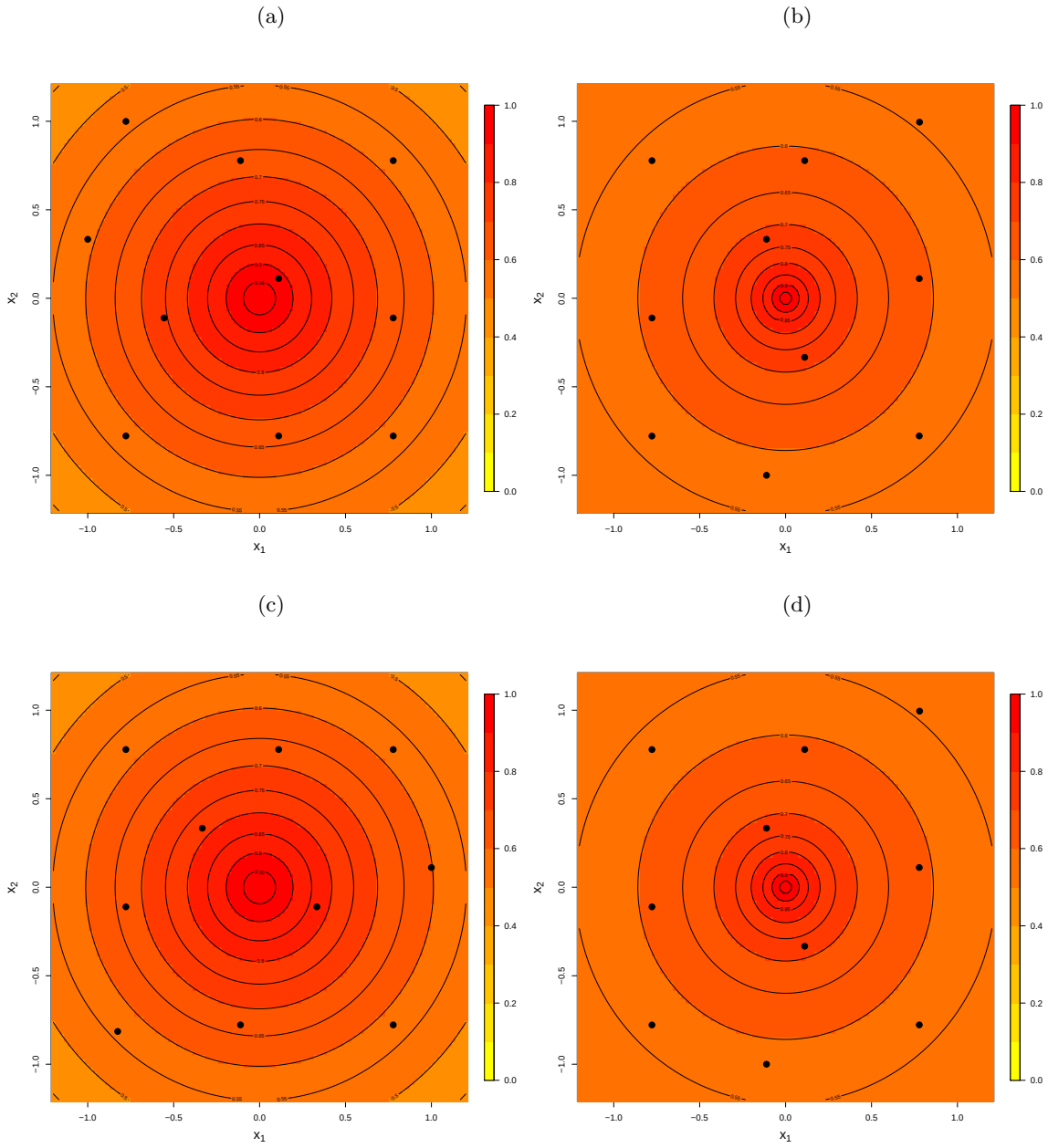


Figure 4.1: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), constant mean ($F_2 = 0$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

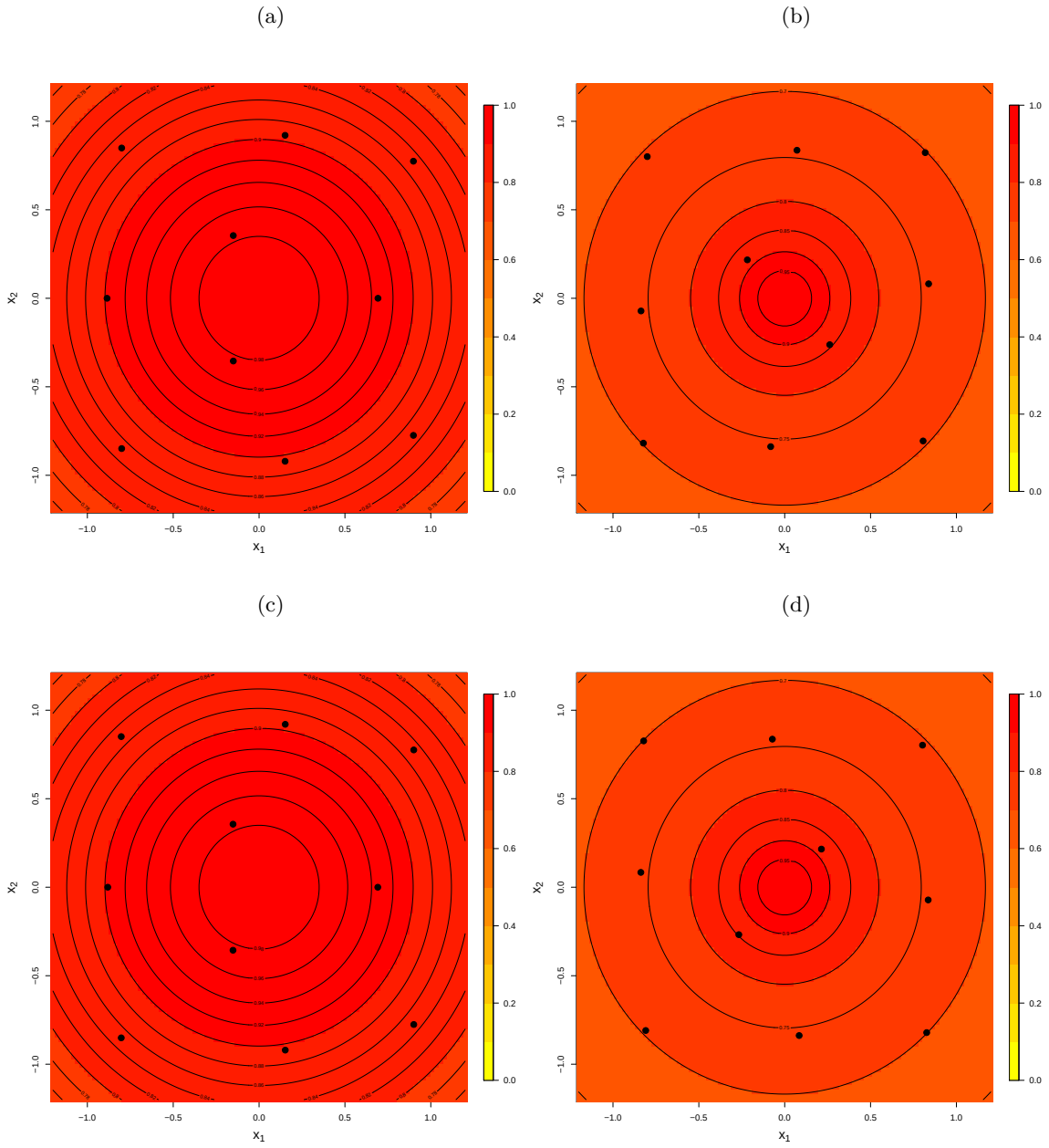


Figure 4.2: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), constant mean ($F_2 = 0$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

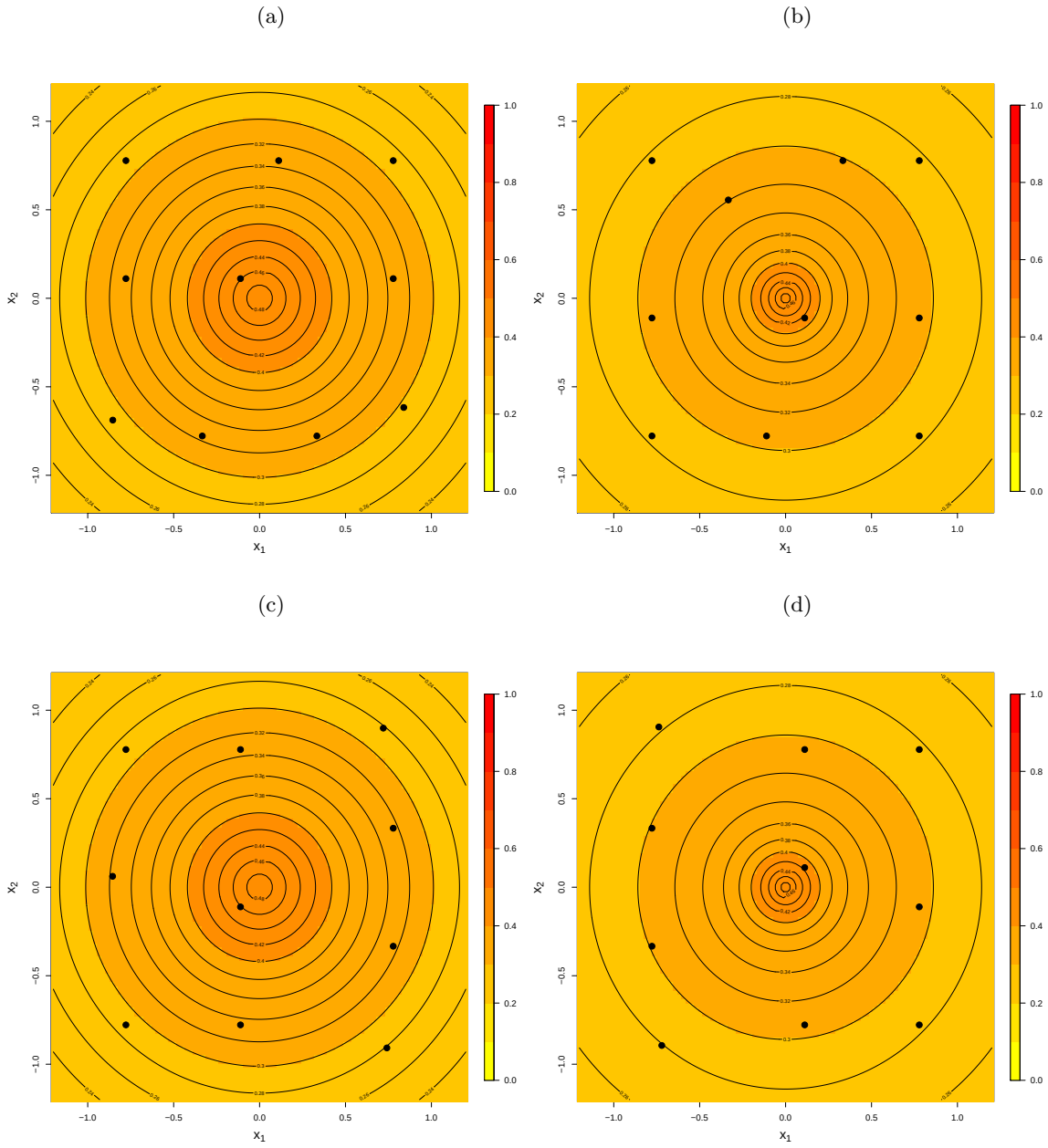


Figure 4.3: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), constant mean ($F_2 = 0$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

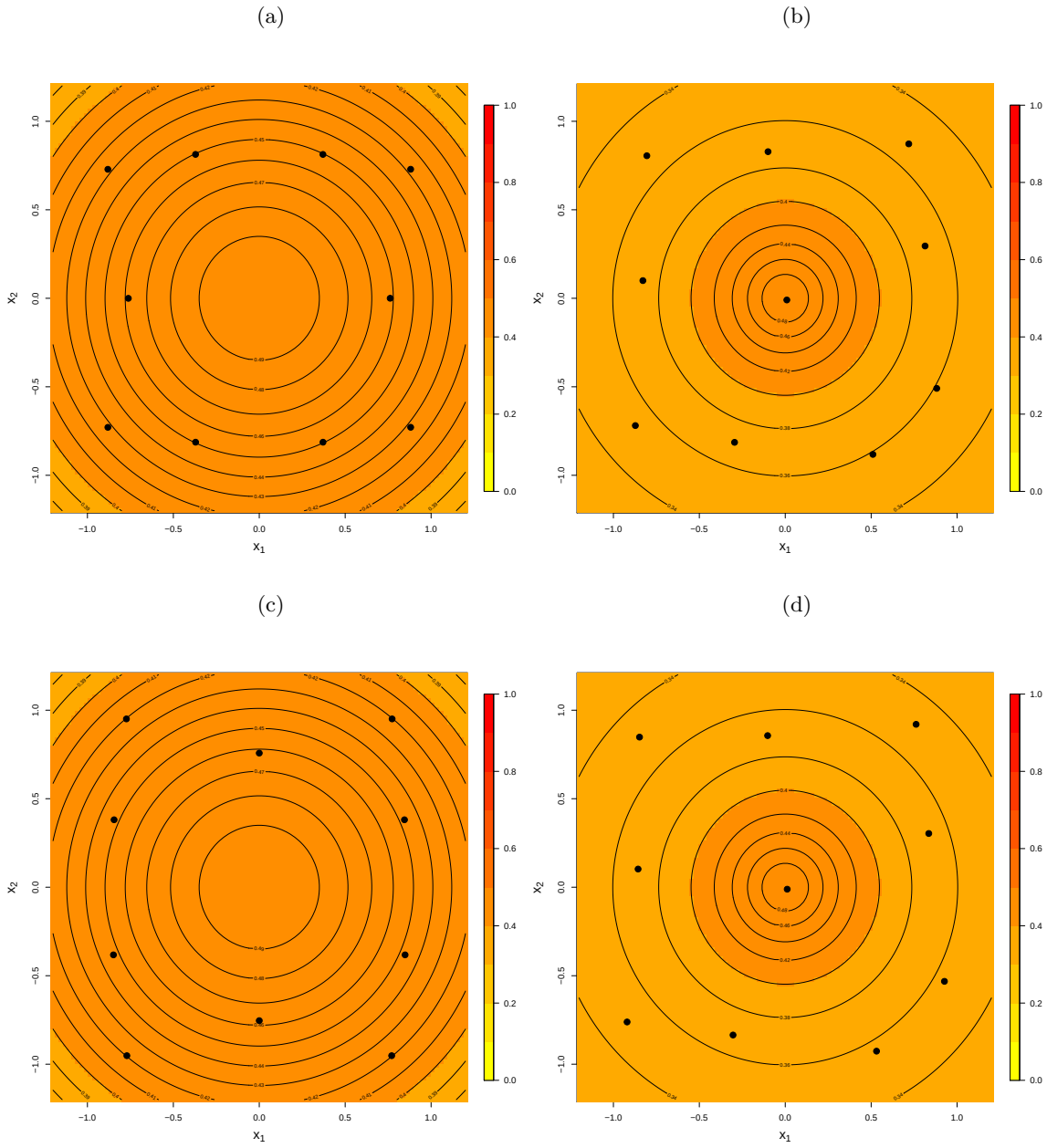


Figure 4.4: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), constant mean ($F_2 = 0$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

to cover the region with points allocated at the centre of the region. This is also supported from Table 4.2 where the average inter-point distances are similar to those for $F_4 = 0$. However, we can spot a variation in the case of $F_3 = 1$ and $F_4 = 1$ where the average inter-point distances are larger than the corresponding cases of $F_3 = 1$ and $F_4 = 0$, indicating the interaction between factors F_3 and F_4. The sum of squares of the interaction of F_3 and F_4 is 0.0564 and the interaction between these two factors has a small effect (5%) on the spread of the design points, see Table 4.3.

Figures 4.5-4.8 correspond to the 16 combinations when the mean function is the linear trend, i.e. $F_2 = 1$. We can condition on the values of all the other factors and compare the plots in Figure 4.1 and 4.5, Figure 4.2 and 4.6, Figure 4.3 and 4.7, Figure 4.4 and 4.8, respectively, to assess the effect of changing mean function.

From the figures and Table 4.3, we conclude that the designs are highly sensitive to the choice of factors F_2 and F_4, and their interaction. The interaction between F_2 and F_4 introduced 13% variability to the spread of the design points and in fact when $F_2 = 0$, F_4 has almost no effect. On the other hand, when $F_2 = 1$ the largest difference in the spread of the points is for $F_4 = 1$. Factor F_4 has the largest effect on the spread of the design points and introduces 35% variability to this response and F_2 16%. The importance of mean function is also indicated in Table 4.2 where the average inter-point distance of 16 combinations with $F_2 = 0$ increases compared to those for $F_2 = 1$ (see columns $F_1 = 0, F_2 = 0$ and $F_1 = 0, F_2 = 1$ in Table 4.2).

We conclude that the impact of changing the mean function depends on the choice of $\delta^2(F_4)$:

(i) When $\delta^2 = 0$ ($F_4 = 0$), the Ψ-optimal designs in Figures 4.5 and 4.6 show that the optimal designs are very similar to their corresponding designs for $F_2 = 0$, see Figures 4.1 and 4.2. All the designs spread out the points in the design region but with some points located at the centre. The average inter-point distance for designs with $F_2 = 1$ and $F_3 = 0$ varies from $1.282 - 1.357$, and we can notice that for $F_3 = 0$, changing F_6 results in higher average inter-point distances (Table 4.3) while F_6, $F_2F_3F_4$ and F_2F_3 have small effect on the spread of the designs, $(5\%, 0.5\%, 0.4\%)$.

(ii) When $\delta^2 = 1$ ($F_4 = 1$), the pattern changes. The designs now are strongly influenced by the choice of mean function. The designs spread out the points towards to the boundaries, influenced by the need to estimate the trend parameters, see Figures 4.7 and 4.8. The corresponding average inter-point distance in Table 4.2 varies from $1.442-$ 1.758, larger than for the case of $F_2 = 0$. Similarly to the case $F_2 = 0$, when $F_3 = 1$, the average inter-point distance increases compared to $F_3 = 0$, see Table 4.2.

The conclusions for $F_1 = 1$ are similar to those for $F_1 = 0$, see Figures A.1-A.8. In general, for constant mean function, $F_2 = 0$, the Ψ-optimal designs cover the region for both $F_4 = 0$ and $F_4 = 1$. For linear mean function, $F_2 = 1$, then the points spread out for $F_4 = 1$, and the Ψ-optimal designs contains repeated points.

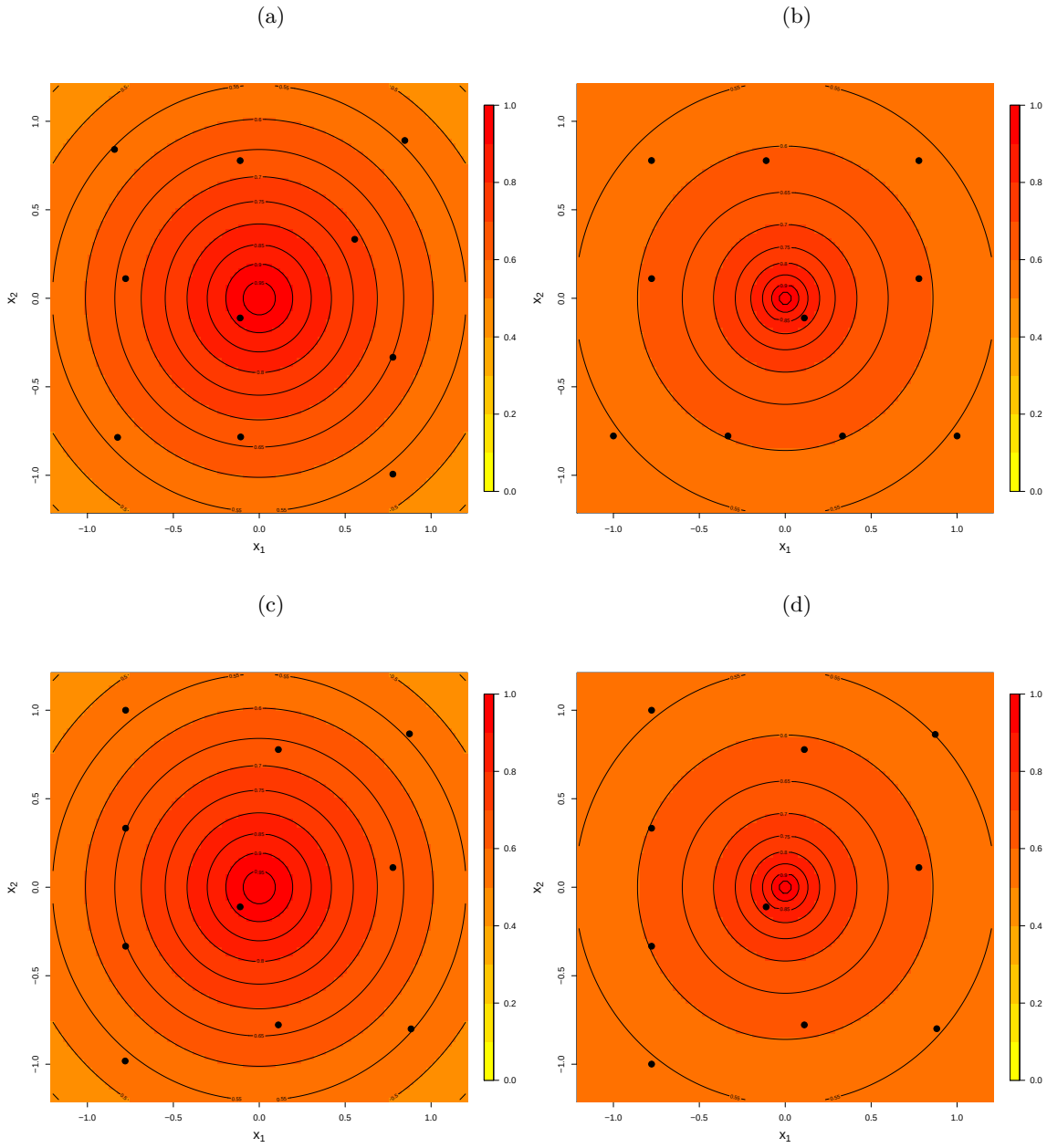


Figure 4.5: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), linear mean ($F_2 = 1$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

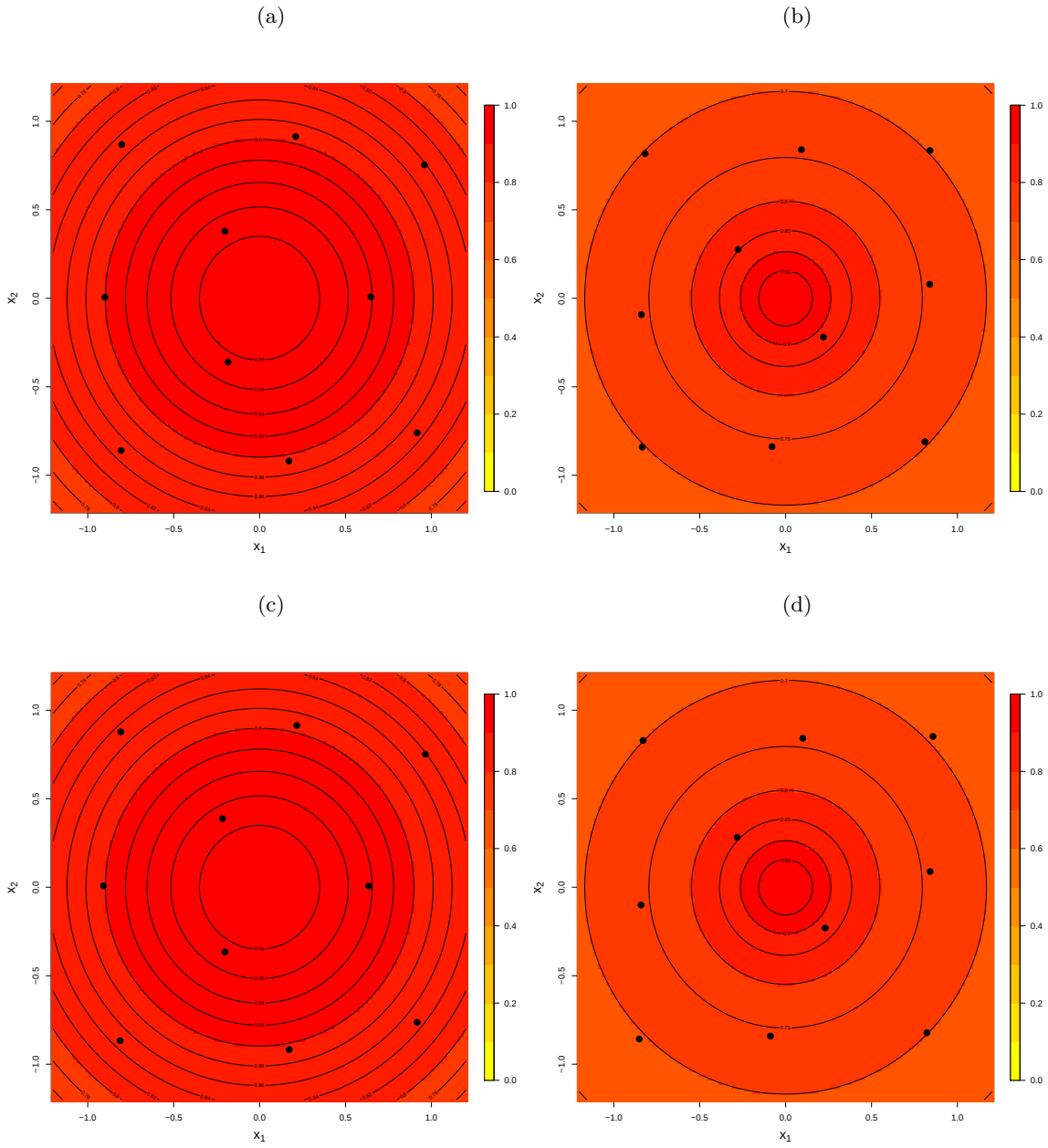


Figure 4.6: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), linear mean ($F_2 = 1$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

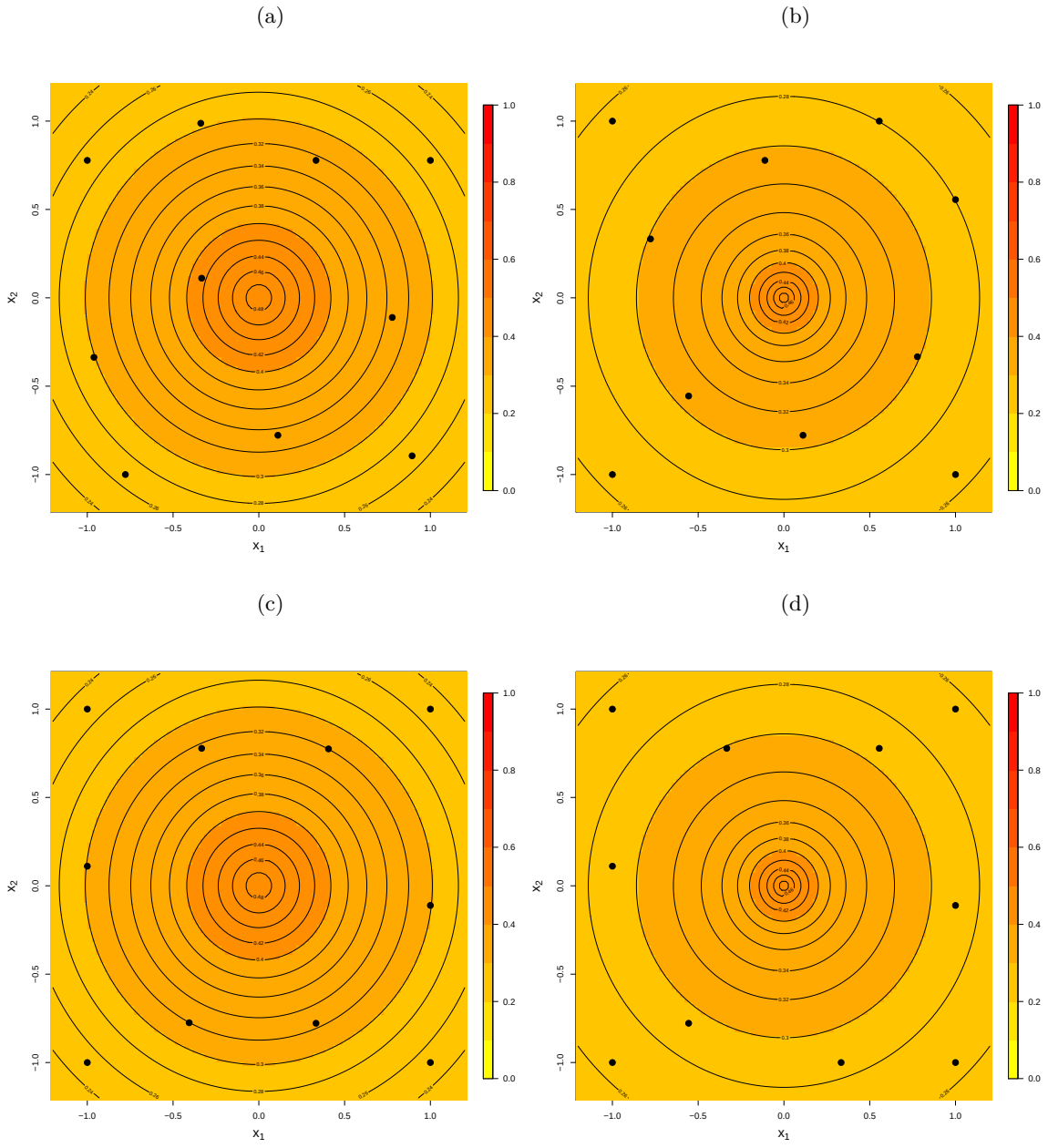


Figure 4.7: $\Psi-$optimal designs for $n = 10$ runs ($F_1 = 0$), linear mean ($F_2 = 1$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

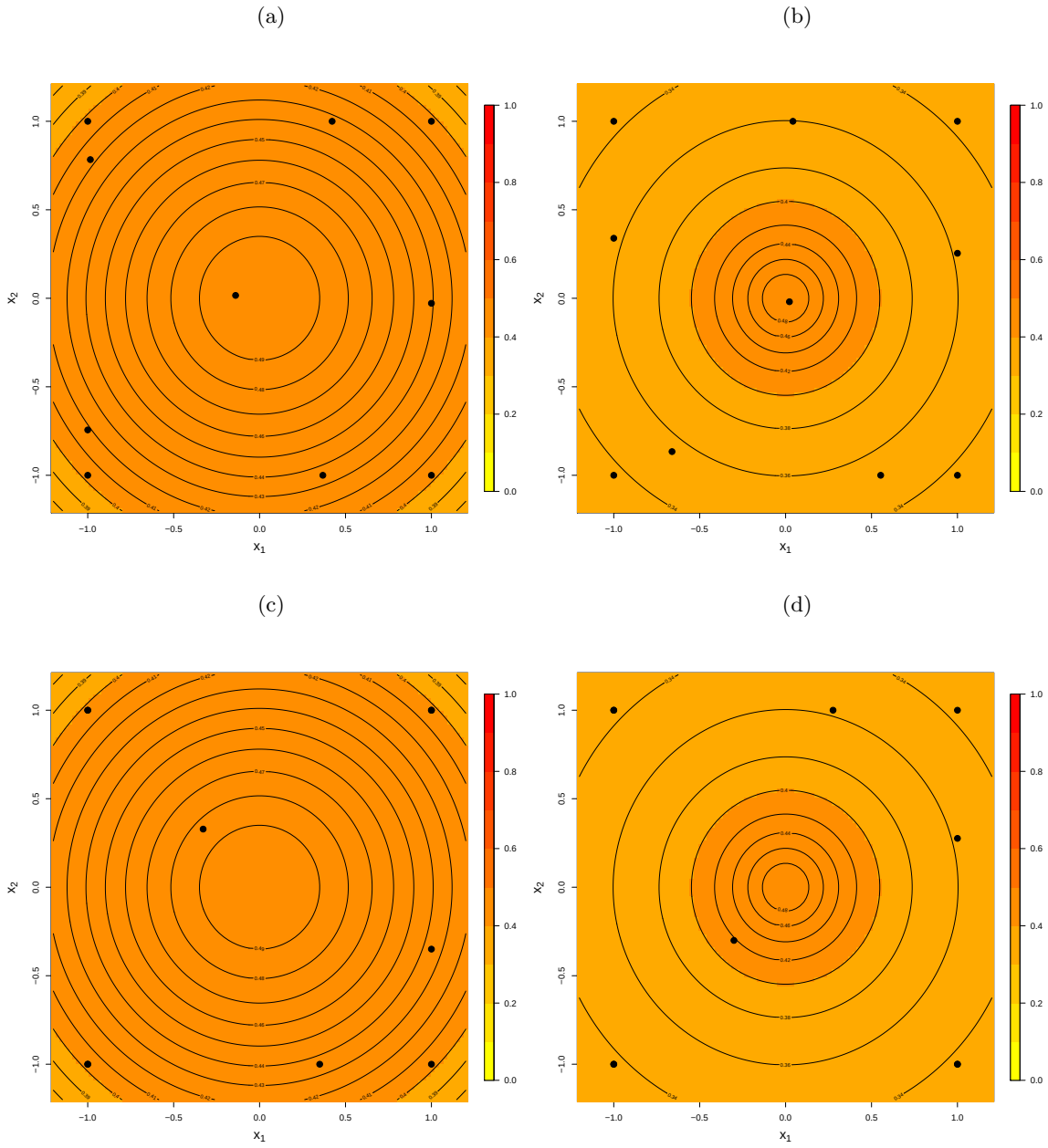


Figure 4.8: Ψ-optimal designs for $n = 10$ runs ($F_1 = 0$), linear mean ($F_2 = 1$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ. In plots (c) and (d) three points are repeated.

4.3.2 Robustness of design efficiency

The particular values of δ^2, ν, $\mathbf{R^{-1}}$ and ϕ assumed when finding an Ψ-optimal design may not be appropriate, and different choices may be made when collecting the data. For example, an optimal design found by assuming a uniform prior distribution for ϕ might be less efficient when, in fact, a log-normal prior distribution is assumed. We investigate this issue using the results for the 64 combinations of parameter values, which we split into four sets according to the number of runs (n, F_1) and the mean function (M, F_2).

Tables 4.4 and 4.5 give the relative efficiencies for 10 runs ($F_1 = 0$) constant mean ($F_2 = 0$) and linear mean ($F_2 = 1$) function, respectively; the corresponding tables for 30 runs ($F_1 = 1$) are Tables A.1 and A.2, and can be found in Appendix A.2.

The rows of each table correspond to the different Ψ-optimal designs found for each of the 16 combination of values of the remaining factors ($F_3 - F_6$), and we use the index $i_1 = 1, \ldots, 16$ to denote $\xi_{i_1}^{j_1 k_1}$ design across rows. Each column also corresponds to one of the 16 combinations of $F_3 - F_6$ and we use the index $i_2 = 1, \ldots, 16$ to denote $\xi_{i_2}^{j_2 k_2}$ design across columns.

Suppose that $\xi_{i_1}^{j_1 k_1}$ where ($i_1 = 1, \ldots, 16$, $j_1 = 0, 1$ and $k_1 = 0, 1$) is an optimal design for the i_1th combination of values F_3, F_4, F_5, F_6 and j_1, k_1 the levels of F_1 and F_2, respectively. We can calculate the efficiency of $\xi_{i_2}^{j_2 k_2}$ relative to $\xi_{i_1}^{j_1 k_1}$ with respect to $\Psi_1(\xi_{i_1}^{j_1 k_1})$ ($i_1 = 1, \ldots, 16; i_1 \neq i, j_1 = 0, 1$ and $k_1 = 0, 1$), see Chapter 3 for a definition of relative efficiency.

$$\text{Eff}(\xi_{i_1}^{j_1 k_1}, \xi_{i_2}^{j_2 k_2}) = \frac{\Psi_1(\xi_{i_1}^{j_1 k_1})}{\Psi_1(\xi_{i_2}^{j_2 k_2})} \quad i_1, i_2 = 1, \ldots, 16, \quad j_1, j_2 = 0, 1 \quad k_1, k_2 = 0, 1.$$

The column i_2 gives efficiencies of a design ξ_{i_2} under each of the other combinations $F_3 - F_6$, i.e. row is the numerator and column the denominator. Therefore, by looking along each row, we can assess the variability in performance of the Ψ-optimal designs for a given row of factor setting. Looking across each column allows us to assess the different performance of each design under the combinations of factors. Study of these tables supports the conclusions made in Section 4.3.1 based on the space-filling properties.

Table 4.4 shows the relative efficiencies for designs found for constant mean, $F_2 = 0$. The Ψ-optimal designs for 8 combinations of factors $0000 - 0111$, (factor $F_4 = 0$ for all 8 combinations, $i_1 = 1, \ldots, 16$ and $i_2 = 1, \ldots, 8$), are quite robust to the choice of factors $F_3 - F_6$ as the relative efficiencies of these designs are very high, $0.92 - 1$. Also, the robustness of these designs is supported by the corresponding interquartile range (IQR), last row of Table 4.4, which is very small, 0.01 for all 8 combinations.

When the nugget effect is included in the Gaussian process model (2.6), $F_4 = 1$ the

majority of the designs $1000-1111$ are robust, i.e. the IQR varying from $0.01-0.09$. That is, some combinations of factors results in lower efficiencies, i.e. $0.75-1$ (columns $i_2 = 13, 15$), especially when the efficiencies are found with respect the Ψ-optimal designs for $F_3 = 0$ and $F_4 = 0$. These results are in line with those obtained in Table 4.3 where the interaction between $F_3 = 0$ and $F_4 = 0$ has a small effect (5%).

Table 4.5 shows the relative efficiencies for designs found for linear mean function, $F_2 = 1$. Here, the robustness of the Ψ-optimal designs vary substantially mainly according to δ^2, (F_4), indicating the strong relationship between the mean function and noise-to-signal ratio, (see Table 4.3 for interaction between F_2 and F_4). When $F_4 = 0$, the Ψ-optimal designs for combinations $0000-0111$, ($i_1 = 1, \ldots, 16$ and $i_2 = 1, \ldots, 8$), are robust with efficiencies varying from $0.95-1$ and very small IQR from $0.01-0.05$. However, for $F_4 = 1$ the efficiencies of the designs $1000-1111$, ($i_1 = 1, \ldots, 16$ and $i_2 = 9, \ldots, 16$), are smaller, i.e. $0.54-0.9$ and the IQR is larger, i.e. $0.03-0.48$. Also the Ψ-optimal designs for combinations 1101 and 1111 have zero efficiencies with respect to the combinations $0100-0111$ as they have repeated points (see Figure 4.8 (c) and (d)). This results in singular correlation matrices when these designs are evaluated with respect the objective with $F_4 = 0$, i.e. no nugget, $\delta^2 = 0$.

In general for $n = 30$ ($F_1 = 1$) the results are very similar to those obtained for $n = 10$ ($F_1 = 1$). However, the efficiency of the Ψ-optimal designs is more sensitive to varying the values of the study factors, see Table A.1 and Table A.2. This is in line with ANOVA Table 4.3 where the number of runs is an important factor explaining 11% of the variation. There are specific combinations which result in designs which have high efficiency for $n = 10$ but for $n = 30$ have much lower efficiency, for example the rows for 0011 and 0101 in both tables. Also the number of designs that cannot be evaluated under some combinations increases as there are more designs with repeated points compared to $n = 10$ and cannot be evaluated under the objective function with $F_4 = 0$.

4.4 Summary

We have assessed how a Ψ-optimal design changes when features of the experiment, model and prior distributions vary. The aim was to perform a sensitivity study to investigate the robustness of the choice of a Ψ-optimal design when six factors are assumed.

We conclude that a Ψ-optimal design is sensitive to the choice of mean function, the degree and the range of the correlation, and if a nugget effect is included in the model. More specifically:

1. when the constant mean function is chosen for the Gaussian process model, the optimal designs are in general coverage designs (see Section 3.2.1) and the points

are quite uniformly spread over the study region, with no two points close together. This is true regardless of the choice of the correlation function, the nugget effect, the prior distribution of the decay parameter and the prior distribution of the trend parameter. However, a combination of $\delta^2 = 1$ ($F_4 = 1$) and $\nu = 1.5$ ($F_3 = 1$), which results in higher degree and narrower range of the correlation compare to $\nu = 0.5$ ($F_3 = 0$), results in design points moving outward to the boundaries of the study region. This indicates the sensitivity of the designs to the degree and the range of the correlation.

2. when a linear mean function is chosen, then the Ψ-optimal designs are highly influenced by the range and the degree of the correlation, and the presence of the nugget effect in the model. The designs compromise between coverage and spread of the design points. Particularly, when there is no nugget effect in the model the designs are quite similar to coverage designs. When the range of the correlation is smaller, this is controlled by the choice of ν ($F_3 = 1$ results to a correlation function with narrower range of correlation compare to $F_3 = 0$), the points generally move towards the boundaries, as with a constant mean, they still spread points over the study region. On the other hand, when the nugget effect is included in the model, $\delta^2 = 1$ ($F_4 = 1$), the designs change considerably and are more similar to spread designs. The points are concentrated at the corners and the boundaries of the region, with very few points at the centre, and also some points are repeated. The designs with linear mean function are affected by the need to estimate the regression coefficients.

Zimmerman (2006) performed a small sensitivity study to investigate how the mean function, the nugget effect and the degree of correlation affects the choice of an optimal design found by minimising the maximum prediction variance (called K-optimal designs). The main difference from our work is that he assumed known and fixed decay parameter, ϕ, in the exponential correlation function ($\nu = 0.5$). Initially, he investigated two choices of mean function, constant and linear, three values of $\delta^2 = 0, 0.25, 0.5$ and exponential correlation function with known and fixed values of $\phi = 0.62, 1.44, 4.54$. For these combinations Zimmerman (2006) concluded that the locations of design points for K-optimal designs were mainly affected by the choice of the mean function; constant mean resulted in points allocated to the study region quite uniformly regardless of the choice of the decay and the noise-to-signal ratio parameters, whereas use of the linear mean function gave rise to designs which concentrated the points near the boundaries of the region and, especially for $\delta^2 = 0.25, 0.5$, at the corners of the region.

Also Zimmerman (2006) found optimal designs by minimising the average prediction variance, i.e. $\Psi(\xi)$ (3.7) for known and fixed decay parameter as for the maximum prediction variance objective function and concluded that, although the location of the points were not exactly the same as those found by minimising the maximum prediction variance, their performance were very similar.

Moreover, in order to assess the uncertainty resulting from the estimation of the covariance parameter, Zimmerman (2006) proposed a criterion which minimises the maximum of the prediction variance with known covariance parameters plus a term which takes into account the covariance parameter estimation. The designs were still locally optimal, requiring a known value for ϕ and δ^2. This term was obtained by a first-order expansion of the prediction variance at the true value and the optimal designs are called EK-optimal designs. Zimmerman (2006) concluded that EK-optimal designs had points located in a similar fashion to K-optimal designs but with some additional clustering of points. In general, he concluded that designs depend on the strength of the correlation, the mean function employed and the size of the nugget.

Our approach differs from the Zimmerman (2006) study as it is a Bayesian approach and we consider ϕ unknown and investigate the sensitivity of the Ψ-optimal design with respect to its prior distribution and also we investigate the impact on the choice of correlation function, which is controlled through ν. When there is no nugget in the model and ϕ is unknown, the Ψ-optimal design is strongly influenced by the small values of ϕ in the support of its prior distribution (which correspond to high correlation). Especially when $\nu = 1.5$, our approach results in spreading out the points in the region. We agree with Zimmerman (2006) that the designs are strongly influenced by the strength, and we could also add the ranges of the correlation, the mean function and the nugget.

In the sensitivity study in this chapter, we only investigated the case of known and fixed δ^2. However, we found that this parameters plays a crucial role in the choice and the performance of the Ψ-optimal design, so in the next chapters we will investigate the case of unknown δ^2.

$\xi_{i_1}^{j_1 k_1}$ $i_1 = 1, \cdots 16$	Settings for factors $F_4F_3F_5F_6$ $\xi_{i_2}^{j_2 k_2}, i_2 = 1, \cdots 16$																IQR
	0000	0010	0001	0011	0100	0110	0101	0111	1000	1010	1001	1011	1100	1110	1101	1111	
0000	1	1.00	1.00	1.00	0.98	0.99	0.98	0.99	0.99	1.00	1.00	1.00	0.93	0.98	0.93	0.97	0.02
0010	1.00	1	1.00	1.00	0.98	0.99	0.98	0.99	0.99	1.00	1.00	1.00	0.93	0.97	0.93	0.97	0.02
0001	1.00	1.00	1	1.00	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.93	0.97	0.93	0.97	0.02
0011	1.00	1.00	1.00	1	0.98	0.99	0.98	0.99	0.99	1.00	1.00	1.00	0.93	0.97	0.93	0.97	0.02
0100	0.92	0.92	0.94	0.92	1	0.99	1.00	0.99	0.86	0.89	0.91	0.91	0.75	0.91	0.76	0.92	0.05
0110	0.96	0.96	0.98	0.96	0.99	1	0.99	1.00	0.94	0.95	0.96	0.96	0.85	0.96	0.84	0.95	0.03
0101	0.92	0.92	0.94	0.92	1.00	0.99	1	0.99	0.86	0.89	0.91	0.90	0.75	0.90	0.77	0.92	0.05
0111	0.96	0.96	0.98	0.96	0.99	1.00	0.99	1	0.94	0.95	0.96	0.96	0.85	0.95	0.84	0.95	0.03
1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.002
1010	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	1.00	1.00	0.002
1001	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	1.00	0.001
1011	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	0.001
1100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.001
1110	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.0004
1101	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	0.001
1111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	0.001
IQR	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.09	0.03	0.09	0.04	

Table 4.4: Relative efficiencies for $F_1 = 0, j_1, j_2 = 0$ and $F_2 = 0, k_1, k_2 = 0$ together with interquartile range (IQR).

$\xi_{i_1}^{j_1k_1}$ $i_1 = 1, \cdots 16$	Settings for factors $F_4F_3F_5F_6$ $\xi_{i_2}^{j_2k_2}, i_2 = 1, \cdots 16$																IQR
	0000	0010	0001	0011	0100	0110	0101	0111	1000	1010	1001	1011	1100	1110	1101	1111	
0000	1	1.00	1.00	1.00	0.98	0.99	0.98	0.99	0.98	0.95	0.90	0.86	0.80	0.86	0.67	0.68	0.13
0010	1.00	1	1.00	1.00	0.98	0.98	0.97	0.98	0.98	0.96	0.90	0.88	0.82	0.87	0.70	0.70	0.11
0001	0.99	0.99	1	1.00	0.98	0.99	0.98	0.99	0.98	0.96	0.90	0.87	0.81	0.86	0.68	0.69	0.12
0011	1.00	0.99	1.00	1	0.98	0.98	0.98	0.98	0.99	0.97	0.91	0.89	0.83	0.88	0.70	0.71	0.10
0100	0.92	0.93	0.94	0.94	1	0.99	1.00	0.99	0.90	0.79	0.64	0.57	0.54	0.62	0	0	0.35
0110	0.92	0.94	0.95	0.95	1.01	1	1.01	1.00	0.91	0.80	0.64	0.57	0.54	0.62	0	0	0.35
0101	0.92	0.93	0.95	0.94	1.00	0.99	1	0.99	0.90	0.79	0.64	0.57	0.54	0.62	0	0	0.35
0111	0.92	0.94	0.95	0.95	1.01	1.00	1.01	1	0.91	0.80	0.64	0.57	0.54	0.62	0	0	0.35
1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.99	0.99	0.99	0.97	0.97	0.004
1001	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.99	1.00	0.98	0.98	0.003
1010	0.98	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.99	1.00	1	1.00	0.99	1.00	0.98	0.98	0.02
1011	0.97	0.97	0.98	0.98	0.97	0.97	0.98	0.97	0.99	0.99	1.00	1	1.00	1.00	0.98	0.98	0.02
1100	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1	1.00	1.00	1.00	0.01
1110	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.01
1101	0.96	0.95	0.96	0.96	0.96	0.95	0.96	0.96	0.98	0.98	0.99	0.99	1.00	0.99	1	1.00	0.04
1111	0.96	0.95	0.96	0.97	0.96	0.95	0.96	0.96	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1	0.04
IQR	0.05	0.05	0.04	0.04	0.02	0.01	0.02	0.01	0.03	0.08	0.17	0.21	0.26	0.20	0.48	0.48	

Table 4.5: Relative efficiencies for $F_1 = 0, j_1, j_2 = 0$ and $F_2 = 1, k_1, k_2 = 1$ together with interquartile range (IQR).

Chapter 5

Designs for Spatial Processes

The objective of Chapter 5 is to provide a coherent and complete coverage of the decision theoretic approach for finding Bayesian optimal designs for continuous spatial data. Firstly we review the existing approaches for optimal designs for spatial data. Then, we introduce Bayesian optimal designs when the covariance parameters are known, followed by our new methodology for optimal design when all the parameters are unknown. A numerical study is given to validate our new closed-form design criterion and examples of spatial designs found by this criterion are demonstrated. In Chapter 4 we concluded that the Bayesian optimal designs are sensitive to the choice of the mean and correlation function, and also to the value of the noise-to-signal ratio. For this reason, in this chapter we further investigate the impact of the noise-to-signal ratio on the optimal design. Although the resulting designs are optimal for prediction at unobserved locations, we also perform a simulation study to assess design performance for inference about the unknown model parameters. Finally we compare our optimal designs with designs from the literature.

5.1 Introduction

Modern problems in climate science, such as pollution damage to the natural environment, have led to increased interest in the spatial design problem, i.e. the spatial configuration of the monitoring stations where the data are collected (Zidek and Zimmerman, 2010). In practice, the collected data are correlated and therefore we have to take account of the strength and structure of the correlation in developing optimal designs for setting up monitoring networks, see for example Section 1.1.1.

The geostatistical approach for these objectives is to assume there are n data points of the form $(\mathbf{x}_i, y(\mathbf{x}_i)), i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathcal{X}$, denotes the ith sampling location or point within the study region, $\mathcal{X} \subseteq \mathbb{R}^2$ and $y(\mathbf{x}_i)$ denotes an observation taken at $\mathbf{x}_i$ on a single, random realisation of a spatial stochastic process. In practice, observations $y(\mathbf{x}_i)$ are noisy versions of the used spatial stochastic process and are described by the

Gaussian process model (2.6).

In Gaussian process model (2.6), the large scale spatial variation, i.e the trend, is modelled through the mean function $\mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta}$ and the small scale spatial variation is modelled through the Gaussian process $Z(\mathbf{x}_i)$. The mean function can be a constant intercept term only or a polynomial function of the geographic coordinates, known as a trend surface model.

5.2 Literature Review

In this review we discuss both frequentist and Bayesian approaches for optimal design for spatial data collection.

The design problem is often approached using one of two main schools of thought: probability-based and model-based. The first approach is a model-free methodology which does not rely on any knowledge on the distribution of the response. Usually, this technique uses methods from sampling theory. By contrast, the aim of the second, model-based, approach is to draw inferences about the structure of the model, i.e. estimate the unknown parameters, and obtain predictions, using a highly efficient design. The majority of environmental monitoring networks rely on the model-based approach for estimation and prediction of characteristics of interest. A comprehensive review of the two main approaches for design is presented by Dobbie et al. (2008) and Zidek and Zimmerman (2010). Our research is focused on model-based designs achieved through optimal design methods. Throughout this thesis we adopt the model-based approach since statistical inference is the main goal for data collection, and hence the former approach will not be discussed any further.

The main aims of optimal design for spatial data described by the model (2.6), are to find optimal designs for (i) estimation of unknown parameters and (ii) spatial prediction at an unmonitored location. Designs for estimation can be separated into two categories: those for estimation of covariance parameters, σ^2, ϕ and τ^2, and those for estimation of the trend parameter, $\boldsymbol{\beta}$.

5.2.1 Designs for estimation of covariance parameters

It is generally agreed that the covariance structure in model (2.6) has an important role in the analysis of spatial data. Usually the values of the covariance parameters are unknown, and their estimates are influenced by the locations where the data are observed. Ad-hoc estimation of the covariance parameters by examining the variogram has been discussed by many authors, see Müller (2007) and references therein for a comprehensive review. The variogram is related to the correlation function and visual inspection of the empirical variogram can be used to suggest a possible parametric model for the variogram function. Then, using least squares or generalised least squares,

the parameters of the variogram can be estimated. Müller and Zimmerman (1999) considered the problem of finding optimal design for variogram estimation and proposed a modification of D-optimality as a design criterion. They found that their designs had points or locations close to each other and were different from random and regular designs.

More recently, a more rigorous model-based approach to covariance parameter estimation using inferential procedures has been developed, with Zhu and Stein (2005), Zimmerman (2006) and Xia et al. (2006) being the most recent contributions. We now give further details on each of these papers.

Zhu and Stein (2005) investigated optimal design for maximum likelihood (ML) estimation of the covariance parameters for model (2.6) with mean function equal to zero, when the inverse information matrix approximates the covariance of the ML estimators of the parameters. They used a Matérn correlation function given by (2.4). They were the first to address the problem of unknown covariance parameters, i.e. ϕ, ν, σ^2 and τ^2, in the objective function for the D-criterion and proposed using three types of designs:

- a locally optimal design where estimates or guesses for the covariance parameters are plugged into the objective function;
- a maximin design using the relative efficiency of a design that achieves the maximum value, over the set of all possible locally optimal designs, of the minimum efficiency over all possible parameter values;
- a pseudo-Bayesian design where a prior distribution is assigned to the parameters. They used, as utility function, the relative efficiency of the performance of a design with respect to the locally optimal design averaged over the prior distribution. This is equivalent to Bayesian D-optimality only if we consider the relative efficiency on the log scale, using the log ratio of the determinant.

In all cases, an optimal design was found using a simulated annealing algorithm and a discrete design region. The authors concluded that the covariance parameters are estimated more precisely by a locally optimal design rather by a regular or random design. However, the locally optimal designs are more sensitive to mis-specification of parameters values and the locally optimal design can change dramatically. The maximin and Bayesian designs gave more accurate estimates of the unknown parameters and outperformed the regular designs. The Bayesian designs were found to be more computationally expensive.

Zimmerman (2006) obtained similar results by maximising the determinant of the inverse of information matrix $M(\xi)$. He called this the CP-criterion to indicate the dependency on the covariance parameters. He again considered model (2.6) and compared the cases of constant mean ($k = 1$) and linear mean function (planar mean, $k = 3$). He used the exponential correlation function and found optimal designs for estimating the unknown parameters σ^2 and ϕ when the nugget is equal to $\tau^2 = 0$,

$0.25\sigma^2$ and $0.50\sigma^2$.

From this study, he concluded that when the aim of experiment is the estimation of the covariance parameters, an optimal design has a larger number of small distances between the points than a regular or random design. The designs also have large distances and an appropriate distribution of distances is achieved by regularly spaced clusters, lying mostly around the edge of the design space. He observed that including a nugget effect in the model changed the strength of spatial correlation and, as a result, the optimal design.

The difference between Zimmerman (2006) and Zhu and Stein (2005) is that former author considered the case of constant or linear mean function instead of setting the mean function zero. Also Zimmerman (2006) assumed an exponential correlation function whereas Zhu and Stein (2005) considered the more general Matérn with unknown smoothness parameter.

Xia et al. (2006) used likelihood-based methods to find optimal designs that allow both covariance and trend estimation. Their criterion was to maximise the trace of the information matrix which has block diagonal form corresponding to trend and covariance parameters. The authors considered algorithms such as sequential selection, block selection and stochastic search. They concluded that block selection gives different designs compared to sequential selection approach.

Entropy-based designs are very popular for estimating the unknown covariance parameters. Shannon (1948) introduced the entropy to measure the amount of available information. In the field of design of experiments Lindley (1956) used this measure to determine the information provided by the experiment. The better understanding of the process corresponds to the lower values of entropy. Entropy is defined as the gain of information between prior and posterior distribution. Maximum Entropy designs were suggested by Shewry and Wynn (1987), who showed that maximising the information about the unknown parameters is equivalent to maximising the information for prediction at unobserved locations. The maximum entropy designs correspond to the D-optimal designs for the case of the linear model with correlated errors. Also Sebastiani and Wynn (2000) showed that the experiment which maximises the entropy of the marginal entropy of the data will be most informative for the estimation of the parameters.

The theory of optimal design for the linear regression model with uncorrelated errors has influenced the development of model-based designs through the work of Müller (2007, Ch.5) and Spöck and Pliz (2010). The idea is to approximate the spatial model with a linear model having uncorrelated errors using a linear approximation to random fields such as the Karhunen-Loeve approximation and the polar spectral representation of an isotropic random field. Then classical experimental design theory is applied to this regression model. The problem is then to choose the design points for efficient estimation of the trend parameter $\boldsymbol{\beta}$ which now incorporates the correlation parameters.

This approach is difficult to use in realistic models such as model (2.6) because of the difficulty in finding an infinite expansion to approximate the model, see Zidek and Zimmerman (2010).

5.2.2 Designs for prediction at unmonitored sites

The ultimate objective for analysing spatial data is often prediction at unmonitored sites based on the data that are taken at monitored sites. The choice of an optimal design for prediction depends on the spatial covariance function and whether or not the covariance parameters are known or unknown. Several authors, for example, McBartney et al. (1981) and Su and Cambanis (1993), considered the case of known covariance parameters and model (2.6) with constant mean ($k = 1$). They concluded that an optimal design for prediction minimising either the average or the maximum prediction variance forms a fairly regular grid.

More recently, Zimmerman (2006) investigated the influence of the mean function on the choice of optimal designs for prediction where the covariance parameters (σ^2, ϕ, τ^2) are assumed known. He used nine combinations of values (ϕ, τ^2), namely, $(\phi_i, 0)$, no nugget, $(\phi_i, 0.25\sigma^2)$ and $(\phi_i, 0.5\sigma^2)$ for $\phi_1 = 0.62$, $\phi_2 = 1.44$ and $\phi_3 = 4.54$. He found designs for model (2.6) with constant mean ($k = 1$) and with planar mean ($k = 3$) by minimising the maximum prediction variance, and compared them graphically. He concluded that both the strength of the correlation and the presence or absence of the nugget effect have much less impact on the design points than the choice of mean function. He observed that the designs were uniformly dispersed over the study region for the constant mean model, whilst most of the points were located around the edge of the design for the linear mean function. These results are in line with our findings, see Section 4.4.

Generally, different designs are obtained when the aim is prediction and the covariance parameters are unknown, compared with designs for estimating the covariance parameters. Zhu and Stein (2006) and Zimmerman (2006) combined these two goals in a single design criterion with an objective function formed as a linear combination of the two separate functions, one that measures the quality of the design with respect to prediction with known covariance parameters, and one with respect to covariance parameter estimation. They considered model (2.6) with an isotropic correlation function and both Zhu and Stein (2006) and Zimmerman (2006) proposed criterion for prediction that takes into account the additional prediction uncertainty due to estimation of the unknown covariance parameters. Zhu and Stein (2006) considered the problem of redesigning an existing network, while Zimmerman (2006) the problem of adding a location to existing network. The best linear predictor is a function of the responses at the observed sites and of the unknown parameters. For this reason the unknown parameters are estimated using ML or Restricted maximum likelihood and the estimates are plugged into the best linear predictor. Then the prediction variance is adjusted to

incorporate the uncertainty due to the estimation of the unknown parameters. Both demonstrated the behaviour of their optimal designs for numerous simulations and real examples. All the designs found were locally optimal, for given values of covariance parameters.

Zimmerman (2006) also proposed a criterion to compromise between the optimal estimation of the unknown covariance parameters and optimal prediction. His criterion is the maximum value of the asymptotic approximate prediction error variance of the estimated best linear unbiased predictor (E-BLUP) over all sites in the design region, known as the empirical kriging EK-criterion. His simulation studies showed that an EK-optimal design is similar overall to an optimal design for prediction with known covariance parameters but contains a few small clusters enabling compromise between opposing objectives.

Zhu and Stein (2006) aimed to find optimal designs that minimise a combination of the kriging variance and the uncertainty in the estimated mean squared prediction error (MSPE) in order to incorporate the uncertainty due to unknown covariance parameters. Their criterion was a weighted linear combination of Zimmerman (2006) and the variance of the plug-in kriging variance estimator. The uncertainty in minimising MSPE was considered by approximating the variance of the plug-in kriging variance using a second order Taylor expansion of the kriging variance. This criterion is preferred if we are interested in estimating the MSPE of the best linear predictor more accurately. They also introduced an alternative criterion that is a weighted linear combination of the kriging variance and an approximation of the Kullback divergence of the plug-in conditional density from the conditional density evaluated at the covariance parameters. For a specific value of the weight, the two criteria are almost equivalent and for this value the criterion was called estimation adjusted, EA-criterion. Similarly to the work of Zhu and Stein (2005), Zhu and Stein (2006) used EA-criterion to find:

- locally optimal designs where the covariance parameters are assumed fixed or estimates for the covariance parameters are plugged into the objective function;
- maximin designs designs that maximise the minimum relative efficiency criterion for the EA-criterion. The relative efficiency of EA-criterion measures the relative performance of a design with respect to the locally optimal design.
- pseudo-Bayesian designs where they average the EA-criterion over the prior distribution of the unknown parameters. They did not follow a full Bayesian approach which makes inference from the posterior predictive distribution because they found it computationally infeasible to carry out a brute force Bayesian calculation in this context;

The resulting designs have some clustered points rather than being regularly spaced. Moreover, they concluded that finding minimax and Bayesian designs is computationally expensive and for large sample size they introduced a two-step algorithm to find

optimal design instead of using simulating annealing algorithm.

As we have mentioned entropy designs are very popular for designs for estimation of the unknown model parameters. However, based on the entropy of the posterior predictive distribution, entropy designs can be used for prediction problems. Fuentes et al. (2007) proposed a new entropy-based design criterion based on evaluating the posterior predictive entropy, which maximise the determinant of the covariance matrix between locations to be added to the design. They followed a Bayesian approach to incorporate the uncertainty about the covariance parameters. Fuentes et al. (2007) considered non-stationary correlation function, which is a mixture of a family of stationary process, and used simulated annealing to obtain an optimal subnetwork design. More discussion about entropy based design can be found in Zidek and Zimmerman (2010).

The development and rapid utilisation of computer algorithms and MCMC techniques have contributed to the introduction of the Bayesian spatial design in recent years. The Bayesian method for spatial data modelling was firstly introduced by Kitanidis (1986) who examined the effect of parameter uncertainty in a Bayesian framework and used the posterior distribution to gain an estimate for the unknown parameters.

Diggle and Lophaven (2006) investigated Bayesian optimal designs for two cases. The first one concerned how to add or remove locations from an existing network by minimising the average prediction variance, known as "the retrospective design problem". The second is how to design before any data are available by minimising the expectation of the average prediction variance with respect to the marginal distribution of the data, known as " prospective design problem". They considered model (2.6) with constant mean, i.e. $k = 1$, and exponential correlation function.

For the retrospective design problem, they found Bayesian designs with a diffuse prior distribution for $\boldsymbol{\beta}, \sigma^2$, a uniform prior distribution for ϕ, and either a known value for $\delta^2 = \tau^2/\sigma^2$, the noise-to-signal ratio $\delta^2 = 0, 0.3$ and 0.6 or unknown δ^2 with uniform prior distribution. They also compared these designs with locally optimal designs where all the parameter values are assumed known. These designs had points that were well separated compared with the Bayesian designs which had some close pairs of points. The Bayesian optimal designs changed according to the value of δ^2 and whether or not the ratio δ^2 was considered known or unknown. They also compared the posterior predictive variance for the nine different optimal designs evaluated under the Bayesian criterion and found that the Bayesian designs to be 5 to 10 times better than the locally optimal design whether δ^2 is known or not.

For the prospective design problem they did not find Bayesian optimal designs but they compared the performance of a regular lattice with a lattice plus close pairs design and a lattice plus infill designs which are designs with irregularly spaced locations. They evaluated the design criterion for each one of the three designs by assuming diffuse prior distribution for $\boldsymbol{\beta}, \sigma^2$ and uniform prior distributions for ϕ and δ^2. They concluded that a lattice plus close pairs design results in lower values of the design criterion and the

lattice plus infill design is slightly better than the regular lattice see Section 5.6.

In general we have seen that if we assume that the covariance parameters are known the optimal design for prediction is a design with more regular spacing. However, when the covariance parameters are assumed unknown then the optimal designs contain clusters of points to incorporate the estimation of the unknown parameters.

Our approach for optimal designs for spatial data is Bayesian and we concentrate on the prediction problem but we take into account the uncertainty due to unknown model parameters. Our proposed approach is different from the approaches in the existing literature because we follow a decision theoretic approach which is natural in Bayesian approach. In contrast with Diggle and Lophaven (2006), we assume that we do not have any data available before the experiment and we find designs that minimise the average prediction variance.

5.3 Optimal Design With Known Covariance Parameters

In this section, we apply the Bayesian decision theoretic approach described in Chapter 3 when the aim is to find the optimal designs to maximise predictive accuracy. Here we consider the simplest case where the covariance parameters, ϕ and δ^2 are known. The objective function to be minimised is given by (3.7) and as we have mentioned this integral is tractable and the objective function can be evaluated analytically.

The posterior distribution of a future observation $\pi(y(\mathbf{x}_p)|\mathbf{y})$ is a t-distribution given by (2.30). Therefore if we assume that ϕ and δ^2 are known the second part of the objective function (3.9) vanishes, and the objective function is:

$$\begin{aligned}
\Psi(\xi) &= \int_{\mathcal{X}_\mathcal{P}} \Sigma^* \int_{\mathcal{Y}} \frac{b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top [\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)}{2a + n - 2} \pi(\mathbf{y}) d\mathbf{y} d\mathbf{x}_p \\
&= \int_{\mathcal{X}_\mathcal{P}} \Sigma^* \frac{2b + \frac{2bn}{2a-2}}{2a + n - 2} d\mathbf{x}_p \\
&= \frac{b}{a-1} \int_{\mathcal{X}_\mathcal{P}} \{1 + \delta^2 - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\omega} + \\
&\qquad (\mathbf{f}_p^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R})^{-1}(\mathbf{f}_p^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})^\top \} d\mathbf{x}_p. \qquad (5.1)
\end{aligned}$$

The integral with respect to the unknown data is calculated using the quadratic form (3.13), where $\boldsymbol{\mu}$ and K are the mean vector and variance-covariance matrix of $\boldsymbol{\varepsilon}$ respectively. Here we apply $\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)$, $\Lambda = [\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}$, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{\varepsilon}) = 0$ and $K = \frac{2b}{2a-2}[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]$.

To illustrate, we assume the Gaussian process model (2.6) with mean function which is taken to be a linear function of the geographic coordinates only and the exponential correlation function

$$\rho(d_{ij}; \phi) = \exp^{-\phi d_{ij}},$$

where d_{ij} is the Euclidean distance between two sampling locations $\mathbf{x}_i, \mathbf{x}_j \in [-1,1]^2$. The choice of optimal design is affected by the strength of the correlation between two observations at their corresponding sampling locations. Therefore the parameters which play a crucial rule here are the decay parameter ϕ and the fixed noise-to-signal ratio δ^2.

Initially, 50 random starting designs each having 10 sampling locations in the square study region $\mathcal{X} = [-1,1]^2$ are generated, i.e the design points $\mathbf{x}_i : (i = 1, \ldots, 10)$ form an independent random sample from a uniform distribution on $\mathcal{X}$. The prior distributions for $\boldsymbol{\beta}|\sigma^2$ and σ^2 are chosen to be $N(\mathbf{0}, \sigma^2\mathbf{I})$ and $IG(3,1)$, since we assume conjugate prior distributions for evaluating the objective function (5.1) and also the prior hyperparameters are chosen in order to have finite variance for the inverse gamma. We assume four values for $\phi = 0.1, 1, 10$ and 100 and three values for $\delta^2 = 0, 0.5, 1$. We choose the smallest value of ϕ to be 0.1 because any smaller value makes the correlation matrix, $\mathbf{C}(\phi)$, numerically singular. We choose $\phi = 100$ as the largest value because any larger value makes the correlation matrix almost equal to the identity matrix, so we have almost uncorrelated observations.

In order to find the optimal designs, we employ the coordinate exchange algorithm (Subsection 3.6.1) and from the 50 random starting designs we choose the design with the minimum value of $\Psi(\xi)$ (5.1). We seek a design to predict at a regular 10×10 grid of points, i.e. $|\mathcal{X}_P| = 100$, and the objective function (5.1) is averaged across these points. We investigate how choices for ϕ and δ^2 affect the design by finding optimal designs for each of the 12 combinations of values ϕ and δ^2.

(i) $\delta^2 = 0$. There is no nugget effect in the model, i.e. $\tau^2 = 0$, and the variance-covariance matrix is equal to $\mathbf{C}(\phi)$. Then an optimal design depends only on the decay parameter which describes how the correlation decreases. The correlation between a corner point and the centre of the region is $0.86, 0.25, 10^{-7}$ and $\simeq 0$ for $\phi = 0.1, 1, 10$ and 100 respectively. When ϕ is small, the observations become more highly correlated. Figure 5.1 shows optimal designs for $\delta^2 = 0$. The two plots in the first row are for small values of ϕ which correspond to strong correlation between the observations. It can be seen that, for these small values of ϕ, the design points are scattered throughout the study region with no two points close together.

In contrast, for larger values of ϕ all the design points are concentrated at the periphery of the study region $\mathcal{X}$, for example as in Figure 5.1(d). For these values of ϕ, the data tend to be less correlated. An optimal design for prediction takes into account the estimation of the unknown coefficients in the trend parameter and hence is strongly influenced by the linear trend. The optimal design for uncorrelated data would only include the four corner points.

(ii) $\delta^2 = 0.5$ and $\delta^2 = 1$. Figures 5.2 and 5.3 show optimal designs for $\delta^2 = 0.5$ and $\delta^2 = 1$, respectively. For these cases the correlation between a corner point and the centre point of the region is 0.57 for $\phi = 0.1$ and $\delta^2 = 0.5$, and 0.43 for $\phi = 0.1$ and

$\delta^2 = 1$. When $\phi = 1$ the corresponding correlations drops to 0.16 and 0.12 respectively.

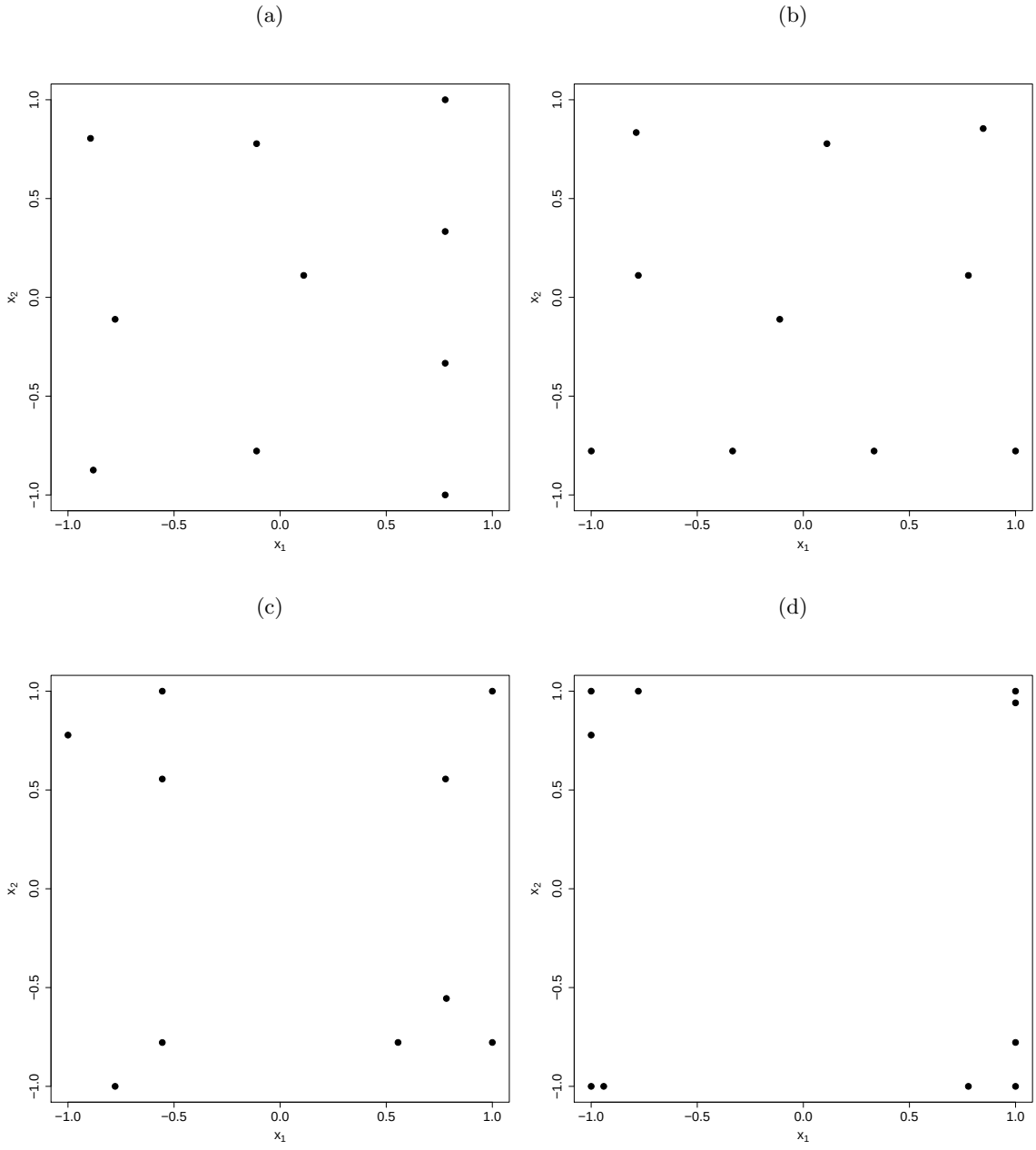


Figure 5.1: Ψ-optimal designs for prediction, found by minimising (5.1), when $\delta^2 = 0$ and (a) $\phi = 0.1$, (b) $\phi = 1$, (c) $\phi = 10$, and (d) $\phi = 100$.

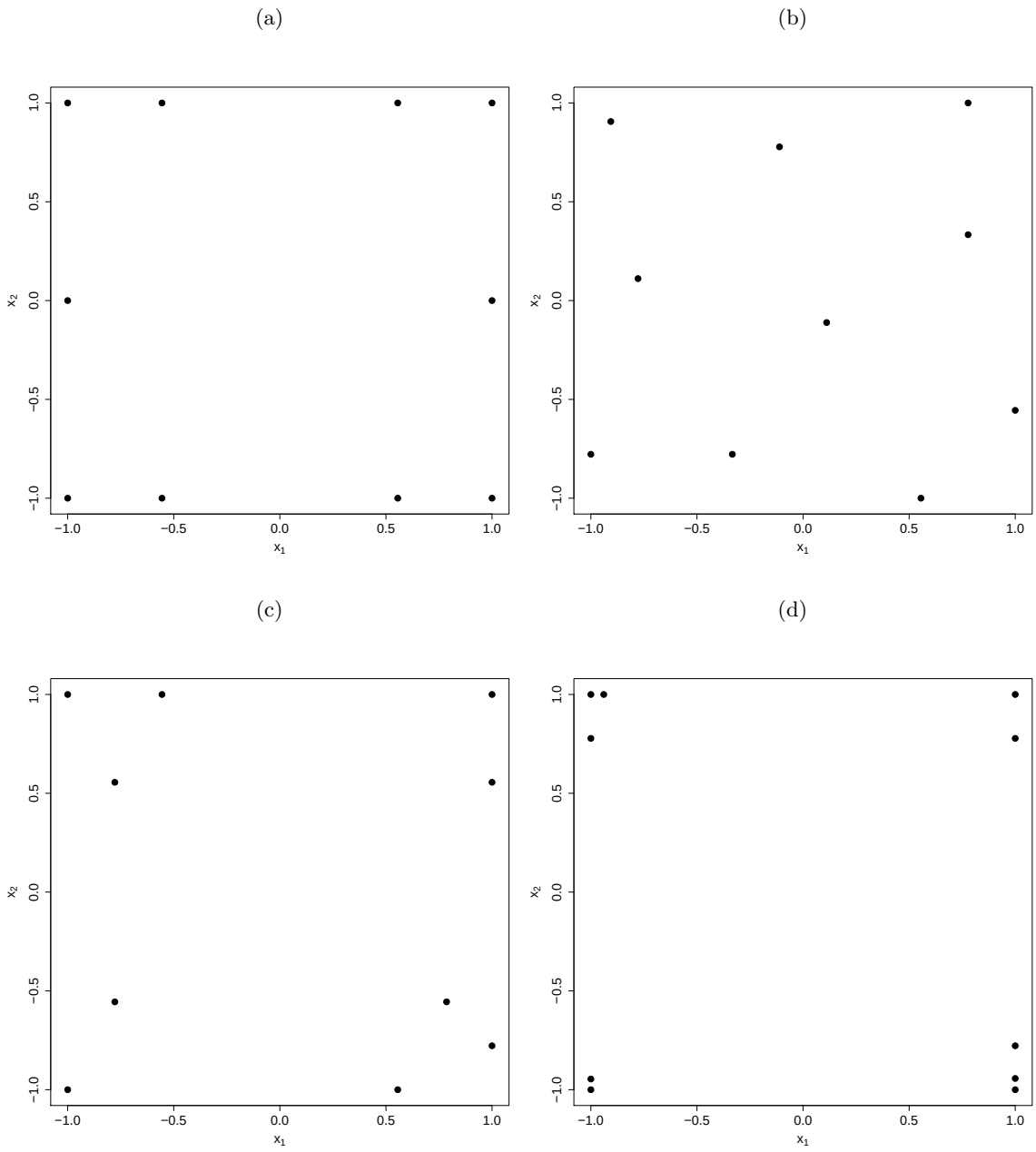


Figure 5.2: Ψ-optimal designs for prediction, found by minimising (5.1), when $\delta^2 = 0.5$ and (a) $\phi = 0.1$, (b) $\phi = 1$, (c) $\phi = 10$, and (d) $\phi = 100$.

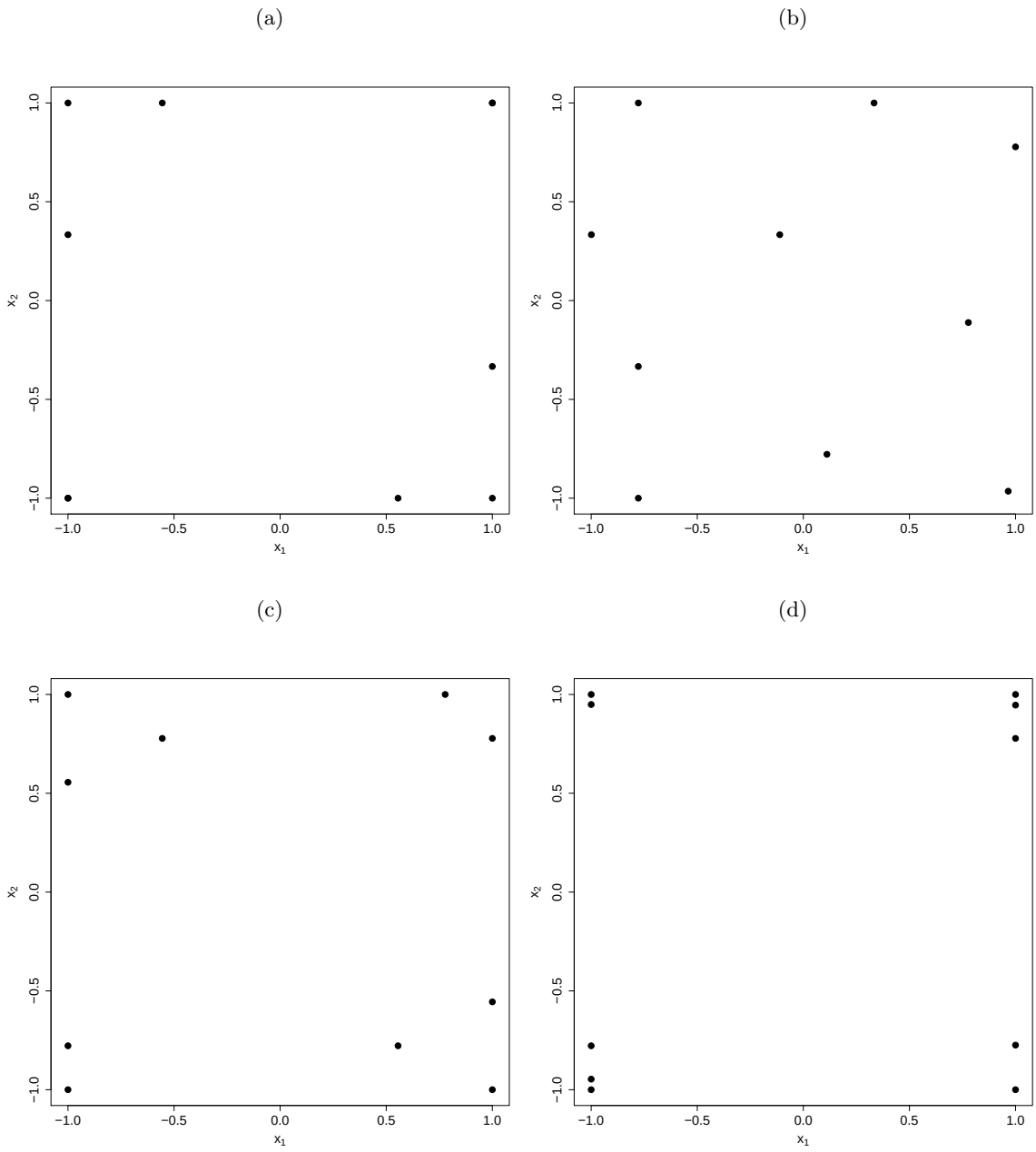


Figure 5.3: Ψ-optimal designs for prediction, found by minimising (5.1), when $\delta^2 = 1$ and (a) $\phi = 0.1$, (b) $\phi = 1$, (c) $\phi = 10$ and (d) $\phi = 100$.

For each combination of ϕ and δ^2, the designs are similar for the two values of δ^2. It can be observed from the plots that, when the value of ϕ is very small, the optimal design is mainly influenced by the noise-to-signal ratio, δ^2. Hence, for very small values of ϕ and a non-zero value of δ^2, the correlation is a constant function of the distance and any point in the region $[-1, 1]^2$ has similar values of correlation. For this reason the sampling locations move towards the boundaries. Similar patterns can be seen for very large value of the decay parameter, e.g $\phi = 100$, and non-zero δ^2, where again the distance between the design points does not affect the correlation. Figures 5.2 (d) and 5.3 (d), corresponding to $\phi = 100$ and $\delta^2 = 0.5$ and $\delta^2 = 1$ respectively, show that the

points tend to concentrate at the four corners.

These results are in line with Zimmerman (2006). He considered this kind of problem from the frequentist point of view and indicated that as correlation decreases the points move to the corners for the case of a linear trend.

5.4 Optimal Design With Unknown Covariance Parameters

The optimal designs discussed in the preceding section require the assumed covariance parameters are known and fixed. However, in practice, we will not know the values of the decay, ϕ, and noise-to-signal, δ^2, parameters and more realistically we need to allow for uncertainty in the values of all of the model parameters.

In this section we develop optimal designs for spatial data when the objective of the design is efficient prediction assuming that the values of the covariance parameters are unknown. As described in Chapter 3, in this case the optimality criteria become more complicated as the posterior and predictive distributions cannot be expressed in closed-form and subsequently objective function $\Psi(\xi)$ (3.7) cannot be evaluated analytically.

5.4.1 Assessment for closed-form approximation for spatial experiments

Here, a numerical study is presented to explore the relationship between $\Psi(\xi)$ and $\Psi_1(\xi)$ and to study how the choice of the parameters in the experiment and model affects the accuracy of the approximation from Conjecture 3.1. We perform a factorial study similar to that in Chapter 4 but here we consider five crossed factors and two nested factors, each with either two or three levels. For each combination of factor levels, we evaluate the objective function $\Psi(\xi)$.

There are five crossed factors given in Table 5.1 together with their levels and coded values.

	Levels		
Factors	0	1	2
F_1: Number or runs	$n = 5$	$n = 10$	
F_2: Mean function	$M = \beta_0$	$M = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	
F_3: Correlation function	$\nu = 0.5$	$\nu = 1.5$	
F_4: Noise-to-signal ratio	$\delta^2 = 0$	$\delta^2 = 1$	$\delta^2 \sim \text{Unif}(0, 1)$
F_5: Decay parameter	$\phi \sim \text{Unif}(0.1, 1)$	$\phi \sim \text{log-normal}(-1.1, 1)$	

Table 5.1: Five crossed factors together with their levels and coded values.

There are also two nested factors. For the regression parameters $\boldsymbol{\beta}$ we assume a normal prior with prior mean $\boldsymbol{\beta}_0$ and matrix $\mathbf{R}^{-1}$.

1. Factor F_6 determines the prior mean of trend parameters, and is nested within factor F_2 (form of mean function) and has two levels:

$$F_6|(F_2 = 0) = \begin{cases} 0, & \Rightarrow \ \beta_0 = 0 \\ 1, & \Rightarrow \ \beta_0 = 1 \end{cases}$$

$$F_6|(F_2 = 1) = \begin{cases} 0, & \Rightarrow \ \boldsymbol{\beta}_0 = (0,0,0) \\ 1, & \Rightarrow \ \boldsymbol{\beta}_0 = (1,1,1). \end{cases}$$

2. Factor F_7 determines the prior precision of trend parameters, and is nested within factor F_2 (form of mean function) and has three levels:

$$F_7|(F_2 = 0) = \begin{cases} 0 & \Rightarrow \ R^{-1} = 0.25 \\ 1 & \Rightarrow \ R^{-1} = 1 \\ 2 & \Rightarrow \ R^{-1} = 4 \end{cases}$$

$$F_7|(F_2 = 1) = \begin{cases} 0 & \Rightarrow \ \mathbf{R^{-1}} = 0.25\mathbf{I}_3 \\ 1 & \Rightarrow \ \mathbf{R^{-1}} = \mathbf{I}_3 \\ 2 & \Rightarrow \ \mathbf{R^{-1}} = 4\mathbf{I}_3. \end{cases}$$

All possible combinations of the levels of these factors are considered. The total number of combinations investigated is 288. For each combination, we generate 100 random designs with $n = 5$ and $n = 10$ points in $\mathcal{X} = [-1, 1]^2$ and assume prediction is required across $\mathcal{X}_P = 10 \times 10$ grid. For each design we evaluate $\Psi(\xi)$ and each of $\Psi_1(\xi)$ and $\Psi_2(\xi)$.

We consider the results separately for uniform and log-normal prior distributions on ϕ.

(i) Uniform prior on ϕ. When δ^2 is assumed known ($\delta^2 = 0$ or $\delta^2 = 1$), the objective function $\Psi(\xi)$ is given by (3.10), with $\Psi_1(\xi)$ and $\Psi_2(\xi)$ given (3.14) and (3.15), respectively, and approximated numerically by (3.33) and (3.34). When δ^2 is unknown and a uniform prior is assumed, then the objective function $\Psi(\xi)$ is given by (3.9) with $\Psi_1(\xi)$ and $\Psi_2(\xi)$ given by (3.11) and (3.12), respectively, and approximated by (3.29) and (3.30) respectively.

(ii) Log-normal prior on ϕ. We approximate the integrals using Gauss-Hermite quadrature methods with $\Psi_1(\xi)$ and $\Psi_2(\xi)$ evaluated by (3.33) and (3.34). For unknown δ^2, $\Psi(\xi)$ is given by (3.9) and the two parts, $\Psi_1(\xi)$ and $\Psi_2(\xi)$, are approximated by (3.31) and (3.32), respectively.

Tables 5.2 and 5.3 show the correlation between the values of $\Psi(\xi)$ and $\Psi_1(\xi)$ for $n = 5$ ($F_1 = 0$) and $n = 10$ ($F_1 = 1$), respectively. For each combination of the factor

levels in Table 5.1 we found the values of $\Psi(\xi)$ and $\Psi_1(\xi)$ for 100 randomly generated designs. For each combination we then calculate dthe correlation between $\Psi(\xi)$ and $\Psi_1(\xi)$ using those 100 values. As can be seen, for all 288 combinations of study factors, the correlation between $\Psi(\xi)$ and $\Psi_1(\xi)$ is very high, almost equal to one. This evidence suggests that ordering of designs is preserved under $\Psi(\xi)$ and $\Psi_1(\xi)$.

In general Tables 5.2 and 5.3 show that the two objective functions yield very similar results and we can conclude that regardless the choice of mean function, the value of the precision matrix, the correlation function, (exponential or Matérn $\nu = 1.5$), and either known or unknown δ^2, the closed-form approximation $\Psi_1(\xi)$ is a good approximation for the objective function and can be used as a design selection criterion for Bayesian optimal designs. The same conclusions can be drawn for either $n = 5$ or $n = 10$.

From substantial numerical evidence, we conclude that $\Psi_1(\xi) \simeq \Psi(\xi)$ and in fact always $\Psi_2(\xi) << \Psi_1(\xi)$. We have strong evidence to assume that $\Psi_2(\xi) \simeq 0$ and approximate the objective function $\Psi(\xi)$ (3.9) or (3.10), by $\Psi_1(\xi)$.

5.4.2 Discussion of analytical results

This numerical evidence is also supported by Lemma 3.1 in Section 3.4.1 and the connection between the objective function $\Psi(\xi)$ and the integrated likelihood. Since $\Psi_2(\xi)$ is mainly dependent on the posterior density of the unknown covariance parameters, which is linked with the $L^I(\phi, \delta^2)$, we can have a more thorough understanding about our proposed approximation if we understand $L^I(\phi, \delta^2)$.

As we have seen, the values of $L^I(\phi, \delta^2)$ are always smaller than a function which depends on ϕ and δ^2 and $L^I(\phi, \delta^2)$ decreases much faster than any chosen values of ϕ and δ^2. This is true regardless of δ^2 being known and fixed or unknown. In Section 3.4.1 we made the connection between $L^I(\phi, \delta^2)$ and the second part, $\Psi_2(\xi)$, of the objective function $\Psi(\xi)$. Since the value of $L^I(\phi, \delta^2)$ tends to get very small very quickly, then the second part $\Psi_2(\xi)$ always yields small values, much smaller in magnitude than the values of $\Psi_1(\xi)$.

5.4.3 Theoretical insight into $\Psi_2(\xi)$

To theoretically investigate the closed-form approximation and provide intuition about the domination of $\Psi_2(\xi)$ by $\Psi_1(\xi)$, we derive a linear approximation to the integrand of $\Psi_2(\xi)$. We investigate the general case where both covariance parameters ϕ and δ^2 are unknown; similar results can be derived when only ϕ is unknown.

Function $\Psi_2(\xi)$ depends on the mean, μ^*, of the predictive posterior distribution given by (2.27). We define $\boldsymbol{\theta} = (\phi, \delta^2)$ and $\mu^*(\boldsymbol{\theta}) = \mathbb{E}\left[y(\mathbf{x}_p)|\mathbf{y}, \boldsymbol{\theta}\right]$.

			F_1 F_2 F_3 F_6							
F_4	F_5	F_7	0000	0001	00010	0011	0100	0101	01011	0111
0	0	0	0.999979	0.999976	0.999960	0.999973	0.999963	0.999949	0.999829	0.999863
0	0	1	0.999964	0.999936	0.999925	0.999884	0.999874	0.999890	0.999749	0.999786
0	0	2	0.999992	0.999991	0.999979	0.999980	0.999945	0.999935	0.999985	0.999956
1	0	0	0.999954	0.999967	0.999253	0.998894	0.999995	0.999992	0.999998	0.999999
2	0	0	0.999582	0.999667	0.996103	0.996035	0.999874	0.999869	0.999700	0.999752
1	0	1	0.999726	0.999803	0.999240	0.998954	0.999972	0.999958	0.999963	0.999959
2	0	1	0.999390	0.999588	0.992186	0.997101	0.999593	0.999331	0.998726	0.997900
1	0	2	0.999989	0.999984	0.999330	0.999396	0.999999	0.999998	1.000000	1.000000
2	0	2	0.999795	0.999426	0.997576	0.996940	0.999971	0.999971	0.999788	0.999869
0	1	0	0.999957	0.999949	0.999962	0.999959	0.999743	0.999794	0.998887	0.999061
0	1	1	0.999890	0.999841	0.999193	0.999251	0.999837	0.999774	0.997260	0.999018
0	1	2	0.999971	0.999957	0.998891	0.999301	0.999736	0.999858	0.999480	0.998731
0	1	2	0.999836	0.999826	0.999107	0.999569	0.999996	0.999995	1.000000	1.000000
1	1	0	0.999762	0.998779	0.996575	0.998709	0.999835	0.999333	0.999602	0.999500
2	1	0	0.999823	0.999578	0.999282	0.998496	0.999982	0.999967	0.999983	0.999988
2	1	1	0.999591	0.997443	0.998400	0.998525	0.997912	0.999296	0.997447	0.997934
1	1	2	0.999917	0.999912	0.999508	0.999699	0.999999	0.999999	1.000000	1.000000
2	1	2	0.999633	0.996794	0.997708	0.997805	0.999912	0.999960	0.999894	0.999828

Table 5.2: Correlation of 100 random designs under objective functions Ψ and Ψ_1 for factor level combinations $F_1, F_2, F_3, F_4, F_5, F_6, F_7$.

			F_1 F_2 F_3 F_6							
F_4	F_5	F_7	1000	1001	10010	1011	1100	1101	11011	1111
0	0	0	0.999994	0.999991	0.999986	0.999990	0.999970	0.999956	0.999989	0.999990
0	0	1	0.999983	0.999976	0.999988	0.999981	0.999965	0.999934	0.999981	0.999974
0	0	2	0.999996	0.999996	0.999991	0.999991	0.999994	0.999994	0.999999	0.999999
1	0	0	0.999910	0.999891	0.999022	0.999040	0.999994	0.999994	0.999994	0.999996
2	0	0	0.999580	0.998898	0.995985	0.998554	0.999951	0.999875	0.999907	0.999978
1	0	1	0.999668	0.999710	0.997759	0.998655	0.999970	0.999978	0.999951	0.999954
2	0	1	0.999086	0.999154	0.982064	0.993719	0.999905	0.999681	0.999258	0.999078
1	0	2	0.999928	0.999951	0.998549	0.998177	0.999998	0.999999	0.999999	0.999998
2	0	2	0.999276	0.999199	0.996621	0.998957	0.999986	0.999984	0.999966	0.999946
0	1	0	0.999962	0.999952	0.998991	0.999041	0.999960	0.999959	0.999788	0.998942
0	1	1	0.999951	0.999935	0.998891	0.999248	0.999914	0.999927	0.998864	0.999411
0	1	2	0.999955	0.999981	0.998764	0.999193	0.999984	0.999993	0.999129	0.999678
1	1	0	0.999759	0.999741	0.996167	0.997546	0.999992	0.999995	0.999998	0.999998
2	1	0	0.998939	0.998273	0.987187	0.998558	0.999904	0.999928	0.999914	0.999922
1	1	1	0.999561	0.999549	0.993949	0.996990	0.999975	0.999969	0.999975	0.999975
2	1	1	0.997954	0.997893	0.993433	0.991596	0.999761	0.999441	0.999852	0.999596
1	1	2	0.999800	0.999599	0.998408	0.998469	0.999999	0.999999	0.999999	1.000000
2	1	2	0.998192	0.998933	0.998492	0.985207	0.999980	0.999933	0.999981	0.999950

Table 5.3: Correlation of 100 random designs under objective functions Ψ and Ψ_1 for factor level combinations $F_1, F_2, F_3, F_4, F_5, F_6, F_7$.

Recall that $\Psi_2(\xi)$ is given by

$$\begin{aligned}\Psi_2(\xi) &= \int_{\mathcal{X}_P}\int_{\mathcal{Y}} \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[(\mu^* - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*))(\mu^* - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*))^\top]\pi(\mathbf{y})d\mathbf{y}d\mathbf{x}_p \\ &= \int_{\mathcal{X}_P}\int_{\mathcal{Y}}\int_{\boldsymbol{\theta}} [(\mu^* - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*))(\mu^* - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*))^\top]\pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y})d\boldsymbol{\theta}d\mathbf{y}d\mathbf{x}_p. \quad (5.2)\end{aligned}$$

The objective function $\Psi_2(\xi)$ is the average, with respect to the data $\mathbf{y}$, of the variance of μ^* with respect to the posterior distribution of $\boldsymbol{\theta}$, where the mean, μ^*, of the predictive distribution depends on both $\boldsymbol{\theta}$ and the data $\mathbf{y}$. We employ a linear approximation to $\mu^*(\boldsymbol{\theta})$ about the prior mean of $\boldsymbol{\theta}$, i.e. $\overline{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}]$.

Our aim is to show that the predictive mean μ^* does not depend on $\boldsymbol{\theta}$, the posterior distribution (2.34) does not affect the value of μ^*, and as a result does not give rise to large values of $\Psi_2(\xi)$.

A first order Taylor expansion about $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$ gives

$$\mu^*(\boldsymbol{\theta}) \simeq \mu^*(\overline{\boldsymbol{\theta}}) + \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}). \quad (5.3)$$

Initially, we approximate the term $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*)$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*) &= \int_{\boldsymbol{\theta}} \mu^*(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &\simeq \int_{\boldsymbol{\theta}} \mu^*(\overline{\boldsymbol{\theta}})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} + \int_{\boldsymbol{\theta}} \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &\simeq \mu^*(\overline{\boldsymbol{\theta}}) + \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} \int_{\boldsymbol{\theta}} \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} - \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} \overline{\boldsymbol{\theta}}. \quad (5.4)\end{aligned}$$

Hence, we have

$$\mu^* - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\mu^*) \simeq \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})). \quad (5.5)$$

We substitute (5.5) into (5.2) to obtain:

$$\begin{aligned}\Psi_2(\xi) &\simeq \int_{\mathcal{Y}}\int_{\boldsymbol{\theta}} \left(\left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}))\right)\left(\left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}))\right)^\top \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y})d\boldsymbol{\theta}d\mathbf{y} \\ &\simeq \int_{\mathcal{Y}}\int_{\boldsymbol{\theta}} \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}))(\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}))^\top \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}^\top \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y})d\boldsymbol{\theta}d\mathbf{y} \\ &\simeq \int_{\mathcal{Y}} \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} \mathrm{Cov}[\boldsymbol{\theta}|\mathbf{y}] \left|\frac{\partial \mu^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}^\top \pi(\mathbf{y})d\mathbf{y}. \quad (5.6)\end{aligned}$$

The partial derivatives of μ^* with respect to the ϕ and δ^2 are then calculated.

$$\begin{aligned}\frac{\partial \mu^*}{\partial \phi} =& \left[-\frac{\partial \boldsymbol{\omega}^\top}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F} + \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F}\right](\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 + \\ & (\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 \\ & +\frac{\partial \boldsymbol{\omega}^\top}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{y} - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{\partial \boldsymbol{\omega}^\top}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ & +\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y} \\ & +(\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y} \\ & -(\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \end{aligned} \tag{5.7}$$

$$\begin{aligned}\frac{\partial \mu^*}{\partial \delta^2} =& \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 + \\ & (\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 \\ & -\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ & +(\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y} \\ & -(\mathbf{f}_p - \boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \end{aligned} \tag{5.8}$$

To proceed, we now make the assumption of exponential correlation function (2.4) with $\nu = 0.5$ to provide analytical tractable derivatives for $\frac{\partial \boldsymbol{\Sigma}}{\partial \phi}$ and $\frac{\partial \boldsymbol{\Sigma}}{\partial \delta^2}$, involved in both (5.7) and (5.8).

Then, we substitute (5.7) and (5.8) into (5.6) to obtain

$$\Psi_2(\xi) \simeq \int_{\mathcal{X}_P}\int_{\mathcal{Y}} \left(\frac{\partial \mu^*}{\partial \phi}, \frac{\partial \mu^*}{\partial \delta^2}\right) \mathrm{Cov}(\phi, \delta^2|\mathbf{y}) \left(\frac{\partial \mu^*}{\partial \phi}, \frac{\partial \mu^*}{\partial \delta^2}\right)^\top \pi(\mathbf{y}) d\mathbf{y} d\mathbf{x}_p. \tag{5.9}$$

We now evaluate (5.9) for a variety of combinations of $\boldsymbol{\beta}_0$ and $\mathbf{R}^{-1}$, the prior hyperparameters for $\boldsymbol{\beta}$, by the following steps:

- assign prior distributions to ϕ and δ^2
- generate a sample from the marginal posterior of $\pi(\mathbf{y})$
- evaluate the derivatives (5.7) and (5.8) at the prior mean of ϕ and δ^2 respectively
- generate a sample from the posterior distribution $\pi(\phi, \delta^2|\mathbf{y})$ (2.34) and evaluate the $\mathrm{Cov}(\phi, \delta^2|\mathbf{y})$
- evaluate (5.9) using Monte Carlo integration.

$\mathbf{R}^{-1}$	Prior mean $\boldsymbol{\beta}_0$		
	$(0,0,0)$	(1,1,1)	(10,10,10)
diag(0.25)	0.0021	0.0025	0.0378
diag(1)	0.0010	0.0012	0.0166
diag(4)	0.0008	0.0009	0.0034
diag(10)	0.0007	0.0008	0.0013

Table 5.4: Values of the linear approximation to $\Psi_2(\xi)$.

We investigate the case of a linear trend as it is the most interesting. The value of $\Psi_2(\xi)$ is evaluated for $\boldsymbol{\beta}_0 = (0,0,0)$, $\boldsymbol{\beta}_0 = (1,1,1)$ and $\boldsymbol{\beta}_0 = (10,10,10)$ and $\mathbf{R}^{-1} = 0.25\mathbf{I}_3$, $\mathbf{R}^{-1} = \mathbf{I}_3$, $\mathbf{R}^{-1} = 4\mathbf{I}_3$ and $\mathbf{R}^{-1} = 10\mathbf{I}_3$. We assign uniform prior distributions for both ϕ and δ^2, Unif$(0.1, 1)$ and Unif$(0, 1)$ respectively. We investigate 50 random designs with 10 points and for all cases the value of the linear approximation of $\Psi_2(\xi)$ is always less than 5%. In Table 5.4 we summarised the results for one design.

To summarise, the choices of the prior mean and precision matrix of the trend parameters do not affect the value of $\Psi_2(\xi)$. The value is always very small ($\Psi_2(\xi) < 0.01\Psi_1(\xi)$) and this is in line with results from our numerical study (Section 5.4.1).

Moreover, in order to find out how much information is contributed by the posterior distribution to the linear approximation for $\Psi_2(\xi)$, we evaluate 10 samples from the posterior distribution $\pi(\phi, \delta^2|\mathbf{y})$ given by (2.34) and we evaluate the posterior covariaince $\text{Cov}(\phi, \delta^2|\mathbf{y})$ for each one of these samples. Always, $\text{Cov}(\phi, \delta^2|\mathbf{y}) < 0.05$ and as a result yields very small values for the second part of the objective function.

5.4.4 Examples of optimal designs

We now demonstrate the methodology for the general problem of sensor placement. For example, if the objective of the experiment is to predict at unobserved locations in the geographical region of interest, we address the problem of finding optimal locations to place sensors. We find Bayesian optimal designs for prediction using minimisation of the closed-form approximation $\Psi_1(\xi)$ as a selection criterion.

Optimal designs are found with $n = 10$ or $n = 20$ sampling locations in the study region $\mathcal{X} = [-1, 1]^2$. The Gaussian process model (2.6) depends on both the mean function and the correlation function, and we investigate the effect of both on the Bayesian optimal design. The optimal designs are found by minimising the average of $\Psi_1(\xi)$ across a 10×10 regular prediction grid, $\mathcal{X}_\mathcal{P} \subseteq [-1, 1]^2$.

In Chapter 4, we demonstrated that the form of the optimal design is affected most strongly by the choice of mean function, correlation function and covariance parameters. Hence, here we find examples of optimal designs for constant and linear mean function, the Matérn correlation function, with $\nu = 0.5$ and $\nu = 1.5$, and (a) four assumed values of δ^2 with unknown ϕ or (b) unknown δ^2 and ϕ. We also consider a normal prior for the

regression coefficients with zero mean and variance-covariance matrix $\sigma^2\mathbf{I}$; in Chapter 4 it was shown that the optimal design is robust to the choice of precision matrix and prior mean of $\boldsymbol{\beta}$.

Specifically, designs are found for:

- $n = 10$ and $n = 20$ points
- mean function:

  (1) $\mathbf{f}^\top(\mathbf{x})\boldsymbol{\beta} = \beta_0$

  (2) $\mathbf{f}^\top(\mathbf{x})\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Matérn correlation function with either $\nu = 0.5$ and $\nu = 1.5$.
- prior distributions:
  - $\sigma^2 \sim \text{IG}(3, 1)$, recall that the hyperparmeters a, b of the inverse gamma prior do not affect the choice of design as a, b only influence the objective function through a multiplicative constant, see Section 3.5.1.
  - **(1)** $\beta \sim \text{N}(0, \sigma^2)$

    (2) $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$
  - Two different prior distributions for ϕ are chosen, having the same prior mean:

    (1) $\pi(\phi) \sim \text{Uniform}(0, 1, 1)$

    (2) $\pi(\phi) \sim \text{log-normal}(-1, 1, 1)$
  - Two different cases for δ^2 are considered

    (1) known and fixed values $\delta^2 = 0, 0.5, 1$ and 2.5

    (2) unknown with $\delta^2 \sim \text{Uniform}(0.1, 1)$.

In total, we investigate 80 combinations of these individual settings. For each combination we generate 50 random designs selected from the design region. For each of these starting designs, the coordinate exchange algorithm (Section 3.6.1) is used to find a design that minimises $\Psi_1(\xi)$. The final choice of design is that which has the smallest value of $\Psi_1(\xi)$ among these 50 designs.

Here, we will present the result for $n = 10$; similar results and conclusions are obtained for $n = 20$ and can be found in Appendix A.4. Figures 5.4–5.9 show the Bayesian optimal designs for $n = 10$ with contours displaying the average, with respect to the prior values on δ^2 and ϕ, correlation between each point in the study region, $\mathcal{X} = [-1, 1]^2$, and the centre point.

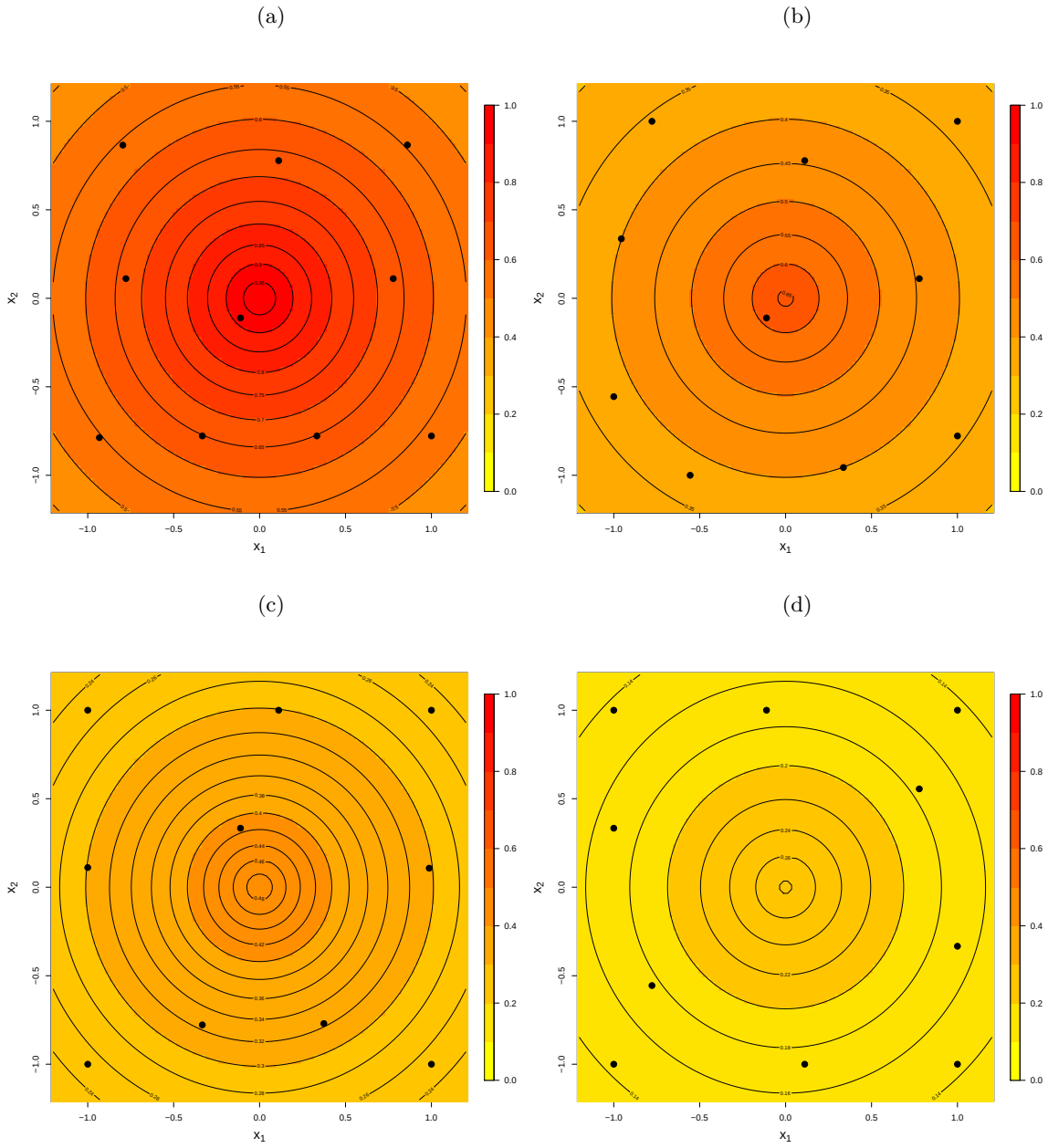


Figure 5.4: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 0.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

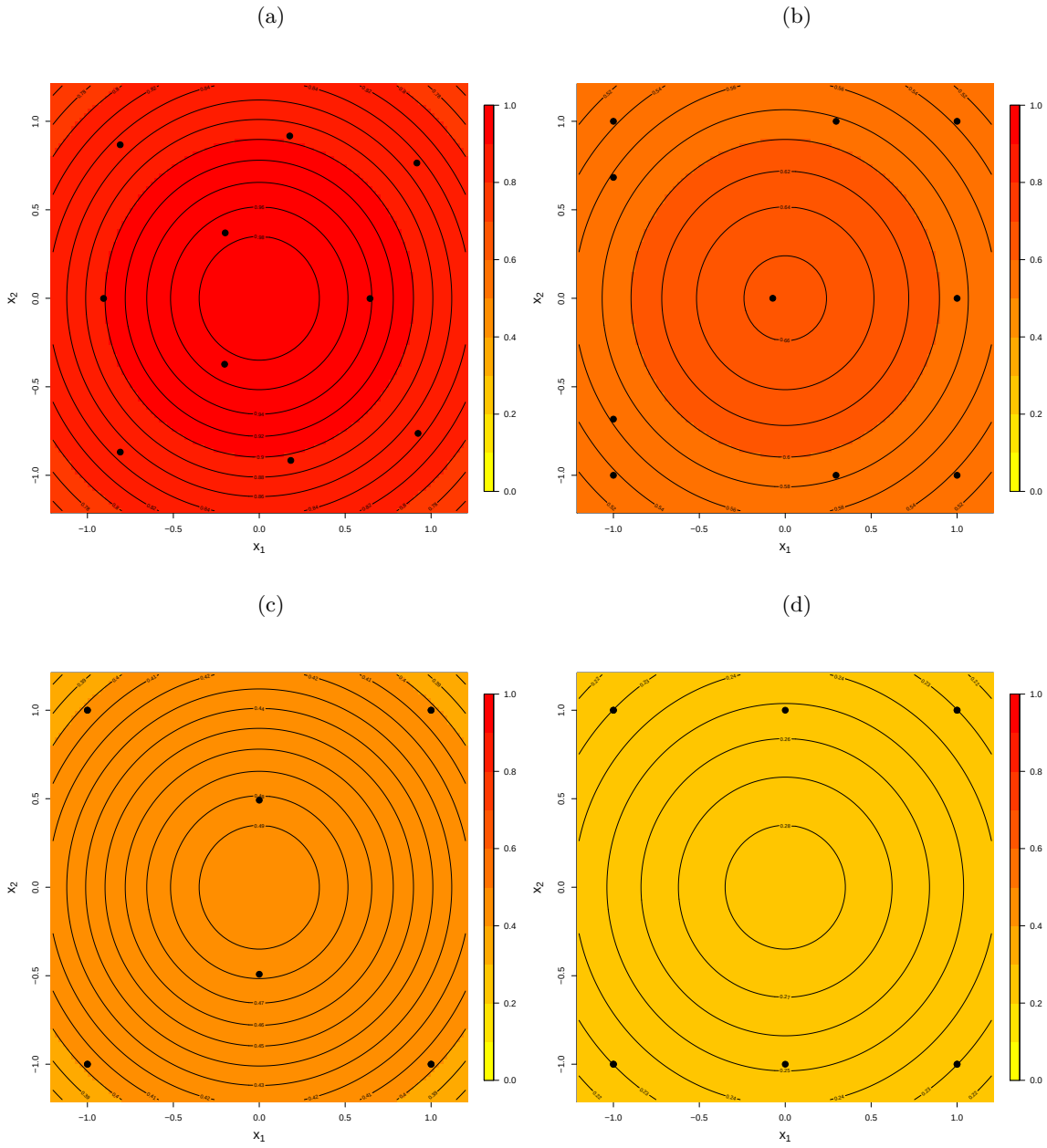


Figure 5.5: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 1.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ. In plots (c) and (d) four points are repeated.

Constant mean function

When we assume constant mean function, Ψ-optimal designs are not much influenced by the values of parameters. Table 5.5 shows that Ψ-optimal designs for constant mean have similar coverage and spread values. The results here are in line with those in Chapter 4. Figures A.9–A.14 correspond to the Ψ-optimal designs for a Gaussian process with constant mean function.

Linear mean function

The linear mean function models the large scale variation in a spatial process. Figures 5.4–5.9 correspond to the Ψ-optimal designs for a Gaussian process with linear mean function, and different combinations of prior distributions on ϕ, δ^2 and correlation functions. The contour plots show the average correlation between each point in the study region and the centre, averaged across the prior distributions for ϕ and δ^2. As the correlation between observations is not affected by the mean function, these contours are identical to those in Figures A.9–A.14.

The optimal designs are influenced by the range and the strength of the correlation.

(i) $\nu = 0.5$: Figures 5.4 and 5.6 correspond to Ψ-optimal designs for $\nu = 0.5$ and uniform prior distribution for ϕ, and $\nu = 0.5$ and log-normal prior distribution for ϕ, respectively. When the Gaussian process model (2.6) does not include a nugget effect (i.e. $\delta^2 = 0$), the optimal design spreads points throughout the study region, see Figures 5.4 (a) and 5.6 (a) for uniform and log-normal prior distributions. However, when a nugget is included in the model, and potentially as the value of δ^2 increases, the correlation between observations decreases. Therefore, the choice of optimal design points are strongly influenced from the mean function. For large values of δ^2, i.e. $\delta^2 = 2.5$, (Figures 5.4 (d) and 5.6 (d)), the points move to the corners of the region mimicking the optimal design for problems assuming a linear model and uncorrelated errors.

(ii) $\nu = 1.5$: Figures 5.5 and 5.7 correspond to Ψ-optimal designs for $\nu = 1.5$ and uniform prior distribution for ϕ, and $\nu = 1.5$ and log-normal prior distribution for ϕ, respectively. For this correlation function, large values of δ^2 correspond to designs with repeating points and especially, with points at the corners, see for example Figure 5.5 (d) and Figure 5.7 (d) and for uniform and log-normal prior on ϕ respectively. Moreover, if we compare Figure 5.4 with Figure 5.5 and Figure 5.6 with Figure 5.7 for uniform and log-normal prior distributions respectively, we conclude that the mean function is even more influential as the range of the correlation is smaller.

In general, if we compare the designs obtained here with the corresponding designs in Chapter 4, we see that they are not exactly the same. This is because the optimal design is not unique and every time we obtain a different optimal design. The optimal designs are not unique because we are finding exact designs and efficient or near-optimal

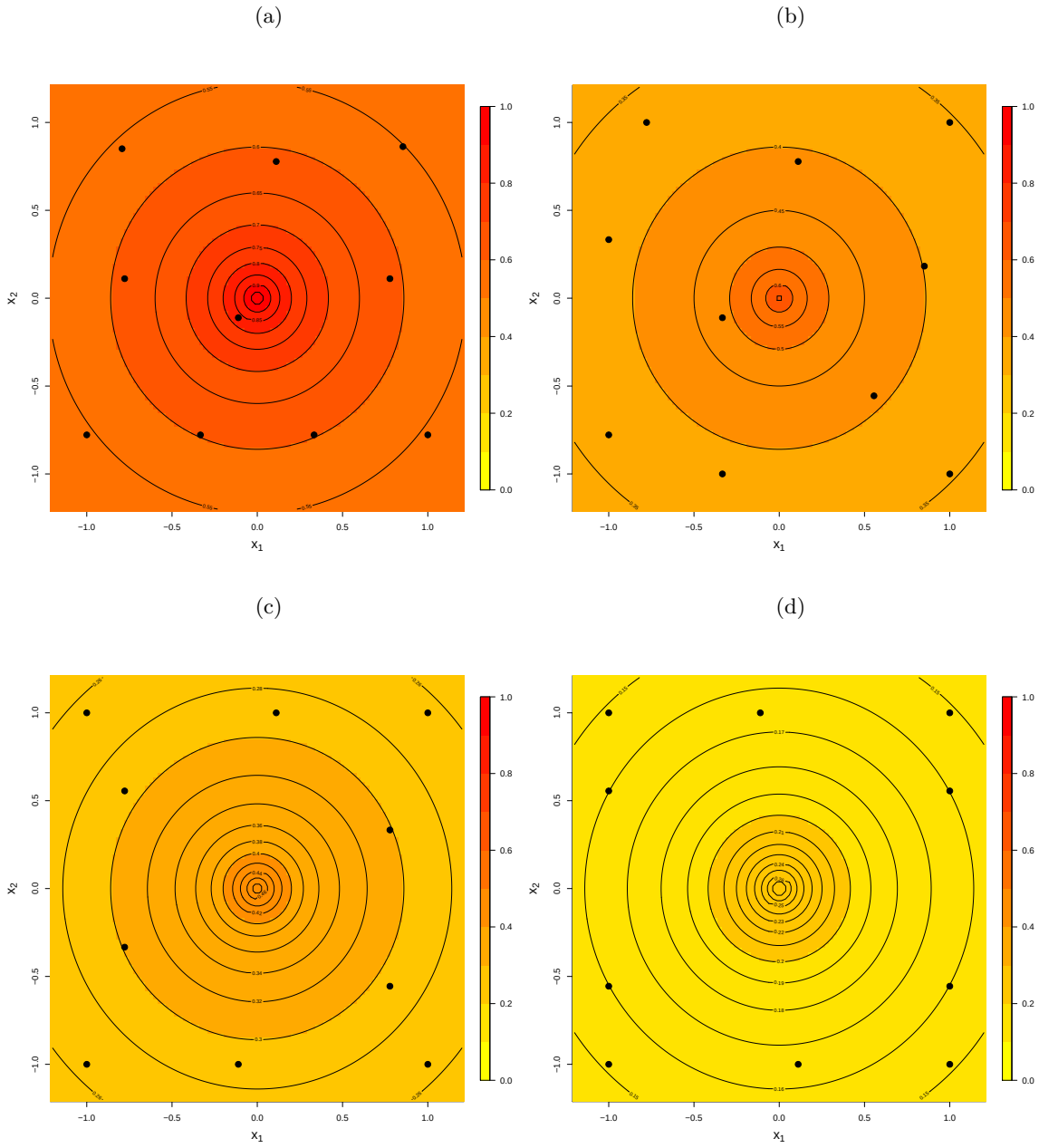


Figure 5.6: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 0.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

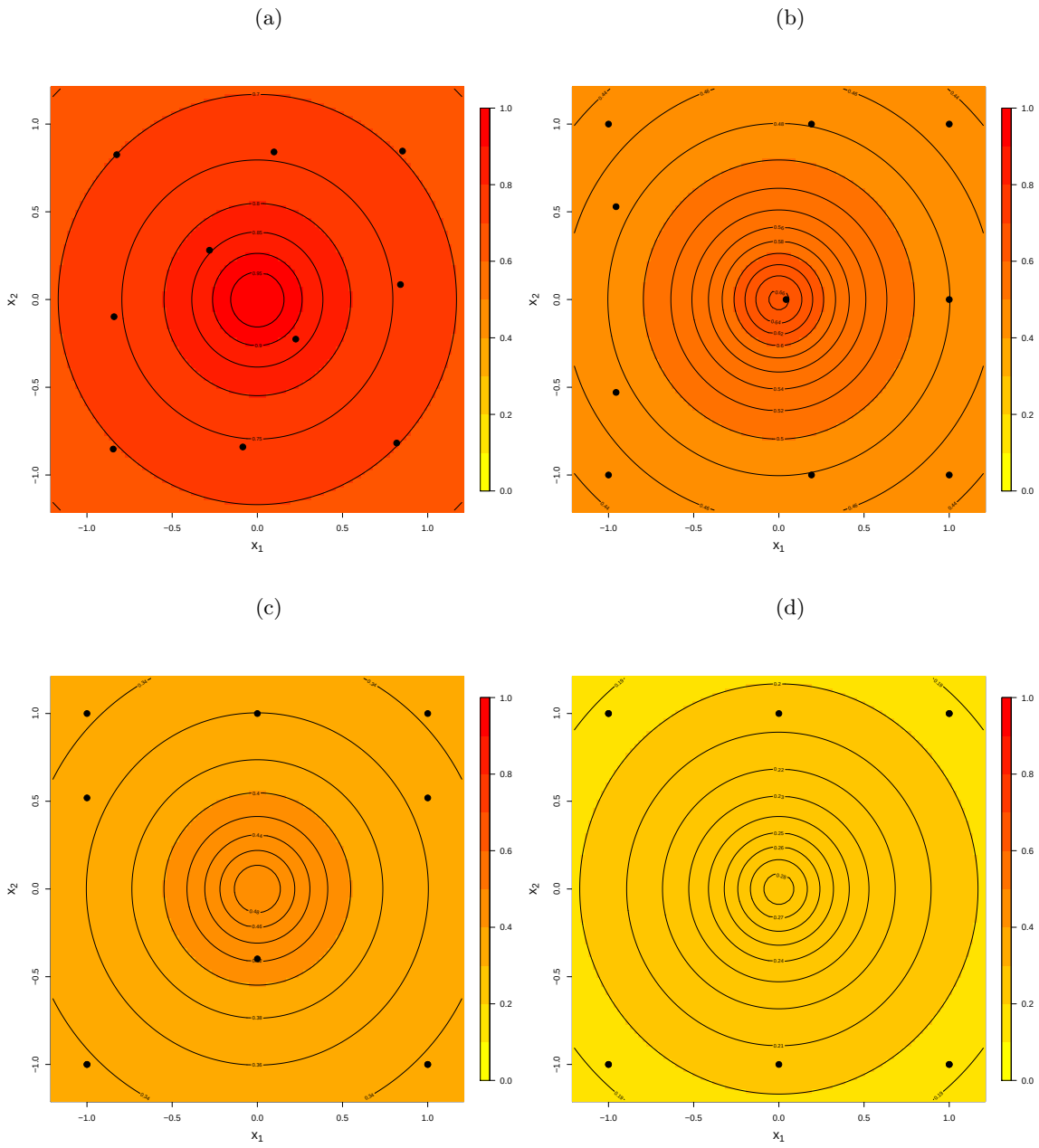


Figure 5.7: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 1.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ. In plot (c) two points are repeated and in plot (d) four points are repeated.

(a) (b)

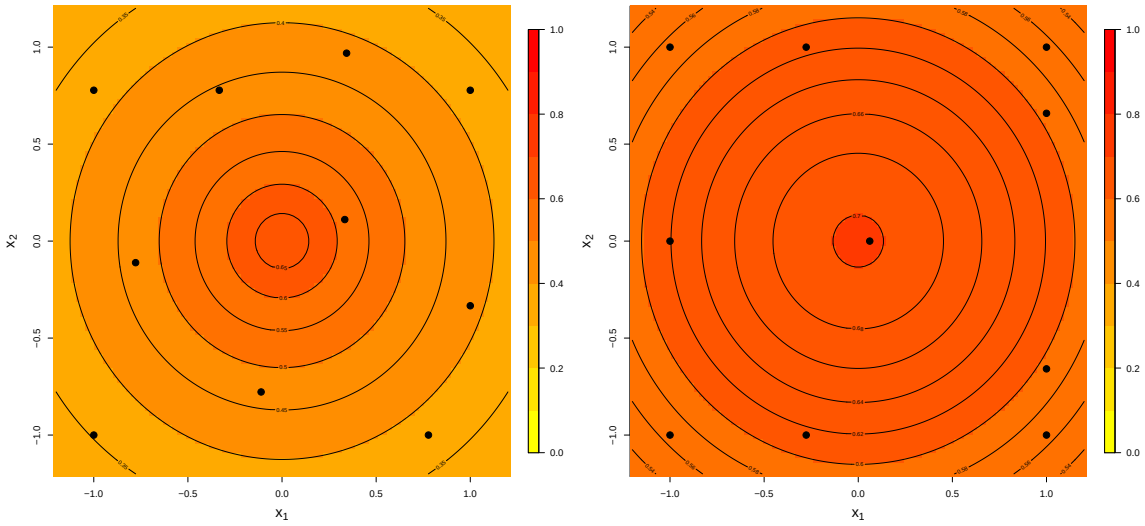


Figure 5.8: Ψ-optimal designs for a linear mean function, uniform prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

(a) (b)

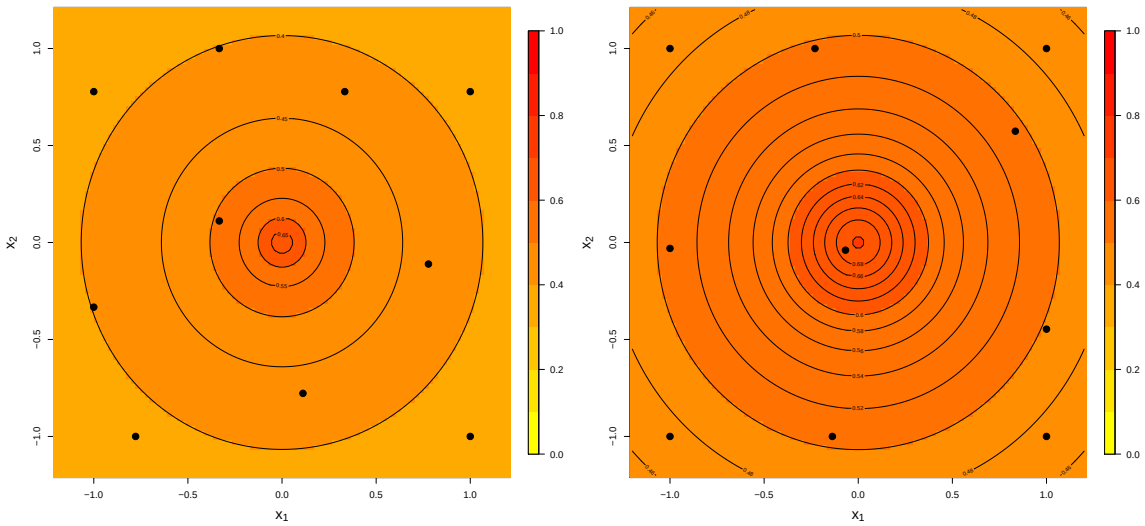


Figure 5.9: Ψ-optimal designs for a linear mean function, log-normal prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

Case	Coverage		Spread	
	Linear	Constant	Linear	Constant
	Uniform prior distribution ϕ			
$\nu = 0.5$				
$\delta^2 = 0$	0.2756	0.2643	0.7228	0.7090
$\delta^2 = 0.5$	0.3136	0.2672	0.7669	0.6430
$\delta^2 = 1$	0.3436	0.2632	0.7703	0.6196
$\delta^2 = 2.5$	0.3999	0.2574	0.6432	0.6361
δ^2 unknown	0.3084	0.2675	0.7656	0.6440
$\nu = 1.5$				
$\delta^2 = 0$	0.2771	0.2730	0.7352	0.7339
$\delta^2 = 0.5$	0.3879	0.3131	0.6150	0.7023
$\delta^2 = 1$	0.4659	0.3022	1.094	0.5901
$\delta^2 = 2.5$	0.5934	0.2871	0.9999	0.4405
δ^2 unknown	0.3854	0.3081	0.6313	0.7203
	Log-normal prior distribution ϕ			
$\nu = 0.5$				
$\delta^2 = 0$	0.2767	0.2557	0.7062	0.6697
$\delta^2 = 0.5$	0.3162	0.2584	0.7410	0.6362
$\delta^2 = 1$	0.3607	0.2261	0.6575	0.6389
$\delta^2 = 2.5$	0.4553	0.2584	0.5333	0.6351
δ^2 unknown	0.3103	0.2585	0.7693	0.6361
$\nu = 1.5$				
$\delta^2 = 0$	0.2717	0.2658	0.7211	0.7140
$\delta^2 = 0.5$	0.3675	0.2878	0.7040	0.7122
$\delta^2 = 1$	0.4335	0.2795	0.9858	0.6551
$\delta^2 = 2.5$	0.5934	0.2660	0.9999	0.5261
δ^2 unknown	0.3626	0.2850	0.7150	0.7135

Table 5.5: Coverage and spread of designs in Figures 5.4–5.9 and Figures A.9–A.14. Best design for coverage is the one with the smallest value and the best design for spread is the one with the largest value.

designs are found using computer search.

However, the key point here is that all the designs yield very similar values of the objective function and they are all highly efficient.

Next, we assign a uniform prior distribution on δ^2 and find the optimal design for $\nu = 0.5$ and $\nu = 1.5$ and both prior distributions of ϕ. The resulting designs are those indicated in Figures 5.8 and 5.9.

By comparing Figure 5.8 to Figures 5.4 and 5.5, and Figure 5.9 to Figures 5.6 and 5.7, it is clear that the designs for unknown δ^2 are strongly influenced by small values of δ^2. That is, the choice of points for designs with unknown δ^2 resembles those designs for low δ^2. As seen before, designs for $\nu = 1.5$ have points near the corners of the study region, due to higher and more equal correlation across the region.

To further demonstrate the impact of the range of the correlation on the choice of opti-

mal design, in Figures 5.10 and 5.11 we provide density plots of the average correlation between observations at the centre point of the study region and each other point (on a 100×100 grid) for the four values of δ^2 used in the quadrature scheme to approximate $\Psi_1(\xi)$ (3.29). In these plots, the correlation is averaged with respect to the uniform prior distribution on ϕ.

From Figure 5.10, when $\delta^2 = 0$ and the correlation function is Matérn with $\nu = 0.5$, the range of the correlation is wider compared with the corresponding range for Matérn with $\nu = 1.5$ (Figure 5.11); the latter case has much higher mean. This pattern is the same for all other values of δ^2 considered. In general the range of the correlation is much smaller for $\nu = 1.5$ but the mean is higher.

To summarise, the optimal designs compromise between minimax designs, i.e. the design points tend to cover the study region and maximin designs, i.e. the design points are spread out, according to the choice of correlation and mean function and the prior information on the decay parameter ϕ and noise-to-signal parameter δ^2. Table 5.5 shows the coverage and spread values for 20 Ψ-optimal designs.

Especially for the case of linear mean function, the designs for $\delta^2 = 0$ give good coverage properties but are not good for spread. The designs with the best spread are those for $\delta^2 = 1, 2.5$. The optimal designs are strongly influenced by the degree and the range of correlation, i.e. the spread of the design points depends crucially on the distribution of the correlations with less uniform correlations across the study region producing designs with better coverage properties. Also for more uniform correlations, we see more influence of the mean function.

Diffuse prior distribution on regression parameters

Design criterion (3.9) is formulated assuming conjugate prior distributions for $\boldsymbol{\beta}$ and σ^2. If a diffuse prior is used for the trend and variance, namely,

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2},$$

we are able to derive analytically the posterior distributions $\boldsymbol{\beta}|\mathbf{y}, \phi, \delta^2$ and $\sigma^2|\mathbf{y}, \phi, \delta^2$, and the predictive distribution $y_{n+1}|\mathbf{y}, \phi, \delta^2$. However, the analytical derivation of the marginal distribution of the data is not feasible. That is, the distribution $\mathbf{y}|\phi, \delta^2$ does not have a closed form and this prevents the derivation of the objective function.

However, we can approximate a diffuse distribution if we assign a matrix, $\mathbf{R}^{-1}$, to the prior distribution of $\boldsymbol{\beta}$ with very large diagonal elements. We find Bayesian optimal designs minimising $\Psi_1(\xi)$ where a normal prior distribution for the trend parameters is assumed with zero mean and matrix $\mathbf{R}^{-1} = 1000\mathbf{I}_3$.

Also both decay, ϕ, and noise-to-signal, δ^2, parameters are considered unknown with uniform prior distribution for both. The correlation function is chosen to be the Matérn,

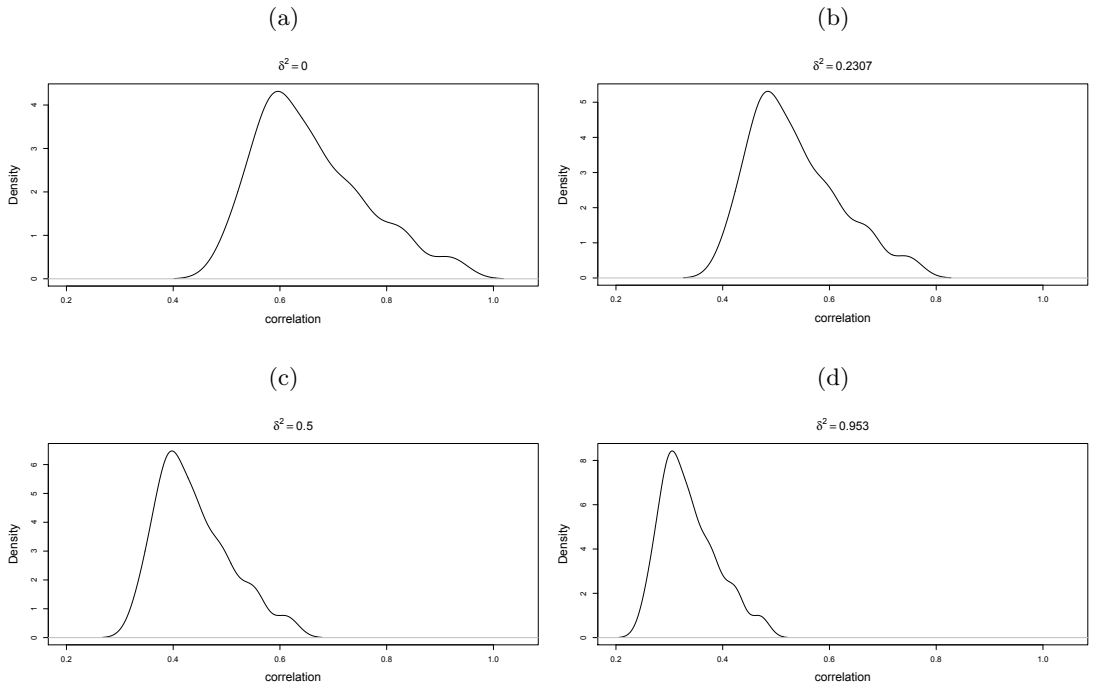


Figure 5.10: Density plots of the average correlation between observations at the centre of the study region and all other points for the Matérn correlation function $\nu = 0.5$ and $\delta^2 = 0$, (b) $\delta^2 = 0.2307$, (c) $\delta^2 = 0.5$ and (d) $\delta^2 = 0.953$. The correlation is averaged with respect to the uniform prior distribution on ϕ.

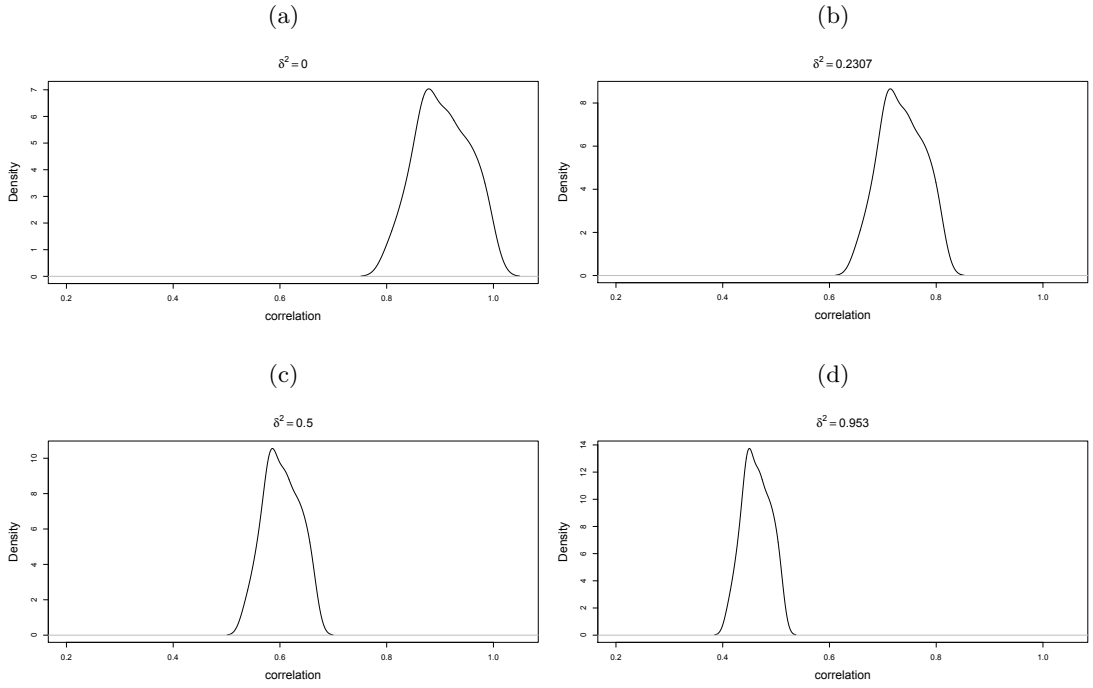


Figure 5.11: Density plots of the average correlation between observations at the centre of the study region and all other points for the Matérn correlation function $\nu = 1.5$ and $\delta^2 = 0$, (b) $\delta^2 = 0.2307$, (c) $\delta^2 = 0.5$ and (d) $\delta^2 = 0.953$.The correlation is averaged with respect to the uniform prior distribution on ϕ.

(a) (b)

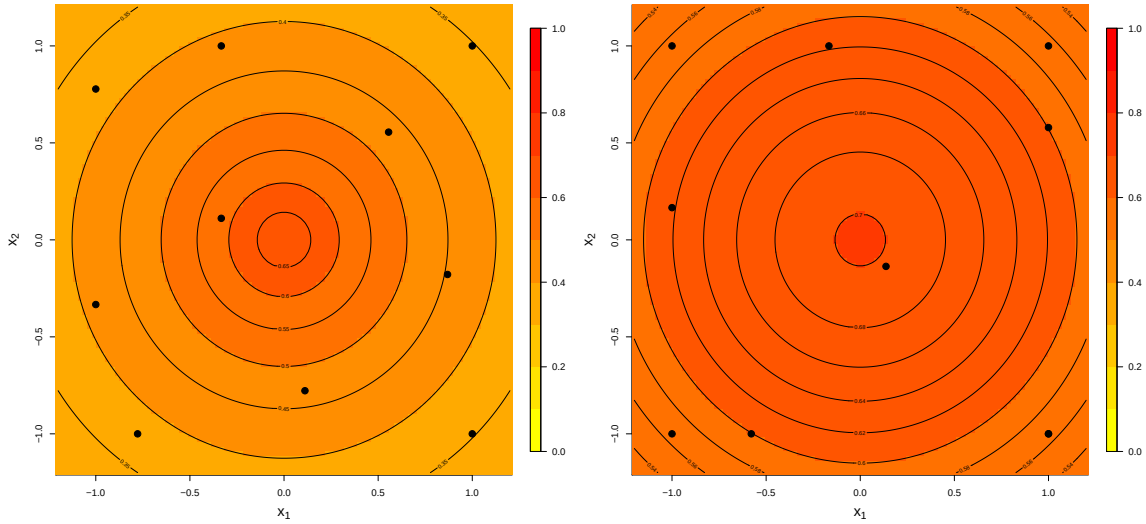


Figure 5.12: Optimal designs for linear mean function, diffuse prior distribution on $\boldsymbol{\beta}$, uniform prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

we consider both $\nu = 0.5$ and $\nu = 1.5$, corresponding to two different ranges of correlation.

In general, we have seen that the choice of precision matrix does not affect the design. Based on that we can assume that our Bayesian optimal designs are not sensitive on the choice of prior for the regression coefficients and we would obtain the same result if we were assuming a non-informative prior.

Figure 5.12 shows the Bayesian optimal designs for $\nu = 0.5$ and $\nu = 1.5$. If we compare these two designs with those in Figure A.14, we can see that very similar designs are obtained when $\mathbf{R}^{-1} = 1000\mathbf{I}_3$ and $\mathbf{R}^{-1} = \mathbf{I}_3$. We evaluate the efficiencies of the designs in Figure 5.12 with respect to the optimal designs found by assuming $\mathbf{R}^{-1} = \mathbf{I}_3$ (Figure A.14) and are very high, 0.997 and 0.990, for $\nu = 0.5$ and $\nu = 1.5$ respectively. This evidence supports our findings in Chapter 4 that our optimal designs are robust to the choice of prior distribution for $\boldsymbol{\beta}$.

5.5 Inference About the Unknown Model Parameters

In general our aim in this thesis is to develop optimal designs for prediction at unobserved points. However, a secondary objective in many experiments is inference about the unknown parameters. In this section, we assess the performance of optimal designs for prediction in terms of information gained about unknown model parameters. Throughout, we use Markov Chain Monte Carlo methods to make inference about the posterior distributions.

5.5.1 Markov Chain Monte Carlo methods

Markov Chain Monte Carlo (MCMC) methods are popular for sampling from posterior distributions which do not have a standard form, see, for example, Gelman et al. (2003). The main idea of MCMC is to generate a Markov chain whose stationary distribution is the posterior distribution of interest and then collect samples from that chain. The two most popular MCMC algorithms are the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampling (Geman and Geman, 1984) algorithms. We start this section with these two techniques.

1. **Metropolis-Hastings Algorithm**
   The Metropolis-Hasting algorithm samples from a non-standard posterior distribution through an acceptance-rejection mechanism. A proposal distribution is used to suggest an arbitrary next step in the chain and the accept-reject step controls the moves of the chain. Assume that we want to obtain a sample from the density $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{y})$, for some unknown model parameters $\tilde{\boldsymbol{\theta}}$. The MH algorithm proceeds as follows:

   i. Choose a starting value $\tilde{\boldsymbol{\theta}}_0$, at $t = 1$

   ii. At iteration t draw a candidate value $\tilde{\boldsymbol{\theta}}^*$ from the proposal distribution $q(\tilde{\boldsymbol{\theta}}^*|\tilde{\boldsymbol{\theta}}_{t-1})$

   iii. Calculate the acceptance probability, $\alpha = \min\left\{1, \frac{\pi(\tilde{\boldsymbol{\theta}}^*|\mathbf{y})q(\tilde{\boldsymbol{\theta}}_{t-1}|\tilde{\boldsymbol{\theta}}^*)}{\pi(\tilde{\boldsymbol{\theta}}_{t-1}|\mathbf{y})q(\tilde{\boldsymbol{\theta}}^*|\tilde{\boldsymbol{\theta}}_{t-1})}\right\}$

   iv. Sample $U \sim \text{Unif}(0, 1)$

   v. If $U < \alpha$ then accept $\tilde{\boldsymbol{\theta}}^* = \tilde{\boldsymbol{\theta}}_t$ else assign $\tilde{\boldsymbol{\theta}}_{t-1} = \tilde{\boldsymbol{\theta}}_t$

   vi. set $t = t + 1$, go to ii.

   According to the choice of the proposal distribution there are some special cases of the MH algorithms, the *Random-Walk Metropolis* and *Independent Metropolis-Hastings* algorithms. For the former, we assume that the proposal distribution is symmetric, i.e. $q(\tilde{\boldsymbol{\theta}}^*|\tilde{\boldsymbol{\theta}}_{t-1}) = q(\tilde{\boldsymbol{\theta}}_{t-1}|\tilde{\boldsymbol{\theta}}^*)$, and depends on the previous state, while for the latter case the proposal distribution is independent of $\tilde{\boldsymbol{\theta}}_{t-1}$.

2. **Gibbs Sampler**
   The Gibbs sampler samples $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ from full conditional posterior distributions. Unlike the MH algorithms, the Gibbs sampler updates the chain one component at a time. The Gibbs sampler proceeds as follows:

   i. Choose a starting value $\tilde{\theta}_1^0, \ldots, \tilde{\theta}_k^0$, at $t = 1$

   ii. Repeat draws:
   $\tilde{\theta}_1^t \sim \pi(\tilde{\theta}_1|\tilde{\theta}_2^{t-1}, \tilde{\theta}_3^{t-1}, \ldots, \tilde{\theta}_k^{t-1}, \mathbf{y})$
   $\tilde{\theta}_2^t \sim \pi(\tilde{\theta}_2|\tilde{\theta}_1^t, \tilde{\theta}_3^{t-1}, \ldots, \tilde{\theta}_k^{t-1}, \mathbf{y})$
   $\tilde{\theta}_3^t \sim \pi(\tilde{\theta}_3|\tilde{\theta}_1^t, \tilde{\theta}_2^t, \ldots, \tilde{\theta}_k^{t-1}, \mathbf{y})$

$$\vdots$$
$$\tilde{\theta}_k^t \sim \pi(\tilde{\theta}_k|\tilde{\theta}_1^t, \tilde{\theta}_2^t, \ldots, \tilde{\theta}_{k-1}^t, \mathbf{y})$$

Typically, we monitor the performance of an MCMC algorithm by inspecting the value of the acceptance rate and using diagnostic plots and statistics to decide about the mixing, i.e. has the chain sufficiently explored the entire posterior distribution, and the convergence. As a matter of practice, we throw out a certain number of the first draws, known as the burn-in, in order to make sure that our sample does not depend on the starting point and is closer to the stationary distribution. Another issue is the choice of simulated sample size and since iterateions in an MCMC algorithm are not independent, we can use the effective sample size. That is an estimate of the equivalent number of independent iterations that the chain represents. The formula for the effective sample size is given by:

$$ESS = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}, \tag{5.10}$$

where N is the original sample size and ρ_t is the autocorrelation at lag t. Autocorrelation of lag t is the correlation between samples that are t time steps apart.

5.5.2 Example using a Ψ-optimal design

In this section we make inference about the unknown parameters of the model using optimal designs for prediction for the following three cases:

(i) both ϕ and δ^2 known,

(ii) ϕ unknown and δ^2 known,

(iii) both ϕ and δ^2 unknown.

For all three cases, we assume the Gaussian process model (2.6) with linear mean function and exponential correlation function with Euclidean distance.

We select informative normal-inverse gamma prior distributions for $\boldsymbol{\beta}$ and σ^2, with

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}) \quad \text{and} \quad \sigma^2 \sim \mathrm{IG}\ (3, 1).$$

Initially, we find the Ψ-optimal design for prediction when there is no prior data available. We find a design with $n = 10$ runs for prediction on a 10×10 grid. For each design, a simulated dataset is generated in the region $\mathcal{X} = [-1, 1]^2$ from the Gaussian process model (2.6), with zero mean, i.e. $\boldsymbol{\beta} = \mathbf{0}$ and covariance parameters $\sigma^2 = 1$, $\phi = 0.2$ and $\delta^2 = 0$ or $\delta^2 = 1$. We refer to this model as the *simulation model*. We choose this simulation model since Diggle and Lophaven (2006) used it in their simulation studies.

Case (i) Known ϕ and δ^2: When the covariance parameters are known, the posterior distributions of the unknown parameters can be expressed analytically. The posterior

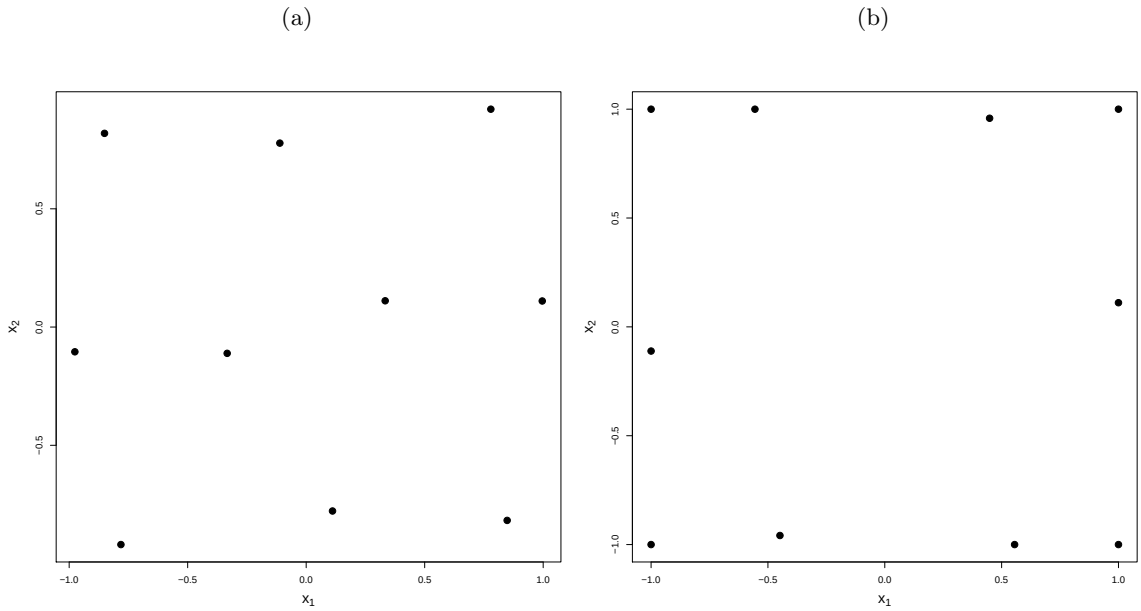


Figure 5.13: Ψ-optimal designs for prediction: (a) $\phi = 0.2$ and $\delta^2 = 0$, (b) $\phi = 0.2$ and $\delta^2 = 1$.

distribution $\pi(\boldsymbol{\beta}|\mathbf{y})$ is a multivariate t-distribution (2.18), with inverse gamma (2.19) for $\pi(\sigma^2|\mathbf{y})$, see Chapter 2.

The objective function $\Psi_1(\xi)$ when ϕ and δ^2 are known is derived in Section 5.3. Initially we find the optimal design by minimising Ψ (5.1). We considered two combinations of known values of ϕ and δ^2, $(\phi, \delta^2) = (0.2, 0)$ and $(\phi, \delta^2) = (0.2, 1)$. The first combination correspond to a correlation of 0.56 for observations at two points separated by the maximum Euclidean distance, $\sqrt{8}$. The second combination corresponds to a low correlation, i.e. 0.23 for distance $\sqrt{8}$. Figure 5.13 shows the optimal designs for the two combinations.

Firstly, using the optimal design for $(\phi, \delta^2) = (0.2, 0)$ (Figure 5.13 (a)), we simulate 100 independent data sets $\mathbf{y}_k : (k = 1, \ldots, 100)$ from the simulation model. For each simulated data set, we directly simulate 1000 values from the posterior distributions of the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Figures 5.14 and 5.15 show histograms of posterior samples, with the posterior densities, from the simulated data sets for $\delta^2 = 0$ and $\delta^2 = 1$, respectively. In this case, only the regression coefficients and the variance of the Gaussian process are unknown and are presented in plots (a), (b), (c) and (d).

Table 5.6 summarises these results in terms of marginal posterior means and 95% Highest Posterior Density (HPD) intervals for the unknown model parameters. For each parameter, the 95% HPD intervals include the true values from the simulation model.

		$\delta^2 = 0$	$\delta^2 = 1$
Parameter	True value	95% CI	95% CI
β_0	0	-0.6863, 0.8473	-1.5310, 1.3691
β_1	0	-0.1679, 0.5944	-0.9560, 1.0371
β_2	0	-0.5684, 0.2359	-1.0561, 0.8149
σ^2	1	0.3382, 1.5411	0.2528, 1.1316

Table 5.6: 95% Highest Posterior Density intervals for the parameters of the model fitted using the Ψ-optimal design for $(\phi, \delta^2) = (0.2, 0)$ and $(\phi, \delta^2) = (0.2, 1)$.

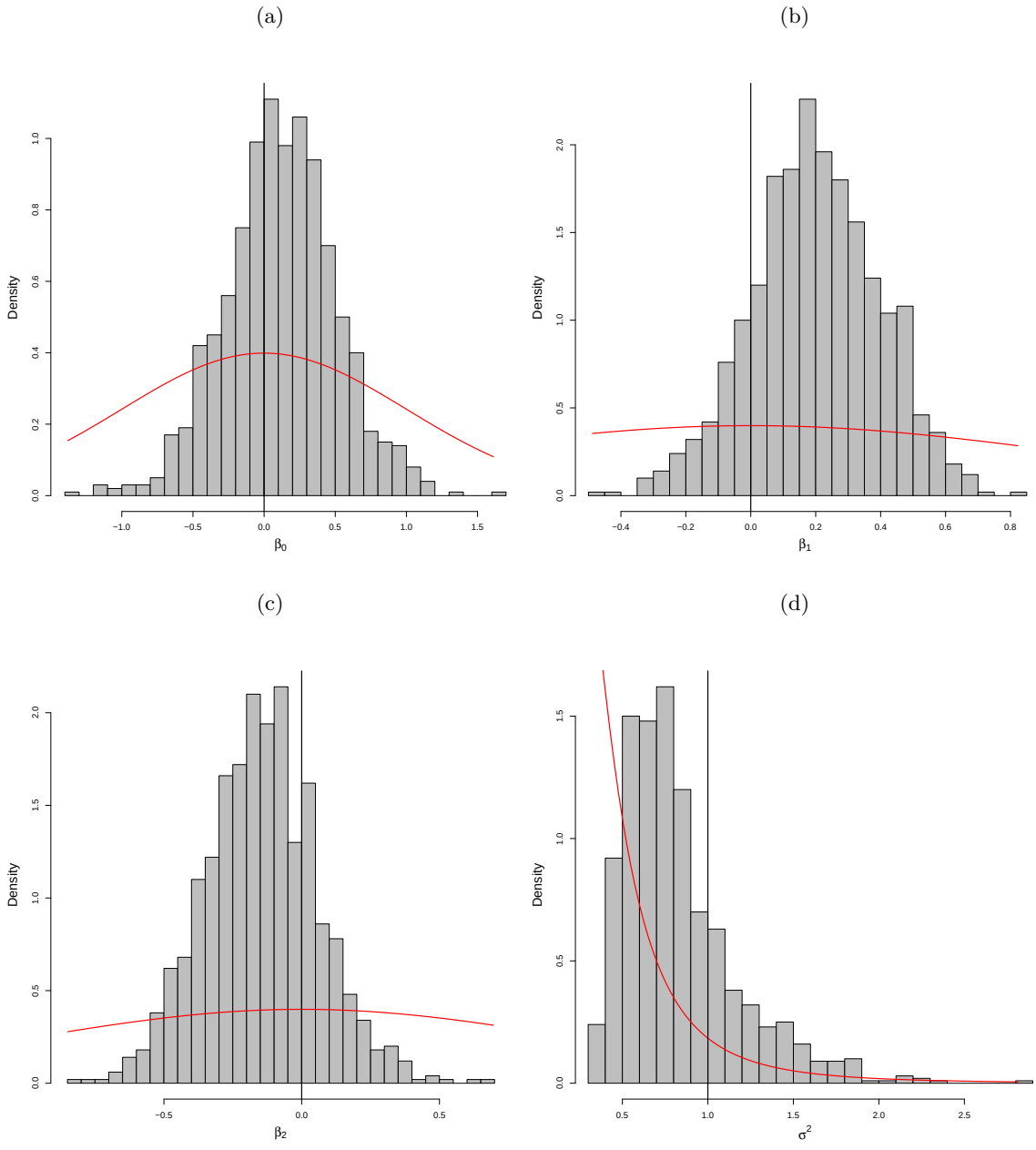


Figure 5.14: Posterior histograms of parameters for the Ψ-optimal design for $(\phi, \delta^2) = (0.2, 0)$ (a) β_0, (b) β_1, (c) β_2, (d) σ^2. In each figure, the red line represents the prior density and vertical black line the true value.

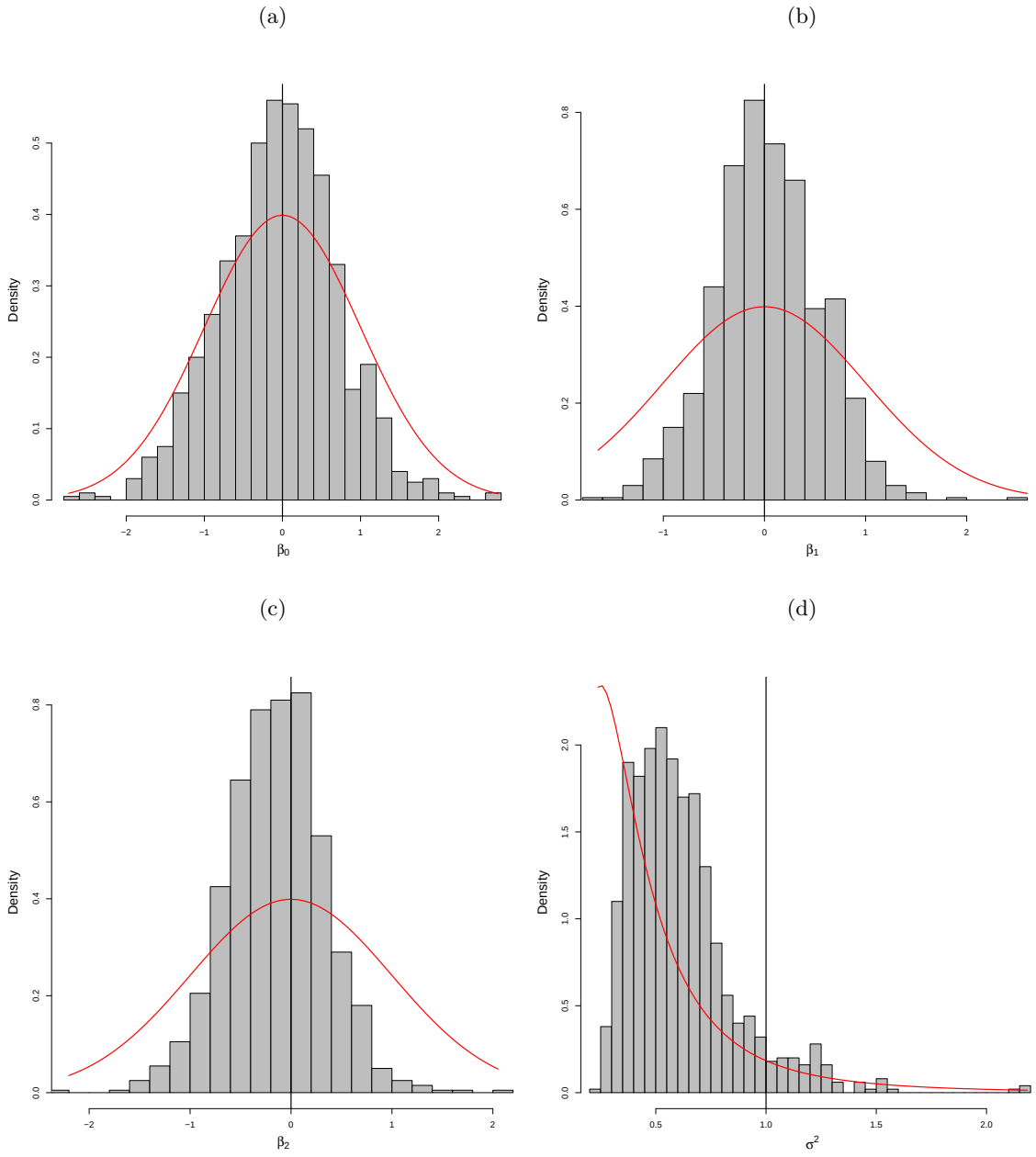


Figure 5.15: Posterior histograms of parameters for the Ψ-optimal design for $(\phi, \delta^2) = (0.2, 1)$ (a) β_0, (b) β_1, (c) β_2, (d) σ^2. In each figure, the red line represents the prior density and vertical black line the true value.

For each one of the 100 simulated data sets we found the posterior distributions for $\boldsymbol{\beta}$ and σ^2. Table 5.7 shows the average posterior mean and variance across the data sets for $\delta^2 = 0$ and $\delta^2 = 1$. We can see that the average posterior mean are close to the true value of the parameters. In particular, the posterior mean of the Gaussian process variance is substantially closer to the true value ($= 1$) than the prior mean ($= 0.5031$). However, the spread is large, as can be seen from the average posterior variance.

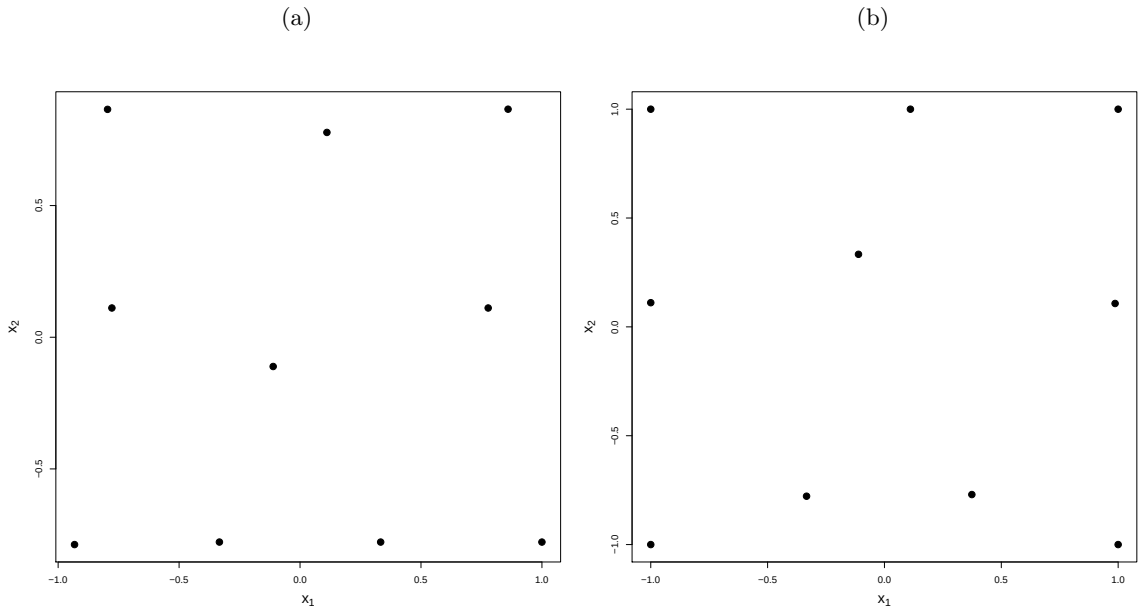


Figure 5.16: Ψ-optimal designs for prediction for unknown ϕ: (a) $\delta^2 = 0$ and (b) $\delta^2 = 1$.

Case (ii) Unknown ϕ and known δ^2: The Ψ-optimal design is found by minimising the objective function $\Psi_1(\xi)$ given by (3.33). The optimal designs for $\delta^2 = 0$ and $\delta^2 = 1$ are displayed in Figure 5.16. Again we generate 100 simulated data sets for each design using the simulation model and find the marginal posterior distributions of the unknown parameters for each data set.

For this case, there is no analytical form for the posterior distributions and hence inference must be performed by simulations. A uniform distribution on $[0.1, 1]$ is assumed for ϕ and samples from the posterior distributions of the unknown parameters are obtained using MCMC methods, employing similar procedure to the Metropolis with Gibbs algorithm described by Diggle and Ribeiro (2007):

1. Using a MH algorithm with log-normal proposal distribution (as ϕ is positive), we generate from the posterior distribution (2.32).

2. Given the sampled value of ϕ, we generate from the conditional distributions of $\boldsymbol{\beta}$ and σ^2, a t-distribution (2.18) and inverse gamma (2.19) distribution respectively.

3. We repeat the procedure until 1000 samples are taken from the marginal posterior distribution of ϕ, $\boldsymbol{\beta}$ and σ^2.

Case	Average posterior mean				Average posterior variance			
	β_0	β_1	β_2	σ^2	β_0	β_1	β_2	σ^2
$\delta^2 = 0$	-0.0313	0.0354	0.0258	0.7192	3.6329	0.9475	0.94801	0.1004
$\delta^2 = 1$	0.0149	0.0430	0.0215	0.6997	4.6013	2.0205	1.8986	0.09479

Table 5.7: Average posterior mean and variance across the 100 simulated data sets when ϕ and δ^2 are known.

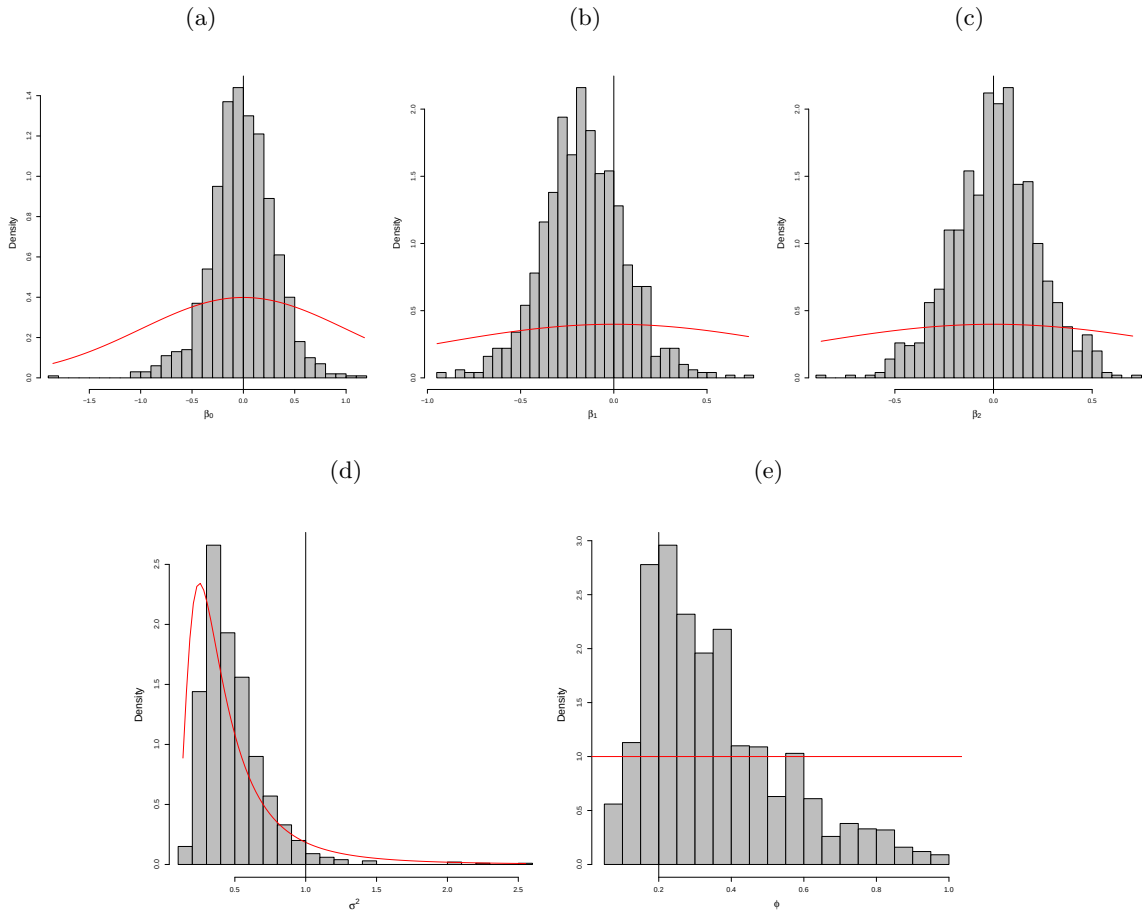


Figure 5.17: Posterior histograms of parameters for the Ψ-optimal design for unknown ϕ and $\delta^2 = 0$ (a) β_0, (b) β_1, (c) β_2, (d) σ^2, (e) ϕ. In each figure, the red line represents the prior density and vertical black line the true value.

Figures 5.17 and 5.18 show the posterior densities for the for $\delta^2 = 0$ and $\delta^2 = 1$. Once again, the plots of these posterior samples are based only one simulated dataset.

The shapes of the posterior densities for $\boldsymbol{\beta}$ and σ^2 resemble that of normal and inverse gamma densities respectively. Table 5.8 summarises the results in terms of 95% HPD intervals for the unknown model parameters. For both values of δ^2, the 95% HPD intervals include the true values of the unknown trend parameters. However, it is generally known that when a nugget effect is included in the model, it is difficult to estimate σ^2.

Table 5.9 shows the average posterior mean and variance across the 100 data sets. Again, the difficulty in estimating σ^2 when δ^2 is non-zero is clear; when $\delta^2 = 0$ the average posterior variance is 0.0362 whereas for $\delta^2 = 1$ is 0.0937. The noise-to-signal ratio and the variance are strongly related, as $\delta^2 = \tau^2/\sigma^2$, and for this reason there is a difference between the cases.

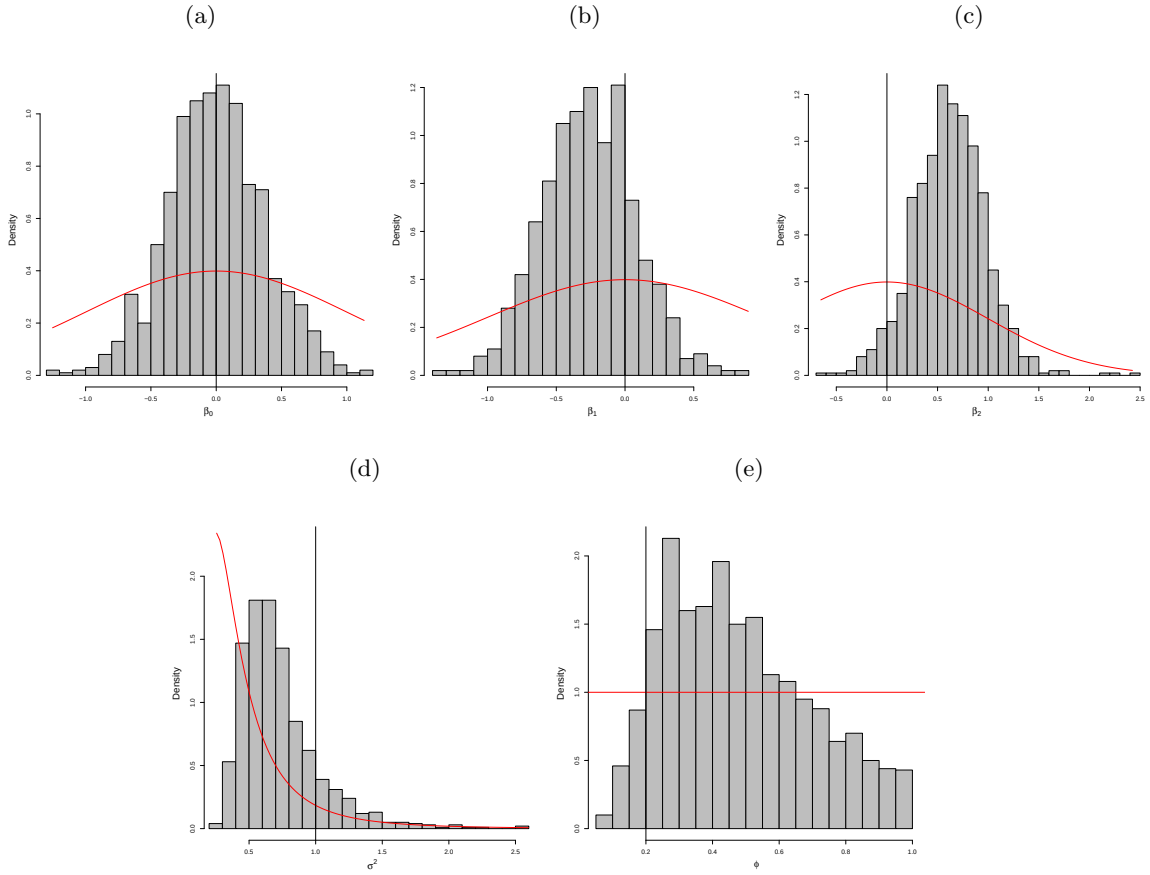


Figure 5.18: Posterior histograms of parameters for the Ψ-optimal design for unknown ϕ and $\delta^2 = 1$ (a) β_0, (b) β_1, (c) β_2, (d) σ^2, (e) ϕ. In each figure, the red line represents the prior density and vertical black line the true value.

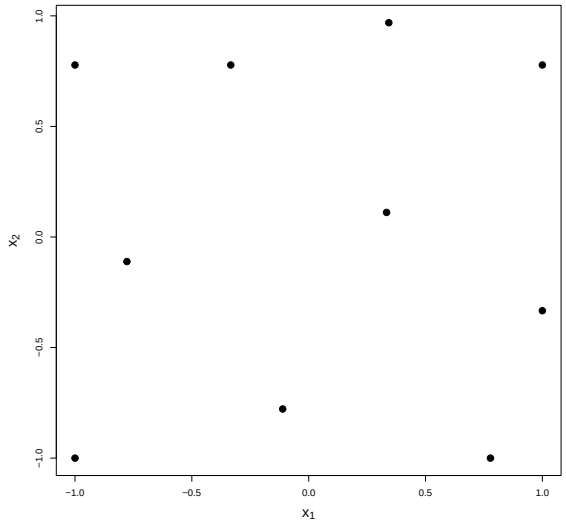


Figure 5.19: Ψ-optimal designs for prediction for unknown ϕ and δ^2.

Case (iii) Unknown ϕ and δ^2: Here we assign to δ^2 a uniform prior distribution on the interval $[0, 1]$. The Ψ-optimal design in Figure 5.19 was obtained by minimising the objective function $\Psi_1(\xi)$ given by (3.29). Using this design, we generate 100 data sets from the simulation model, as before.

To sample from the posterior distribution and make inference about the unknown parameters we follow a similar algorithm to that described for unknown ϕ and known δ^2. The only difference is that we employ a MH algorithm to sample from the joint posterior distribution of ϕ and δ^2, given by (2.34). We use a log-normal proposal distribution for ϕ and a uniform distribution for to propose values of δ^2. We then sample from the conditional distributions for $\boldsymbol{\beta}$ and σ^2. We repeat this procedure in order to obtain 1000 samples of $(\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)$.

Table 5.10 summarises the results in terms of posterior means and 95% HPD intervals for all the parameters of the model. With regard the covariance parameters, σ^2, ϕ and δ^2, the width of the credible intervals underline the difficulty of estimating these parameters precisely.

		$\delta^2 = 0$	$\delta^2 = 1$
Parameter	True value	95% CI	95% CI
β_0	0	-0.6470, 0.6055	-0.7508, 0.7128
β_1	0	-0.5646, 0.3164	-0.9195, 0.3665
β_2	0	-0.4809, 0.4069	-0.1329, 1.2383
σ^2	1	0.3270, 1.3714	0.2433, 1.2682
ϕ	0.2	0.0797, 0.7789	0.0977, 0.8775

Table 5.8: 95% Highest Posterior Density intervals for the parameters of the model fitted using the Ψ-optimal design for unknown ϕ and $\delta^2 = 0$ and $\delta^2 = 1$.

Case	Average posterior mean				
	β_0	β_1	β_2	σ^2	ϕ
$\delta^2 = 0$	-0.0322	-0.0461	0.0248	0.3895	0.5614
$\delta^2 = 1$	- 0.0104	0.0348	-0.0792	0.6703	0.5405
	Average posterior variance				
	β_0	β_1	β_2	σ^2	ϕ
$\delta^2 = 0$	0.13523	0.0788	0.0688	0.0362	0.0403
$\delta^2 = 1$	0.2654	0.1961	0.19115	0.0937	0.0527

Table 5.9: Average posterior mean and variance across the 100 simulated data sets when ϕ is unknown and δ^2 is known.

		$\delta^2 = 0$	$\delta^2 = 1$
Parameter	True value	95% CI	95% CI
β_0	0	-0.7145, 0.3855	-0.5693, 0.7548
β_1	0	-0.2838,0.4599	-0.3895, 0.6649
β_2	0	-0.4659, 0.2715	-0.1847, 0.8709
σ^2	1	0.1785, 0.9384	0.5235, 2.5182
ϕ	0.2	0.1130, 0.8813	0.2233, 0.9834
δ^2	0/1	0.0213, 0.9431	0.2943, 0.9989

Table 5.10: 95% Highest Posterior Density intervals for the parameters of the model fitted using the Ψ-optimal design for unknown ϕ and δ^2.

Case	Average posterior mean					
	β_0	β_1	β_2	σ^2	ϕ	δ^2
$\delta^2 = 0$	-0.0944	-0.0671	-0.0049	0.3059	0.4655	0.3844
$\delta^2 = 1$	0.0733	-0.0163	-0.0424	0.8361	0.5650	0.6799
	Average posterior variance					
	β_0	β_1	β_2	σ^2	ϕ	δ^2
$\delta^2 = 0$	0.1172	0.0583	0.0587	0.0228	0.0470	0.0664
$\delta^2 = 1$	0.3123	0.1998	0.1998	0.2068	0.0524	0.0528

Table 5.11: Average posterior mean and variance across the 100 simulated data sets when ϕ and δ^2 are unknown.

Histograms of the posterior samples for each parameters for one simulated data set are presented in Figures 5.20 ($\delta^2 = 0$) and 5.21 ($\delta^2 = 1$) respectively.

Table 5.11 shows the average of the posterior mean for all the unknown parameters across the 100 generated data sets. Similarly to the previous case of unknown ϕ and fixed δ^2, we again see that the covariance parameters are difficult to estimate.

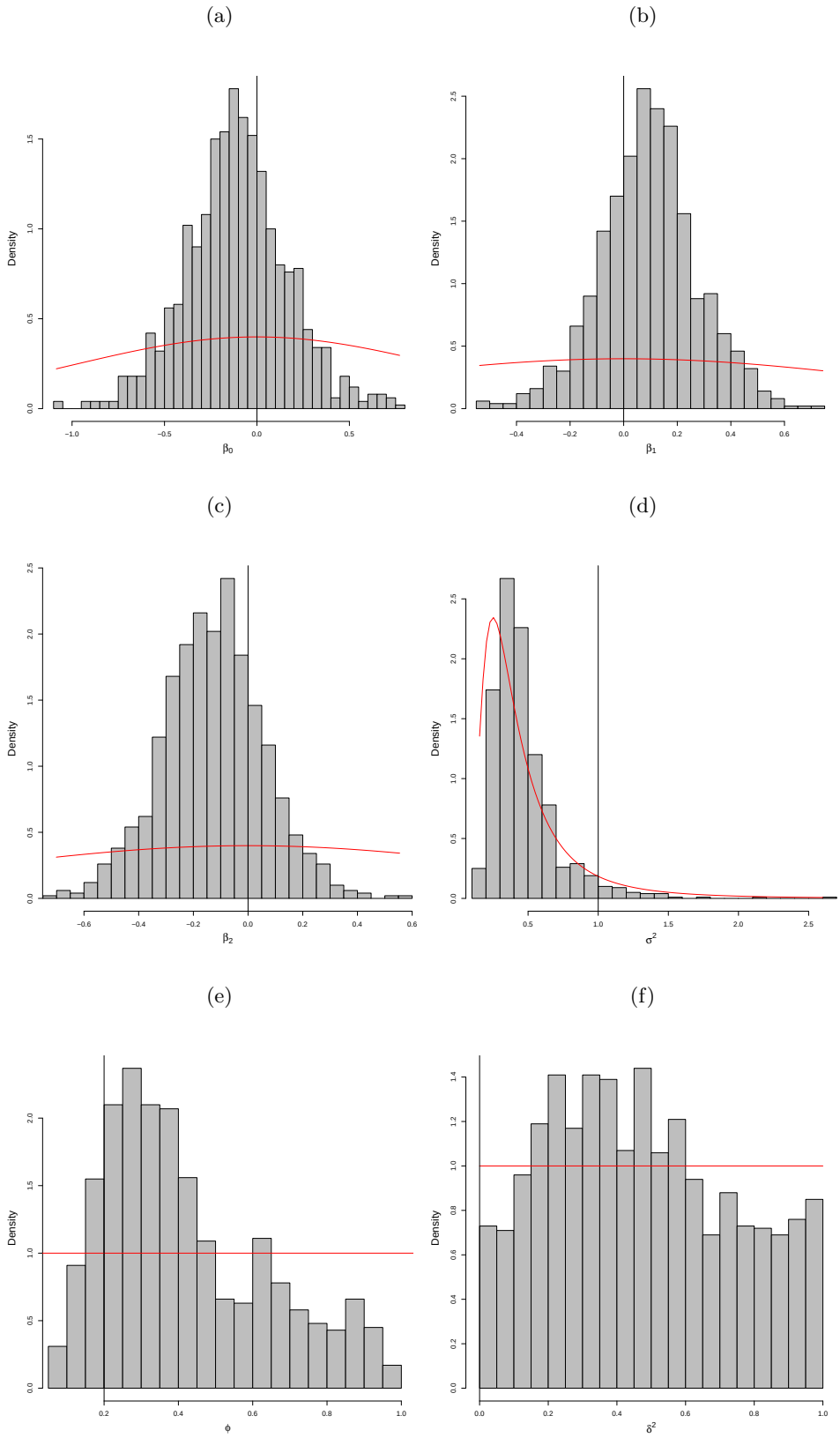


Figure 5.20: Posterior histograms of parameters for the Ψ-optimal design for unknown ϕ and δ^2 (a) β_0, (b) β_1, (c) β_2, (d) σ^2, (e) ϕ, (f) δ^2. In each figure, the red line represents the prior density and vertical black line the true value.

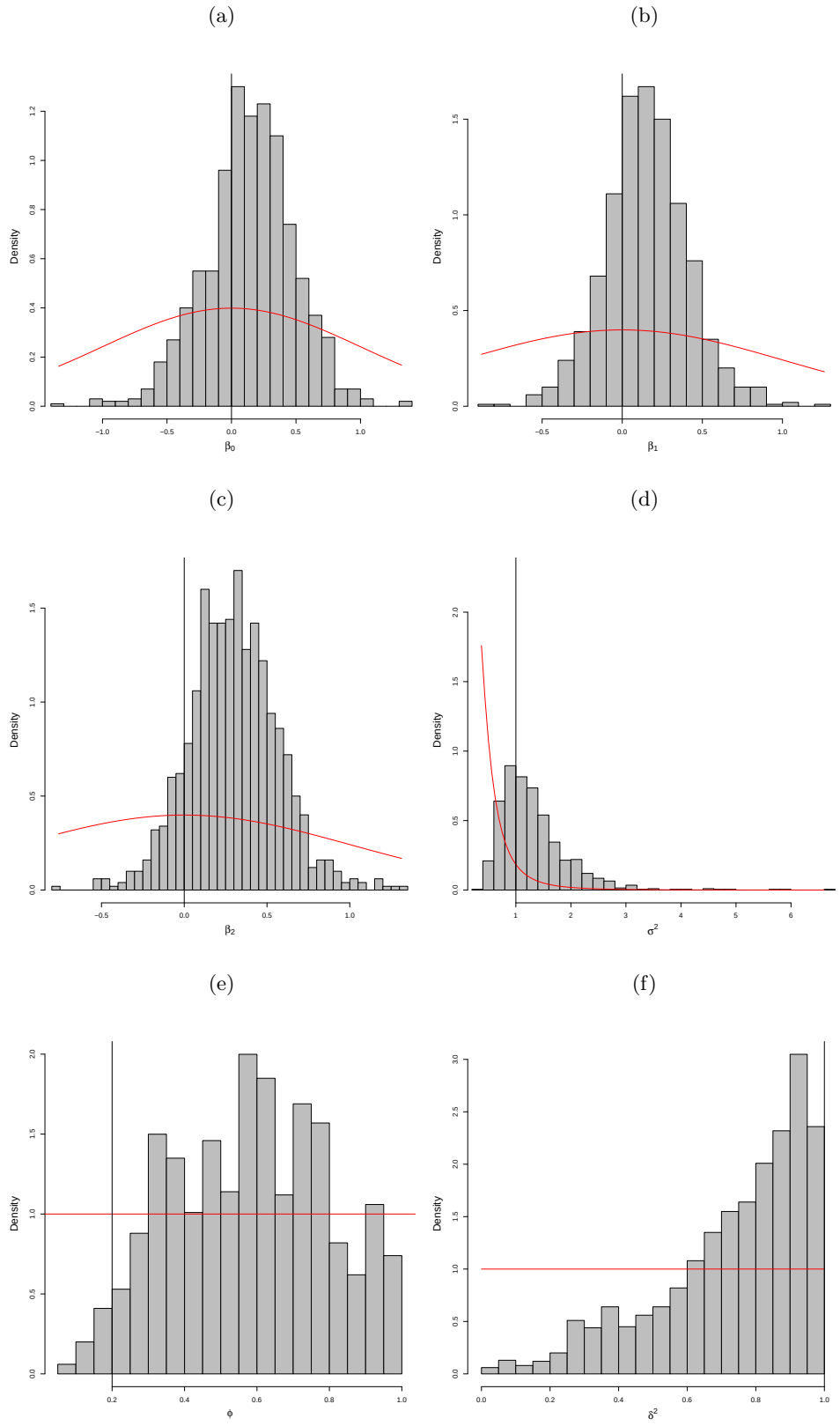


Figure 5.21: Posterior histograms of parameters for the Ψ-optimal design for unknown ϕ and δ^2 (a) β_0, (b) β_1, (c) β_2, (d) σ^2, (e) ϕ, (f) δ^2. In each figure, the red line represents the prior density and vertical black line the true value.

5.6 Comparison With Existing Designs

The majority of designs for spatial data in the existing literature were found using a frequentist approach with either the covariance parameters assumed known a priori or estimated using likelihood methods and then their estimated values plugged into the design objective function.

The Bayesian approach has not been very popular due to computational issues. However, Diggle and Lophaven (2006) proposed a Bayesian design criterion which minimised the averaged prediction variance, $\Psi(\xi)$ (3.7), similar to our approach. The difference between their approach and our approach is that we propose an approximation to the objective function $\Psi(\xi)$ (3.7) and, hence if we use conjugate priors we are able to integrate out the data. The approximation allow us to optimise the objective function and find optimal designs. Rather that find optimal designs under $\Psi(\xi)$, Diggle and Lophaven (2006) assumed two classes of designs, regular lattice, i.e. a set of points that are equally spaced in the study region, augmented with close pairs or infill points.

1. They defined the lattice plus close pairs design as a design which consists of locations in a regular $p \times p$ lattice together with a further m points, each of which is located uniformly at random within a disc of radius a whose centre is a randomly selected point of the lattice. They use the notation $(p \times p, m, a)$.

2. They defined the lattice plus infill design as a regular $p \times p$ lattice together with further locations in a more finely spaced $r \times r$ lattice within m randomly chosen cells of the primary lattice. Hence, $r^2 - 4$ additional points are added in the initial lattice. Their notation for design is $(p \times p, m, r \times r)$.

Diggle and Lophaven (2006) stated that the exact choice of close pairs or infill pairs has only a small impact on the Bayesian objective function.

In this section we illustrate the efficiency of our Bayesian Ψ-optimal designs relative to (i) a regular lattice, (ii) a lattice plus close pairs designs (LPCPD) and (iii) lattice plus infill designs (LPIFD), when the total number of points is $n = 36$. Specifically, we use the regular lattice 6×6, the $(4 \times 4, 20, 0.5)$ lattice plus close pairs design and the $(4 \times 4, 4, 3 \times 3)$ lattice plus infill design with our Bayesian optimal design found by minimising the objective function $\Psi_1(\xi)$, (3.9). All the designs were constructed on the unit square, i.e. $\mathcal{X} = [0, 1]^2$ with $\mathcal{X}_{\mathcal{P}}$ a 10×10 regular grid. We compare the performance of these four designs under our objective function.

The lattice plus close pairs design $(4 \times 4, 20, 0.5)$ and the lattice plus infill design $(4 \times 4, 4, 3 \times 3)$ vary because of the random selection of the additional locations and for this reason we average the objective function over five independent replicates. Figure 5.22 shows examples of a $(4 \times 4, 20, 0.5)$ lattice plus close pairs design and a $(4 \times 4, 4, 3 \times 3)$ lattice plus in-fill design.

We consider the Gaussian process model (2.6) with two cases of mean function, constant

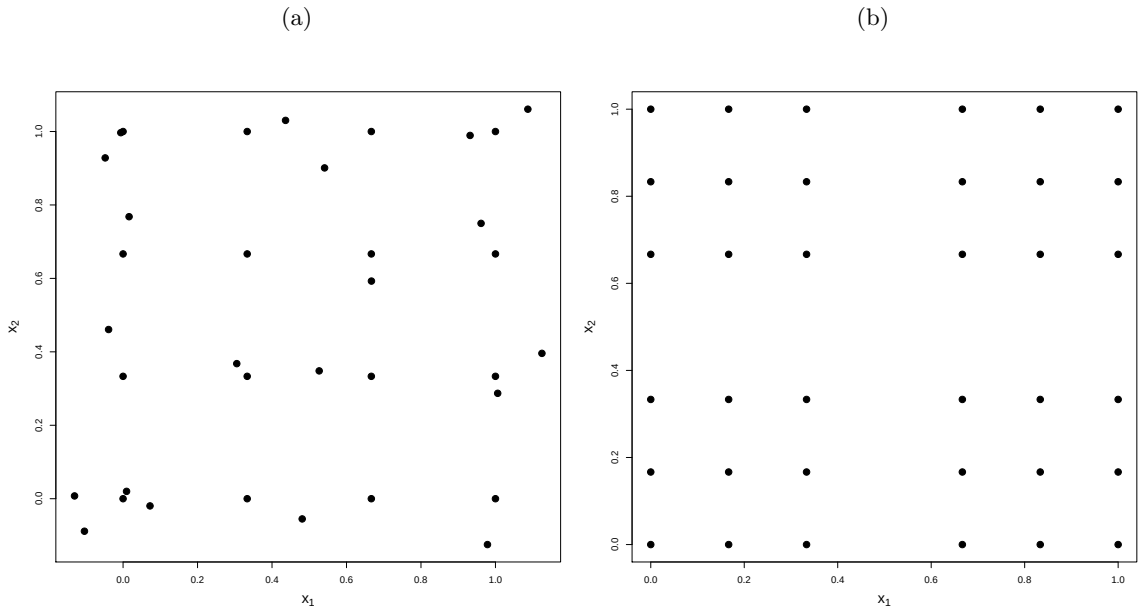


Figure 5.22: Examples of a (a) $(4 \times 4, 20, 0.5)$ lattice plus close pairs design, and (b) $(4 \times 4, 4, 3 \times 3)$ lattice plus in-fill design.

and linear, and exponential correlation function. The exponential correlation function is chosen as it is the function used by Diggle and Lophaven (2006). In addition, the correlation in the region $\mathcal{X} = [0, 1]^2$ is very high for the Matérn $\nu = 1.5, 2.5$, resulting in near singular correlation matrices and computational issues. The prior specification for the model parameters is as follows. For ϕ we used a uniform prior distribution $(0.1, 1.5)$ which corresponds to range of correlation from 0.22 to 0.90 for a pair of points at the maximum distance. We also chose a uniform prior distribution $(0.07, 0.4)$ which correspond to a narrower range of correlation from 0.67 to 0.99. Diggle and Lophaven (2006) considered for ϕ a uniform prior distribution $(0.4, 10000)$ which corresponds to a range of correlation from 0.65 to $\simeq 0$. For δ^2 we used a uniform prior distribution on $(0, 1)$, following Diggle and Lophaven (2006). For σ^2 and $\boldsymbol{\beta}$ a normal-inverse gamma conjugate prior distribution is assumed, in contrast to the diffuse prior which was used by Diggle and Lophaven (2006). In order to approximate their diffuse prior, we assume $\beta \sim N(0, 1000\sigma^2)$ for a model with a constant mean and $\boldsymbol{\beta} \sim N(0, \mathbf{R}^{-1}\sigma^2)$ with $\mathbf{R}^{-1} = 1000\mathbf{I}_3$ for a model with linear trend. Similar to preceding sections, we choose $\sigma^2 \sim \text{IG}(3, 1)$.

Initially, we found the Bayesian optimal designs for each prior specification by minimising the objective function $\Psi_1(\xi)$ in (3.9). The method for selecting an optimal Bayesian design for this example is as follows. We generate 30 randomly selected starting designs from $\mathcal{X} = [0, 1]^2$ with 36 points. Then for each starting design we use the coordinate exchange algorithm to find a design that minimises $\Psi_1(\xi)$. From the 30 designs obtained we select the design with the minimum $\Psi_1(\xi)$ value.

In order to compare the designs we evaluate the objective function $\Psi(\xi)$, (5.1), for each

of the four designs (Ψ-optimal, regular lattice, LPCPD, LPIFD) for all combinations of $\phi = 0.2, 0.4, 0.6, 0.8, 1$ and $\delta^2 = 0, 0.2, 0.4, 0.6, 0.8, 1$.

Figures 5.23 and 5.24 compare the performance of the four designs for constant and linear trend respectively, with the regular 6×6 lattice, $(4 \times 4, 20, 0.5)$ lattice plus close pairs, and $(4 \times 4, 4, 3 \times 3)$ lattice plus infill design. For both constant and linear mean function the Bayesian Ψ-optimal design performs better than the other three designs, giving the smallest value of the objective function for all combinations of ϕ and δ^2.

If we compare Figures 5.23 and 5.24 with similar results from Diggle and Lophaven (2006), we see that we do not obtain the same ordering of designs. Those authors found that the LPCPD gave lower values of the objective function compared with LPIFD and regular lattice designs. The reason for the difference to our results is due to the assumed correlation parameters. In the region $[0, 1]^2$ the maximum distance between points is $\sqrt{2}$ and for the values of $\phi = 0.2, 0.4, 0.6, 0.8, 1$ the correlation is $0.753, 0.568, 0.428, 0.326, 0.243$, averaged across δ^2, at this distance. The correlation parameters assumed by Diggle and Lophaven (2006) correspond to very low correlation, i.e. $0.008, 0.029, 0.094, 0.17, 0.243$.

The highest correlation considered in Diggle and Lophaven (2006) corresponds roughly to the lowest we considered. We use different correlation parameters compared to Diggle and Lophaven (2006) to get more interesting designs, as the main objective is to study how high correlation affects the choice of optimal design points.

These results are in line with our findings in previous sections. If we assume a constant mean, a Ψ-optimal design has points uniformly spread across the study region, see Section 5.4.4 and Figures A.9-A.14. Hence in Figure 5.23, the Bayesian Ψ-optimal design and the regular lattice give lower values of the objective function than the other two designs. However, when a linear trend is assumed, the designs are influenced by the need to estimate the trend parameters and design points can be close together, especially when the correlation is lower. For this reason, in Figure 5.24 (d) and (f), the LPCPD which has points very close together, gives smaller values of the objective function that the LPIFD.

Based on our evidence that the Bayesian optimal design is sensitive to the range of the correlation, (Section 5.4.4) we also considered a second prior on $\phi \sim$Unif$(0.07, 0.4)$, which corresponds to a smaller range of prior values. The comparison between the Bayesian optimal design and the other three designs is displayed in Figure 5.25 for constant mean and in Figure 5.26 for the linear mean function.

The ranks of the LPCPD, LPIFD and lattice design is the same here as for the first prior distribution on ϕ. The difference between Figures 5.23–5.26 is only for the Bayesian Ψ-optimal design. In the second case the Bayesian Ψ-optimal design yields objective function values very close values to these of lattice design because of smaller range of correlation. As this prior distribution indicates higher correlation, the best design in

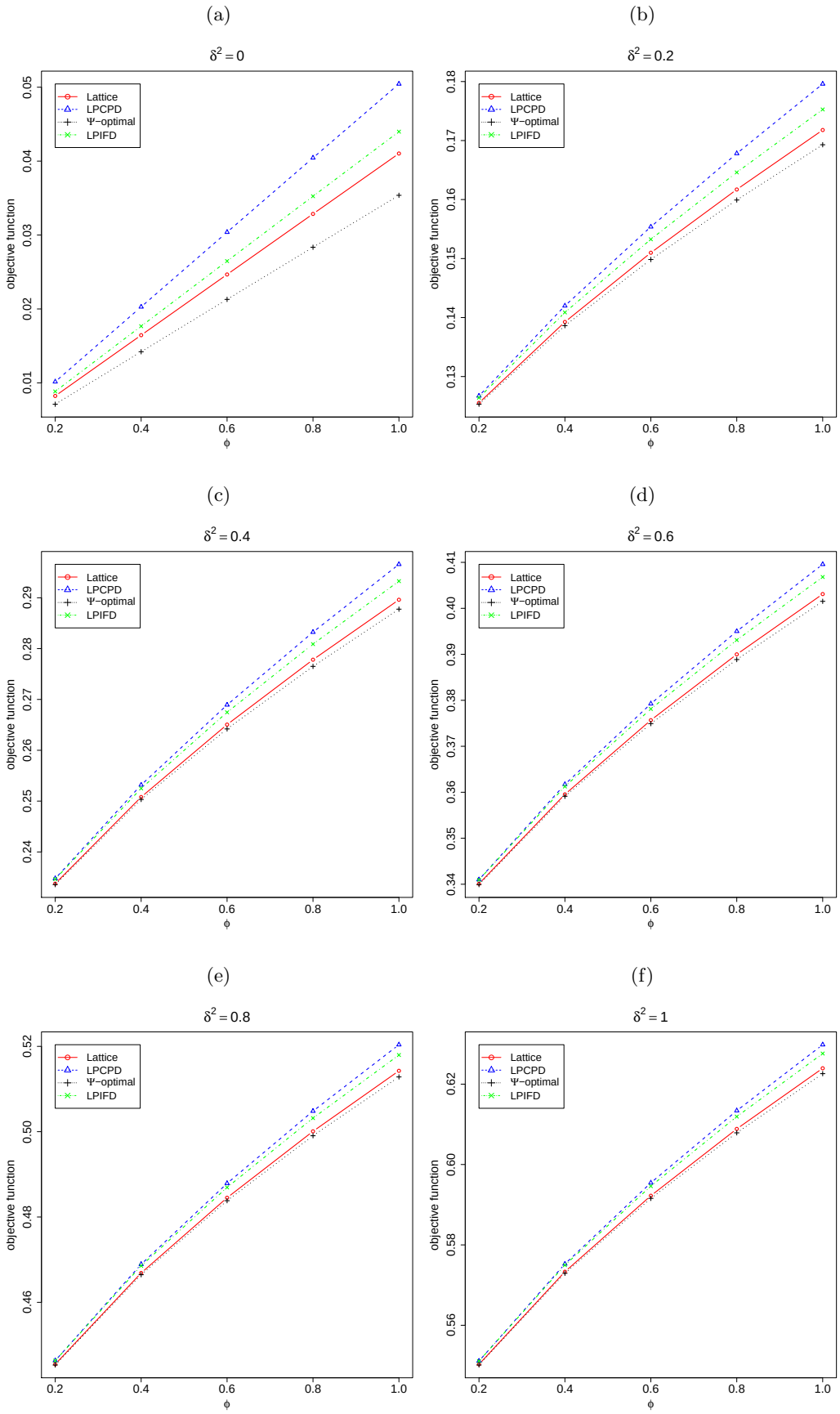


Figure 5.23: Comparison under objective function Ψ (5.1) of the Ψ-optimal design for constant mean, the regular 6×6 lattice, the $(4 \times 4, 20, 0.5)$ LPCPD, the $(4 \times 4, 4, 3 \times 3)$ LPIFD for $\phi \sim$Unif$(0.1, 1.5)$.

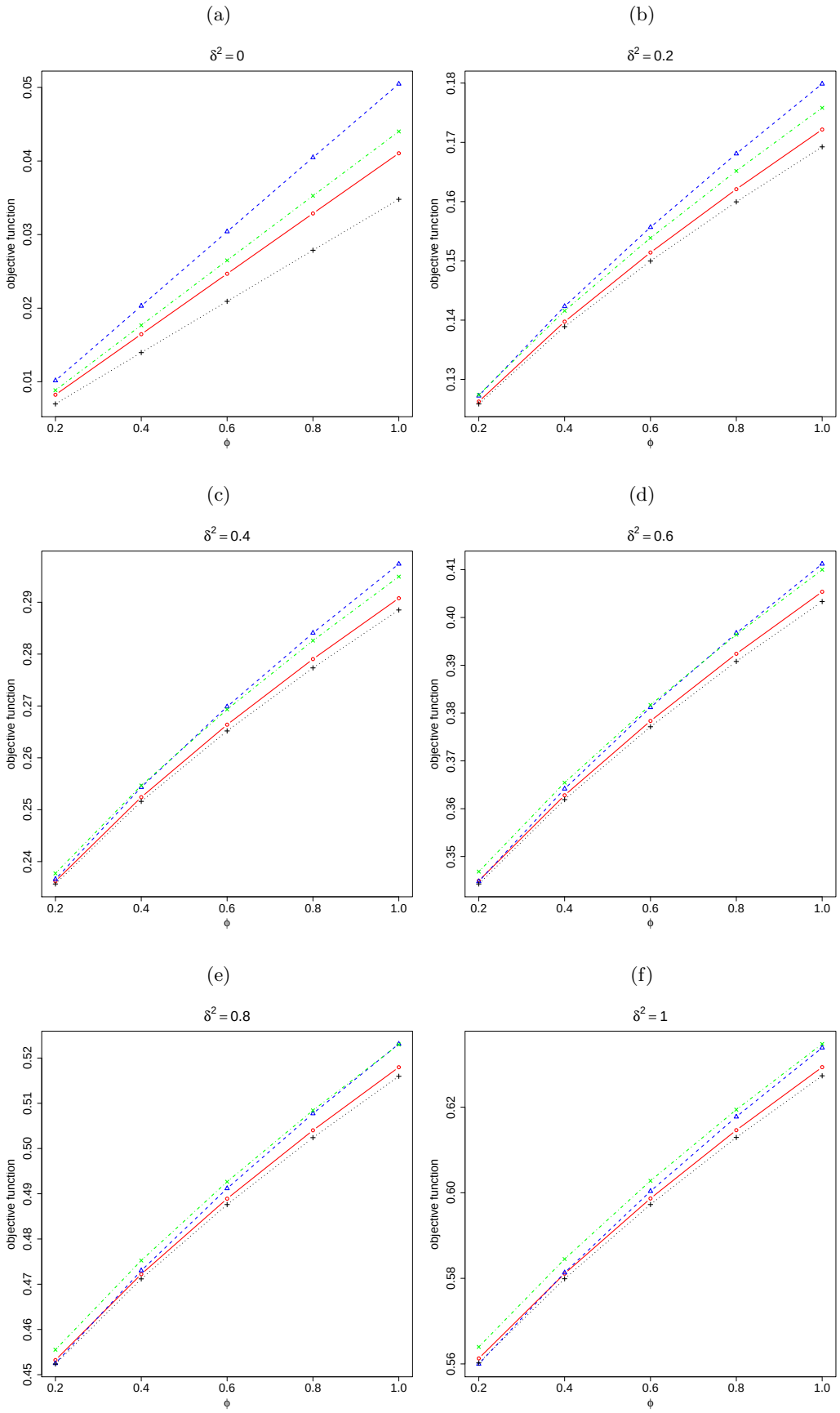


Figure 5.24: Comparison under objective function Ψ (5.1) of the Ψ-optimal design for linear mean, the regular 6×6 lattice, the $(4 \times 4, 20, 0.5)$ LPCPD, the $(4 \times 4, 4, 3 \times 3)$ LPIFD for $\phi \sim$Unif$(0.1, 1.5)$.

Design	$\phi \sim \text{Unif}(0.1, 1.5)$		$\phi \sim \text{Unif}(0.07, 0.4)$	
	Constant mean	Linear mean	Constant mean	Linear mean
Ψ-optimal	0.3267	0.3287	0.2872	0.2907
Lattice	0.3283	0.3308	0.2880	0.2922
LPCPD	0.3337	0.3357	0.2893	0.2926
LPIFD	0.3316	0.3345	0.2890	0.2940

Table 5.12: Evaluation of the objective function $\Psi_1(\xi)$ (3.29) when ϕ and δ^2 are unknown for Ψ-optimal, lattice, LPCPD, LPIFD designs.

this case pushes points as far apart as possible as two points close together will provide highly correlated observations and provide very similar information for predictions at a nearby location.

Up to this point, we have compared designs by evaluating the average prediction variance at combination of values for ϕ and δ^2. In Table 5.12 we evaluate the objective function $\Psi_1(\xi)$ (3.29) for ϕ and δ^2 both unknown. This objective function is approximated by quadrature methods as it described in Section 3.5.1. Again, the Bayesian Ψ-optimal design for both constant and linear trend results to lower values of the objective function $\Psi_1(\xi)$ (3.29) compared to lattice, LPCPD and LPIFD designs.

5.7 Summary and Discussion

In this chapter we applied a Bayesian optimality criterion for spatial experiments that minimises the average prediction variance. Numerical search is used to find the optimal design, employing the coordinate exchange algorithm. Our main contribution to the area of the spatial design is to consider all the model parameters unknown and follow a fully Bayesian approach to design.

Although our approach is less computationally expensive than, for example, Monte Carlo evaluation of the objective function, searching for the optimal design can still be very hard. The coordinate exchange algorithm and the optimisation method used in this thesis allows some design points to be very close together, which leads to an almost singular correlation matrix, especially when there is no nugget in the model. A possible solution to avoid this numerical complication is to add an extra step in our coordinate exchange algorithm to prevent points being placed within a distance of specific radius from another point.

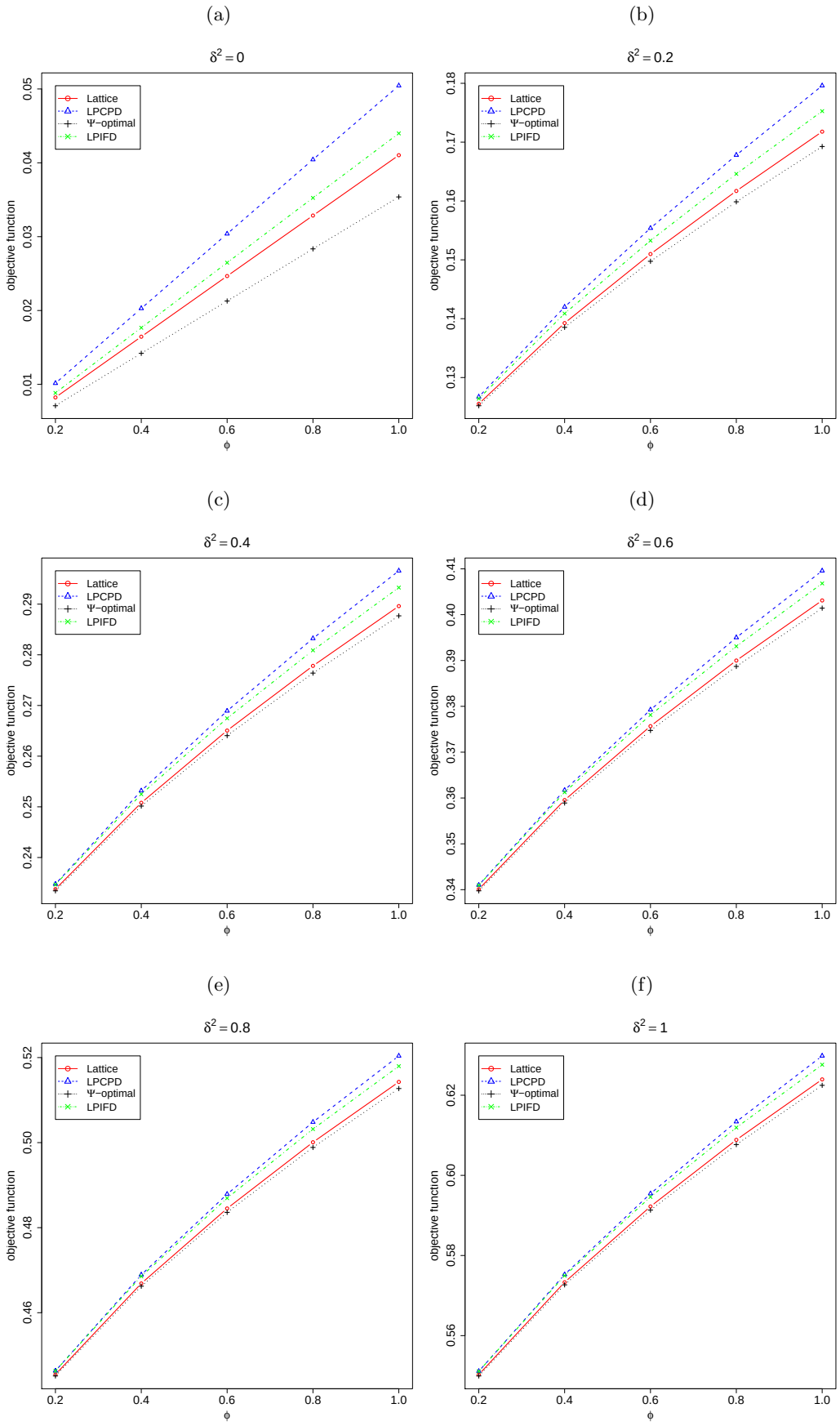


Figure 5.25: Comparison under objective function Ψ (5.1) of the Ψ-optimal design for constant mean, the regular 6×6 lattice, the $(4 \times 4, 20, 0.5)$ LPCPD, the $(4 \times 4, 4, 3 \times 3)$ LPIFD for $\phi \sim$Unif$(0.07, 0.4)$.

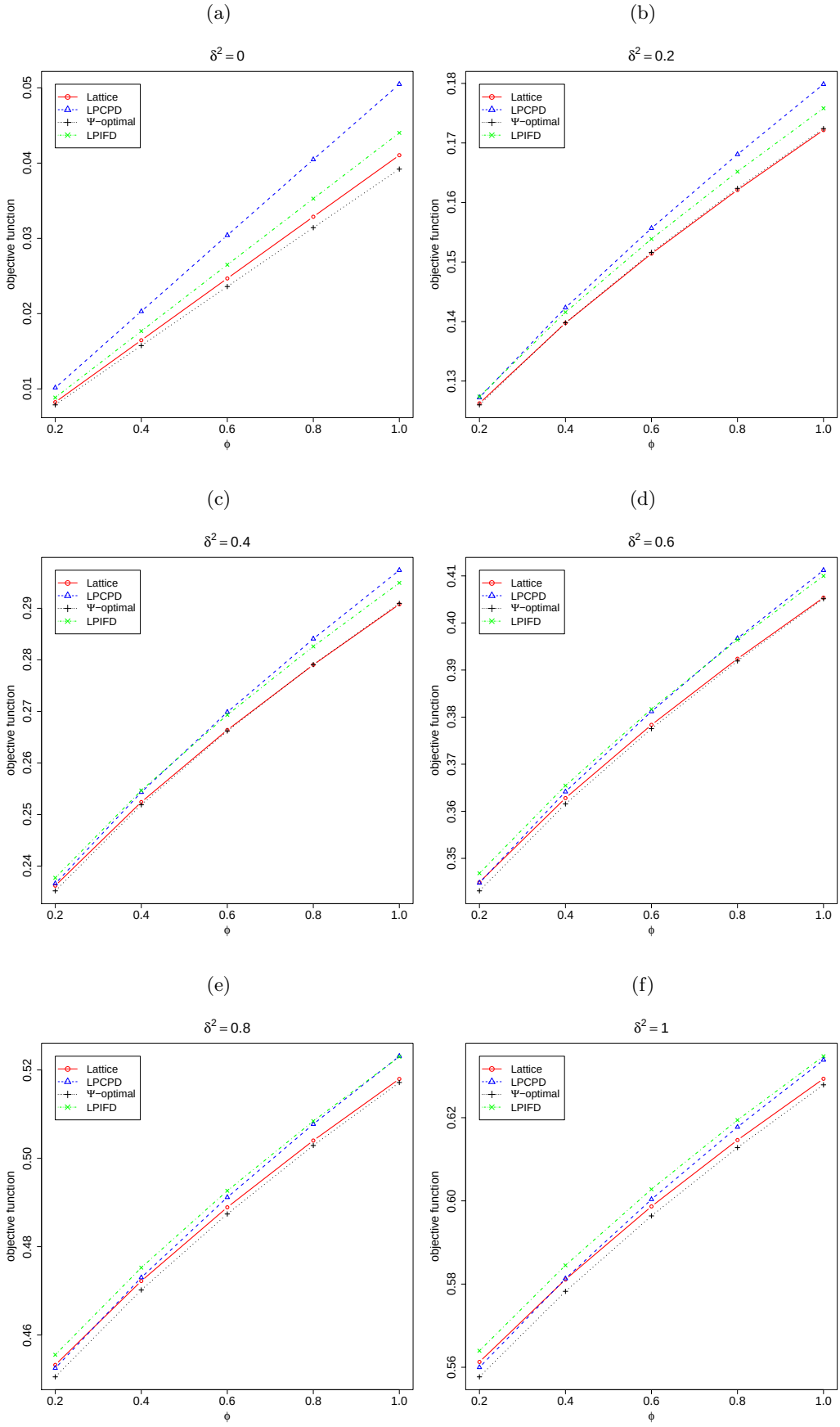


Figure 5.26: Comparison under objective function Ψ (5.1) of the Ψ-optimal design for linear mean, the regular 6×6 lattice, the $(4 \times 4, 20, 0.5)$ LPCPD, the $(4 \times 4, 4, 3 \times 3)$ LPIFD for $\phi \sim$Unif$(0.07, 0.4)$.

Large spatial datasets pose computational challenges to the application of a Gaussian process model. In particular, estimation and prediction involve inversion of an $n \times n$ covariance matrix for a dataset of size n, which can be computationally intractable for large datasets. We briefly review three recently developed approaches for modelling large spatial data sets that have shown promise as general methodologies to overcome this obstacle.

The first approach is based on a reduced rank approximation of the Gaussian process. Banerjee et al. (2008) proposed a method to reduce the dimensionality of a Gaussian process model when the aim is spatial prediction. The idea here is the spatial information available from the dataset observed at all the locations can be summarised from a smaller but representative set of locations, called the knots. The lower dimensional subspace is chosen by the user by selection a set of knots where the parent process is defined instead of the original process. The parent process is realisations of the Gaussian process at the knots. The best linear unbiased prediction of the Gaussian process at any fixed point based on the parent process is the predictive process, which is defined as a kriging interpolator and has a covariance function that is completely specified by the parent covariance function. A drawback of this approach is that it fails to capture the small scale dependence accurately. Finley et al. (2009) discussed this limitation of the reduced rank method and also indicated the importance of knot design, that is how the set of knots is chosen.

The second approach for large datasets was proposed by Kaufman et al. (2008) where the covariance function is tapered and a sparse covariance matrix approximation to covariance matrix is constructed. This method introduces zero covariance for distant pairs of observations and then efficient sparse matrix techniques can be applied. In contrast to reduced rank methods, which account for the large scale variation effectively, the covariance tapering method of Kaufman et al. (2008) may fail to capture the large scale variation and may limit modelling flexibility. For this reason it is important to choose the mean function since now all the large scale variation is captured through the mean function. However, this method is very effective for small scale variation.

Sang and Huang (2012) proposed a third approach which combined the two approaches of Banerjee et al. (2008) and Kaufman et al. (2008) called full scale approximation of the covariance function. Using both a reduced rank representation and tapered covariance function, it captures both the large scale and small scale variations. For this approach we need to specify two parameters, the number of knots and the taper range.

These three approaches have been applied in the statistical analysis and modelling of large spatial datasets. Inversion of the covariance matrix is repeated many times in our exchange algorithm when finding an optimal design, and hence extension of work in this chapter could be to apply these three methods to the problem of Bayesian optimal designs to reduce the computational burden.

Another, related avenue of future work could be to extend our method to choose the

optimal set of knots for the reduced rank method. Banerjee et al. (2008) and Finley et al. (2009) indicated that the selection of knots is a challenging problem, and considered standard space filling designs and the model-based approach introduced by Diggle and Lophaven (2006). Recently, Gelfand et al. (2013) indicated that a model-based approach for knot selection is preferable to standard space-filling techniques as it incorporates the dependence structure into the knot design. Both Finley et al. (2009) and Gelfand et al. (2013) proposed an algorithm to find the optimal knot design using as a design criterion the minimisation of the average prediction variance.

Gelfand et al. (2013) performed simulation studies to explore how the size and the configuration of the set of knots affected model fitting. They investigated a space filling design, the lattice plus close pairs designs and lattice plus infill designs of Diggle and Lophaven (2006) and an optimal design minimising the average prediction variance. They concluded that the optimal design provide improved model fit and prediction over the other designs. Therefore an extension of our work with regard to knot selection is to use our closed-form approximation $\Psi_1(\xi)$ to find the optimal set of knots. The advantage of our approach is that instead of using a sequential search algorithm to find an optimal designs, we are going to employ the coordinate exchange algorithm and optimise.

One of the assumptions we have made about the Gaussian process model is that a stationary isotropic correlation function describes the correlation between observations at two locations. However, we can extend the stationary correlation functions to anisotropic correlation functions where the spatial correlation between two observations depends upon the separation vector and not only on its length. In general anisotropy is difficult to deal with but there are some cases which it is tractable see for example Banerjee et al. (2004) and reference therein. The most popular case is the geometric anisotropy where the coordinate space can be linearly transformed to an isotropic function, more details can be found in Banerjee et al. (2004). Using an anisotropic correlation function we have to define more parameters, and then from the Bayesian prospective to assign prior to those parameters as well. Thus the complexity of the Gaussian process model increases and the problem of finding Bayesian optimal designs is more complicated. In Chapter 7 we use separable anisotropic correlation functions and allow different correlation parameters in each dimension.

In many real examples of collecting spatial data the stationarity assumption may also violated. Banerjee et al. (2004) presented approaches for nonstationary spatial process models. In the context of spatial design, Fuentes et al. (2007) employed a non-stationary covariance function and proposed an entropy-based design criterion based on evaluating the posterior predictive entropy. An alternative solution to this problem can be the use of Bayesian treed Gaussian process models, proposed by Gramacy and Lee (2008), where the region of interested is partitioned and a stationary Gaussian process model is fitted to the data in each partition. The latter approach can directly apply to our Bayesian design criterion and is an area for future research.

Finally, moving away from the Gaussian framework increases the complexity of finding optimal designs. Recently, Evangelou and Zhu (2012) considered the case of augmenting an existing design for discrete data by minimising the average prediction variance. They proposed an approximation to the posterior predictive variance as the prediction variance is not analytically tractable even for the case of known correlation parameters. Questions related to the design problems when the correlation parameters are unknown have yet to be addressed and also the problem has yet to be viewed from a Bayesian perspective.

Chapter 6

Application of Spatial Design to Monitoring Chemical Deposition

The main objective of this chapter is to demonstrate the design methodology from Chapter 3 on a real environmental example. We start with the monitoring network in the eastern USA as described in Section 1.1.1 and consider the deletion of locations from the network of 122 stations. The network measures chemical deposition, including pollutants such as sulphur dioxide, nitrogen oxides and heavy metals. These chemicals are emitted to the air, transformed to acid and return to the Earth through wet deposition. In the eastern USA, this chemical deposition is mainly because of large fossil fuel power plants.

We distinguish between two design situations for reducing the existing network: prospective design, where we find the optimal design in advance of data collection, and retrospective design where we incorporate data from the existing monitoring network to find an optimal design.

6.1 Introduction

The monitoring dataset contains deposition data measured at 122 stations irregularly placed over the eastern USA. A map of the region with the respective locations of the 122 measurement stations is displayed in Figure 6.1 together with the 10 sites at which prediction is required. The available data give the measurements of weakly deposition for the 52 weeks in the year 2001.

Boxplots of the weekly sulphate deposition levels (kilograms per hectare) are plotted in Figure 6.2. The plots confirm the intuitively obvious fact that deposition levels are higher on average for the wetter spring and summer months than the dryer winter months; see for example, Brook et al. (1995).

The average weekly wet sulphate deposition at these stations for the year 2001 yields

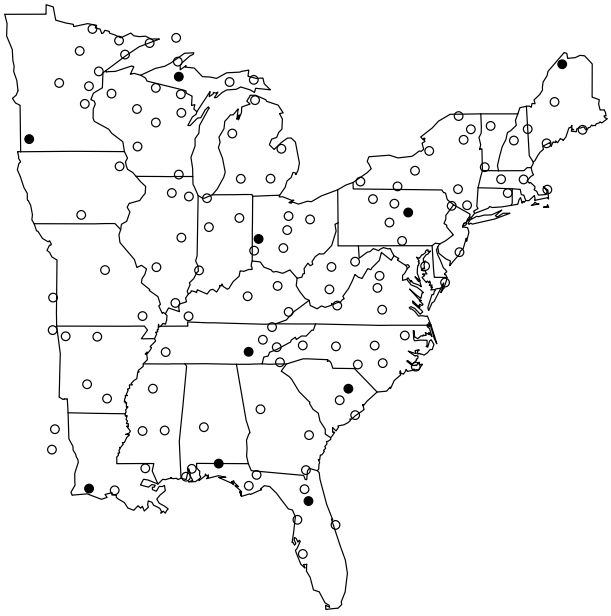


Figure 6.1: Monitoring locations (○) of the chemical deposition dataset and prediction locations (●) within the region of the eastern USA.

the total sulphate deposition in eastern USA for this year. At this point, we consider only the spatial correlation between the data collected at the stations, and for this reason we remove the temporal correlation by averaging the deposition across time and consider as the response the annual total sulphate deposition measured at each station. This annual deposition is displayed in Figure 6.3.

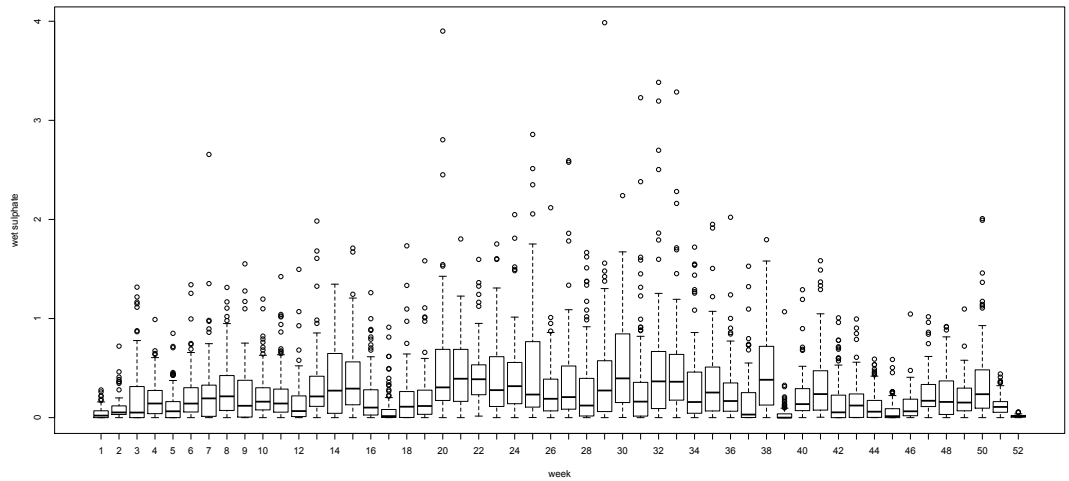


Figure 6.2: Boxplots of weekly deposition: wet sulphate.

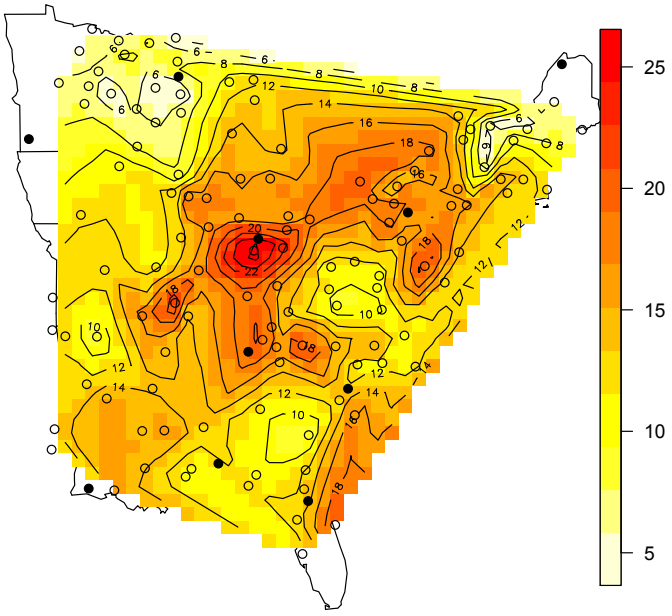


Figure 6.3: Contour plot of the total annual sulphate deposition in 2001, together with sampling (○) and prediction (●) locations.

The prediction sites are chosen in order to be spread across the study region without forming clusters. Also they are chosen to be in areas which are exposed to all levels of sulphate deposition, see for example Figure 6.3. Finally, we chose some sites to be in areas where there are no close monitoring stations.

In this chapter, we assume the Gaussian process model (2.6) for the chemical deposition data and consider both constant mean, $k = 1$, and linear trend, $k = 3$. The correlation between observations is modelled by the Matérn correlation function (2.4) with $\nu = 0.5$, resulting in the commonly applied exponential correlation function.

We use the geodesic distance between two locations with given latitudes, λ_1 and λ_2, and longitudes, γ_1 and γ_2, i.e. the geodesic distance d is the distance at the surface of the Earth considered as a sphere of radius $R = 6371$km. The geodesic distance is the length of the arc of a great circle joining the two locations, defined as

$$d = R \arccos\{\sin(\lambda_1)\sin(\lambda_2) + \cos(\lambda_2)\cos(\gamma_1 - \gamma_2)\}.$$

The design problem for this environmental application concerns which monitoring stations to remove if it is necessary to reduce the network to 40 stations. We want to find the optimal way to reduce the number of the stations with minimum loss of prediction accuracy. We address both the prospective and the retrospective design problems. The

former approach does not take into account any available data, while the latter approach uses available data to update the prior information assumed when finding the optimal configuration of the stations.

6.2 Design Search via the Modified Fedorov Point Exchange Algorithm

In order to find a Ψ-optimal design, we employ the modified Fedorov point exchange algorithm, Atkinson et al. (2007, Ch. 3). This algorithm is a natural choice if we consider the 122 stations as a candidate list of possible design points. This small list of possible design points mitigates many of the advantages of the coordinate exchange algorithm described in Section 3.6.1 and employed in Chapter 5.

The basic steps for the modified Fedorov algorithm are given below:

1. Construct a candidate list with N points.
2. Pick a starting design composed of n points, chosen, for example, randomly from the candidate list and calculate the design performance measure (objective function) for the chosen criterion.
3. Set $i = 1$, $j = 1$.
4. Exchange the ith design point with the jth candidate point. Keep the exchange if it improves the objective function, otherwise, reverse the set up.
5. If $j < N$, set $j = j + 1$ and go to 4, otherwise, if $i < n$, set $i = i + 1$, $j = 1$ and return to 4.
6. If $i = n$, return to 3 and repeat until no improvement can be made.

In this chapter, the starting design is composed of 40 points which are randomly selected from the 122 locations of monitoring stations, and then these points are swapped with the 122 points in the candidate list. This algorithm does not allow repeat points in the design. In each iteration, the objective function Ψ_1 (3.29) is evaluated. The same algorithm is applied for both prospective and retrospective designs.

6.3 Prospective Design

In this section we apply our methodology for Bayesian optimal designs when no prior data are available. The correlation between the observations at two stations is modelled by the exponential correlation function, (2.4) with $\nu = 0.5$, using the geodesic distance between two monitoring stations. The correlation function depends upon the unknown decay parameter ϕ and three different ranges of correlation are considered:

(i) very low correlation, resulting in almost uncorrelated observations, using a uniform prior distribution for ϕ on the interval $[0.1, 1]$, i.e. $\pi(\phi) \sim \text{Unif}(0.1, 1)$,

(ii) medium correlation between the observations, using a uniform prior distribution for ϕ on the interval $[10^{-4}, 10^{-3}]$, i.e. $\pi(\phi) \sim \text{Unif}(10^{-4}, 10^{-3})$,

(iii) high correlation, using a uniform prior distribution for ϕ on the interval $[10^{-6}, 10^{-5}]$, i.e. $\pi(\phi) \sim \text{Unif}(10^{-6}, 10^{-5})$.

The noise-to-signal ratio δ^2 is considered unknown with $\pi(\delta^2) \sim \text{Unif}(0, 1)$.

6.3.1 Constant mean function

In this section, we find Ψ-opitmal designs under a Gaussian process model (2.6) with constant mean. Conjugate normal inverse gamma prior distributions are assumed for the constant mean parameter, $\beta|\sigma^2 \sim \text{N}(0, \sigma^2)$, and variance $\sigma^2 \sim \text{IG}(3, 1)$.

Initially, 30 random designs are generated, each with 40 sampling locations. Then for each one of these designs we use the modified Fedorov point exchange algorithm to find a Ψ-optimal design minimising the objective function $\Psi_1(\xi)$ (3.29). From the 30 resulting designs, we select the design with the smallest objective function. Figure 6.4 gives the Ψ-optimal design for low correlation and Figure 6.5 the designs for medium and high correlation.

With constant mean and very low correlation, we find that any choice of locations gives equal average prediction variance. This equivalence of designs is because (i) the sets

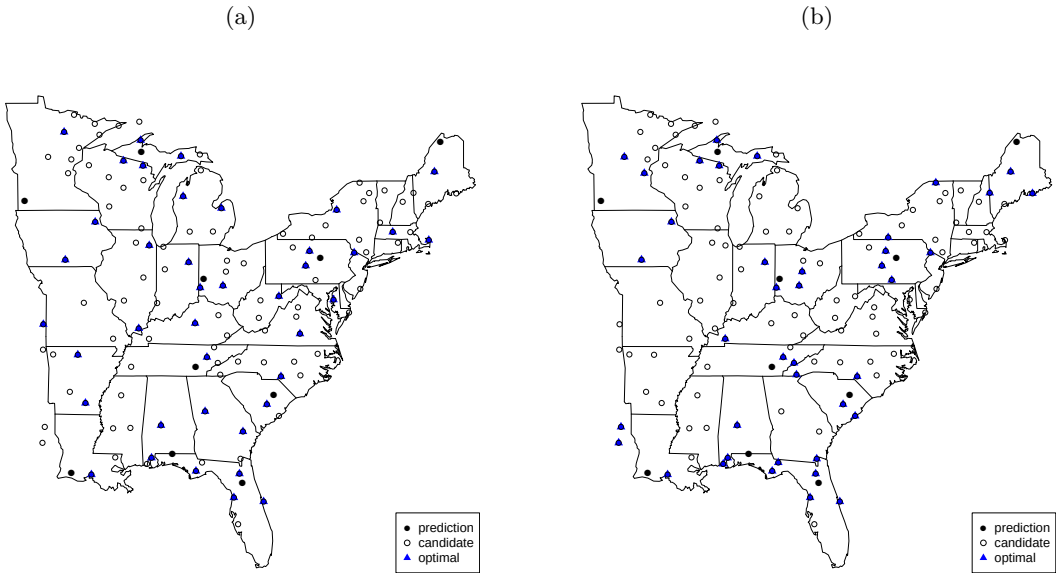


Figure 6.4: Two Ψ-optimal designs with constant mean for $n = 40$ sites with low spatial correlation correlation.

of candidate and predictions points are disjoint, and (ii) observations taken at any of the candidate points are essentially independent of observations at the prediction sites. That is, there is no advantage to choosing any particular set of location over any other set. The maps in Figure 6.4 show two designs, which both of them yield the same value for the objective function.

However, when the spatial correlation is stronger we find a unique Ψ-optimal design for each of 30 random starts of the point exchange algorithm. Figure 6.5 (a) shows the optimal design for medium correlation, i.e the prior range for correlation is $[0.1-0.8]$ between two prediction points at the maximum distance (2778km). The optimal sampling locations are those closest to the prediction locations. Mainly the optimal locations are clustered around the prediction locations. This is natural as the correlation function is now a non-constant function of distance, and hence precise prediction will be provided by a design with locations close to the prediction points.

The map in Figure 6.5 (b) gives Ψ–optimal design for high correlation, i.e. $\phi \sim \text{Unif}(10^{-6}, 10^{-5})$ and the corresponding prior range of the correlation between two prediction locations at the maximum distance is $[0.97 - 0.99]$. Compared to case of medium correlation, here the strength of the correlation is higher and also the range of the correlation is narrower. For this case we can see that the Ψ-optimal design has locations that are close to the prediction locations, similar to the medium range of correlation. The difference between Figure 6.5 (a) and Figure 6.5 (b) is that there are fewer sampling locations at the centre of the region for Figure 6.5 (b). The optimal locations form larger clusters.

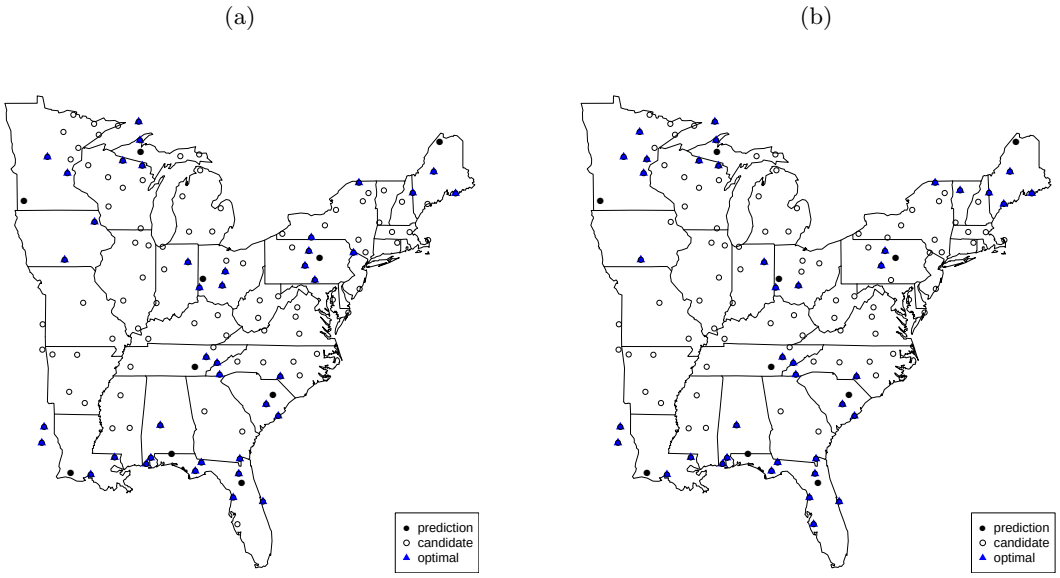


Figure 6.5: Ψ-optimal designs with constant mean for $n = 40$ sites with (a) medium spatial correlation and (b) high spatial correlation.

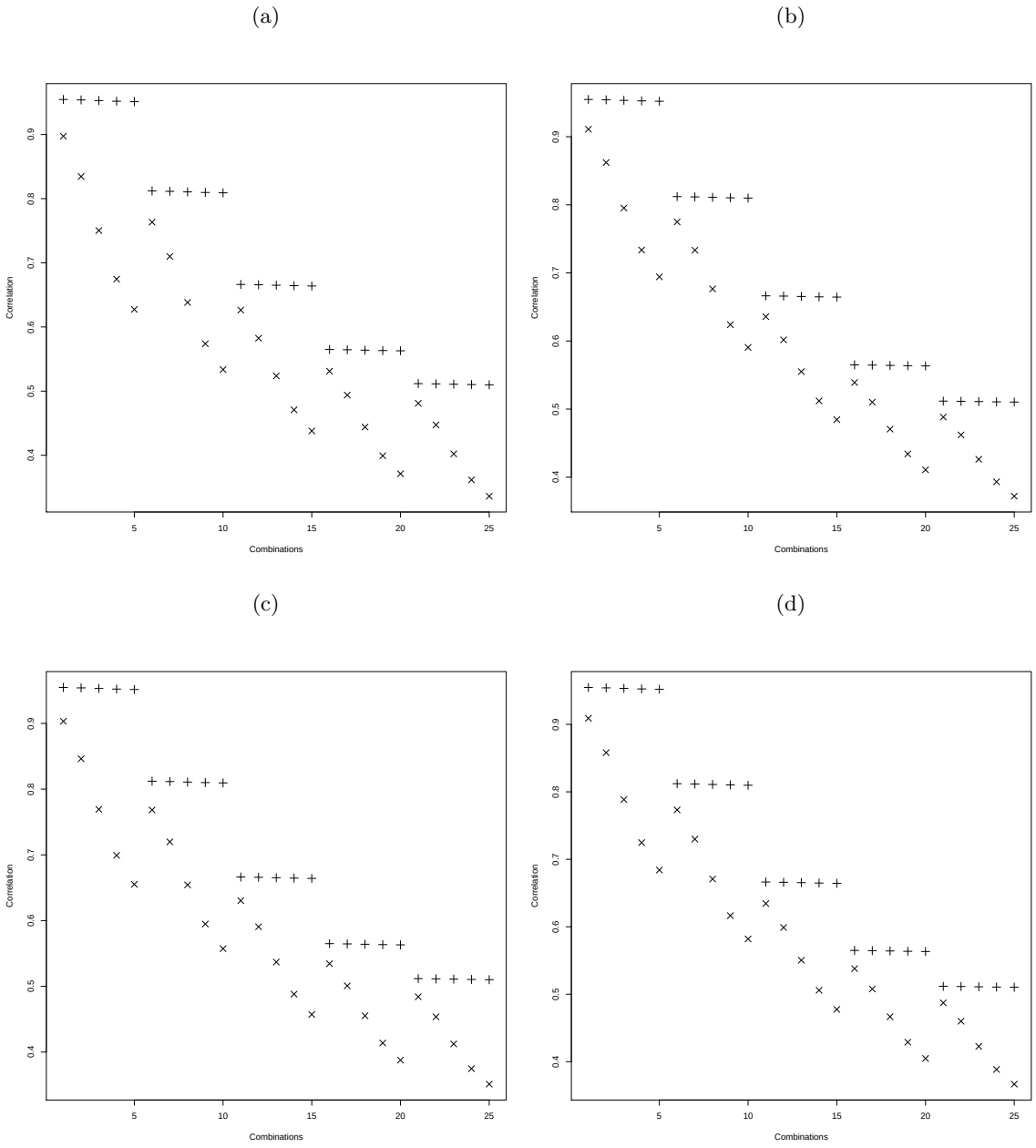


Figure 6.6: Correlations between two prediction and two optimal points for 25 combinations of ϕ and δ^2 obtained from the quadrature points used to approximated the objective function Ψ_1 for medium and high correlation, ($\times$) corresponds to medium correlation (case (ii)) and ($+$) corresponds to high correlation (case (iii)): (a) prediction point 1 and design point 1, (b) prediction point 1 and design point 2, (c) prediction point 1 and design point 2 and (d) prediction point 2 and design point 2.

In order to have a better understanding why for the case of high correlation the optimal sampling locations form some clusters, for example at the top left corner of Figure 6.5 (b), we evaluate the correlation between the prediction points and those optimal design points that differ between the two Ψ-optimal designs from Figure 6.5 (a) and (b). Figure 6.6 shows the correlations between the prediction points at the top left corner of maps of Figure 6.5 (a) and the two additional optimal points in the top left corner of Figure

6.5 (b). The correlation between the prediction and optimal points are evaluated for the 25 combinations of the quadrature points of ϕ and δ^2 used to approximate the objective function $\Psi_1(\xi)$ (3.29) for both medium and high correlation, (cases (ii) and (iii) respectively). From Figure 6.6 we conclude that for case (ii), for all 25 combinations of ϕ and δ^2, the correlations between the prediction points and the optimal points are always smaller than the corresponding correlations for case (iii). These plots indicate that the points that do not belong to the optimal design for medium correlation, case (ii), are not informative for the two prediction points; however, for the high correlation, case (iii), are more highly correlated with both prediction points.

6.3.2 Linear mean function

In this section, we find designs for a Gaussian process model (2.6) with a linear mean function, i.e. we make the assumption that there is a linear trend with respect to the longitude and latitude of a point within the geographical region. We model this trend using a first order polynomial, i.e. $k = 3$ with the regression functions $\mathbf{f}(\mathbf{x_i}) = [1, x_1, x_2]$ with x_1 and x_2 corresponding to the coordinates of the sampling locations of the stations.

The conditional prior distribution for the regression coefficients is assumed to be a normal distribution with zero mean and variance covariance matrix $\sigma^2\mathbf{I}$, i.e. $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. For σ^2 we assume the conjugate inverse-gamma prior $\sigma^2 \sim \mathrm{IG}(3, 1)$. The Ψ-optimal designs are found by minimising objective function $\Psi_1(\xi)$ (3.29) using the modified Fedorov point exchange algorithm to choose $n = 40$.

We find designs for three cases of spatial correlations described in Section 6.3. The maps in Figure 6.7 correspond to the Ψ-optimal designs for each of the three cases of correlation. Figure 6.7 (a) gives the Ψ-optimal design when the spatial correlation is very low (case (i)). In contrast to the findings for a constant mean function, now all the 30 random starts of the point exchange algorithm gave the same Ψ-optimal design. The optimal locations for the stations are those close to the boundaries of the region, and there are no locations in the interior. The Ψ-optimal design is strongly influenced by the linear mean function and the requirement to estimate the unknown regression coefficients.

However, when the spatial correlation is higher, then the pattern for the Ψ-optimal designs changes. Figure 6.7 (b) shows the optimal locations for the stations when the prior correlation is between $[0.1 - 0.8]$, case (ii). The Ψ-optimal design chooses stations which are close to the prediction locations, and is very similar to the corresponding optimal design for the constant mean function, Figure 6.5 (a).

Finally, Figure 6.7 (c) gives the Ψ-optimal design corresponding to high correlation between the observations at two stations, case (iii). The optimal choice locates the stations at the boundaries of the region, with no stations in the centre. This selection

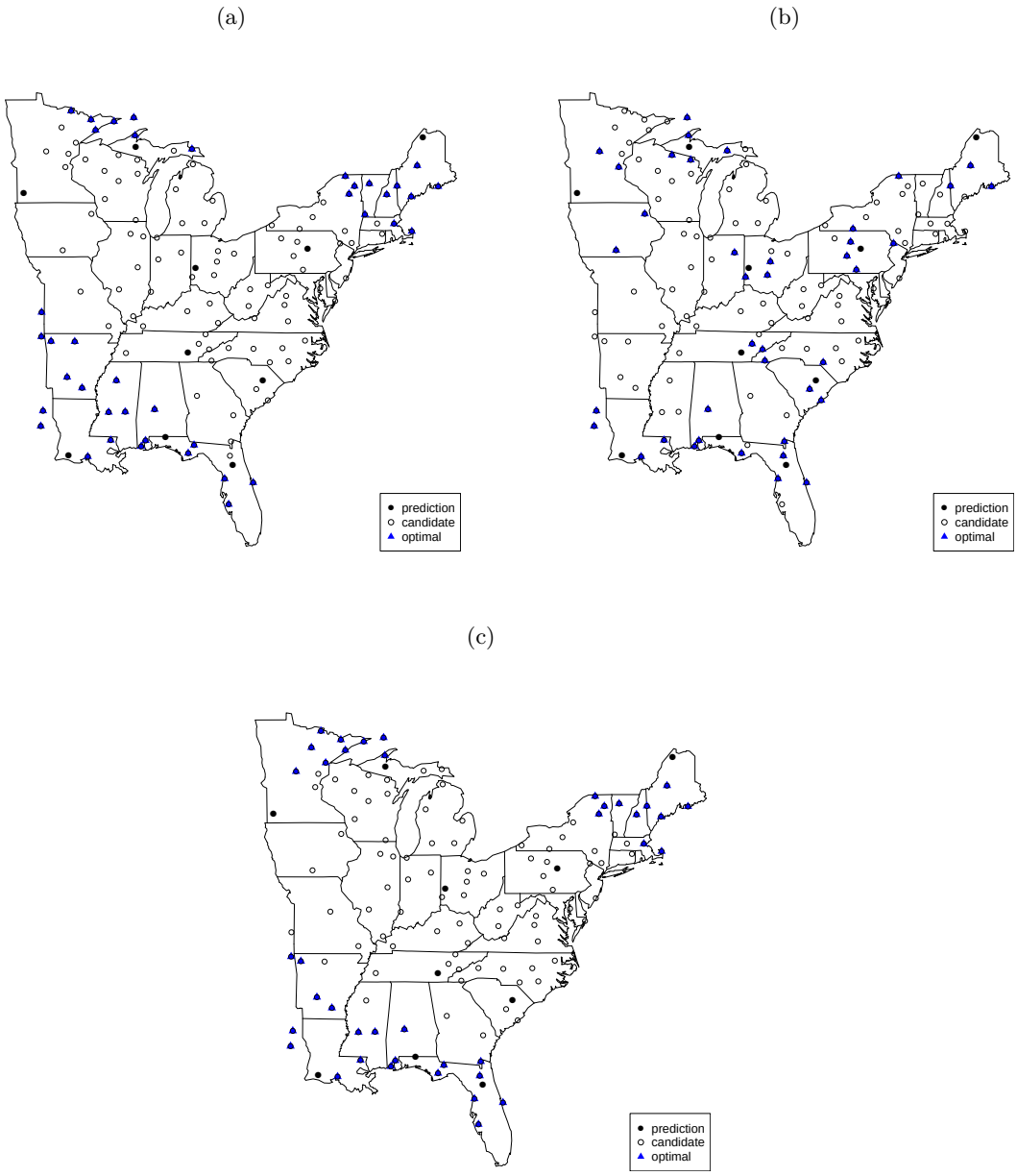


Figure 6.7: Ψ-optimal designs with a linear trend for $n = 40$ sites with (a) low spatial correlation, (b) medium spatial correlation and (c) high spatial correlation.

is strongly influenced by the prior range of the correlation, which corresponds to $[0.97 - 0.99]$. Hence, most variation between observations at different stations is explained via the mean function.

6.4 Retrospective Design

In this section, we consider the reduction of 122 stations to a set of 40, chosen using a Ψ-optimal design that now takes into account available prior data. This is achieved by constructing posterior distributions for all the model parameters using this data, and

then using these distributions as priors for design selection. As before, we employ the closed-form approximation $\Psi_1(\xi)$ (3.29) and find designs for both constant and linear mean function.

As already discussed in Chapter 2, when the correlation parameters and the noise-to-signal ratio are unknown there are no closed form solutions for the posterior distributions of any of the parameters. However, conditional on these unknown parameters and by employing MCMC procedures, we are able to obtain samples from the posterior distributions. The procedure to obtain posterior samples is the following:

1. Set prior distributions for ϕ and δ^2.
2. Using a MH algorithm with log-normal proposal distribution for ϕ and a uniform distribution proposal for δ^2 we generate a sample from the posterior distribution of $\pi(\phi, \delta^2|\mathbf{y})$ (2.34).
3. Given the sampled values of ϕ and δ^2 we generate from the conditional distributions $\pi(\boldsymbol{\beta}|\mathbf{y}, \phi, \delta^2)$ and $\pi(\sigma^2|\mathbf{y}, \phi, \delta^2)$, which are a t-distribution (2.18) and an inverse gamma distribution (2.19), respectively.
4. We repeat the procedure until sufficient samples are taken from the marginal posterior distributions of $\boldsymbol{\beta}, \sigma^2, \phi, \delta^2$.

6.4.1 Constant mean function

Initially we assume a Gaussian process model with constant mean function. The prior distributions for the regression coefficients and the variance are normal and inverse gamma respectively, $\beta|\sigma^2 \sim \mathrm{N}(0, \sigma^2)$ and $\sigma^2 \sim \mathrm{IG}(3, 1)$. For the noise-to-signal ratio we assume a uniform distribution $\delta^2 \sim \mathrm{Unif}(0, 1)$ and for the decay parameter $\phi \sim \mathrm{Unif}(10^{-5}, 0.01)$ which corresponds to a wide prior range for the correlation, i.e. between 0.1 to 0.97. This choice of prior distribution of ϕ results in medium-high spatial correlation and covers case (ii) and case (iii) in Section 6.3.

Figure 6.8 gives trace plots and empirical density plots for samples from the posterior distributions $\pi(\phi|\mathbf{y})$ and $\pi(\delta^2|\mathbf{y})$. Initially, we generated a sample of size 3000 but both chains mixed very poorly, and for this reason we doubled the sample size. The plots are based on sampling after a burn-in of 3000 iterations. The mixing of the chains is fairly poor because there is high correlation and it is difficult to explore the parameter space. However, for the purpose of this study the mixing of the chain is considered sufficient. The effective sample size (5.10), is larger than 350 for both ϕ and δ^2, which we consider sufficient for this study because the aim is not to make any inference using the posterior distributions of ϕ and δ^2. In particular, we use these MCMC samples to approximate the values of posterior densities for each node of the quadrature points that we use to approximate the objective function.

Figure 6.9 shows the posterior distributions for the regression coefficient β and the

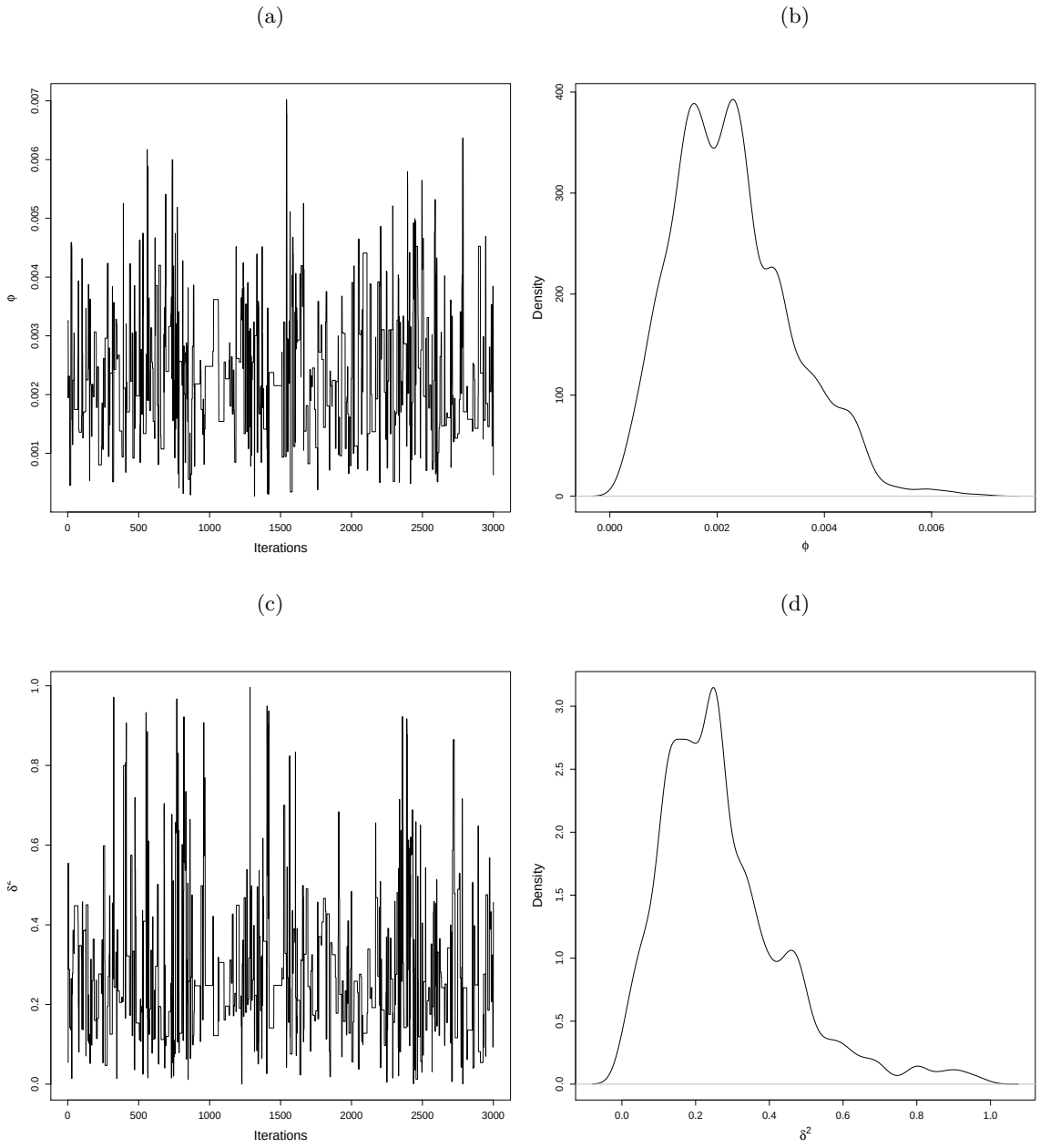


Figure 6.8: (a) Trace plot and (b) empirical density plot from MCMC samples for decay parameter ϕ, (c) trace plot and (d) empirical density plot from MCMC samples for noise-to-signal ratio δ^2.

variance of the Gaussian process σ^2. The mixing of these chains is very good and the effective sample size for both posterior distributions is larger than 3000. To find designs using approximation $\Psi_1(\xi)$ we require conjugate conditional distributions, We approximate the posterior distribution for β by a normal distribution $\beta \sim \mathrm{N}(9.70, 5.11)$ and for σ^2 with an inverse gamma distribution, $\sigma^2 \sim \mathrm{IG}(7.5, 133.5)$. The hyper-parameters of the normal and inverse gamma distributions were found via matching moments of the distributions, using the mean and the variance of the posterior sample of β and σ^2 obtained from the MCMC procedures.

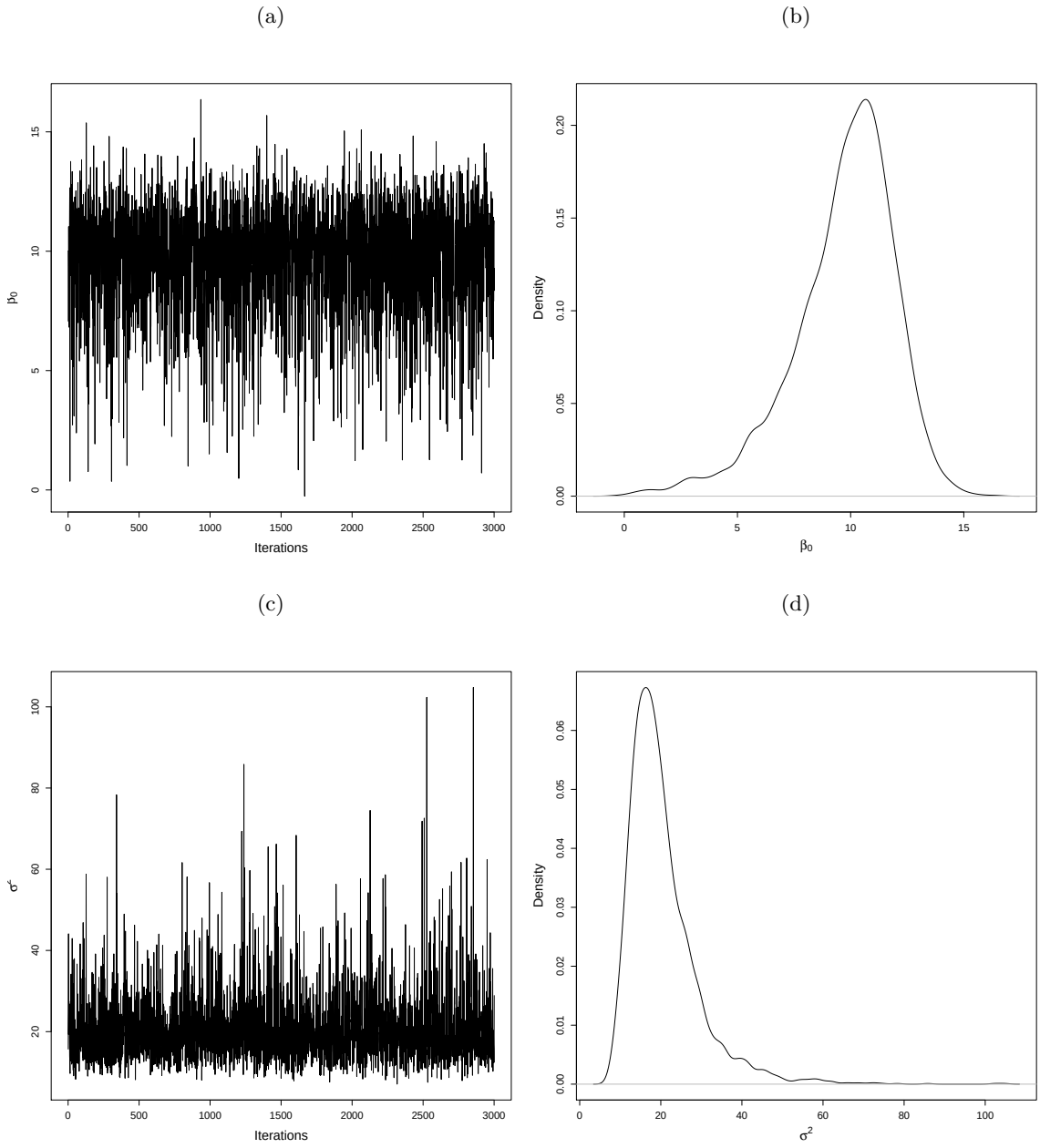


Figure 6.9: (a) Trace plot and (b) empirical density plot from MCMC samples for regression coefficients β, trace plot and (d) empirical density plot from MCMC samples for variance σ^2.

As it can be seen from the density plots in Figure 6.8 the posterior densities of ϕ and δ^2 do not resemble densities from known distributions. The objective function $\Psi_1(\xi)$ (3.11) is approximated using Gauss-Legendre quadrature method (Section 3.5) by (3.29). The weights w_i^1 and nodes α_i^1 are obtained from the Legendre polynomials from the prior distribution of ϕ, and w_i^2 and nodes α_i^2 from prior distribution for δ^2. The Gauss-Legendre quadrature method requires uniform distributions for ϕ and δ^2 and as the posterior densities obtained from MCMC do not resemble uniform distributions we approximate for each node α_i^1 and α_i^2 the values $\pi(\phi_{\alpha_i^1})$ and $\pi(\delta^2_{\alpha_i^2})$ from the empirical

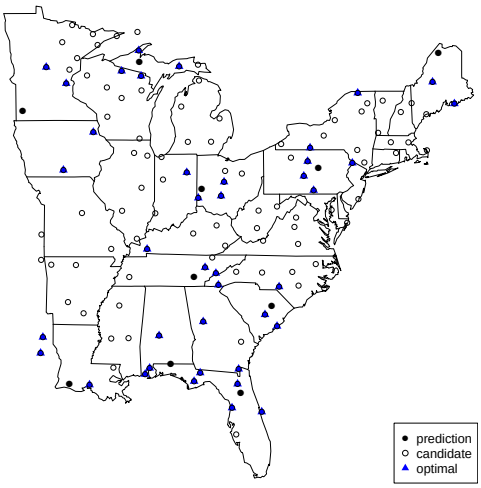


Figure 6.10: Retrospective Ψ-optimal designs for $n = 40$ sites for constant mean function.

densities and multiply them by w_i^1 and w_i^2. The values are:

$$\pi(\phi_{\alpha_i^1}) = (91.131, 392.272, 20.178, 0.0077, 0.0077)$$

$$\pi(\delta^2_{\alpha_i^2}) = (1.058, 2.982, 0.756, 0.095, 0.075).$$

The modified Fedorov point exchange algorithm is employed, with 30 random starts. The resulting Ψ-optimal design for a retrospective study is presented in Figure 6.10; all 30 random starts gave the same Ψ-optimal design. The optimal locations are those close to the prediction locations and also there are some locations spread in the interior of the geographical region.

We now compare this retrospective design to the three prospective designs, for different priors on ϕ found in Section 6.3. In order to do this comparison, we evaluate the value of the objective function $\Psi_1(\xi)$ (3.29) under each prior distribution and then find the efficiency

$$\text{eff}(\xi^*, \xi) = \frac{\Psi_1(\xi^*)}{\Psi_1(\xi)},$$

where $0 \leq \text{eff}(\xi^*, \xi) \leq 1$. An efficiency close to 1 means that the design ξ is as good as the optimal design ξ^*. Table 6.1 shows the efficiencies of the four designs evaluated under each prior distribution for the prospective and retrospective studies.

As it can be seen, the efficiency is always 1 when the designs are evaluated under the case of low correlation. This is in line with our discussion in Section 6.3 for very low correlation and constant mean, here any choice of designs will be Ψ-optimal. When the designs are evaluated for medium and high correlation, then the retrospective design is highly efficient (0.9986). Similarly, when we assess all four designs under the retro-

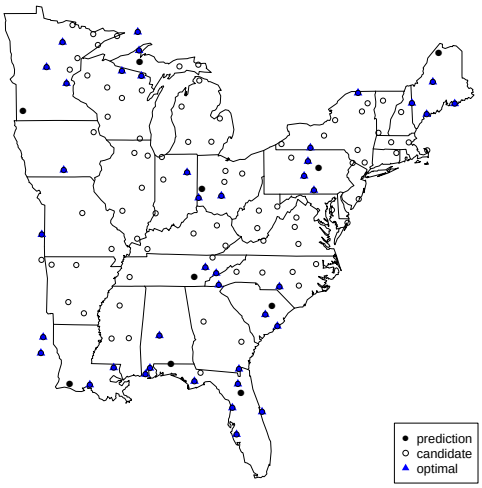


Figure 6.11: Retrospective Bayesian optimal design for 40 sites for linear mean function.

spective prior distributions we see that the optimal designs under medium and high correlation are both highly efficient, with the designs for medium correlation having efficiency just 1% higher than that for the high correlation.

These comparisons are supported by the density plot of decay parameter ϕ, see Figure 6.8, where the density of the posterior distribution is higher for ϕ in the interval $[0 - 0.003]$. This interval is included in the support of the uniform prior distribution assumed for ϕ in the prospective design for medium correlation, $\pi(\phi) \sim \text{Unif}(0.0001, 0.001)$. Once again our conclusions is that the range of the correlation is crucial in influencing the Ψ-optimal design.

6.4.2 Linear mean function

Similar to the procedure followed for constant mean function, a Ψ-optimal design has been found for the Gaussian process model (2.6) with linear mean function, $k = 3$. The considered mean function is a linear function of the longitude and latitude of a point within the geographical region. The prior distributions for the unknown parameters are assumed the same as for the case of constant mean function, i.e. normal inverse-

	Prospective			Retrospective
Optimal Design	Low	Medium	High	High - Medium
Retrospective	1	0.9986	0.9998	1
Prospective 1	1	0.9707	0.9995	0.9531
Prospective 2	1	1	0.9999	0.9970
Prospective 3	1	0.9951	1	0.9836

Table 6.1: Efficiencies of Ψ-optimal designs for constant mean function. Low, medium and high column headings correspond to the degree of spatial correlation.

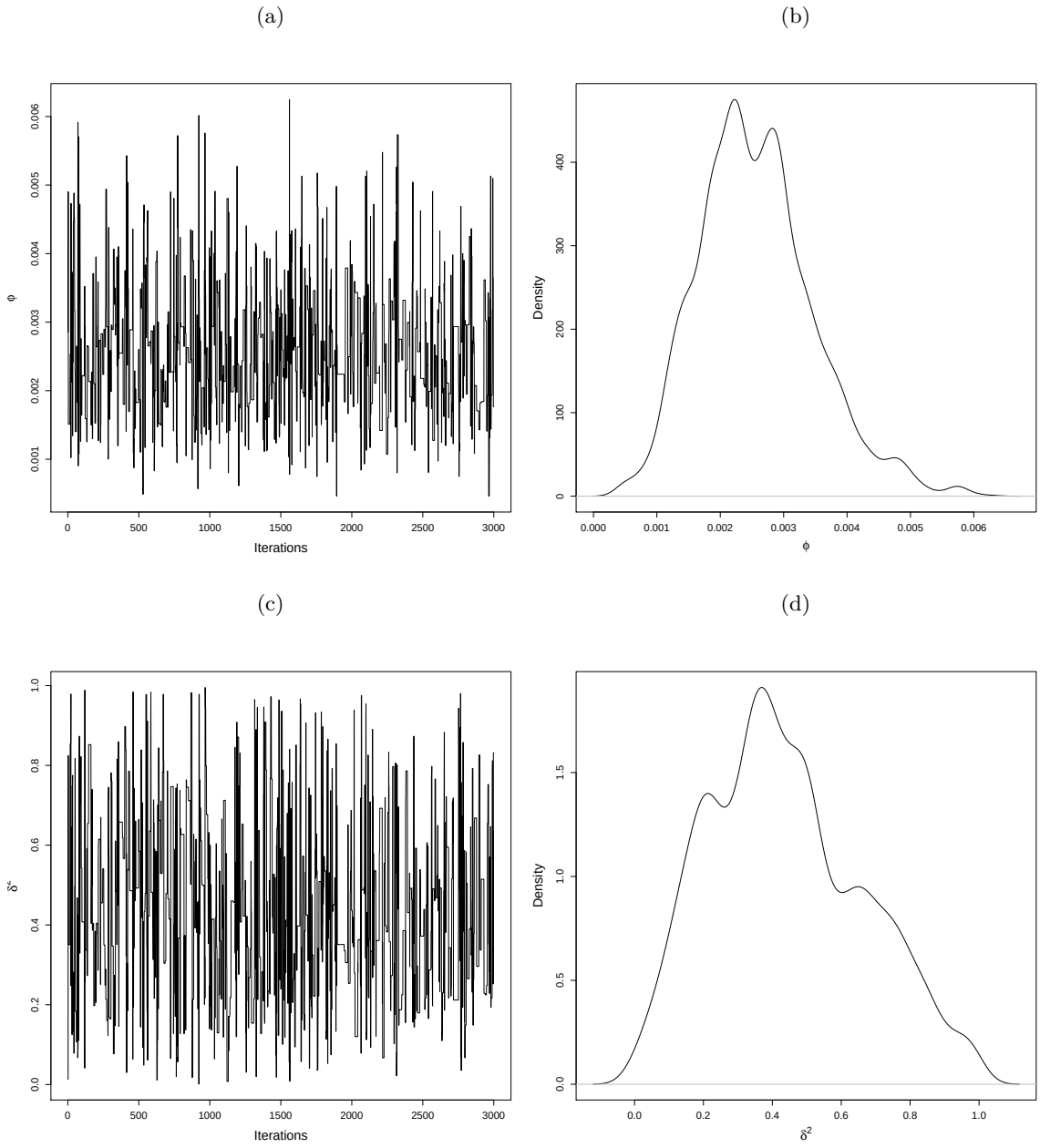
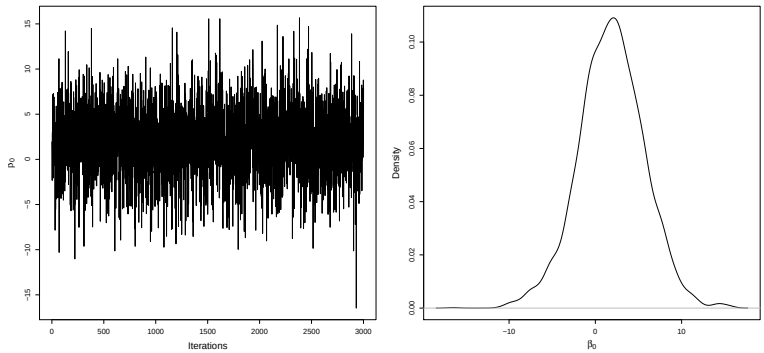


Figure 6.12: (a) Trace plot and (b) empirical density plot from MCMC samples for decay parameter ϕ, (c) trace plot and (d) empirical density plot from MCMC samples for noise-to-signal ratio δ^2.

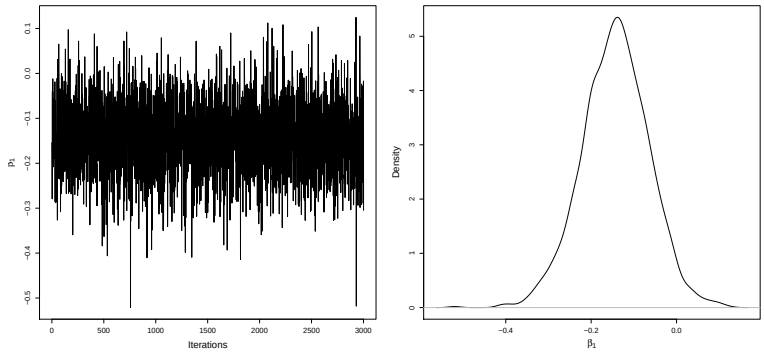
gamma prior distributions for trend parameters and Gaussian process variance, $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\sigma^2 \sim \mathrm{IG}(3,1)$, and uniform prior distributions for both δ^2 and ϕ, $\mathrm{Unif}(0,1)$ and $\mathrm{Unif}(10^{-5}, 0.01)$, respectively.

Figures 6.12 and 6.13 give the trace plots and empirical density plots for samples from posterior distributions of the unknown model parameters. A sample of size 6000 is generated and after the burn-in we keep 3000. The trace plots and the effective sample size for $\boldsymbol{\beta}, \sigma^2, \phi$ and δ^2 indicate sufficient mixing of the chains for this study.

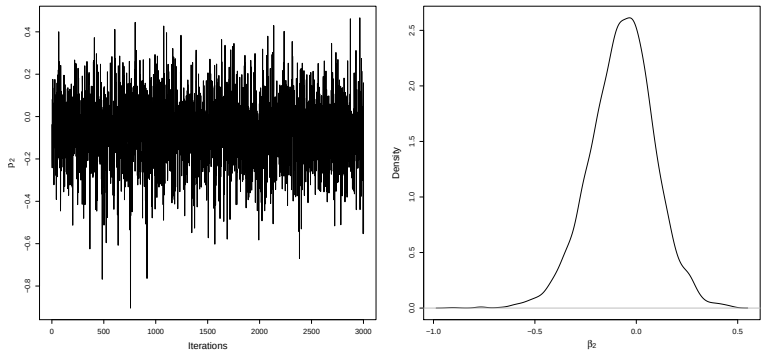
(a) (b)



(c) (d)



(e) (f)



(g) (h)

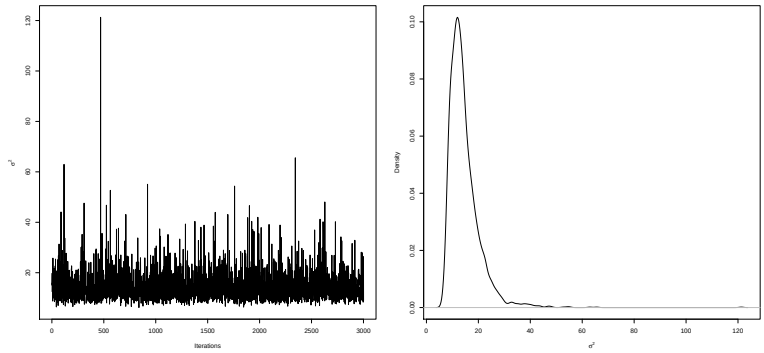


Figure 6.13: (a) Trace plot and (b) empirical density plot from MCMC samples for regression coefficients β, trace plot and (d) empirical density plot from MCMC samples for variance σ^2.

In line with the approach followed for the case of constant mean, Section 6.4.1 we approximate the posterior distribution for $\boldsymbol{\beta}$ by a normal distribution with mean a 3×1 vector $(0.862, -0.143, -0.069)^\top$ and a 3×3 variance covariance matrix with diagonal elements $14.506, 0.006, 0.025$, and for σ^2 we use an inverse gamma distribution with $a = 7.73$ and $b = 98$; all hyper-parameters are obtained by matching moments.

The density plots for ϕ and δ^2 (Figures 6.12) show that the posterior densities do not resemble uniform distributions and to approximate the objective function $\Psi_1(\xi)$ (3.11) using the Gauss-Legendre quadrature method (3.29) we approximate for each node α_i^1 and α_i^2 the values $\pi(\phi_{\alpha_i^1})$ and $\pi(\delta^2_{\alpha_i^2})$ from the empirical densities and multiply them by weights w_i^1 and w_i^2. The values are:

$$\pi(\phi_{\alpha_i^1}) = (16.779, 460.922, 30.474, 0.009, 0.009)$$

$$\pi(\delta^2_{\alpha_i^2}) = (0.390, 1.379, 1.524, 0.741, 0.238).$$

Again, the modified Fedorov point exchange algorithm is employed, with 30 random starts and the design which minimises the objective function $\Psi_1(\xi)$ is saved as the optimal choice. The resulting Ψ-optimal design is displayed in Figure 6.11. The optimal locations are those near the prediction locations.

The final step is to compare the retrospective designs to the three prospective designs obtained for different priors on ϕ found in Section 6.3. Table 6.2 shows the efficiencies of the four designs evaluated under each prior distribution for the prospective and retrospective studies. From to this table, we conclude that the prospective design with very high spatial correlation is efficient for low spatial correlation and this is true because for both cases the linear trend strongly influences the design. When the designs are evaluated for medium and high correlation, then the retrospective design is highly efficient and when we assess all four designs under the retrospective prior distributions we see that the optimal design under medium correlation gives the largest efficiency (0.9989) compare to those for low and high correlation.

	Prospective			Retrospective
Optimal Design	Low	Medium	High	High - Medium
Retrospective	0.9917	0.9994	0.9923	1
Prospective 1	1	0.8715	0.9989	0.9077
Prospective 2	0.9894	1	0.9896	0.9989
Prospective 3	0.9996	0.88271	1	0.9180

Table 6.2: Relative efficiencies of Ψ-optimal designs for linear mean function. Low, medium and high column headings correspond to the degree of spatial correlation.

6.5 Summary

To summarise, in this chapter we employed the decision theoretic approach for finding Bayesian optimal designs on a real environmental example of a monitoring network in the eastern USA. We distinguished between two design approaches, the prospective and retrospective designs and we compared these two approaches.

We concluded that if the chosen prior range of the spatial correlation, described by the prior distribution of the decay parameter, corresponds to medium spatial correlation then both retrospective and prospective designs give designs with similar efficiency. The Ψ-optimal designs obtained from the retrospective approach have very similar efficiencies to those obtained from the prospective approach and medium correlation. In general, the designs are strongly influenced by the degree and the range of the correlation, the choice of the mean function and the assumed prediction grid.

In the retrospective design, we did not consider any missing data. In fact, some stations had missing data for some weeks out of the 52. We did not take into account the variability that may be introduced to the posterior analysis due to the missing data. An alternative approach would be to obtain the dataset with observations per week, impute the missing data and then average across the time.

Moreover, in a future investigation, the temporal correlation should be taken into account. In this chapter, we only considered the spatial correlation between the sampling location but we could investigate how the design changes according to the time that observations are taken. Although we propose a Bayesian methodology for spatio-temporal optimal designs in Chapter 8, the set up is somewhat limited and future investigation is needed in order to be applied to a real dataset.

Chapter 7

Design for Computer Experiments

The objective of this chapter is to apply the decision theoretic approach to develop Bayesian optimal designs for prediction in computer experiments. We start by introducing the field of computer experiments and reviewing the existing approaches for design. We then introduce Bayesian Ψ-optimal designs for computer experiments found by our new methodology. A numerical study is performed to validate the approximation necessary to find a closed-form objective function for design selection, and examples of Ψ-optimal designs are demonstrated.

7.1 Introduction

Many physical phenomena are difficult to investigate via physical experimentation, which may be financially prohibitive, dangerous, unethical or impossible to pursue. Computer experiments are becoming an alternative to traditional physical experiments, where a computer model provides a representation of the real physical system that can be investigated and explored.

Scientists and engineers make use of computer models to study relationships between the input and output variables of a system or process, and explore the entire experimental region. Although computer power has significantly increased during the last years, the mathematical models underlying the computer simulations are often very complex; for example there is no simple explicit mathematical formula which describes the relationship between the input and the output for the finite element model mentioned in Chapter 1. As a result, computer codes that implement these relationships may have very long run times, taking minutes, hours or days to produce a single response. Therefore, computer simulations can be time consuming and very computationally expensive to run, and there is need to find a computationally inexpensive surrogate model, or

metamodel or emulator as it often called, that can replace to a lesser degree the computer model. Such a surrogate model allows fast prediction of the outputs at untested input points.

A very popular surrogate model, introduced to the field of computer experiments in the pioneering paper of Sacks et al. (1989), is the Gaussian process because it is an adaptive and flexible non-parametric interpolator/smoother. A Gaussian process surrogate model can be used to gain insight into the computer model over the whole design region and makes tasks such as sensitivity analysis, uncertainty analysis, validation and calibration feasible (see Santner et al. (2003) and Fang et al. (2006)).

The main focus of this chapter is on the selection of the points at which to run the computer model to obtain good predictive performance of the Gaussian process surrogate model. In practice, time and computational resources are limited, so an experimental design plays a crucial role by identifying a set of inputs at which the underlying computer code will be evaluated. Such a set of points is called the design of the computer experiments. The prediction quality of the surrogate model is influenced by both the type of model used and the design points where the computer model is evaluated. Hence, in order to increase the quality of the predictions form the surrogate model, an optimal choice of the design points is crucial.

7.2 Statistical Surrogate

The computer model, implemented in code, can be considered as a function f with inputs $\mathbf{x} \in \mathcal{X}$ and the output $\mathbf{y} \in \mathcal{Y}$. The surrogate model treats the computer code as a black box, with no assumed knowledge about the function f. We model the computer output using the Gaussian process model (2.6). Figure 7.1, as it is presented in Fang et al. (2006), shows the idea of the surrogate model.

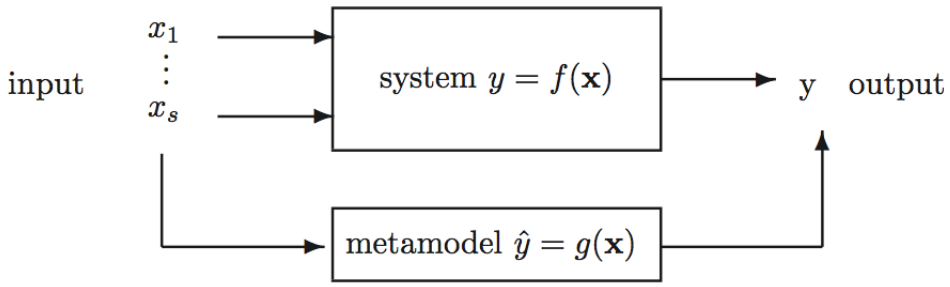


Figure 7.1: Role of the surrogate model, or metamodel, for computer experiments (taken from Fang et al. (2006)).

For deterministic computer experiment applications, it is natural to consider the Gaussian process model from a Bayesian perspective, where it represents our prior uncertainty about the function f. Following the Bayesian approach, the predictions are achieved by combing the prior information about the function with information obtained from the data, see for example Chapter 2 and Section 2.4 for a detailed derivation of the posterior predictive distribution. Currin et al. (1991) presented the Bayesian interpretation of prediction for computer experiments and adopted a plug-in approach with the parameters estimated via maximum likelihood.

For deterministic computer models, it may be natural to set the nugget $\tau^2 = 0$ to provide an interpolating Gaussian process. To show this interpolation property, we consider the MSPE predictor for y_{n+1}, $\hat{y}_{n+1}$, given by (2.7), we set $\tau^2 = 0$ and assumed that all model parameters are known. Then, if we choose a prediction point to be a point from the design, i.e. $\mathbf{x}_{n+1} = \mathbf{x}_i$ where $i = 1, \cdots, n$ we have that $\mathbf{f}_{n+1} = \mathbf{f}^\top(\mathbf{x}_i)$ and $\boldsymbol{\omega}^\top$ now corresponds to ith row of the correlation matrix $\mathbf{C}$. Hence, $\boldsymbol{\omega}^\top\mathbf{C} = [0, \cdots, 1, \cdots, 0]^\top = \mathbf{e}_i^\top$. Substituting $\mathbf{e}_i^\top$ into (2.7) with $\tau^2 = 0$ we have $\hat{y}_{n+1} = \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta} + y_i - \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta} = y_i$, where y_i is the observation for $\mathbf{x}_i$.

However, several authors have proposed including the nugget in the surrogate model to ensure numerical stability and overcome computational issues due to near-singularity of the correlation matrix, see Ababou et al. (1994) and Neal (1997). Apart from computational issues, Gramacy and Lee (2012) indicated more reasons why a nugget should be included in modelling for computer experiments. They discussed the role of the nugget in improving the adequacy of the statistical surrogate, even for deterministic computer models, via mitigating incorrect surrogate modelling assumptions, such as stationarity. Increasingly, computer models may include intrinsic stochastic elements, for example, Monte Carlo simulators, and hence produce (pseudo) random output requiring the use of a nugget term.

Gramacy and Lee (2012) also argued that a nugget effect captures the computer model bias, i.e. that the mathematical model of the physical process is not a perfect description of reality, and that uncertainty about the true function is best modelled by a random process that smooths rather than interpolates. For all these reasons, in our approach for finding optimal designs for computer experiments we include a nugget effect.

7.2.1 Parametric correlation functions

In this section, we outline the key difference between Gaussian process models for spatial data and computer experiments. Recall in the Gaussian process model (2.6), the correlation matrix $\mathbf{C}$, the key element of a Gaussian process $Z(\mathbf{x}_i)$, is typically determined by a stationary and isotropic parametric correlation function. Examples include the isotropic power and Matérn correlation functions described in Section 5.4.4. However, in many computer experiments applications, the correlations due to some

inputs is stronger than for other inputs, and for this reason we need to use a anisotropic correlation function.

It is common to use a separable correlation function, i.e. we take the product of correlation functions across each dimension, each of which are stationary:

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\phi}) = \prod_{k=1}^{d} \rho_k(x_{ik}, x_{jk}; \phi_k).$$

Here, the decay parameter ϕ_k controls the correlation in the kth direction. The resulting correlation function is not isotropic as each dimension has a different decay parameter. In general, the isotropic correlation function in Section 5.4.4 is a special case of separable correlation function when $\phi_k = \phi$ for $k = 1, \ldots, d$.

In this chapter we focus on two popular choices of product correlation functions: the Power Exponential and the Matérn correlation function which are extensions of the isotropic functions described in Section 5.4.4.

(a) **Matérn separable correlation function**
The product Matérn exponential correlation function has the form

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\phi}, \nu) = \prod_{k=1}^{d} \frac{1}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu}|x_{ik} - x_{jk}|\phi_k)^{\nu} K_{\nu}(2\sqrt{\nu}|x_{ik} - x_{jk}|\phi_k) \quad \nu > 0, \quad \phi_k > 0. \tag{7.1}$$

The parameter ν controls the smoothness of the Gaussian process.

(b) **Power exponential separable correlation function**
The product power exponential correlation function has the form

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\phi}, \nu) = \prod_{k=1}^{d} \exp\left(-\phi_k(|x_{ik} - x_{jk}|)^{\nu}\right) \quad 0 < \nu \leq 2, \quad \phi_k > 0. \tag{7.2}$$

For $\nu = 1$, we obtain the product exponential correlation function and for $\nu = 2$, the product Gaussian correlation function.

The decay parameters in the correlation functions can be interpreted as measuring the importance or activity of the input. Therefore for large ϕ_k, the kth variable is not important as the correlation in the kth direction is largely independent of the distance between the points $\mathbf{x}_i$ and $\mathbf{x}_j$; for small values of ϕ_k, the correlation depends on the distance and hence the variable is important, see Linkletter et al. (2006).

7.3 Literature Review

The area of design for computer experiments has received great attention in recent years and there is a considerable literature that indicates its rapid development, see for

example Santner et al. (2003), Fang et al. (2006) and Kleijnen (2008).

Two classes of designs for computer experiments have been considered: the model-free and model-based approaches. The former approach does not make use of any assumptions about the statistical model that approximates the computer code, whereas the latter approach explicitly accounts for the statistical surrogate model. The most popular model-free approach is space-filling designs whereas, the model-based approach is separated into designs for prediction and designs for estimation. Pronzato and Müller (2012) gave a general overview of both approaches and described many of the existing methods for finding a design for a computer experiment.

7.3.1 Space-filling designs

There are several ways to define "space-filling" via a distribution of the design points which covers the input space. Three common approaches use :

- measures of distance between the design points, i.e geometric criteria (Johnson et al., 1990), see Chapter 3,
- sampling methods (McKay et al., 1979),
- statistical measures of uniformity (Fang et al., 2000).

Another way of selecting the design points makes use of sampling methods, in particular, simple random sampling, stratified random sampling and Latin hypercube sampling. A comparison of these three methods of selecting a design was given by McKay et al. (1979). Simple random sampling selects the n points of the design at random from the design region $\mathcal{X}$ with respect to a uniform distribution. Although simple to apply, in high dimensions this method exhibits clustering of points and poorly covered areas of the design space. To overcome this problem, stratified random sampling was proposed to select the points. The design region is divided into n equally spread strata and then one point from each strata is randomly selected. Using stratified sampling promotes coverage of the entire experimental region.

In general, it is desirable for a design for a computer experiment to have good projection properties, i.e. when the output is influenced by only some input variables we want the points evenly spaced across the projections onto these significant inputs. This requirement led McKay et al. (1979) to introduce Latin Hypercube Designs (LHD), which have the property that the points are evenly spaced across the one-dimensional projections. The idea is to divide the design region into equally-sized cells and then randomly select n cells under the restriction that the projections of the selected cells on to each axis do not overlap. McKay et al. (1979) compared the three methods of sampling, and concluded that Latin Hypercube Designs gave more precise, as in low variance, estimator for the mean and the variance. Since this pioneering paper, Latin Hypercube Designs have became the most popular sampling method for computer ex-

periments; they are easy to generate, computationally simple and have good projection properties. Due to this popularity, several authors have proposed different extensions of Latin Hypercube Designs (see for example Handcock (1991) and Tang (1993)).

The last approach for space-filling designs considers the problem of finding a design that mimics a uniform distribution on the design space, i.e. the distribution of the points of the design is comparable to our expectation from a uniform distribution. This comparison is made through calculation of a discrepancy measure. Discrepancy, and the criteria that rely on it, were introduced by Fang (1980). An optimal uniform design ξ^* minimises

$$D(\xi) = \max_{\mathbf{x}\in\mathcal{X}} |F_n(\mathbf{x}) - U(\mathbf{x})|,$$

where $F_n(\cdot)$ is the empirical distribution function of design ξ and $U(\cdot)$ is the empirical distribution function of the uniform distribution. This discrepancy measures how the distribution of the design differs from the uniform distribution. The lower this discrepancy is, the more uniform the designs points are scattered over the design region. Fang et al. (2000) proposed algorithms to construct nearly uniform designs and explored the possibility of uniform designs being orthogonal.

Several authors considered the case of combining aspects of these three methods of finding designs for computer experiments. For example, a discrepancy criterion can be applied to the class of Latin Hypercube Design to find the most uniformly distributed design. Alternatively, a geometric criterion can be used to find, for example, a maximin Latin Hypercube Design, see Morris and Mitchell (1995), Santner et al. (2003, Ch. 5) and references therein.

7.3.2 Model-based designs

The model-based approach for the design of computer experiments assumes a Gaussian process model for the response, and finds designs for either prediction or estimation of the unknown model parameters.

When the aim of the experiment is prediction at untested inputs, the most popular design criteria for computer experiments are functions of the Mean Square Prediction Error (MSPE), introduced by Sacks et al. (1989) and described in detail in Santner et al. (2003, Ch. 6). Designs are typically found by minimising the Maximum Mean Square Prediction Error (MMSPE) or, more commonly the Integrated Mean Square Prediction Error (IMSPE). The IMSPE averages the mean square prediction error over the design region. Sacks et al. (1989) proposed the use of a quasi-Newton algorithm to find IMPSE optimal designs. An IMPSE optimal design ξ^* minimises the IMSPE objective function

$$\eta(\xi) = \sigma^2 \left\{ 1 + \delta^2 - \int_{\mathcal{X}_p} \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\omega} + (\mathbf{f}_p^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}(\mathbf{f}_p^\top - \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{-1}\mathbf{F})^\top d\mathbf{x}_p \right\},$$

where the elements of this function are defined in Chapter 2.

These MSPE criteria depend on the correlation parameters ϕ (through $\boldsymbol{\Sigma}$ and $\boldsymbol{\omega}$) and δ^2, which are usually assumed known. However, in practice these parameters are unknown and Sacks et al. (1989) proposed a two-stage approach where a LHD is used to collect data to estimate the unknown model parameters, which are then plugged-in to the MSPE objective function. These authors also proposed using robustness studies to assess the performance of given IMPSE optimal designs for a range of values of the correlation parameters for the power exponential correlation function (7.2). A more recent sequential approach was proposed by Picheny et al. (2010).

Harari and Steinberg (2014) used a spectral decomposition (Karhunen-Loéve expansion) of the Gaussian process model, with known constant mean and known correlation parameters, to find IMSPE designs. They investigated the performance of this criterion when a nugget effect is included in the model and, similar to Section 2.4.4 in this thesis reparameterised using the noise-to-signal ratio δ^2 to overcome computational issues. They extended their model to unknown constant mean and compared IMSPE designs for both known and unknown mean and concluded that the two models yielded similar designs.

More recent work has incorporated the unknown correlation parameters by averaging the values of IMSPE weighed according to possible values of the correlation parameters. Leatherman et al. (2014) proposed the Weighted-IMSPE

$$\eta_w(\xi) = \int \eta(\xi)\pi(\phi)d\phi, \tag{7.3}$$

where $\pi(\phi)$ can be viewed as a prior distribution from a Bayesian perspective. However, $\eta_w(\xi)$-optimal designs are still locally optimal with respect to δ^2.

The important differences between the $\Psi_1(\xi)$ (3.10) and the $\eta_w(\xi)$ (7.3) is that the latter is derived from frequentist perspective and is equivalent to applying non-informative prior for $\boldsymbol{\beta}$; for this reason, we have the matrix $\mathbf{R}$ in (3.10) which is the prior variance-covariance matrix for the trend parameters. In (7.3) this matrix is set equal to zero. This connection leads to the results in this chapter demonstrating that the WIMSPE is also a good approximation to the Bayesian decision theoretic design approach.

Another popular design criterion for prediction in computer experiments is the minimisation of the average kriging variance which is equivalent to IMSPE; see Chapter 5 and the literature for spatial experiments.

When the aim of the experiment is the estimation of the unknown correlation parameters, the most popular designs maximise entropy in the posterior distribution (see Shewry and Wynn (1987), and Chapter 5 in this thesis). Currin et al. (1991) proposed an algorithm to find maximum entropy designs for computer experiments when the correlation parameters are known. Also, Johnson et al. (1990) stated that for very weak correlation, the entropy designs are maximin designs.

We do not chose to compare maximin entropy designs to Ψ-optimal designs because the maximum entropy design is tailored to the estimation of the unknown parameters, while our design is optimal for prediction.

The majority of the literature for designs for computer experiments is focused either on Latin hypercube designs or on a model-based approach with known correlation parameters. In this chapter, we apply our model-based approach to find designs for situations with unknown correlation parameters and also unknown trend parameters. Appling a Bayesian methodology allows the incorporation of uncertainty in all these features.

7.4 Bayesian Optimal Design for Computer Experiments

7.4.1 Assessment for closed-form approximation for computer experiments

The decision theoretic framework for optimal designs for prediction, as it is described in Chapter 3, can be applied in the context of computer experiments. Our proposed design criterion minimises the average posterior predictive variance, with objective function $\Psi(\xi)$ is given by (3.10). Throughout this chapter we consider the case of known and fixed noise-to-signal ratio δ^2, taking one of two values, $\delta^2 = 0$ and $\delta^2 = 1$, which correspond to nugget $\tau^2 = 0$ and when $\tau^2 = \sigma^2$.

In this section we demonstrate the relationship between $\Psi(\xi)$ and $\Psi_1(\xi)$, and study how the choice of the parameters in the experiment and model affects the accuracy of the approximation. We perform a factorial study similar to that in Chapter 5 with four crossed factors and two nested factors, each with either two or three levels. For each combination of factor levels, we evaluate the objective function $\Psi(\xi)$. The factors are listed below in Table 7.1, together with their levels and coded values.

The first level of the correlation function ($F_3 = 0$) indicates the same decay parameter for the three dimensions, i.e. an isotropic correlation function, whereas the second level corresponds to different decay parameter in each dimension, a separable correlation function.

	Levels	
Factors	0	1
F_1	$n = 5$	$n = 10$
F_2	$M = \beta_0$	$M = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
F_3	$\rho = \prod_{k=1}^{3} e^{-\phi\lvert x_{ik} - x_{jk}\rvert}$	$\rho = \prod_{k=1}^{3} e^{-\phi_k\lvert x_{ik} - x_{jk}\rvert}$
F_4	$\delta^2 = 0$	$\delta^2 = 1$

Table 7.1: Four crossed factors together with their levels and coded values.

We also have two nested factors which correspond to the prior distributions for the unknown parameters $\boldsymbol{\beta}$ and ϕ, factors F_5 and F_6 respectively. Factor F_5 determines the prior precision of the trend parameters and is nested within factor F_2 (form of mean function):

$$F_5|(F_2 = 0) = \begin{cases} 0 & \Rightarrow \text{ R=1} \\ 1 & \Rightarrow \text{ R=0.25.} \end{cases}$$

$$F_5|(F_2 = 1) = \begin{cases} 0 & \Rightarrow \mathbf{R} = \mathbf{I}_4 \\ 1 & \Rightarrow \mathbf{R} = 0.25\mathbf{I}_4. \end{cases}$$

Factor F_6 determines the prior distribution for the decay parameter ϕ and is nested within factor F_3 (form of correlation function):

$$F_6|(F_3 = 0) = \begin{cases} 0 & \Rightarrow \phi \sim \text{Unif}(0.1, 1) \\ 1 & \Rightarrow \phi \sim \text{log-normal}(-1.1, 1) \end{cases}$$

$$F_6|(F_3 = 1) = 0 \Rightarrow \begin{cases} \phi_1 \sim \text{Unif}(0.1, 1) \\ \phi_2 \sim \text{Unif}(1, 3) \\ \phi_3 \sim \text{Log-Normal}(-1.1, 1). \end{cases}$$

The hyperparameters a and b for the inverse gamma prior distribution for σ^2 are kept constant for all the combinations of F_1 to F_6, $\sigma^2 \sim \text{IG}(3, 1)$, similar to previous chapters.

All 48 possible combinations of the levels of these factors are considered. For each combination, we generate 200 random designs from $\mathcal{X} = [-1, 1]^3$ and assume prediction is required across a grid with $|\mathcal{X}_\mathcal{P}| = 40$ points chosen as a maximin LHD. For each design we evaluate $\Psi(\xi)$ (3.10) using Monte Carlo integration and quadrature. When $F_6|(F_3 = 0)$, $\Psi_1(\xi)$ and $\Psi_2(\xi)$, are approximated by (3.33) and (3.34) respectively. When $F_6|(F_3 = 1)$ then $\Psi_1(\xi)$ and $\Psi_2(\xi)$ approximated by

$$\Psi_1(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{2} \sum_{k=1}^{d} \sum_{i=1}^{m_1} w_{ik}^1 f_1 \left(\frac{b_{1k} - a_{1k}}{2} a_{ik}^1 + \frac{b_{1k} + a_{1k}}{2}, \mathbf{x}_p \right) d\mathbf{x}_p, \tag{7.4}$$

and

$$\Psi_2(\xi) \simeq \int_{\mathcal{X}_\mathcal{P}} \frac{1}{2} \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{d} \sum_{i=1}^{m_1} w_{ik}^1 f_2 \left(\frac{b_{1k} - a_{1k}}{2} a_{ik}^1 + \frac{b_{1k} + a_{1k}}{2}, \mathbf{x}_p, \mathbf{y}_j \right) d\mathbf{x}_p, \tag{7.5}$$

where $f_1(\cdot)$, $f_2(\cdot)$ is defined in 3.5.1. We approximate Ψ_1 and Ψ_2 via independent quadrature rules in each dimension, and w_{ik}^1 and a_{ik}^1 are the corresponding weights and nodes obtained from the Legendre polynomial for the uniform distributions for ϕ_k for the kth direction.

Table 7.2 shows the correlations between the values of $\Psi(\xi)$ and $\Psi_1(\xi)$ for the 48

			F_1 F_2 F_3							
F_4	F_5	F_6	000	001	010	011	100	101	110	111
0	0	0	0.99998	0.99901	0.99995	0.99808	0.99994	0.99995	0.99988	0.99872
0	0	1	0.99994	-	0.99992	-	0.99987	-	0.99994	-
0	1	0	0.99992	0.99903	0.99962	0.99895	0.99999	0.99962	0.99992	0.99987
0	1	1	0.99996	-	0.99613	-	0.99998	-	0.99970	-
1	0	0	0.99999	0.99946	0.99999	0.99997	0.99991	0.99978	0.99876	0.99951
1	0	1	0.99998	-	0.99937	-	0.99999	-	0.99999	-
1	1	0	0.99957	0.99997	0.99996	0.99930	0.99843	0.99858	0.99844	0.99934
1	1	1	0.99890	-	0.99919	-	0.99975	-	0.99260	-

Table 7.2: Correlation between objective functions Ψ and Ψ_1 for 200 random designs for each factor level combinations of $F_1, F_2, F_3, F_4, F_5, F_6$.

combinations of factors F_1 to F_6. The columns correspond to the number of runs (F_1) mean (F_2) and correlation (F_3) functions, and the rows to noise-to-signal ratio (F_4), precision matrix ($F_5|F_2$), and decay parameter ($F_6|F_3$). As can be seen, there is always a very high correlation between $\Psi(\xi)$ and $\Psi_1(\xi)$, i.e. almost equal to one. In fact, $\Psi(\xi) \simeq \Psi_1(\xi)$, with $\Psi_2(\xi)$ always close to zero. This study shows that our numerical results from Chapter 5 extend to models with more than two dimensions and with anisotropic correlation functions.

Therefore, from this substantial numerical evidence, we conclude that it is sufficient to approximate the objective function $\Psi(\xi)$ (3.10) by $\Psi_1(\xi)$ alone. In what follows in the next sections, we use minimisation of $\Psi_1(\xi)$ as a design selection criterion.

In Chapter 3 we gave a theoretical insight about the closed-form approximation. In fact, the assumptions necessary for those results are not restricted to isotropic correlation functions, and can be extended. Recently, Ren et al. (2013), focusing on parameter estimation in spatial modelling, showed that the integrated likelihood for anisotropic correlation functions is also a bounded function of the correlation parameters. Hence, the integrated likelihood tends to zero faster than the correlation parameters (Lemma 3.1 in Chapter 3).

7.4.2 Example

We now present an example of a simple computer experiment, and compare the designs resulting from model-free and model-based approaches. We compare the maximin Latin hypercube design and our Bayesian Ψ-optimal design.

This example uses data from a simple simulator of a helical compression spring (Tudose and Jucan (2007), Forrester et al. (2008, p. 200-202)). The model can be used to determine the correct dimensions and geometry of the input values required to satisfy a given loading condition. The three main characteristics of a helical compression spring that we take into account are the wire diameter, the spring index and the coefficient of the distance between the coils at the maximum load.

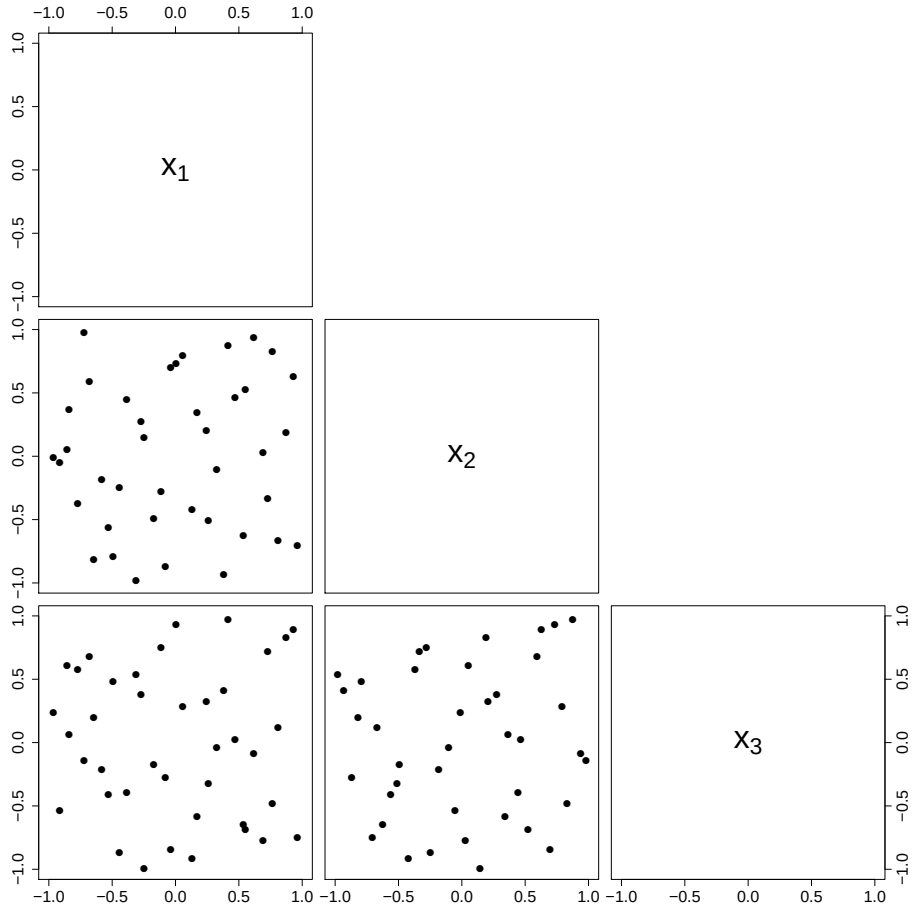


Figure 7.2: Helical spring example: Prediction points, $|\mathcal{X}_{\mathcal{P}}| = 40$, obtained by a maximin Latin Hypercube design.

Available data are modelled by a Gaussian process model given by (2.6), with a linear mean function, and the correlation between the three variables is modelled by

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \phi) = \prod_{k=1}^{3} \exp(-\phi_k |x_{ik} - x_{jk}|) \quad \text{for} \quad \mathbf{x}_i, \mathbf{x}_j \in [-1, 1]^3. \tag{7.6}$$

Our aim is to find Ψ-optimal designs with $n = 10$ runs when we want to predict at 40 untested points, i.e. $|\mathcal{X}_{\mathcal{P}}| = 40$ for two different prior distributions on the correlation parameters. The prediction points are obtained by a maximin LHD and the projections are shown in Figure 7.11.

We consider conjugate normal inverse gamma prior distributions for trend parameters $\boldsymbol{\beta}$ and Gaussian process variance σ^2, and assign two different priors for the correlation parameters ϕ and δ^2:

- prior 1: δ^2=0, $\phi_1 \sim \text{Unif}(1, 3)$, $\phi_2 \sim \text{Unif}(3, 5)$, $\phi_3 \simeq 0$,
- prior 2: δ^2=0.5, $\phi_1 \sim \text{Unif}(1, 3)$, $\phi_2 \sim \text{Unif}(1, 3)$, $\phi_3 = 0$.

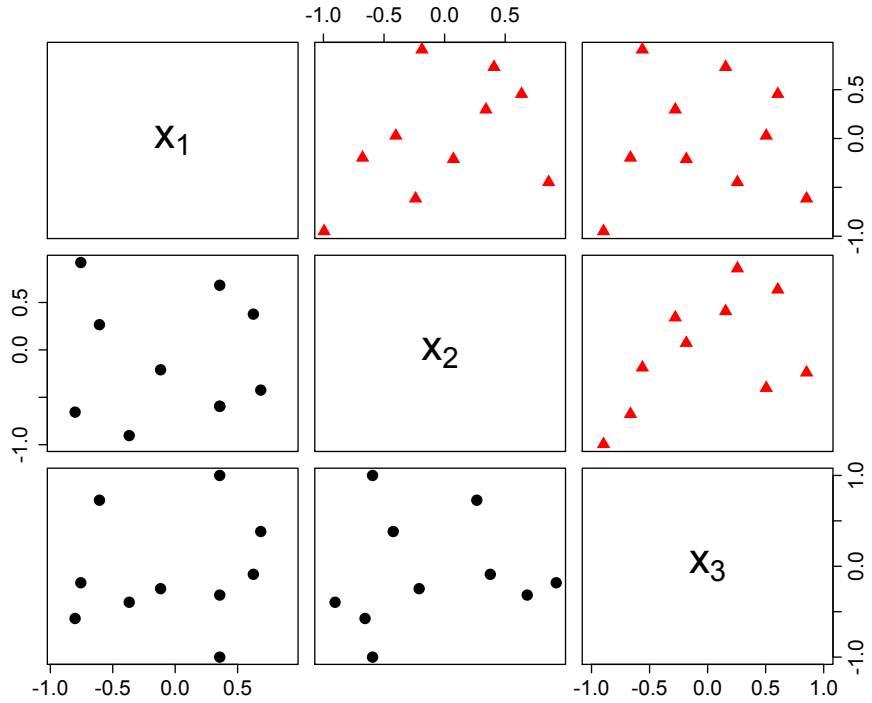


Figure 7.3: Helical spring example: Ψ-optimal design (●) and maximin Latin hypercube design (▲) for prior 1.

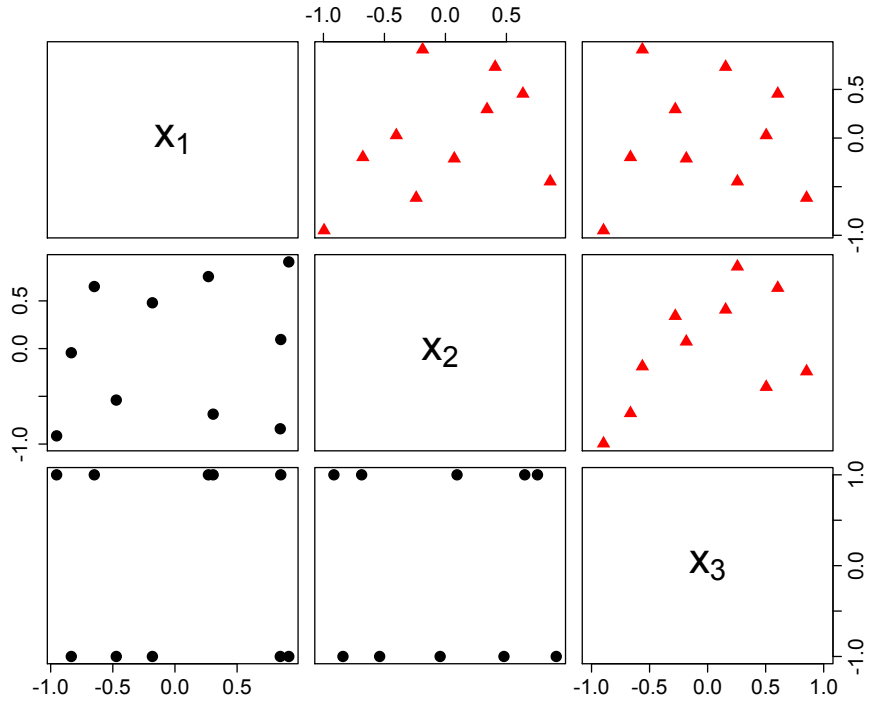


Figure 7.4: Helical spring example: Ψ-optimal design (●) and maximin Latin hypercube design (▲) for prior 2.

For prior 1, we considered the model without a nugget effect, i.e. $\tau^2 = 0$, and assign different priors for the correlation parameter ϕ in each dimension. In prior 2, we introduce the nugget effect in the model, so there is a random error, and also the correlation in the third variable is equal to one for all $|x_{ik} - x_{jk}|$. These prior distributions were obtained by analysing data from a maximin Latin hypercube (Morris and Mitchell, 1995) obtained by Tudose and Jucan (2007).

The Figures 7.3 and 7.4 show Ψ-optimal designs, denoted by black dots, and a maximin Latin hypercube design, denoted by red triangles, for prior 1 and prior 2 respectively.

For prior 1, the Ψ-optimal design has similar space-filling properties as the maximin LHD. The average inter-point distance is 1.43 for the Ψ-optimal design and 1.40 for the maximin LHD. However, the Ψ-optimal design has 30% smaller average posterior predictive variance. For prior 2, where there is a change in correlation strength as a nugget is introduced to the model and the correlation parameter ϕ_3 in x_3 is now zero, we observe that the design points in the third dimension collapse onto the extremes. This happens because there is constant correlation between the observations in the third dimension, and hence the design is heavily influenced by the linear trend. This second design has posterior predictive variance 18% lower than that of the LHD.

To summarise, this example indicates the advantages of using a model-based approach to the design of computer experiments, where sufficient prior information is available, and is in line with the conclusions from Pronzato and Müller (2012).

The designs we have studied are influenced by the degree of correlation, with larger correlation parameters and hence greater changes in correlation leading to designs which are close to space-filling, but still providing lower prediction variance than LHDs. However, in practice model-based optimal designs are more difficult to construct than space-filling designs because they require more prior information, including complete specification of the mean function and the covariance structure of the Gaussian process model.

7.5 Examples of Optimal Design for Computer Experiments

In this section, we present further examples of Ψ-optimal designs and more comparisons to space-filling designs. For all the examples presented in this section, we assume a Gaussian process model (2.6) with either constant or linear mean function. The prior distributions for the unknown correlation parameters are assumed to be either uniform or log-normal. The noise to signal ratio is considered know and fixed at two values $\delta^2 = 0$ and $\delta^2 = 1$. The correlation function is either assumed to be an isotropic or a separable power exponential or Matérn function. If an isotropic correlation function is assumed, then $\Psi_1(\xi)$ is approximated by (3.33), otherwise for a separable function the

numerical approximation when a uniform prior is assumed for each ϕ in each dimension is given by (7.4) .

The method for selecting an optimal design for computer experiments is as follows:

1. Generate 50 randomly selected starting designs from the region $\mathcal{X} = [0,1]^d$, $d = 2$ and $d = 3$, with the number of points required for each example, i.e. $n = 3, 5, 7$ or 10.

2. For each starting design, use the coordinate exchange algorithm (Section 3.6.1) to find a design that minimises $\Psi_1(\xi)$, approximated via (3.33) for an isotropic correlation functions and (7.4) for an anisotropic correlation function.

3. From the 50 designs obtained by the algorithm, select the design with the minimum $\Psi_1(\xi)$ value. In the event of ties, choose at random from the tied designs.

7.5.1 Ψ-optimal designs for $d = 2$

Here, we demonstrate our methodology for $d = 2$ and obtain Bayesian optimal designs with $n = 3, 5, 7$ and the aim is to predict on a 10×10 regular grid. We choose these numbers of points, and the correlation function, in order to compare our results with the minimax and maximin designs of Johnson et al. (1990). Hence, we adopt the exponential correlation function with Euclidean distance, i.e.

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \phi) = \exp\left[-\phi\left\{\sum_{k=1}^{2}(x_{ik} - x_{jk})^2\right\}^{1/2}\right].$$

For one example with $n = 7$, we also consider the exponential correlation function with rectangular distance,

$$\rho(\mathbf{x}_i, \mathbf{x}_j; \phi) = \exp\left[-\phi\sum_{k=1}^{2}|x_{ik} - x_{jk}|\right]. \tag{7.7}$$

(i) **Euclidean distance**

The Ψ-optimal designs for Euclidean distance and constant mean function are similar for $n = 3, 5, 7$; all design points lie in the interior of the study region, $\mathcal{X} = [0,1]^2$, with no points at the edges. When there is no nugget in the model, i.e. $\delta^2 = 0$, there are points near the centre of $\mathcal{X}$ (Figure 7.5, top row (a) and (b)).

When a nugget is added, i.e. $\delta^2 = 1$, the correlation between two observations a given distance apart in $\mathcal{X}$ is smaller than when $\delta^2 = 0$, and the centre point moves towards an edge of $\mathcal{X}$. The design gives less coverage compared to the designs for high correlation. This is illustrated in bottom row of Figure 7.5 (c) and (d).

Figure 7.6 shows the Ψ-optimal designs for a linear mean trend and correlation using

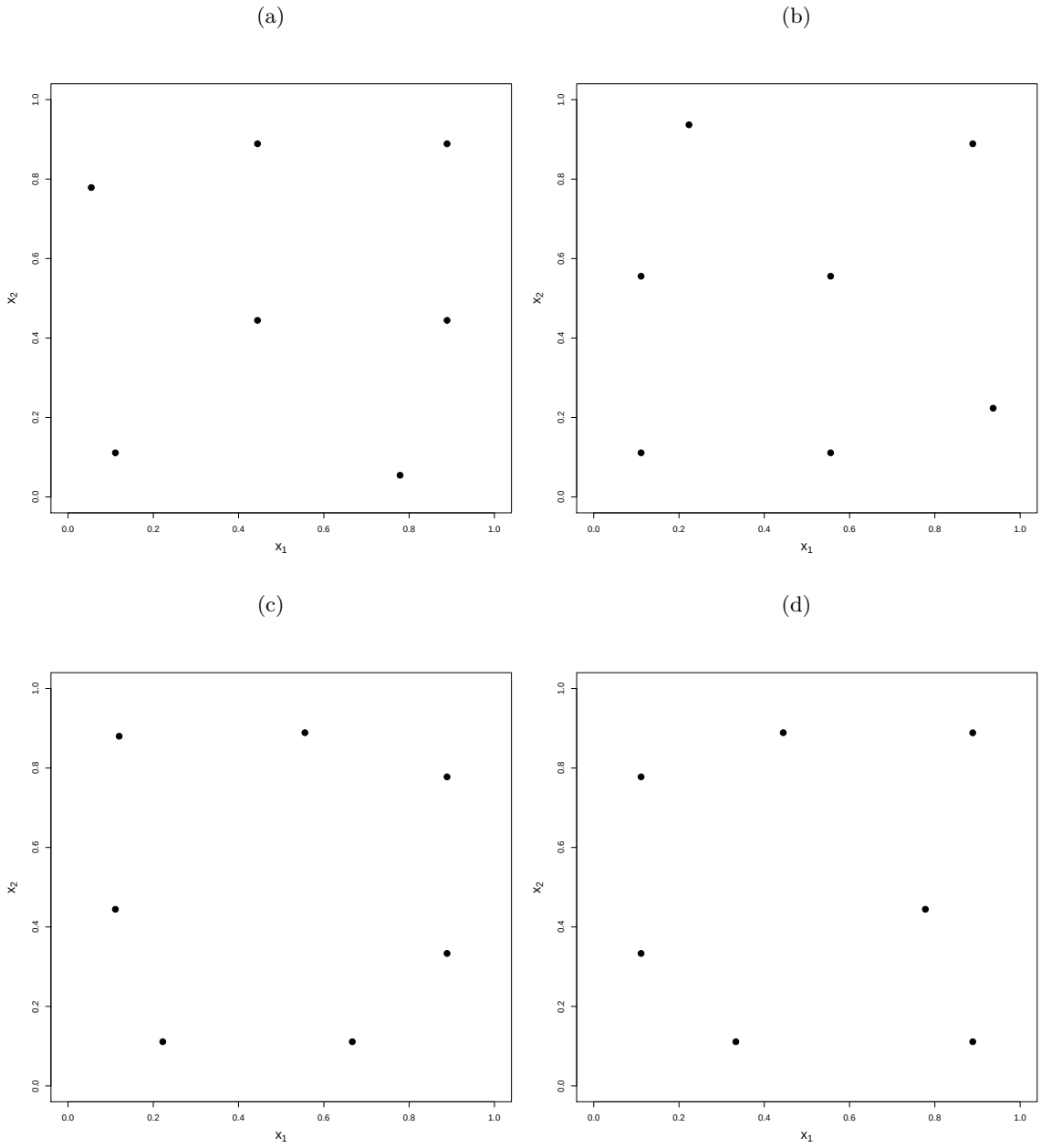


Figure 7.5: Bayesian Ψ-optimal designs with constant mean and correlation using euclidean distance: (a) uniform prior and $\delta^2 = 0$; (b) log-normal prior and $\delta^2 = 0$; (c)uniform prior and $\delta^2 = 1$; (d) log-normal prior and $\delta^2 = 1$.

Euclidean distance and $n = 7$. When $\delta^2 = 0$, the Ψ-optimal design has points distributed across the study region, with points in the interior of $\mathcal{X}$. Once again, when $\delta^2 = 1$, the design points move towards to the boundaries of the region. The bottom row of Figure 7.6, (c) and (d), illustrate this distribution of points further. There are points at the corners of the region, matching the optimal design for prediction from a linear model with uncorrelated observations.

We only present the designs for $n = 7$, however this pattern holds regardless of the number of runs. Similarly, by comparing plots (a) with (b), and (c) with (d), we see

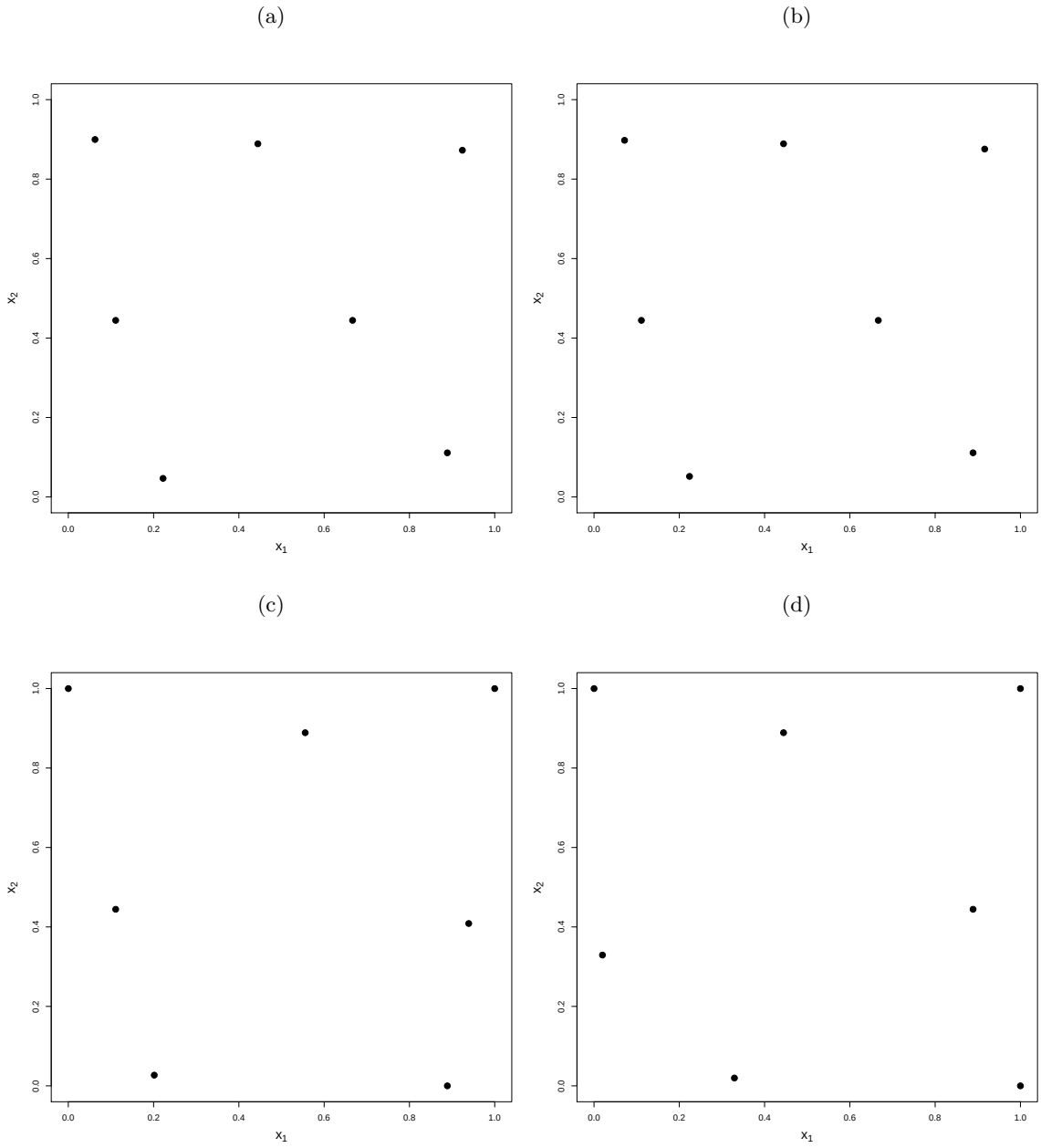


Figure 7.6: Bayesian Ψ-optimal designs with linear mean and correlation using euclidean distance: (a) uniform prior and $\delta^2 = 0$; (b) log-normal prior and $\delta^2 = 0$; (c)uniform prior and $\delta^2 = 1$; (d) log-normal prior and $\delta^2 = 1$.

that the choice of prior distribution for the correlation parameter ϕ also has little input on the design. The designs are affected by the range and degree of correlation, mostly determined by δ^2, and the choice of mean function.

Figure 7.7 gives the minimax and maximin designs with Euclidean distance as found in Johnson et al. (1990). Based on a visual comparison between minimax and maximin designs, Figure 7.7 (a) and (b), and Ψ-optimal designs, Figures 7.5 and 7.6, we conclude that Ψ-optimal designs with constant mean and Ψ-optimal designs with linear mean and $\delta^2 = 0$ distribute the points similarly to the minimax design. When $\delta^2 = 1$, then

(a) (b)

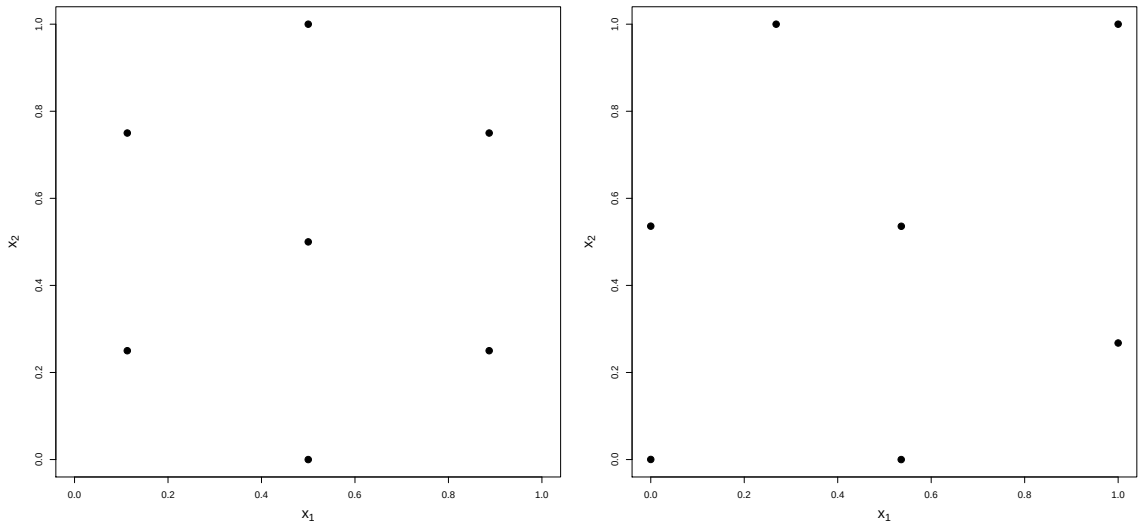


Figure 7.7: (a) Minimax design and (b) Maximin design for 7 points designs with Euclidean distance.

Ψ-optimal designs are more similar to maximin designs, especially when a linear trend is assumed.

In order to numerically compare Ψ-optimal designs with the corresponding minimax and maximin designs, for each design we evaluate (i) the objective function $\Psi_1(\xi)$; (ii) the inter-point Euclidean distance.

		Mean function					
		Constant			Linear		
n		Ψ-optimal	Minimax	Maximin	Ψ-optimal	Minimax	Maximin
3	$\delta^2 = 0$	0.0993	0.1094	0.1157	0.1219	0.1619	0.1272
	$\delta^2 = 0$	0.0904	0.0985	0.1046	0.1127	0.1483	0.1162
	$\delta^2 = 1$	0.7648	0.7687	0.7808	0.8209	1.2775	0.9419
	$\delta^2 = 1$	0.7504	0.7808	0.7642	0.8089	1.2678	0.9311
5	$\delta^2 = 0$	0.0709	0.0752	0.0844	0.0738	0.0873	0.0850
	$\delta^2 = 0$	0.0663	0.0709	0.0782	0.0699	0.0817	0.0793
	$\delta^2 = 1$	0.7038	0.7048	0.7152	0.7415	0.9094	0.7818
	$\delta^2 = 1$	0.6909	0.6917	0.7009	0.7318	0.9033	0.7734
7	$\delta^2 = 0$	0.0584	0.0601	0.0643	0.0605	0.0638	0.0653
	$\delta^2 = 0$	0.0551	0.0568	0.0606	0.0576	0.0606	0.0619
	$\delta^2 = 1$	0.6730	0.6740	0.6777	0.7006	0.7755	0.7424
	$\delta^2 = 1$	0.6612	0.6621	0.6655	0.6922	0.7699	0.7361

Table 7.3: Ψ_1 objective function values for Ψ-optimal designs, minimax and maximin designs for Euclidean distance, and $n = 3, 5, 7$.

Table 7.3 shows the values of the objective function Ψ_1 under of each one of the 24 combinations. As expected, Ψ-optimal designs always give smaller values for the Ψ_1 objective function than either the minimax or maximin designs. In the case of constant mean, the difference between the three designs is small, whereas for linear trend the difference is more obvious, up to 18%, especially, when a nugget is included in the model. The Ψ-optimal design takes into account the uncertainty of the trend parameters and with smaller correlation, the design is more strongly influenced by the need to estimate the trend parameters.

The average inter-point Euclidean distances are 0.72, 0.71 and 0.66 for the Ψ-optimal design, with $n = 3, 5$ and 7 respectively, and constant mean, $\delta^2 = 0$ and either a uniform or a log normal prior distribution for ϕ; when nugget is added to the model, $\delta^2 = 1$, the average inter-point distances are 0.64, 0.66 and 0.65 for $n = 3, 5$ and 7. The corresponding average inter-point distances for maximin design is $1.03, 0.96, 0.79$ and for the minimax design is $0.54, 0.59, 0.64$.

Therefore space-filling properties of the Ψ-optimal design are somewhere between the minimax and maximin designs, i.e. Ψ-optimal designs compromise between minimax and maximin designs.

A similar pattern in the inter-point distances occurs for the designs for a linear trend when $\delta^2 = 0$. However, when there is a nugget in the model the average distance increases for the Ψ-optimal design as the points are influenced by the linear trend and move towards to the boundaries of $\mathcal{X}$.

(ii) **Rectangular distance**

We now find Ψ-optimal designs using the correlation function (7.7) based on rectangular distance. We keep all other model parameters the same. We present designs for $n = 7$ and again compare to maximin and minimax designs found by Johnson et al. (1990).

Figure 7.8 gives Ψ-optimal designs for constant mean function and Figure 7.9 for a mean function being a linear trend. These designs display similar distributions of points to these found using Euclidean distance. Figure 7.9 indicates that, in general, use of the rectangular distance and linear trend results in designs with more points in the interior of the design region than use of Euclidean distance. When a nugget is included, $\delta^2 = 1$, the number of points at or near the boundaries is again larger compare to the designs with $\delta^2 = 0$.

The minimax and maximin designs with rectangular distance, from Johnson et al. (1990), are given in Figure 7.10. The points are distributed throughout the region, and having a similar pattern to the Ψ-optimal designs with $\delta^2 = 0$. The average inter-point rectangular distance for the minimax design is 0.888 and for the maximin design is 1.057. The Ψ-optimal design with constant mean, uniform prior on ϕ and $\delta^2 = 0$ has very similar average rectangular distance, 0.868, to the minimax design, and the other three cases from Figure 7.8 have distances of 0.921, 0.922 and 0.922

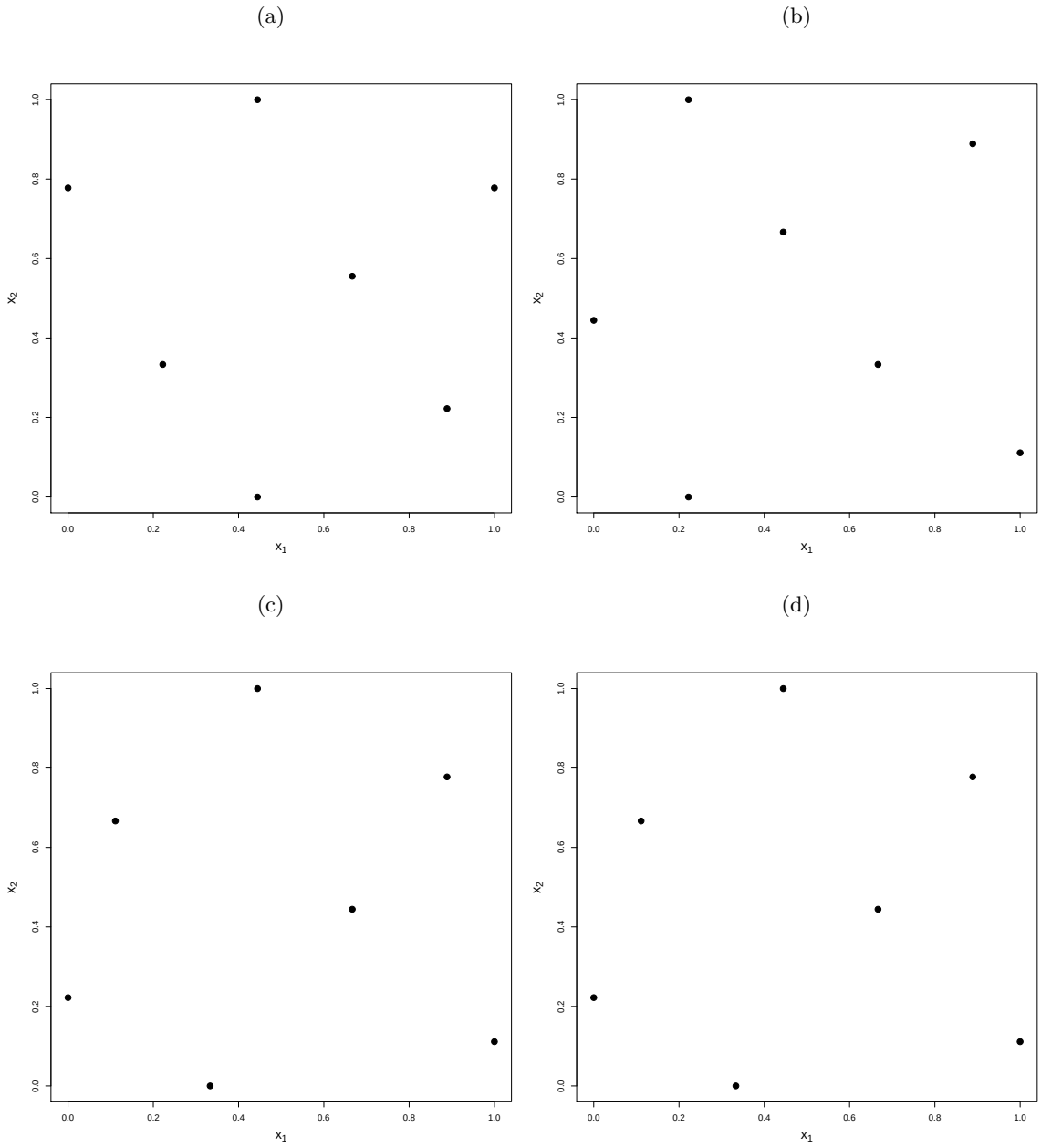


Figure 7.8: Bayesian Ψ-optimal designs with constant mean and correlation using rectangular distance: (a) uniform prior and $\delta^2 = 0$; (b) log-normal prior and $\delta^2 = 0$; (c)uniform prior and $\delta^2 = 1$; (d) log-normal prior and $\delta^2 = 1$.

which are between the average inter-point rectangular distance values for the minimax and maximin designs. For the case of linear mean, see Figure 7.9, the two Ψ-optimal designs from the top row both have average rectangular distance 0.868. When $\delta^2 = 1$, and the points are pushed to the boundaries, Ψ-optimal designs of the bottom row have average rectangular distance close to the maximin design, 1.01.

We again also compare designs by evaluating the objective function $\Psi_1(\xi)$ (Table 7.4). Similar to the previous case of Euclidean distance, all the designs have almost the same average posterior predictive variance and the Bayesian optimal designs give slightly

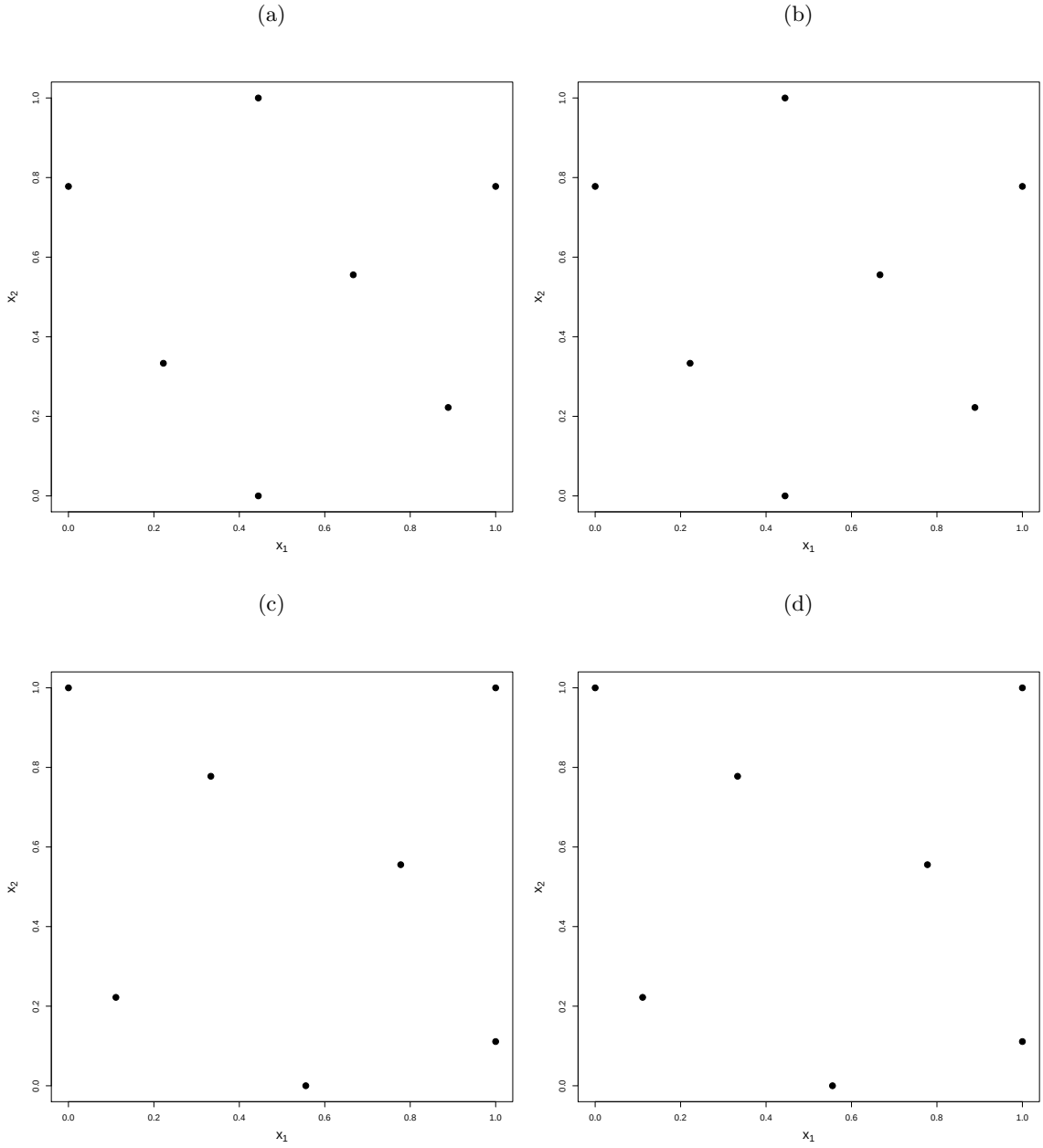


Figure 7.9: Bayesian Ψ-optimal designs with linear mean and correlation using rectangular distance: (a) uniform prior and $\delta^2 = 0$; (b) log-normal prior and $\delta^2 = 0$; (c)uniform prior and $\delta^2 = 1$; (d) log-normal prior and $\delta^2 = 1$.

smaller values compared to the space-filling designs. Also, Ψ-optimal designs always give smaller values for the Ψ_1 objective function, and when a nugget is included in the model the difference is up to 13%.

To summarise, the Ψ-optimal designs for $d = 2$ for both distance metrics give the smallest value for the objective function $\Psi_1(\xi)$ and also they have smaller average prediction variance compared to the very popular maximin and minimax designs. Usually when $\delta^2 = 0$, the Ψ-optimal designs are closer to the minimax designs, and when $\delta^2 = 1$ they are closer to the maximin. This relationship is due to the reduced correlation; $\delta^2 = 1$

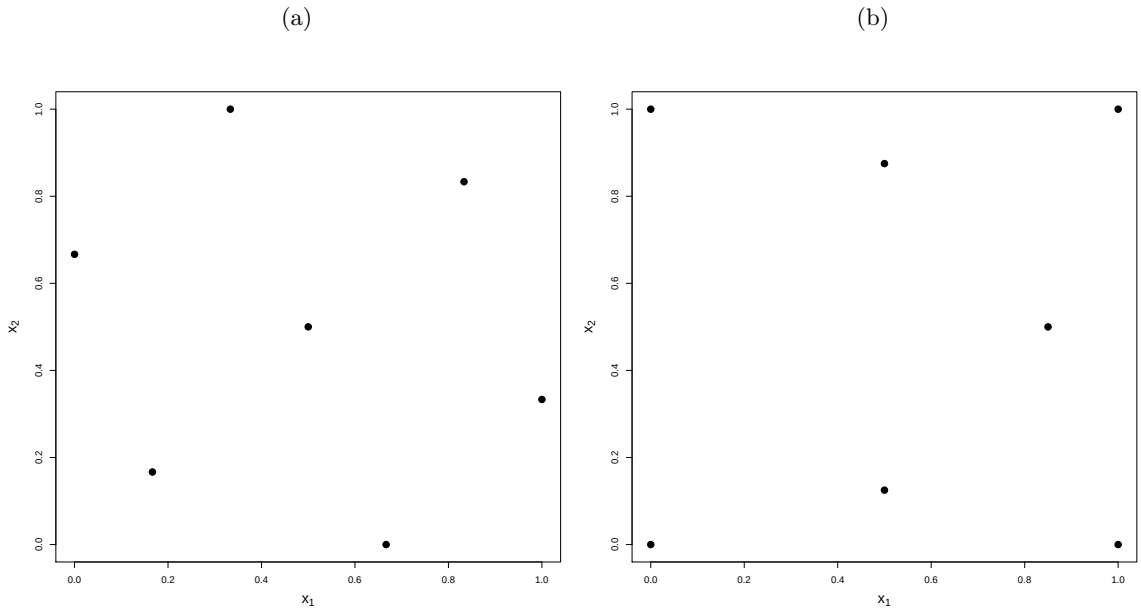


Figure 7.10: (a) Minimax design and (b) Maximin design for 7 points designs with rectangular distance.

leads to the points being moved closer to the boundaries $\mathcal{X}$.

7.5.2 Ψ-optimal designs for $d = 3$

In this section, we demonstrate the methodology for $d = 3$ variables. Ψ-optimal designs are found for $n = 5$ and $n = 10$ points in the study region $\mathcal{X} = [-1, 1]^3$, when the aim is to predict at $|\mathcal{X}_\mathcal{P}| = 40$ points chosen from a random Latin Hypercube sample, see Figure 7.11. We investigate the influence of five crossed factors and one nested factor, shown in Table 7.5, on the designs found. Once again, these factors vary features of the model and design.

Nested factor F_6 determines the prior precision of the trend parameters, and is nested

		Mean function					
		Constant			Linear		
n	$\Psi_1(\xi)$	Ψ-optimal	Minimax	Maximin	Ψ-optimal	Minimax	Maximin
7	$\phi = 0\ \delta^2 = 0$	0.0619	0.0647	0.0753	0.0635	0.0668	0.0754
	$\phi = 1\ \delta^2 = 0$	0.0992	0.1028	0.1119	0.1067	0.1100	0.1163
	$\phi = 0\ \delta^2 = 1$	0.6888	0.6896	0.6985	0.7135	0.7658	0.7483
	$\phi = 1\ \delta^2 = 1$	0.7066	0.7087	0.7162	0.7034	0.8015	0.7817

Table 7.4: Ψ_1 objective function values for Ψ-optimal designs, minimax and maximin designs for rectangular distance, and $n = 7$.

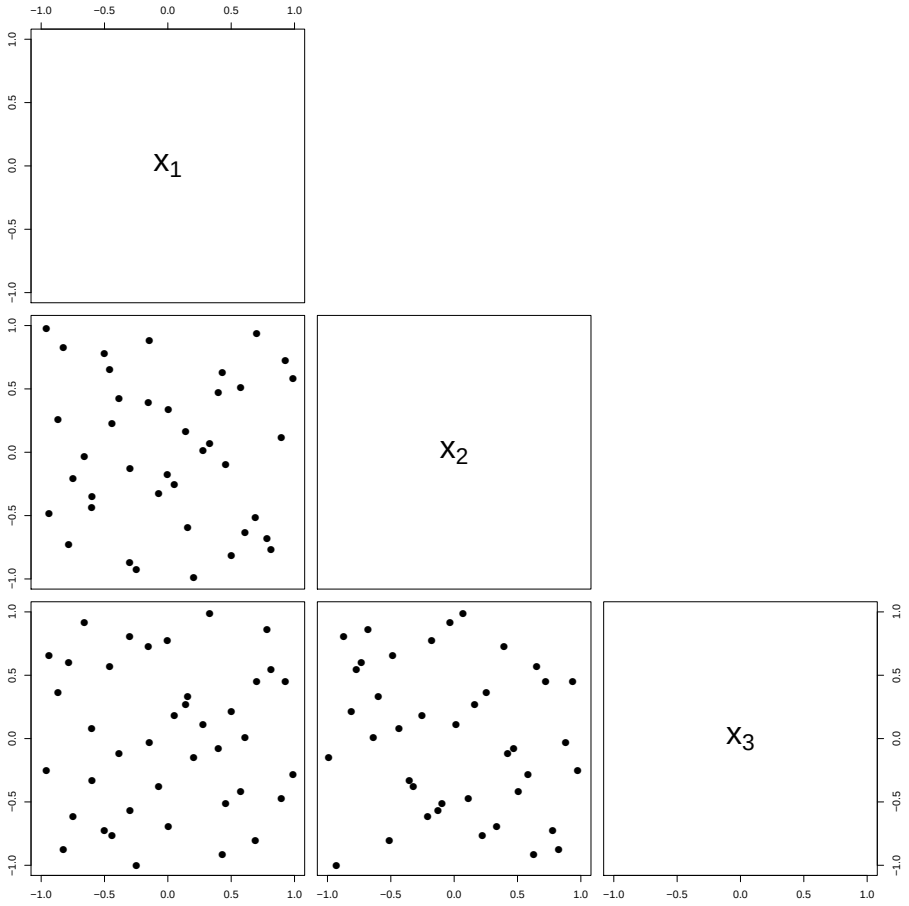


Figure 7.11: Prediction points, $|\mathcal{X}_{\mathcal{P}}| = 40$ obtained by a maximin Latin Hypercube design.

within factor F_2 (form of mean function):

$$F_5|(F_2 = 0) = \begin{cases} 0 & \Rightarrow \text{ R=1} \\ 1 & \Rightarrow \text{ R=0.25.} \end{cases}$$

$$F_5|(F_2 = 1) = \begin{cases} 0 & \Rightarrow \mathbf{R} = \mathbf{I}_4 \\ 1 & \Rightarrow \mathbf{R} = 0.25\mathbf{I}_4. \end{cases}$$

In total, we investigate 32 combinations of these individual settings. For each combination, we generate 20 random designs selected from the design region. For each of these starting designs, the coordinate exchange algorithm (Section 3.6.1) is used to find a design that minimises $\Psi_1(\xi)$. The final choice of Ψ-optimal design is that which has the smallest value of $\Psi_1(\xi)$ among these 20 designs. Here, we will present the result for n = 10; similarly results and conclusions are obtained for $n = 5$ (Appendix 7.3).

To assess and compare the space-filling properties of our designs we will use coverage and spread as defined in Section 3.2.1. These measures are given for all designs in Table 7.6.

	Levels	
Factors	0	1
F_1	$n = 5$	$n = 10$
F_2	$M = \beta_0$	$M = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
F_3	$\rho = \prod_{k=1}^{3} e^{(-\phi_k \lvert x_{ik} - x_{jk} \rvert)}$	$\rho = \prod_{k=1}^{3} e^{(-\phi_k (x_{ik} - x_{jk})^2)}$
F_4	$\delta^2 = 0$	$\delta^2 = 1$
F_5	$\phi_1, \phi_3 \sim \text{Unif}(0.1, 1)$	$\phi_1 \sim \text{Unif}(0.1, 1)$
	$\phi_2 \sim \text{Unif}(2, 3)$	$\phi_2 \sim \text{Unif}(2, 3), \phi_3 \sim \text{Unif}(0.2, 0.5)$

Table 7.5: Five crossed factors together with their levels and coded values.

Each of the plots in Figures 7.12 - 7.13 and A.37-A.42 shows the 2-D projections of Ψ-optimal design for one of the settings of F_1 to F_5; the upper triangle corresponds to design with a nugget effect in the model, i.e. $\delta^2 = 1$ ($F_4 = 1$), and the lower triangle corresponds to the case of $\delta^2 = 0$ ($F_4 = 0$).

From visual inspection of the plots in Figures 7.12 and 7.13, and in Appendix A.6 Figures A.37-A.42, we conclude

i. Including a linear trend changes the designs, and there is interaction between including a linear trend and the value of δ^2 (see 7.12 and 7.13). When the mean function is constant ($F_2 = 0$), then the design points are distributed uniformly across the design region $\mathcal{X}$ for both $\delta^2 = 0$ and $\delta^2 = 1$. The designs have similar coverage values (Table 7.6) and most points lie in the interior of the region. When a linear trend is assumed ($F_2 = 1$), the points are more spread out, and if we compare the spread values between $F_2 = 0$ and $F_2 = 1$, we can see that these values increase up to 15% and 23% for $\delta^2 = 0$ and $\delta^2 = 1$ respectively.

ii. No other factors make a substantial difference to the designs. In fact based on Table 7.6 we can see that the corresponding designs for $F_6 = 0$ and $F_6 = 1$ have similar coverage and spread values; especially when $F_3 = 1$ these values are almost equal. Also for the two specific levels of F_5, the designs are quite robust in the choice of the prior distribution of decay parameters; the coverage and spread values are close with slightly smaller coverage values and slightly larger spread values for $F_5 = 1$ compared to $F_5 = 0$.

Table 7.6 shows the coverage (3.1) and spread (3.2) values for the maximin LHD, maximin design and Ψ-optimal designs found for 32 combinations of F_1 to F_5. From Table 7.6, it can be seen that in general Ψ-optimal designs spread out the points more than the maximin LHD and the maximin designs, but have high coverage values. That is, under coverage, they are not as good as maximin LHD or maximin designs.

Finally, we assess robustness of Ψ-optimal designs to the choice of the prediction grid. The examples up to this point used 40 prediction points from a random LHD. Now for four combinations of F_1 to F_5, $000010, 000110, 010010, 010110$, Ψ-optimal designs were found for 10 different sets of 40 prediction points, each from a different LHD.

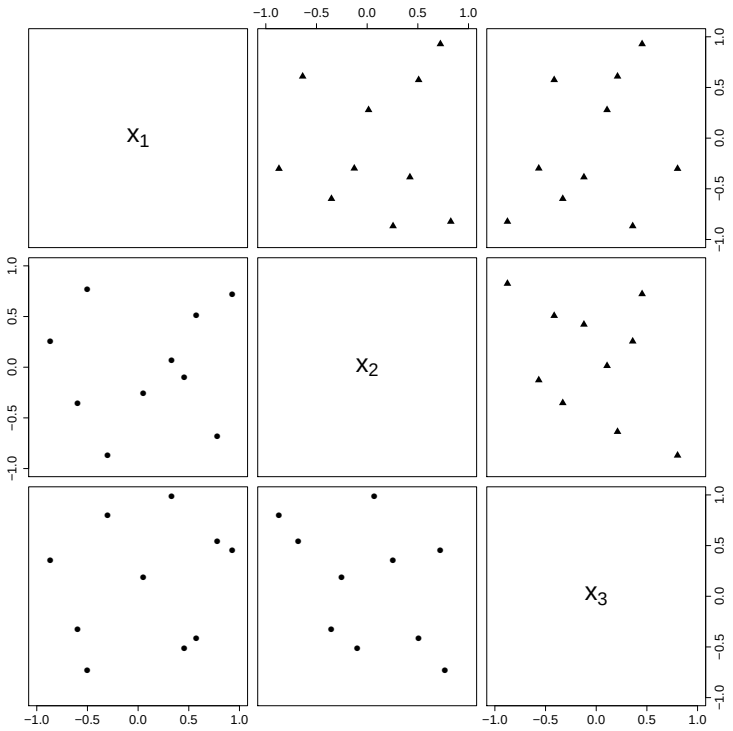


Figure 7.12: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for $F_1 - F_5 = 100000$ (●) and 100100 (▲).

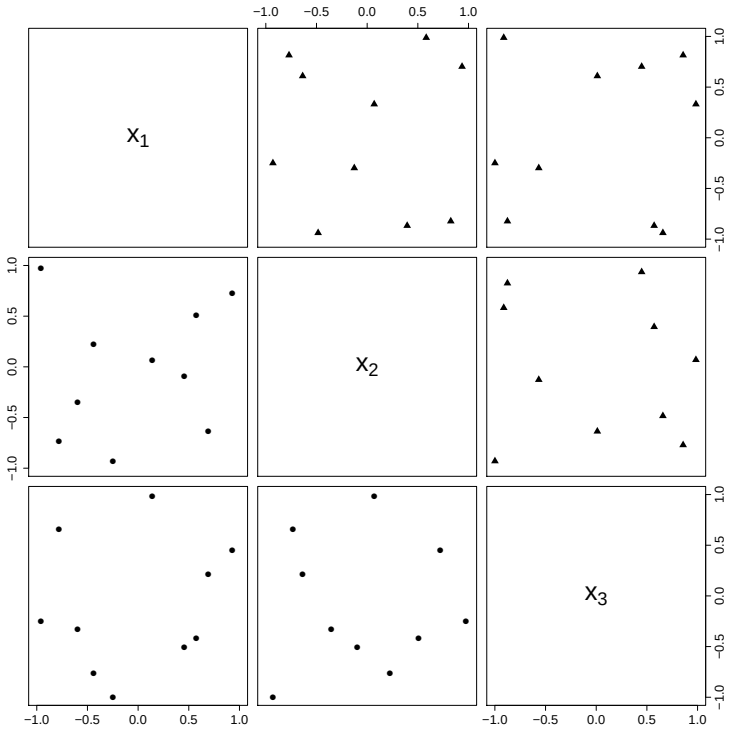


Figure 7.13: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for $F_1 - F_5 = 110000$ (●) and 110100 (▲).

					F_5			
					0		1	
					Coverage	Spread	Coverage	Spread
Maximin LHD					0.257398	0.452268	0.257398	0.452268
Maximin design					0.322719	0.721585	0.322719	0.721585
F_1	F_2	F_3	F_4	F_6				
1	0	0	0	0	0.508650	0.873641	0.474310	0.781800
1	0	0	0	1	0.474311	0.861156	0.504193	0.819752
1	0	0	1	0	0.536063	0.757443	0.495561	0.780442
1	0	0	1	1	0.565556	0.791025	0.495392	0.775948
1	0	1	0	0	0.503221	0.865005	0.505389	0.866427
1	0	1	0	1	0.503968	0.869538	0.506063	0.870295
1	0	1	1	0	0.513142	0.867845	0.514124	0.882879
1	0	1	1	1	0.517175	0.884678	0.505793	0.922701
1	1	0	0	0	0.533666	0.884293	0.505793	0.922701
1	1	0	0	1	0.548022	0.900444	0.527085	0.949808
1	1	0	1	0	0.528585	0.998736	0.521236	1.027515
1	1	0	1	1	0.503219	1.045473	0.529560	1.011731
1	1	1	0	0	0.505972	0.897517	0.503842	0.897859
1	1	1	0	1	0.508097	0.911310	0.505413	0.909315
1	1	1	1	0	0.584152	1.013485	0.505793	0.906742
1	1	1	1	1	0.604165	1.075390	0.580091	1.005707

Table 7.6: Coverage and spread for the maximin LHD, maximin design and Ψ-optimal designs found for 32 combinations of F_1 to F_5. Note that the value of ϕ does not affect the coverage and spread for the maximin LHD and maximin designs; the values are regarded to aid comparisons.

Table 7.7 shows the values of the objective function $\Psi_1(\xi)$ (3.33) and the coverage and spread values for each of the 10 designs; $\Psi_1(\xi)$ is calculated each time using the LHD for which the designs was found.

The value of the objective function varies between the prediction sets but the difference is small, i.e. the maximum difference for the combinations 000010 and 000110 is 5%, for 010010 is 8% and for 010110 is 6%. However, the spread and coverage values of the designs vary substantially. This is may because, for each combination of F_1 to F_6, there are many near Ψ-optimal designs, each having different space-filling properties. When we apply the coordinate exchange algorithm to obtain the optimal designs, in the event of ties between designs found for different runs of the algorithm, we pick up a design at random. The numerical evidence from Table 7.7 indicates that in our future work, it may be beneficial to adjust our algorithm in the event of ties to pick the design with the best space-filling properties.

	F_1 F_2 F_3 F_4 F_5 F_6											
	000010			000110			010010			010110		
	Ψ_1	C	S	Ψ_1	C	S	Ψ_1	C	S	Ψ_1	C	S
1	0.338	1.029	0.993	0.320	0.624	0.915	0.389	0.559	1.347	0.429	1.001	1.496
2	0.334	0.903	0.855	0.321	0.768	1.055	0.411	0.916	1.453	0.421	0.764	1.524
3	0.339	0.734	1.130	0.321	0.732	0.853	0.393	0.755	1.451	0.428	0.507	1.563
4	0.344	0.763	0.857	0.323	0.642	1.038	0.413	0.702	1.368	0.405	0.956	1.538
5	0.328	0.765	1.359	0.309	0.505	0.897	0.408	0.785	1.279	0.426	0.829	1.579
6	0.336	0.899	0.933	0.322	0.833	0.943	0.397	0.819	1.511	0.435	1.156	1.407
7	0.332	0.581	0.870	0.320	0.717	0.903	0.386	0.972	1.372	0.408	1.069	1.532
8	0.338	0.747	0.895	0.307	0.576	1.131	0.383	0.919	1.337	0.425	0.667	1.572
9	0.331	0.585	1.118	0.321	0.590	0.964	0.394	0.858	1.283	0.434	1.028	1.424
10	0.333	0.824	1.159	0.325	0.538	0.889	0.418	0.648	1.383	0.405	0.669	1.584

Table 7.7: Values of the objective function Ψ_1 (3.10), coverage and spread, denoted by C and S respectively, and four different combinations of factors $F_1 - F_6$ for Ψ-optimal designs found using 10 different prediction sets.

7.6 Summary and Discussion

The design of computer experiments is a growing research area with many practical applications. The goal of this chapter was to demonstrate how we can go beyond the standard plug-in approach, which is mainly found in the literature, to a fully Bayesian approach, including the specification of priors for the Gaussian process model parameters. Use of the closed-form approximation to the objective function from Chapter 3 allowed Bayesian optimal designs to be found numerically.

In this chapter, we made the common assumption of stationarity of the Gaussian process model. If this assumption is violated then there are implications in both design and analysis of computer experiments. A way to overcome this problem is to use the Bayesian treed Gaussian process models proposed by Gramacy and Lee (2008), where the input space is partitioned and a different stationary Gaussian process model is fitted in each partition. An additional research problem from the design prospective is how we can optimally select the points to partition the input space and fit the model using the treed partition.

In many computer experiments, at each design point information is available on the partial derivatives of the computer model with respect to the input variables. This derivative information is important for transmission of error and sensitivity analysis, i.e. how uncertainty in the model inputs relates to uncertainty in the model outputs. Moreover, derivatives can help with input screening, to identify which parameters are important. Obtaining derivative information has a computational cost. Therefore, the problem for design is at which design points the response should be observed, and at which points both response and derivatives should be observed. The optimal design, i.e. the optimal choice of the design points where the response and/or the derivatives are observed is crucial, as it allows us to save computational time and resources. The partial derivatives of a Gaussian process are also Gaussian processes and, as a result, joint modelling of the response and derivatives is possible. Morris et al. (1993) applied this

joint modelling and investigated maximin space-filling designs. However they assumed that at every design point you observe both the response and the derivatives and, as a result, there is extra computational cost. It is more efficient to identify at which design points we should observe the response or the derivatives only, and at which design points it is better to observe both.

Also most literature for computer experiments makes the assumption that all the factors are quantitative. However, computer modelling can also include qualitative factors. Qian et al. (2008) proposed a general approach for both design and analysis of computer experiments using Gaussian process model and include both quantitative and qualitative factors. This methodology could be extended from a Bayesian prospective to model-based design selection.

Finally, a common assumption made for the Gaussian process model is that of homosecedasity, i.e. the variance is constant for all inputs. However, in many real application we often face the problem that the variance depends on the input. An interesting future direction is to incorporate heteroscedasity in the Gaussian process model and extend our Bayesian design approach. This approach could also be applied in the spatial applications in Chapter 5. Existing work for design under a heteroscedastic Gaussian process model was presented by Boukouvalas et al. (2014). Their work is an extension of Zhu and Stein (2005) and they developed designs for estimating the unknown correlation parameters.

Chapter 8

Designs for Spatio-temporal Processes

The objective of Chapter 8 is to extend the methodology in earlier chapters to find Bayesian designs for prediction of spatio-temporal processes. We investigate our previously developed closed-form approximation for a particular spatio-temporal correlation structure. We give examples of optimal designs for two situations: (i) when the observations are taken at fixed temporal lags, and (ii) when both space and time are optimised.

8.1 Introduction

A wide variety of scientific areas require understanding and prediction of spatial processes that evolve over time. For example, in many environmental applications, such as the one described in Section 1.1.1, we are interested not only in the spatial nature of the chemical deposition, but also in how this chemical deposition changes over time. Similar to spatial data, an important characteristic of spatio-temporal data is that observations taken nearby in space, and now also time, tend to be more alike than those taken further apart.

From a methodological point of view, two cases of spatio-temporal data are often identified, with time being discrete or continuous: (i) if time is continuous, i.e. $t \in \mathbb{R}^+ = (0, \infty)$, we can employ a Gaussian process model to model spatio-temporal data; (ii) if time is viewed as discrete, discrete time series models can be used, such as conditionally autoregressive regression models or dynamic models. The latter case is beyond of the scope of this thesis and further explanation can be found in Banerjee et al. (2004), Cressie and Wikle (2011) and Mateu and Muller (2012).

In this thesis we focus on Gaussian processes in which every design point $\mathbf{x}$ can be viewed as a point in $\mathcal{X} \subseteq \mathbb{R}^2 \times \mathbb{R}^+$. Here we denote the point $\mathbf{x} = (\mathbf{s}, t)$, corresponding

to the location $\mathbf{s} = (s_1, s_2)$ and time t. Although we can consider the time as an additional coordinate and from a probabilistic point of view we can assume a process on $\mathbb{R}^3$, from a physical perspective time differs from space in that time moves only forward and we cannot compare spatial differences with temporal differences.

In practice, observations $y(\mathbf{x}) = y(\mathbf{s}, t)$ are noisy versions of a spatio-temporal stochastic process and may be described by the Gaussian process model (2.6) which we restate here for completeness. The observations made at each location $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and at each time $t_1, \ldots, t_T$ are collected in an $nT \times 1$ vector $\mathbf{Y_s}^\top = (\mathbf{Y}^\top(\mathbf{s}_1), \ldots, \mathbf{Y}^\top(\mathbf{s}_n))$, where $\mathbf{Y}(\mathbf{s}_i) = (y(\mathbf{s}_i, t_1), \ldots, y(\mathbf{s}_i, t_T))^\top$, $i = 1, \ldots, n$, (see Banerjee et al. (2004)). Our model is then

$$\mathbf{Y_s}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau^2 \sim N(\mathbf{F}_s\boldsymbol{\beta}, \boldsymbol{\Sigma}_{Y_s}), \tag{8.1}$$

where $\boldsymbol{\beta}, \sigma^2, \tau^2$ are trend, Gaussian process variance and nugget parameters as defined in (2.6), $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are spatial and temporal correlation parameters, respectively, $\mathbf{F}_s$ is the $nT \times k$ model matrix and $\boldsymbol{\Sigma}_{Y_s}$ is the $nT \times nT$ spatio-temporal covariance matrix.

Basically, the covariance structure of the Gaussian process $Z(\mathbf{x})$ describes the dependency between observations taken at different points in both space and the time. Therefore, we need to specify a valid spatio-temporal covariance function, i.e. for any set of locations and any set of time points, the resulting covariance matrix is positive definite.

All the properties of the Gaussian process described in Chapter 2 can be extended on the space-time domain $\mathbb{R}^2 \times \mathbb{R}^+$ and in the next section we are going to present all the relevant spatio-temporal concepts. In later sections, we will give a numerical study to validate our closed-form design criterion and examples of spatial-temporal designs found by this criterion. Our aim in this chapter is to find Bayesian optimal designs for spatio-temporal data; the design problem is to find the optimal sampling locations and observation times when the question of interest is how to predict, in a region $\mathcal{X}_\mathcal{P}$ ($\mathcal{X} \subseteq \mathcal{X}_\mathcal{P}$), at given locations at a future time.

8.2 Characteristics of Space-time Covariance Functions

A *weak* or *second order stationary* spatio-temporal process $Z(\mathbf{s}, t) : (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}^+$ has constant mean and covariance function K, defined in $\mathcal{X}_\mathcal{P} \subseteq \mathbb{R}^2 \times \mathbb{R}^+$, such that

$$\text{Cov}[Z(\mathbf{s}, t), Z(\mathbf{s} + \mathbf{h}, t + u)] = K(\mathbf{h}, u) \quad \forall \quad \mathbf{s} \in \mathbb{R}^2 \quad \text{and} \quad \forall \quad t \in \mathbb{R}^+,$$

for any spatial separation vector $\mathbf{h} \in \mathbb{R}^2$ and temporal separation $u \in \mathbb{R}^+$. The covariance function K here is called the space-time covariance function and its margins $K(\cdot, 0)$ and $K(\mathbf{0}, \cdot)$ are purely spatial and purely temporal covariance functions, respectively. A consequence of this assumption is constant variance of $Z(\mathbf{s}, t)$.

Throughout this chapter we employ a second-order stationary Gaussian process with

space-time covariance function $K(\mathbf{h}, u)$. Moreover, we make the assumption that the space-time covariance function is *separable* which means that there exist purely spatial and purely temporal covariance functions K_s and K_t such that

$$K(\mathbf{h}, u) = K_s(\mathbf{h})K_t(u) \quad \forall \quad (\mathbf{h}, u) \in \mathbb{R}^2 \times \mathbb{R}^+.$$

Thus, the space-time covariance function decomposes as the product of individual spatial and temporal covariance functions and indicates that the dependence weakens in a multiplicative manner across space and time.

If the space-time covariance function cannot be expressed as the product of spatial and temporal covariance function, it is called *nonseparable*.

Based on the covariance function K, we can define the spatio-temporal correlation function associated with K as

$$\rho(\mathbf{h}, u) = K(\mathbf{h}, u)/K(\mathbf{0}, 0), \quad \mathbf{h} \in \mathbb{R}^2, u \in \mathbb{R}^+,$$

where $K(\mathbf{0}, 0) = \sigma^2$, the variance of the Gaussian process $Z(\mathbf{x})$. The correlation function represents the spatio-temporal dependence in continuous space and continuous time. The separability assumption of the covariance function implies that the spatio-temporal correlation function satisfies

$$\rho(\mathbf{h}, u) = \rho_s(\mathbf{h}, 0)\rho_t(0, u), \quad \forall \quad \mathbf{h} \in \mathbb{R}^2, u \in \mathbb{R}^+.$$

As in the rest of the thesis, we restrict our choices of correlation functions ρ_s and ρ_t to those from the families of parametric, isotropic correlation functions described in Section 5.4.4.

Separable correlation functions dominate the literature because of their easy interpretation and also the reduction in the computational burden of the necessary matrix calculations; the spatio-temporal covariance matrix can be written as the kronecker product of two smaller dimensional matrices.

Motivated by the high demand for statistical modelling of spatio-temporal data, our aim is to make some preliminary steps towards identifing the optimal locations and/or optimal times to collect such data. By specifying an optimal spatio-temporal design, we answer the questions where and when should we take observations in order to minimise the uncertainty in predicting future observations. We find separable spatio-temporal designs, i.e. one set of sampling locations and one set of time points.

8.3 Literature Review

In this section, we give a general overview on the existing approaches for spatio-temporal design. A recent review of the state of art is given by Mateu and Muller (2012). Most

of the approaches that tackle the problem of optimal allocation of sampling locations and/or optimal time at which to take measurements can be classified into two categories: (i) probability-based and (ii) model-based, as for spatial designs. Our methodology belongs to the model-based approach and we focus on this approach in our review.

The majority of the literature on spatio-temporal designs using a model-based approach assumes a dynamic model, for example Wikle and Royle (1999, 2005). In both these papers, the authors described spatially dynamic designs for ecological and environmental applications, i.e. observations at different spatial locations are taken in discrete time with the locations at time $t+1$ selected using data up to time t. Wikle and Royle (1999) described the spatially dynamic designs for Gaussian processes and Wikle and Royle (2005) for non-Gaussian data.

In this chapter, we develop non-dynamic designs for spatio-temporal process using the Gaussian process model. This topic seems to have received less attention in the literature. Recently, Heuvelink et al. (2013) proposed a design criterion to find spatio-temporal design as an extension of geostatistical applications. As they discussed, from a methodological point of view the extension of spatial data approaches is very possible. However, they indicated two important differences: firstly, in spatio-temporal prediction the assumption of isotropy is violated more often compared to spatial prediction; secondly, the cost of collecting spatio-temporal data may be cheaper if time series of data is collected at fixed spatial locations.

Heuvelink et al. (2013) proposed a design criterion which simultaneously minimises the variance of the estimation error of a linear trend and the interpolation error of the prediction (kriging) residual. They stated that this criterion is equivalent to minimising the average kriging variance. They employed a simulated annealing algorithm and considered three optimisation scenarios: firstly, to reduce the number of sampling locations from an existing static design; secondly, to reduce the number of locations but allowing the location to move within the region of interest; and thirdly, to reduce the number of location at different times. Based on the results of their case study, they concluded that the third optimisation scenario is better compare to the other two in terms of prediction accuracy.

The approach for spatio-temporal designs proposed by Heuvelink et al. (2013) is based on the frequentist approach, where the unknown parameters are estimated from available data and plugged into the objective function. Uncertainty in these parameters is not addressed. Our approach is Bayesian and we do not assume any available data. Moreover, they considered n spatio-temporal points and the correlation in space and/or time is modelled by a space-time covariance function which has three components: the purely spatial covariance function, the purely temporal, and the space-time interaction covariance function. In our approach, we consider a separable form of covariance function.

8.4 Optimal Design for Spatio-temporal Processes

The purpose of this section is to extend the methodology of spatial optimal design described in Chapter 5 to include the time component. The observations made at each of the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and at each of the times $t_1, \ldots, t_T$ are modelled by the Gaussian process model (8.1). We also make the assumption that observations are made at all optimal locations at all time points. Now, the covariance matrix $\mathbf{\Sigma}_{Y_s}$ has dimension $nT \times nT$, potentially making calculations very difficult if n or T is large. Due to this practical limitation, separable covariance functions have become very popular and as we have mentioned in Section 8.2, our approach is limited to this class of covariance functions, i.e. $\mathbf{\Sigma}_{Y_s} = \sigma^2 C_s(\boldsymbol{\theta}_1) \otimes C_t(\boldsymbol{\theta}_2) + \tau^2 \mathbf{I}$ where $C_s(\boldsymbol{\theta}_1)$ is the spatial correlation matrix and $C_t(\boldsymbol{\theta}_2)$ is the temporal correlation matrix.

The posterior and predictive distributions derived in Chapter 2 involve the inverse and/or the determinant of $\mathbf{\Sigma}_{Y_s}$. Using the properties of Kronecker products and the assumption that $\tau^2 = 0$, we are able to evaluate:

$$|\mathbf{\Sigma}_{Y_s}| = |\sigma^2 C_s(\boldsymbol{\theta}_1) \otimes C_t(\boldsymbol{\theta}_2)| = (\sigma^2)^{nT} |C_s(\boldsymbol{\theta}_1)|^n |C_t(\boldsymbol{\theta}_2)|^T,$$

and

$$\mathbf{\Sigma}_{Y_s}^{-1} = [C_s(\boldsymbol{\theta}_1) \otimes C_t(\boldsymbol{\theta}_2)]^{-1} = [C_s(\boldsymbol{\theta}_1)]^{-1} \otimes [C_t(\boldsymbol{\theta}_2)]^{-1}.$$

Hence, we need only the determinant and the inverse of an $n \times n$ and a $T \times T$ matrix instead of an $nT \times nT$ matrix, expediting evaluations of the posterior and predictive distributions and the also evaluation of the design selection objective function.

The posterior predictive distribution can be expressed in closed form, conditional on correlation parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The predictive distribution, found in similar way to those presented in Chapter 2, and replacing $\mathbf{\Sigma}$ with $\mathbf{\Sigma}_{Y_s}$ and setting $\delta^2 = 0$ in all the posterior densities, is given by

$$y(\mathbf{x}_p)|\mathbf{y}_s, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim t_{2a+nT}[1, \mu^*, \frac{2b^*}{2a+nT}\Sigma^*],$$

where $y(\mathbf{x}_p)$ is the observation for a point $\mathbf{x}_p = (\mathbf{s}, t) \in \mathcal{X}_\mathcal{P}$ and

$$\begin{aligned}
\mu^* &= (\mathbf{f}_p^\top - \boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F})(\mathbf{F}^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F} + \mathbf{R})^{-1}\mathbf{R}\boldsymbol{\beta}_0 \\
&\quad + [\boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1} + (\mathbf{f}_p^\top - \boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F})(\mathbf{F}^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F} + \mathbf{R})^{-1}\mathbf{F}^\top \mathbf{\Sigma}_{Y_s}^{-1}]\mathbf{y} \\
\Sigma^* &= (1 - \boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1}\boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top) \\
&\quad + (\mathbf{f}_p^\top - \boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F})(\mathbf{F}^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F} + \mathbf{R})^{-1}(\mathbf{f}_p^\top - \boldsymbol{\omega}_s^\top \otimes \boldsymbol{\omega}_t^\top \mathbf{\Sigma}_{Y_s}^{-1}\mathbf{F})^\top \\
b^* &= b + \frac{1}{2}\left[(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top [\mathbf{\Sigma}_{Y_s} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)\right],
\end{aligned} \tag{8.2}$$

where $\boldsymbol{\omega}_s$, $\boldsymbol{\omega}_t$ are the $n \times 1$ vector of spatial and $T \times 1$ vector of temporal covariances between the response at each of the existing inputs and the response at $\mathbf{x}_p$ respectively, $\mathbf{f}_p$ is the $k \times 1$ vector of regression functions for $\mathbf{x}_p$, and $\mathbf{F}$ is the $nT \times k$ matrix of

regression functions, $\boldsymbol{\beta}_0$ and $\mathbf{R}^{-1}$ are prior hyperparameters for $\boldsymbol{\beta}$, and a and b are prior hyperparameters for σ^2.

The design criterion we employ for spatio-temporal studies is the minimisation of the average posterior prediction variance, as it derived in Chapter 3, and the objective function $\Psi(\xi)$ is given by equation (3.7). Under the assumption of separable correlation functions for space and time and zero nugget, we can again use the laws of total expectation and total variance, (2.20) and (2.21), to decompose the objective function for Ψ-optimal designs:

$$\begin{aligned}\Psi(\xi) =& \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \text{var}(y(\mathbf{x}_p)|\mathbf{y}_s)\pi(\mathbf{y}_s)d\mathbf{y}_s d\mathbf{x}_p \\ =& \int_{\mathcal{X}_\mathcal{P}} \int_{\mathcal{Y}} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}_s}\{\text{var}(y(\mathbf{x}_p)|\mathbf{y}_s,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\} + \text{var}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}_s}\{\mathbb{E}(y(\mathbf{x}_p)|\mathbf{y}_s,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\}\,\pi(\mathbf{y}_s)d\mathbf{y}_s d\mathbf{x}_p \\ =& \Psi_1(\xi) + \Psi_2(\xi). \end{aligned} \tag{8.3}$$

8.4.1 Closed form approximation to the design selection criterion

We now present numerical evidence that the closed-form approximation, $\Psi_1(\xi)$, developed in Chapter 3, is a good approximation for the objective function $\Psi(\xi)$ (8.3) for spatio-temporal problems. We investigate how choices for correlation parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ affect $\Psi_1(\xi)$ by considering two cases:

i. $\boldsymbol{\theta}_1$ unknown and $\boldsymbol{\theta}_2$ known and fixed,

ii. both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ unknown.

For both cases, to evaluate $\Psi_1(\xi)$ and $\Psi_2(\xi)$ we need to employ numerical methods, Monte Carlo and quadrature as introduced in Chapter 3.

Case i: When the temporal correlation parameters $\boldsymbol{\theta}_2$ are fixed, the approximations for $\Psi_1(\xi)$ and $\Psi_2(\xi)$ are similar to those in Section 3.5.2, (3.14), where μ^*, Σ^* and b^* are now given by (8.2). The approximation of $\Psi_1(\xi)$ for a uniform, Unif(a_1, b_1), and log normal, lnN(μ, σ^2), prior distributions for $\boldsymbol{\theta}_1$ are given by (3.29) and (3.31), respectively, and can be expressed for the case of known $\boldsymbol{\theta}_2$ as:

$$\begin{aligned}\Psi_1(\xi) &\simeq \frac{1}{2}\sum_{i=1}^{m_1} w_i^1 f_1\left(\frac{b_1-a_1}{2}a_i^1 + \frac{b_1+a_1}{2}\right) \quad \text{(uniform)} \\ \Psi_1(\xi) &\simeq \frac{1}{\sqrt{\pi}}\frac{1}{2}\sum_{i=1}^{m_1} w_i^1 f_1\left(e^{\mu+a_i^1\sigma\sqrt{2}}\right), \quad \text{(log-normal)}\end{aligned} \tag{8.4}$$

where a_i^1 and w_i^1 are the quadrature abscissae and weights, different for the uniform and log-normal cases and obtained from the Legendre polynomials for uniform distribution, and from the Hermite polynomials for log-normal distribution and $f_1(\cdot)$ is the integrand of $\Psi_1(\xi)$; see Chapter 3.

The approximation for $\Psi_2(\xi)$ is given by

$$\begin{aligned}
\Psi_2(\xi) &\simeq \frac{1}{N}\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{m_1} w_i^1 f_2\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \mathbf{y}_k\right) \quad \text{(uniform)} \\
\Psi_2(\xi) &\simeq \frac{1}{N}\frac{1}{\sqrt{\pi}}\sum_{k=1}^{N}\sum_{i=1}^{m_1} w_i^1 f_2\left(e^{\mu+a_i^1\sigma\sqrt{2}}, \mathbf{y}_k\right), \quad \text{(log-normal)}
\end{aligned} \tag{8.5}$$

where $\mathbf{y}_k$: $k = 1, \ldots, N$ is a random sample, where for each $\mathbf{y}_k$, quadrature is applied to approximate numerically the first integral in $\Psi_2(\xi)$, Monte Carlo integration approximates the second integral, and $f_2(\cdot)$ is the integrand of $\Psi_2(\xi)$; see Chapter 3.

Case ii: When the temporal correlation parameters $\boldsymbol{\theta}_2$ are unknown, the approximations of $\Psi_1(\xi)$ and $\Psi_2(\xi)$ can be extended from those in Section 3.5.1, (3.11), with μ^*, Σ^* and b^* given by (8.2).

For uniform prior distribution for both on $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we can obtain

$$\begin{aligned}
\Psi_1(\xi) &\simeq \frac{1}{2}\frac{1}{2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^3 f_1\left(\frac{b_1-a_1}{2}a_i^1+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_j^3+\frac{b_2+a_2}{2}\right), \\
\Psi_2(\xi) &\simeq \int \frac{1}{4}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_i^3 f_2\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^3+\frac{b_2+a_2}{2}, \mathbf{y}\right)\pi(\mathbf{y})d\mathbf{y} \\
&\simeq \frac{1}{N}\frac{1}{4}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^3 f_2\left(\frac{b_1-a_1}{2}a_i+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^3+\frac{b_2+a_2}{2}, \mathbf{y}_k\right), \\
\mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}}(\mu^*) &\simeq \frac{1}{N}\frac{1}{4}\sum_{k=1}^{N}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i^1 w_j^3 \mu^*\left(\frac{b_1-a_1}{2}a_i^1+\frac{b_1+a_1}{2}, \frac{b_2-a_2}{2}a_i^3+\frac{b_2+a_2}{2}, \mathbf{y}_k\right).
\end{aligned} \tag{8.6}$$

where μ^* is given by (8.2) and is a function of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ and $\mathbf{y}$, a_i^1 and w_i^1 are the abscissae and weights obtained from the Legendre polynomials for the uniform prior distribution for $\boldsymbol{\theta}_1$, a_j^3 and w_j^3 are the abscissae and weights obtained from the Legendre polynomials for the uniform prior distribution for $\boldsymbol{\theta}_2$, and $\mathbf{y}_k$: $k = 1, \ldots, N$ is the random sample necessary for Monte Carlo integration.

We find designs with points from the study region $\mathcal{X} = [-1, 1]^2 \times [0, 1]$. The aim is to predict at a 10×10 regular-spaced grid for a specific new time, $t_0 = 2$. In contrast to Chapters 5 and 7, here we do not perform a factorial study to examine the behaviour of the objective function. Instead we restrict our study to one situation for which we will later find the Ψ-optimal design.

The set up for this problem is as follows:

1. The number of runs is $n = 10$, $T = 3$, i.e. we have $n = 10$ sampling locations from $[-1, 1]^2$ and at $T = 3$ times obtained from $[0, 1]$.

2. The mean function is assumed to be the linear trend $\beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 t$, where

s_1, s_2 are the spatial coordinates and t is time. This model assumes that there is no spatio-temporal interaction.

3. The correlation functions are assumed to be exponential functions (2.4) for both space and time, i.e. for $\boldsymbol{\theta}_1 = (\phi_1, \nu)^\top$ and $\boldsymbol{\theta}_2 = (\phi_2, \nu)^\top$, with $\nu = 0.5$,

$$\rho_s(\mathbf{s}_i, \mathbf{s}_j; \phi_1) = \exp(-\phi_1 \|\mathbf{s}_i - \mathbf{s}_j\|)$$

$$\rho_t(t_i, t_j; \phi_2) = \exp(-\phi_2 |t_i - t_j|).$$

4. We always allow uncertainty in the regression parameters $\boldsymbol{\beta}$, the variance σ^2 and the spatial correlation parameter ϕ_1. The temporal correlation parameter ϕ_2 is in turn considered both known and unknown. Therefore the prior distributions for the unknown parameters are as follows, with two choices for ϕ_1 and four choices for ϕ_2:

   - $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
   - $\sigma^2 \sim \text{IG}(3, 1)$
   - $\phi_1 = \begin{cases} \text{Unif}(0.1, 1), & \text{coded } 0 \\ \text{log-normal}(-1.1, 1), & \text{coded } 1 \end{cases}$
   - $\phi_2 = \begin{cases} 0.1, & \text{coded } 0 \\ 1, & \text{coded } 1 \\ 10, & \text{coded } 2 \\ \text{Unif}(0.1, 1), & \text{coded } 3. \end{cases}$

In a similar fashion to Chapters 5 and 7, to investigate the relationship between $\Psi(\xi)$ and both $\Psi_1(\xi)$ and $\Psi_2(\xi)$, we generate 50 random designs and for each of these designs we evaluate $\Psi(\xi)$ and each of $\Psi_1(\xi)$ and $\Psi_2(\xi)$.

We calculate the correlation between $\Psi(\xi)$ and $\Psi_1(\xi)$ for each set of designs and these correlation values are displayed in Table 8.1. Once again, the correlation between $\Psi(\xi)$ and $\Psi_1(\xi)$ is very high, almost equal to 1.

Regardless the choice of prior distribution for the spatial correlation parameter, ϕ_1, the values of $\Psi_1(\xi)$ and $\Psi(\xi)$ are very similar and $\Psi_2(\xi)$ is much smaller than $\Psi_1(\xi)$. Based on the numerical evidence, for these specific examples, we are able proceed using our closed-form approximation to find Ψ-optimal designs.

In the recent work of Ren et al. (2013), they studied the limiting behaviour of integrated likelihood for non-informative priors and separable correlation functions. Similar to the case of spatial data, the integrated likelihood is bounded by a function of the correlation parameters and can quickly become very small. As a result $\Psi_2(\xi)$ is once again very small in magnitude, as it depends mainly on the integrated likelihood, supporting our numerical results.

ϕ_1	ϕ_2			
	0	1	2	3
0	0.999976	0.999998	0.999992	0.999800
1	0.999976	0.999997	0.999993	0.999801

Table 8.1: Correlation between values of Ψ and Ψ_1 for 50 random designs and 8 combinations of values of ϕ_1 and ϕ_2.

A more thorough factorial study could be conducted in future work to make more general statements about the closed-form approximation, e.g. for different correlation functions. However, for this thesis we only focus on separable correlation functions.

8.4.2 Examples of Bayesian spatio-temporal designs

In this section we find Ψ-optimal designs for four different scenarios, presented in Table 8.2. For all four scenarios, the spatial correlation parameter ϕ_1 is unknown, whereas the temporal correlation parameter ϕ_2 is either known or unknown. We always consider the situation with optimally selected spatial locations, and for the time points we consider situations with fixed observation times and also with optimally selected observation times. The notations V and F in Table 8.2 represent that the variable of interest is non-constant (V) or it is fixed (F) at specific values respectively.

The steps we follow to select a Ψ-optimal design under each scenario are as follows:
(i) When the times are fixed, i.e. Time=F

- generate 30 randomly starting designs from $[-1, 1]^2$ with n sampling locations,
- for each starting design, we use the coordinate exchange algorithm to optimise the sampling locations only in order to find a design that minimises $\Psi_1(\xi)$; we select the design from this list with minimum $\Psi_1(\xi)$ value.

(ii) When the times are allowed to vary and we want to find a set of the optimal times at which to take observations, i.e. Time=V

- generate 30 randomly selected starting designs from $\mathcal{X} = [-1, 1]^2 \times [0, 1]$ with n sampling locations and t time points,
- for each starting design, we use the coordinate exchange algorithm to optimise the

Scenario	Space	Time	ϕ_1	ϕ_2
1	V	F	V	F
2	V	F	V	V
3	V	V	V	F
4	V	V	V	V

Table 8.2: Optimal designs scenarios, where V indicates that the variable is non-constant and F indicates that the variable is fixed.

sampling locations by keeping fixed the times and then we employ the coordinate exchange algorithm to optimise the time,

- we repeat the coordinate exchange algorithm for both space and time until there is no further improvement in the value of $\Psi_1(\xi)$; we select the design from this list with minimum $\Psi_1(\xi)$ value.

For this latter case of non-fixed time points, we do not need to include any constraints in our coordinate exchange algorithm with respect to the time. The time-order does not affect the model (8.1) due to the assumption of separability, and so we are able to use the coordinate exchange algorithm and then the optimal time points can be ordered from the smallest to largest value.

We find optimal spatio-temporal designs with $n = 10$ sampling locations and $T = 3$ or $T = 6$ time points. For all cases, the spatial correlation parameter is unknown and a uniform prior is assigned, $\phi_1 \sim \text{Unif}(0.1, 1)$. When the temporal correlation parameter is assumed known, it takes one of the values, $\phi_2 = 0.01$, which corresponds to high temporal correlation, $\phi_2 = 0.5$, which corresponds to medium correlation and $\phi_2 = 10$, corresponding to very low correlation. Two spatial correlation functions are considered, the exponential and the Matérn, (2.4) with $\nu = 0.5$ and $\nu = 1.5$ respectively, and for the temporal correlation function we assumed the exponential correlation function.

Scenario 1: Initially we find Bayesian optimal designs for Scenario 1 as shown in Table 8.2, with fixed times and the temporal correlation parameter assumed known. These designs found by minimising $\Psi_1(\xi)$ (8.4). We consider two temporal designs:

Times 1: $t_1 = 0.76726, t_2 = 0.84199, t_3 = 0.88814$

Times 2: $t_1 = 0.001, t_2 = 0.1, t_3 = 0.2$.

We chose these times to investigate spatial designs when the time points are close together.

Figure 8.1 (a) shows the Ψ-optimal spatial design for exponential correlation function and $\phi_2 = 0.01$. The design is quite space-filling, with points reasoning equally spread over the study region. The Ψ-optimal design for Matérn correlation function Figure 8.1 (b) has fewer points in the centre of the region, with more points allocated to the boundaries.

For $\phi_2 = 0.5$ and $\phi_2 = 10$ and exponential correlation function, Figure 8.1 (c) and (e), most points are at the boundaries of the design region, whereas using the Matérn correlation function, Figure 8.1 (d) and (f), results in design points moving to the corners. Here, spatial correlation is stronger than the exponential correlation function, and the range of correlation is also smaller.

The pattern of the spatial designs is similar if we assumed the second fixed set of times (*Times 2*); the corresponding designs can be found in Appendix A.7.

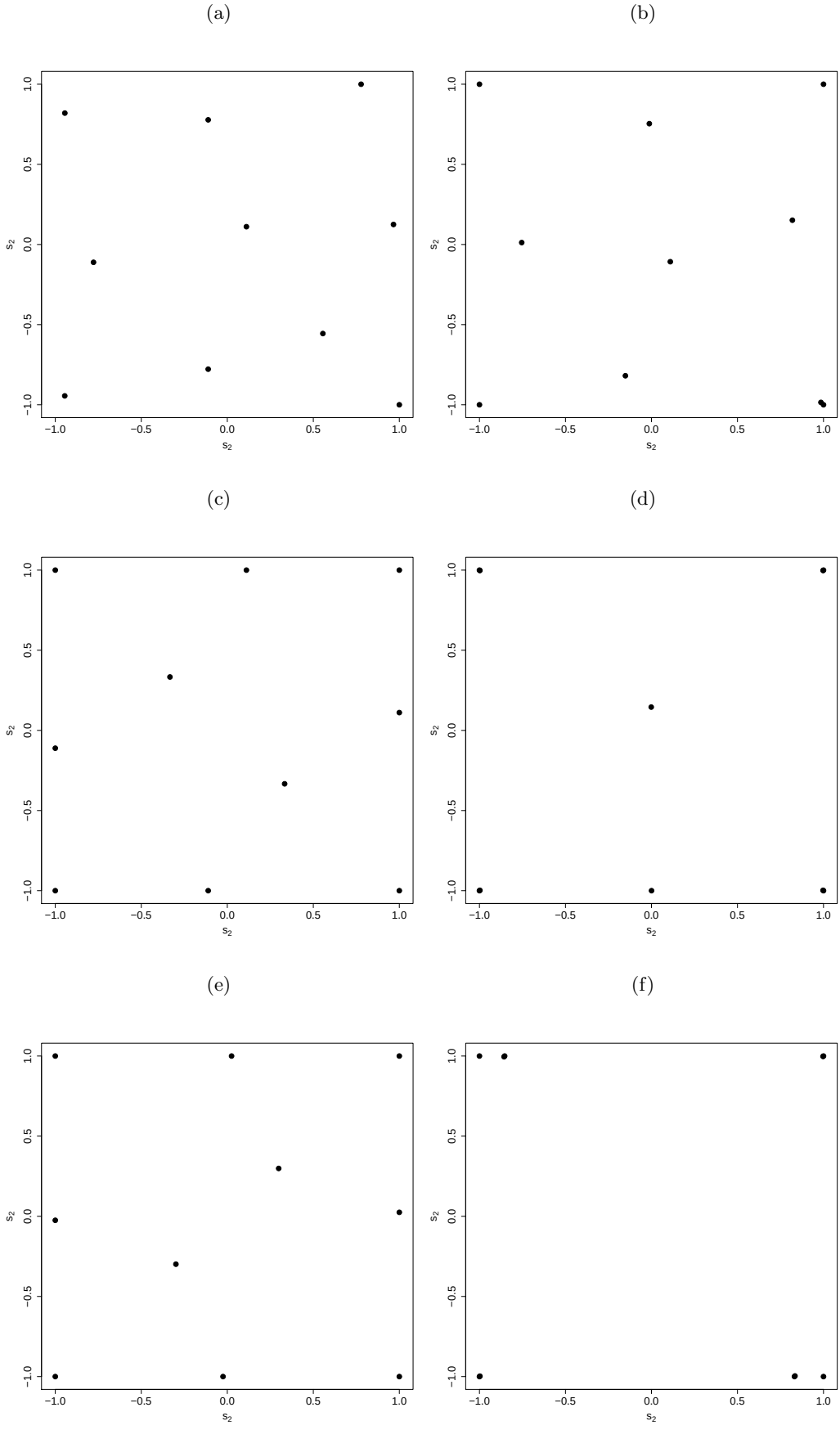


Figure 8.1: Spatial Ψ-optimal designs for fixed times (set *Times 1*) (a) $\nu = 0.5$ and $\phi_2 = 0.01$; (b) $\nu = 1.5$ and $\phi_2 = 0.01$; (c) $\nu = 0.5$ and $\phi_2 = 0.5$; (d) $\nu = 1.5$ and $\phi_2 = 0.5$; (e) $\nu = 0.5$ and $\phi_2 = 10$; (f) $\nu = 1.5$ and $\phi_2 = 10$. In plots (d) and (f) four points are repeated.

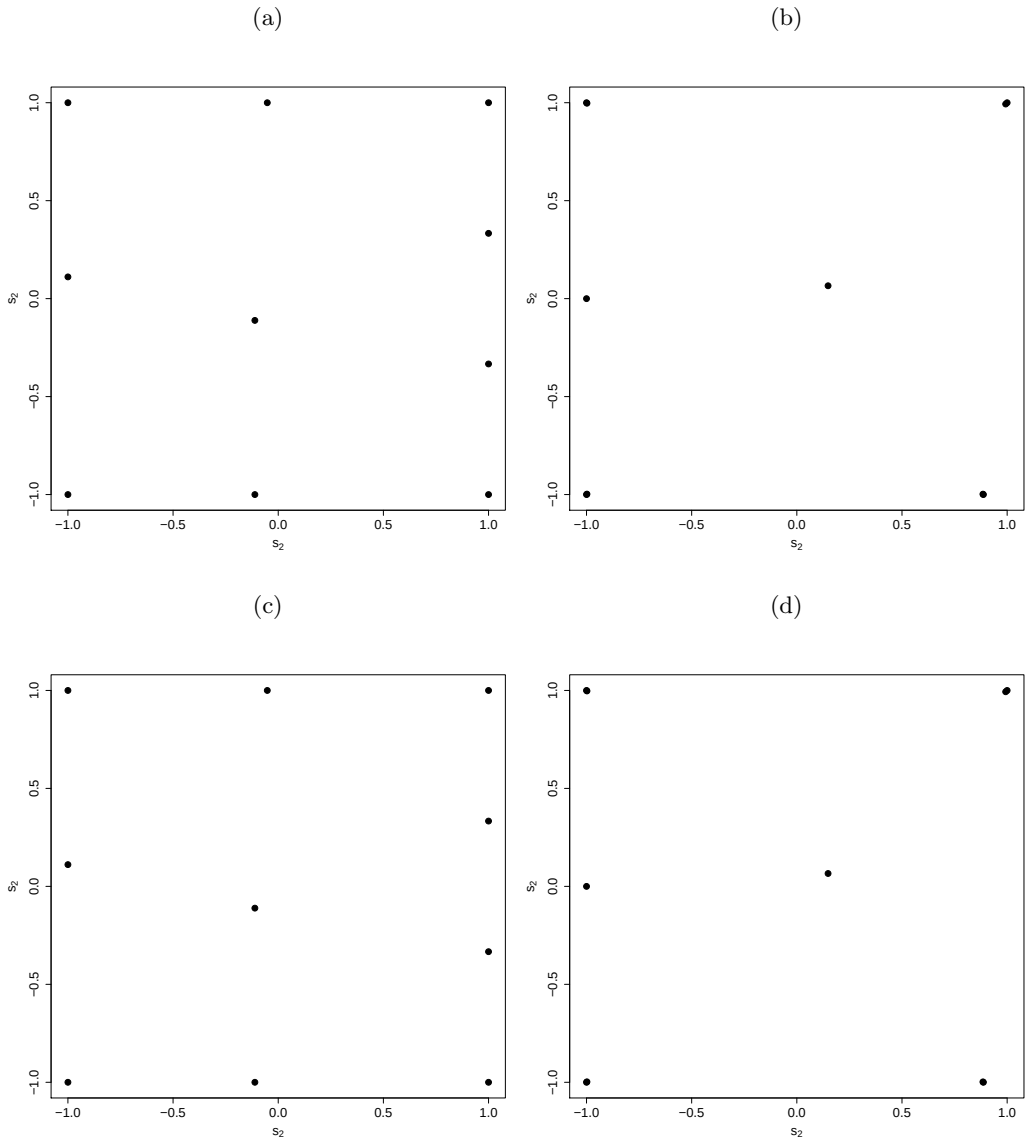


Figure 8.2: Spatial Ψ-optimal designs for randomly selected times: (a) $\nu = 0.5$, $\phi_2 = 0.5$ and $T = 3$; (b) $\nu = 1.5$, $\phi_2 = 0.5$ and $T = 3$; (c) $\nu = 0.5$, $\phi_2 = 0.5$ and $T = 6$; (d) $\nu = 1.5$, $\phi_2 = 0.5$ and $T = 6$. In plots (b) and (d) four points are repeated.

We have also generated random sets of $T = 3$ times from the interval $(0, 1)$, to check if sets of time points result in different spatial designs. We observe that the pattern for the spatial designs is the same regardless of the time points, i.e. if the time points are close together or not, the spatial design is the same. Only the correlation strength of the temporal correlation, which is controlled through the parameter ϕ_2, affects the spatial designs.

We also assess if the number of time points chosen for the temporal design affects the spatial designs. We found designs for $T = 3$ and $T = 6$, both assuming $\phi_2 = 0.5$ and exponential and Matérn correlation functions. Figure 8.2 shows that the same spatial designs result for both $T = 3$ and $T = 6$.

Scenario 2: The second scenario we consider has ϕ_2 unknown and assigned a uniform

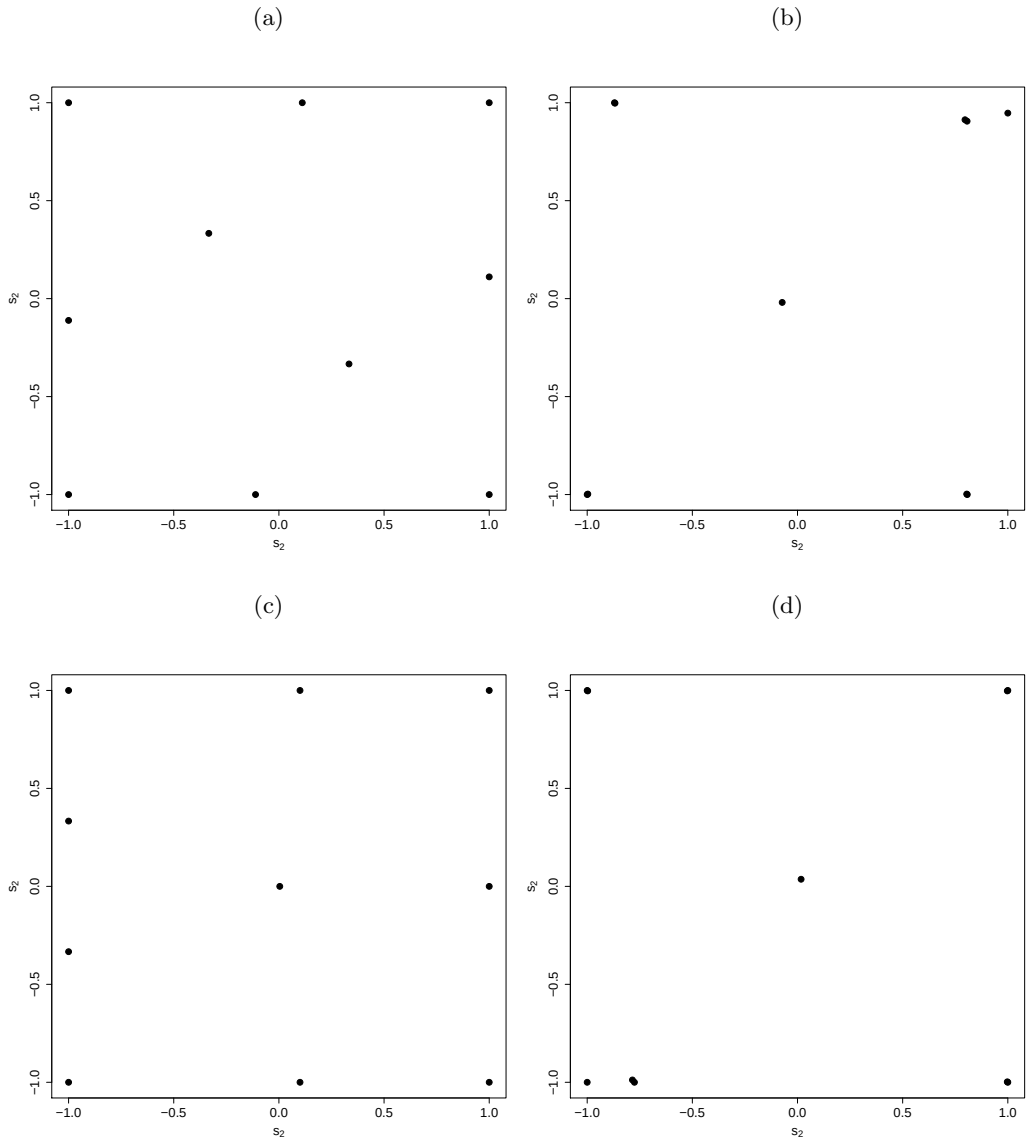


Figure 8.3: Spatial Ψ-optimal designs for fixed times when ϕ_2 is unknown: (a) $\nu = 0.5$ and *Times 1* (b) $\nu = 1.5$ and *Times 1* (c) $\nu = 0.5$ and *Times 2* (d) $\nu = 1.5$ and *Times 2*. In plots (b) and (d) four points are repeated.

prior distribution, Unif$(0, 1)$. The objective function to be minimised is $\Psi_1(\xi)$ (8.6). The temporal designs are fixed at either the set *Times 1* or *Times 2*. Figure 8.3 (a) and (c) show the spatial Ψ-optimal designs for $\nu = 0.5$. We see that the resulting designs are very similar to each other, with points distributed similarly. When the spatial correlation function changes to Matérn $\nu = 1.5$, the spatial designs differ from those obtained by exponential correlation function, but again the two designs for $\nu = 1.5$ are very similar.

If we compare the designs obtained with ϕ_2 unknown to those with known and fixed ϕ_2, Figures 8.1 and 8.3 respectively, we can see that the designs for the latter case are similar with those for $\phi_2 = 0.5$, the mid-point of the support of the prior distribution assumed for ϕ_2.

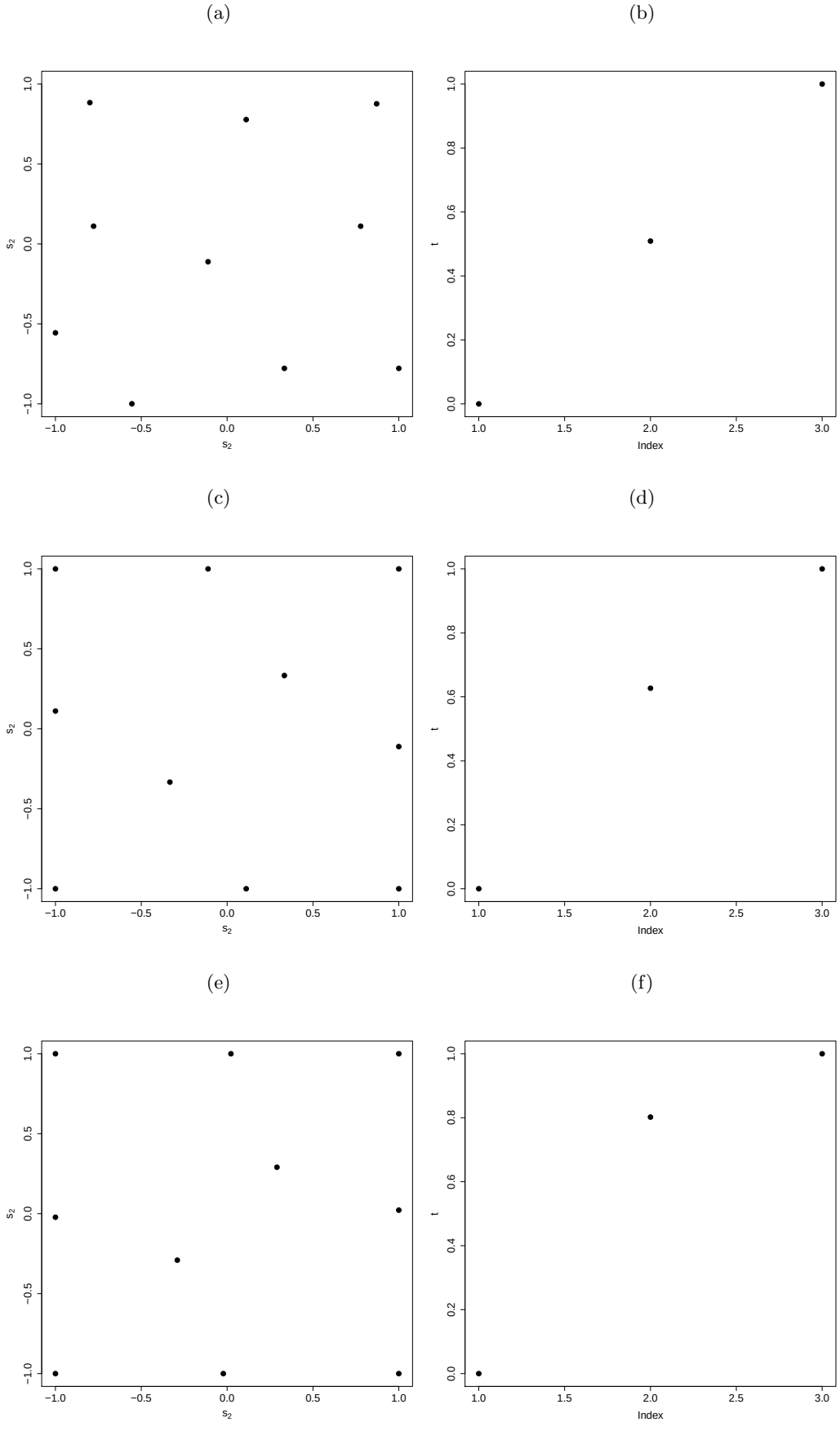


Figure 8.4: Spatial and temporal Ψ-optimal designs for exponential spatial and temporal correlation functions: (a) $\phi_2 = 0.01$; (b) $\phi_2 = 0.01$; (c) $\phi_2 = 0.5$; (d) $\phi_2 = 0.5$; (e)$\phi_2 = 10$; (f) $\phi_2 = 10$.

Until this point, we have only considered fixed temporal designs, with optimisation only performed for the spatial locations. However, in many applications, we additionally need to find out when are the optimal times to observe data. As a result, the last two scenarios investigated here focus on optimising the time as well as spatial locations.

Scenario 3: Figure 8.4 shows the Bayesian Ψ-optimal designs in space and time found by minimising $\Psi_1(\xi)$ (8.4) for $\phi_2 = 0.01, 0.5$ and 10 and the exponential correlation function. The spatial designs are identical to those in the previous two Scenarios, see for example Figures 8.1 and 8.3. The temporal designs in Figure 8.4 indicate that the optimal strategy is to take observations at equally spaced times if the temporal correlation is high and as the temporal correlation decreases, begin to coalesce, see Figure 8.4 (d) and (f). That is, as ϕ_2 increases, the temporal correlation decreases, and the design is influenced by the linear trend in time.

Similar results are obtained when we assumed a Matérn spatial correlation function. The optimal spatial designs are the same as those obtained for fixed sampling times. The optimal temporal designs again are formed from equal spaced points for $\phi_2 = 0.01$, and for larger values of ϕ_2, the second time point moves towards to the upper point, the same as those in Figure 8.4.

Scenario 4: The final, most general case is shown in Figure 8.5, with both spatial locations and time points optimised and all parameters considered unknown and given prior distributions. When the spatial correlation is assumed to be exponential, the Ψ-optimal design is the same as that obtained when $\phi_2 = 0.5$, Figures 8.4 (c) and (d) for space and time respectively. This result agrees with our conclusion that the spatial Ψ-optimal design is not affected by the configuration of the time points but it is affected by the temporal correlation parameter ϕ_2. Similar conclusions are obtained for the Matérn correlation function, see Figure 8.5 (d) where the design is quite similar to that in Figure 8.4 (d).

In order to investigate temporal designs in more detail, we generate 100 random spatial designs and for each found the Ψ-optimal of time points for both $T = 3$ and $T = 6$ with $\phi_2 \sim \text{Unif}(0.1, 1)$. The optimal temporal designs were always equally spaced points in the region $(0, 1)$.

Finally, we compare the spatial Ψ-optimal designs obtained when we minimise the spatio-temporal objective function (8.6) with the spatial Ψ-optimal designs obtained by minimising the spatial objective function $\Psi_1(\xi)$ (3.11) for $n = 10$, $\delta^2 = 0$ and $\phi \sim \text{Unif}(0.1, 1)$. As demonstrated in Chapter 5, when $\nu = 0.5$ the Ψ-optimal design is a coverage design, see Figure 5.4 (a), and when $\nu = 1.5$, Figure 5.5 (a), the points move towards to the boundaries of the design region with few points at the centre. In order to compare the spatial designs obtained from the two different selection criteria, we evaluated the efficiencies of the corresponding optimal designs, see Table 8.3. The efficiencies are very close to 1, indicating that the Ψ-optimal spatial designs obtained by minimising the spatial objective function do not greatly differ in performance from

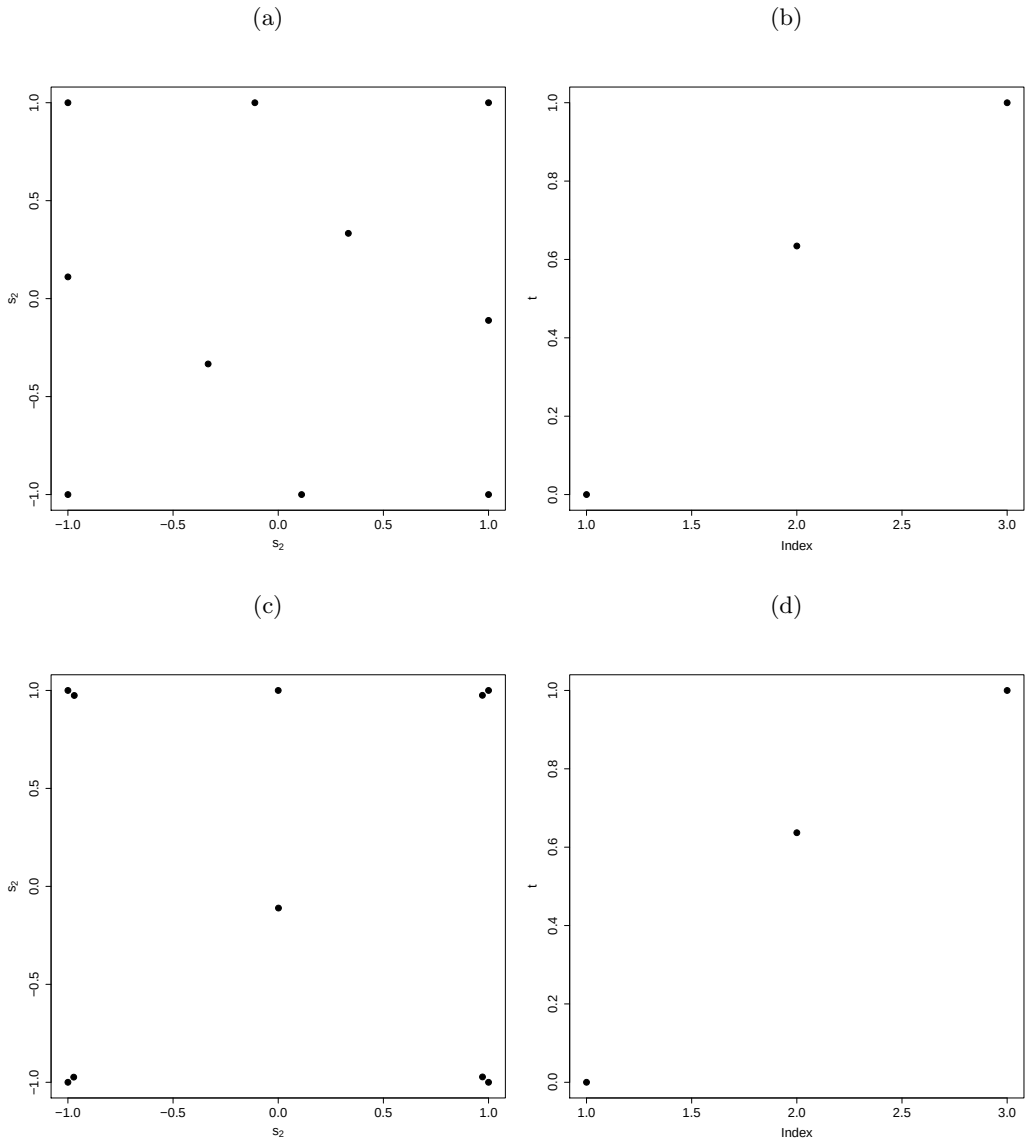


Figure 8.5: (a) Spatial and (b) temporal designs for $\nu = 0.5$ and (c) spatial and (d) temporal designs $\nu = 1.5$.

those obtained from the spatio-temporal approach regardless the choice of time points.

Table 8.3 shows the efficiencies for all the spatial Ψ-optimal designs found under each one of the four scenarios with both $T = 3$ and $T = 6$ time points and either the exponential, $\nu = 0.5$, or the Matérn, $\nu = 1.5$, correlation function and when ϕ_2 is assumed known and fixed is equal to $\phi_2 = 0.5$.

To investigate how the spatio-temporal correlation affects the designs, we evaluate the correlation between any spatial location $\mathbf{s} = (s_1, s_2)$ and time t and the centre of the region $(0, 0)$ at $t = 0$ for each of the times $0, 0.2, 0.4, 0.6, 0.8, 1$. The corresponding spatio-temporal correlation is evaluated by the formula:

$$\rho(\mathbf{s}, \mathbf{0}; t, 0) = \rho_s(\mathbf{s}, \mathbf{0})\rho_t(t, 0).$$

Optimal Design	Spatial		$T=3$ VFVF		$T=6$ VFVF		$T=3$ VFVV		$T=6$ VFVV		$T=3$ VVVF		$T=6$ VFVV		$T=3$ VVVV		$T=6$ VVVV	
	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$	$\nu=0.5$	$\nu=1.5$
ξ_1	1.000	0.940	0.964	0.964	0.964	0.965	0.968	0.975	0.966	0.967	0.968	0.969	0.968	0.968	0.970	0.970	0.971	0.969
ξ_2	0.984	1.000	0.963	0.960	0.963	0.960	0.966	0.970	0.964	0.962	0.966	0.965	0.966	0.964	0.968	0.966	0.968	0.964
ξ_3	0.898	0.633	1.000	0.981	0.999	0.981	0.999	0.992	0.999	0.983	0.999	0.986	0.999	0.9856	0.998	0.986	0.999	0.985
ξ_4	0.633	0.321	0.968	1.000	0.968	1.000	0.957	1.000	0.964	1.001	0.955	1.000	0.956	1.000	0.950	1.000	0.950	1.000
ξ_5	0.898	0.633	1.000	0.981	1.000	0.981	0.999	0.991	0.999	0.984	0.999	0.986	0.999	0.985	0.998	0.986	0.999	0.985
ξ_6	0.634	0.321	0.968	0.999	0.968	1.000	0.957	1.007	0.964	1.001	0.955	1.000	0.956	1.000	0.950	1.000	0.950	1.000
ξ_7	0.935	0.743	0.999	0.982	0.999	0.982	1.000	0.992	1.000	0.984	1.000	0.986	1.000	0.986	1.000	0.987	1.000	0.986
ξ_8	0.616	0.324	0.953	0.991	0.953	0.991	0.942	1.000	0.949	0.993	0.940	0.993	0.940	0.992	0.935	0.993	0.935	0.992
ξ_9	0.935	0.743	0.999	0.982	0.999	0.982	1.000	0.992	1.000	0.984	1.000	0.986	1.000	0.986	1.000	0.987	1.000	0.986
ξ_{10}	0.577	0.265	0.958	0.998	0.958	0.998	0.943	1.005	0.952	1.000	0.941	0.998	0.941	0.998	0.934	0.998	0.938	0.998
ξ_{11}	0.935	0.743	0.999	0.982	0.999	0.982	1.000	0.992	1.000	0.984	1.000	0.986	1.000	0.986	1.000	0.987	1.000	0.986
ξ_{12}	0.642	0.327	0.969	0.999	0.969	0.999	0.958	1.007	0.965	1.000	0.957	1.000	0.957	0.999	0.952	1.000	0.952	0.999
ξ_{13}	0.928	0.721	0.999	0.983	0.999	0.983	0.999	0.993	0.999	0.985	0.999	0.987	1.000	0.986	0.999	0.987	1.000	0.986
ξ_{14}	0.634	0.322	0.968	0.999	0.968	0.999	0.957	1.007	0.964	1.002	0.955	1.000	0.956	1.000	0.950	1.000	0.950	1.000
ξ_{15}	0.935	0.743	0.999	0.982	0.999	0.982	0.999	0.992	1.000	0.984	1.000	0.986	1.000	0.986	1.000	0.987	1.000	0.986
ξ_{16}	0.648	0.332	0.970	0.998	0.970	0.998	0.959	1.006	0.965	1.000	0.958	0.999	0.958	0.999	0.953	1.000	0.953	0.999
ξ_{17}	0.927	0.721	0.999	0.983	0.999	0.983	0.999	0.993	0.999	0.985	0.999	0.987	0.999	0.986	0.999	0.987	1.000	0.986
ξ_{18}	0.640	0.328	0.969	0.999	0.969	0.999	0.958	1.001	0.964	1.001	0.956	1.000	0.957	0.999	0.951	1.000	0.951	1.000

Table 8.3: Relative efficiencies for Ψ-optimal spatial designs for 18 different combinations of fixed or optimised time, fixed or varying correlation parameters and $T = 3$, or $T = 6$ time points and the exponential or Matérn correlation functions.

Figures 8.6–8.7 show the spatio-temporal correlation for two different values of $\phi_2 = 0.5$, with exponential or Matérn correlation functions for the spatial correlation. The contour plots display the spatio-temporal correlation between each spatial point to the centre of the spatial region at time $t = 0$, averaged across the prior distribution for ϕ_1 for each time point $t = 0.2, 0.4, 0.6, 0.8, 1$. In all cases, red indicates high spatio-temporal correlation and light yellow indicates low spatio-temporal correlation.

These figures show how the spatio-temporal correlation affects the choice of the spatial design. The spatio-temporal correlation is given by the Kronecker product of the spatial and temporal correlations; multiplying each entry of the spatial correlation by the temporal correlation may lead to different correlations compared to the purely spatial correlations in Chapter 5. Therefore, a point which has high spatial correlation with the centre point at time $t = 0$ may have very different correlation for $t = 1$.

Figures 8.6 and 8.7 show that the spatio-temporal correlation is a non-constant function of distance in space and time; see for example Figures 8.6 (a) and Figures 8.6 (f), where the degree of the correlation decreases as the time increases. For this reason, spatial designs in Figure 8.1 (c) and (d) have less points in the centre of the region compared to their corresponding spatial designs in Figure 8.1 (a) and (b).

To summarise, for all the cases if we keep ϕ_2 constant and compare plots for $\nu = 0.5$ and $\nu = 1.5$, we can see that the strength of the correlation increases from $\nu = 0.5$ to $\nu = 1.5$. However, the range of the correlation for $\nu = 1.5$ is smaller than $\nu = 0.5$. Moreover, as the value of ϕ_2 increases, i.e. between $\phi_2 = 0.01$ and $\phi_2 = 0.5$, the spatio-temporal correlation decreases and for this reason the corresponding spatial designs for $\phi_2 = 0.01$ tend to cover the spatial region whereas for $\phi_2 = 0.5$ the points move towards to the boundaries of the region, e.g. Figure 8.4 (a) and (c).

8.5 Summary

In this chapter we introduced the problem of finding Bayesian optimal designs for prediction of a spatio-temporal process. Our aim was to introduce some ideas based on design criteria for Gaussian process models. As discussed earlier, this problem has received little attention in the literature. We extended our methodology for spatial design to design for both in space and time together.

We simplify the problem by making specific assumptions about the spatio-temporal correlation function and consider the model without a nugget effect. However, such models are only rarely seen in real applications of spatio-temporal data, due to their limited ability to describe the space-time interactions, see for example Banerjee et al. (2004); Cressie and Wikle (2011). We have also made the assumption that in each spatial location we are able to take observations at all time points. We did not consider the problem of any collecting data in some locations at some time points, i.e. to have

a subset each time. These problems do not have the assumption of separability in the correlation functions.

Future work could be extended by real applications, such as the chemical deposition problem in Chapter 6 where data could be considered to be correlated in time.

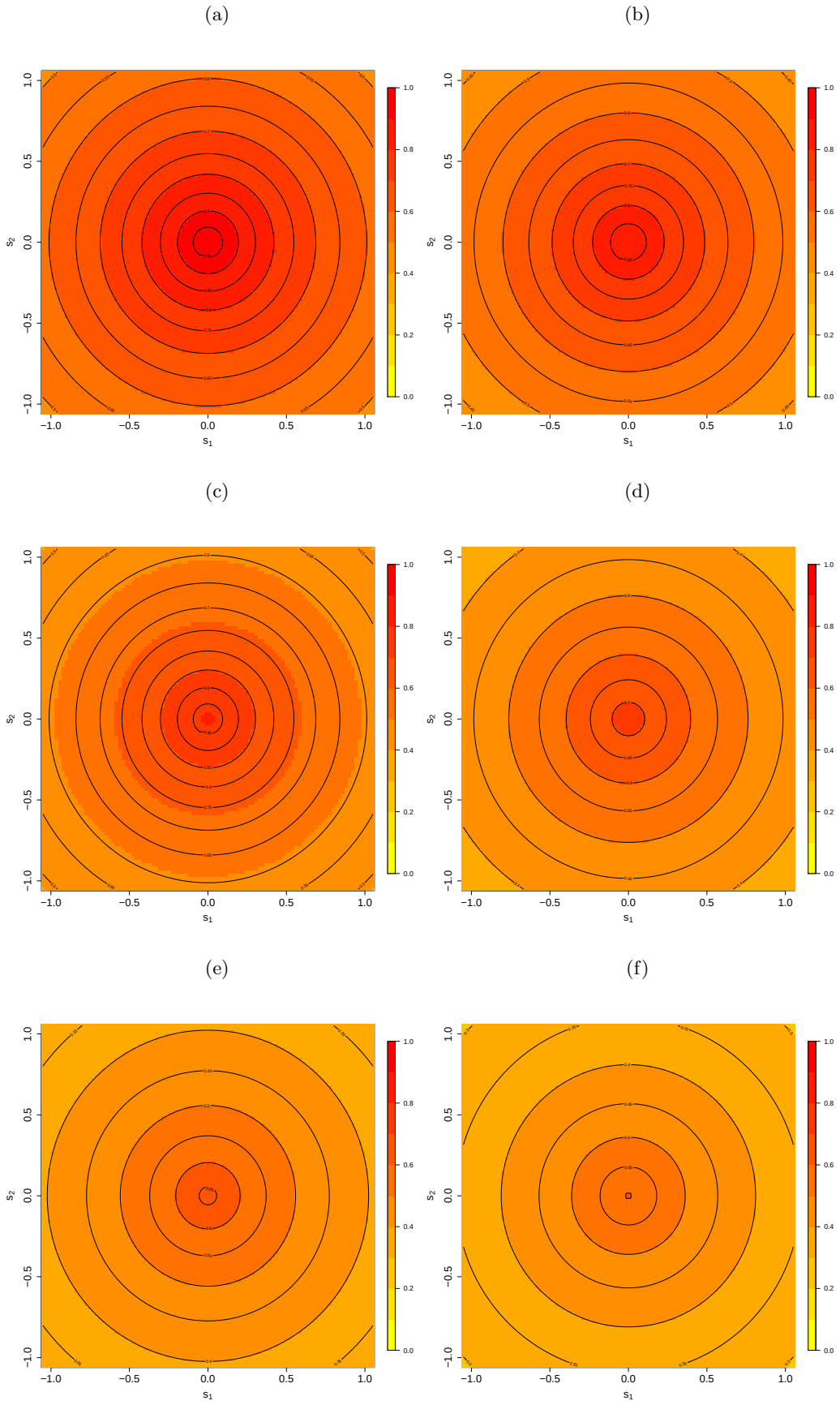


Figure 8.6: Contours displaying correlation between each spatial point and the centre of the design region, across the prior value of ϕ_1 for $\nu = 0.5$ and $\phi_2 = 0.5$: (a) at time=0, (b) at time=0.2, (c) at time=0.4, (d) at time=0.6, (e) at time=0.8, (f) at time=1.

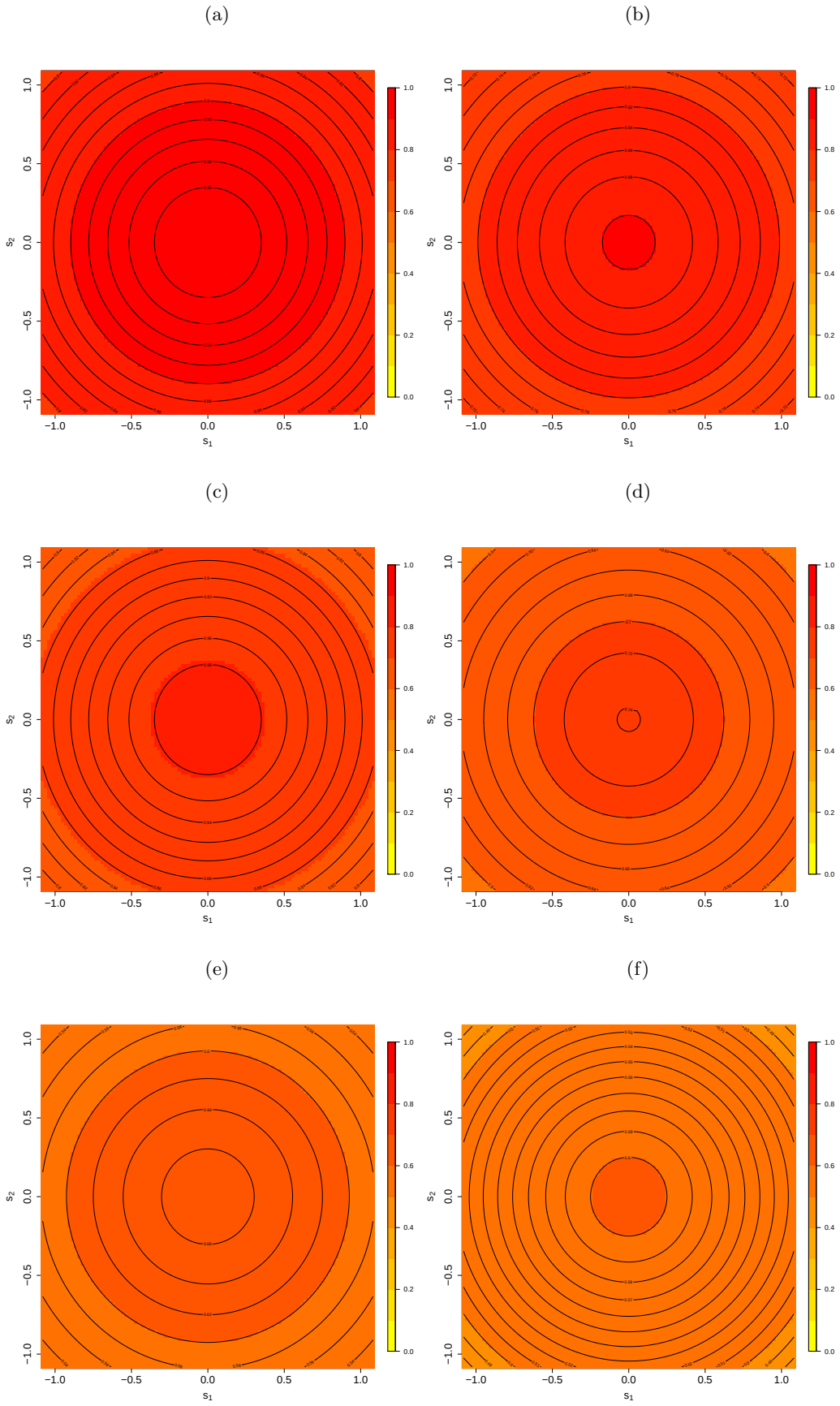


Figure 8.7: Contours displaying correlation between each spatial point and the centre of the design region, across the prior value of ϕ_1 for $\nu = 1.5$ and $\phi_2 = 0.5$: (a) at time=0, (b) at time=0.2, (c) at time=0.4, (d) at time=0.6, (e) at time=0.8, (f) at time=1.

Chapter 9

Conclusions and Future Work

This last chapter contains a summary of the thesis, outlines the major contributions and gives overall conclusions from the work. Also, we discuss some limitations of the work and outline future work to extend the approaches adopted for Bayesian optimal design for Gaussian process models.

9.1 Thesis Summary

In this thesis, we have discussed the problem of Bayesian optimal designs for prediction from the Gaussian process model, as applied in geostatistical science, i.e. in spatial and/or spatio-temporal environmental applications, and computer experiment applications. The objective of the research was to develop methods for optimal design for precise prediction of the response at unobserved points using the Gaussian process model. Throughout the thesis, we have drawn attention to the importance of incorporating the uncertainties introduced into the model due to unknown covariance parameters. To achieve this, our main contribution has been the development, implementation and assessment of a new closed-form approximation to the average expected posterior predictive variance. The methodology has been demonstrated on a variety of diverse applications, i.e. spatial data, computer experiments and spatio-temporal data.

Particular major contributions and conclusions from this work are:

New closed-form approximation to the design selection criterion: A major problem of applying a fully Bayesian approach to design selection for Gaussian process models is the computational burden associated with optimising an analytically intractable function. In Chapter 3, we adopted a decision theoretic approach (Chaloner and Verdinelli, 1995) to find Bayesian optimal designs that minimise the posterior predictive variance. The new proposed approximation allowed us to integrate out the unknown data and avoid the use of Monte Carlo integration which was a key step to overcoming the computational challenges usually associated with Bayesian designs. We

provided theoretical results to give insight into the proposed approximation. Moreover, in each of the application areas, Chapters 5, 7 and 8, we provided numerical studies to justify that the closed-form approximation can form the basis of a good design selection criterion.

Robustness and sensitivity: A major issue for design problems is how robust the choice of an optimal design is to changes in the prior hyperpameters and the sensitivity of the efficiency of an optimal design to these changes. In Chapter 4, we presented a thorough investigation of both the robustness and the sensitivity and concluded that, in general, an optimal design is mainly affected by the choice of the mean function, and the value of the decay parameter and noise-to-signal ratio.

Bayesian optimal designs for spatial data: The proposed closed form approximation criterion was used to find spatial designs, that is an optimal choice of sampling locations in a geographical region of interest. In Chapter 5, the designs we have studied compromise between minimax and maximin designs according to the choice of correlation and mean function and the prior information on the decay and noise-to-signal ratio. In addition our methodology was compared to existing well developed methodology for Bayesian optimal design. Diggle and Lophaven (2006) proposed and evaluated two designs, the lattice plus close pairs design and the lattice plus infill design, and our Bayesian Ψ-optimal design was compared with both. For all cases investigated, our design performs better, i.e. has smaller average prediction variance. The comparisons supported our findings that the designs are strongly influenced by the choice of mean function, and the strength and range of the correlation. This thesis is motivated by the need to obtain optimal sampling locations for environmental monitoring networks, and in this context we used data on chemical deposition in the eastern USA (Chapter 6) to inform the choice of both prospective and retrospective designs using our methodology. These designs indicated which stations should be dropped from the existing monitoring network.

Bayesian optimal design for computer experiments: In recent years design and analysis of computer experiments has received increasing attention, with special emphasis on space-filling designs. After the pioneer paper of Sacks et al. (1989), who proposed the Gaussian process for modelling the deterministic output of computer experiments, much work remains to be done to develop model-based designs for computer experiments. Therefore, in Chapter 7 we applied our methodology to the context of computer experiments and using our closed-form approximation, we found Bayesian optimal designs for prediction in two and three dimensions and compared them with designs in the literature. We also drew attention to the fact that standard space-filling designs may be inefficient for prediction when we take into account uncertainty in the model parameters.

Bayesian optimal designs for spatio-temporal data: As the final part of this thesis, we made some first steps towards extending the design methodology applied

for spatial data to also find some optimal sampling times. A possible model for such data is the Gaussian process model, and for this reason our closed-form approximation can naturally be extended to incorporate both spatial and temporal correlations. Although there is a literature on spatio-temporal statistical modelling, the literature for Bayesian optimal design for spatio-temporal data is limited. The extension from spatial design presents numerical challenges and hence we restricted our studies to separable spatio-temporal correlation functions. Chapter 8 introduced the spatio-temporal design problem and in this final chapter we tried to give a general idea of the problem. We concluded that the degree of the spatio-temporal correlation and the range of the correlation strongly influence the choice of the optimal points.

9.2 Future Work

Throughout this thesis, possibilities for future work and improvements have been highlighted in each application area, i.e in Chapters 5, 7 and 8.

More generally, our methodology can also be applied to machine learning applications. Rasmussen and Williams (2006) proposed the Gaussian process model (2.6) for regression and classifications problems in machine learning. The regression problem concerns the prediction with continuous outputs whereas classification problem addresses discrete output. Our approach can directly be applied in regression problems to identify the best input points to provide precise predictions for the untested input points. For the classification problem, a link function has to used since the output is discrete. Extension of our methodology to this second case could be interesting future research problem.

Clearly, there is also scope to develop our methods to address problems from different application areas, particularly in computer experiments. One example area would be computational chemistry, where computer simulations are used to understand chemical reactions, drug interactions and for molecular discovery. The next step for our research is to develop our methods to design efficient and effective experiments for building surrogates for these problems. A key step will be improving the computational algorithm to enable larger designs to be found, and incorporating physical data to enable the simulator to be calibrated and validated.

Bibliography

Ababou, R., Bagtzoglou, A. and Wood, E. (1994) On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, **26**, 99–133.

Atkinson, A., Donev, A. and Tobias, R. (2007) *Optimum Experimental Designs with SAS*. New York: Oxford University Press, 2nd edn.

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall/CRC.

Banerjee, S., Gelfan, A., Finley, A. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 825–848.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, 2nd edn.

Berger, J., De Oliveria, V. and Sanso, B. (2001) Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**, 1361–1374.

Boukouvalas, A., Comford, D. and Stehlík, M. (2014) Optimal design for correlated processes with input-dependent noise. *Computational Statistics & Data Analysis*, **71**, 1088–1102.

Bowman, V. and Woods, D. (2013) Weighted space-filling designs. *Journal of Simulation*, **7**, 249–263.

Chaloner, K. (1984) Optimal Bayesian experimental design for linear models. *Annals of Statistics*, **12**, 283–300.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.

Cook, R. and Nachtsheim, C. (1980) A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, **22**, 315–324.

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

Cressie, N. and Wikle, C. (2011) *Statistics for spatio-temporal data*. Hoboken, New Jersey: Wiley.

Currin, C., Mitchell, T., Morris, M. and Ylvisaker, D. (1991) Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

Diggle, P. and Lophaven, S. (2006) Bayesian geostatistical design. *Scandinavian Journal of Statistics*, **33**, 53–64.

Diggle, P., Moyeed, R. and Tawn, J. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C*, **47**, 299–350.

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

Dobbie, L., Henderson, L. and Stevens, L. (2008) Sparse sampling: spatial design for monitoring stream networks. *Statistics Surveys*, **2**, 113–153.

Evangelou, E. and Zhu, Z. (2012) Optimal predictive design augmentation for spatial generalised linear mixed models. *Journal of Statistical Planning and Inference*, **142**, 3242 – 3253.

Fang, K. (1980) The uniform design: application of number theoretic methods in experimental design. *Acta Mathematicae Applicatae Sinica*, **3**, 363–372.

Fang, K., Li, R. and Sudjianto, A. (2006) *Design and Modelling for Computer Experiments*. Boca Raton: Chapman & Hall/CRC.

Fang, K., Lin, D., Winker, P. and Zhang, Y. (2000) Uniform design: Theory and application. *Technometrics*, **42**, 237–248.

Fedorov, V. (1972) *Theory of Optimal Experiments*. New York: Academic Press.

Finley, A., Sang, H., Banerjee, S. and Gelfand, A. (2009) Improving the performance of predictive process modelling for large datasets. *Computational Statistics and Data Analysis*, **53**, 2873–2884.

Forrester, A., Sobester, A. and Keane, A. (2008) *Engineering Design via Surrogate Modelling*. Chichester: Wiley.

Fuentes, M., Chaudhuri, A. and Holland, D. (2007) Bayesian entropy for spatial sampling design of environmental data. *Journal of Environmental and Ecological Statistics,*, **14**, 323–340.

Gelfand, A., Banerjee, S. and Finley, A. (2013) Spatial design for knot selection in knot-based dimension reduction models. In *Spatio-temporal design: Advances in efficient data acquisition* (eds. J. Mateu and W. Muller), chap. 7, 142–169. Chichester: Wiley.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003) *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gbibs distribution and the Bayesian restoration of images. *IEEE Ttransactions on Pattern Analysis and Machine Iintelligence*, **6**, 721–741.

Gotwalt, C., Jones, B. and Steinberg, D. (2009) Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics*, **51**, 88–95.

Gramacy, B. and Lee, H. (2008) Bayesian treed Gaussian process models with an application to compute modelling. *Journal of the American Statistical Association*, **103**, 1119–1130.

Gramacy, R. and Lee, H. (2012) Cases for the nugget in modelling computer experiments. *Statistics and Computing*, **22**, 713–722.

Handcock, M. and Stein, M. (1993) A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Handcock, M. S. (1991) On cascading latin hypercube designs and additive models for experiments. *Communications Statistics-Theory Methods*, **20**, 417–439.

Harari, O. and Steinberg, D. (2014) Optimal designs for Gaussian process models via spectal decomposition. *Journal of Statistical Planning and Inference*, **154**, 87–101.

Harville, D. (2008) *Matrix Algebra From a Statistician's Perspective*. New York: Springer.

Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrica*, **57**, 97–109.

Heuvelink, G., Griffith, D., Hengl, T. and Melles, S. (2013) Sampling design optimization for space-time kriging. In *Spatio-temporal design: Advances in efficient data acquisition* (eds. J. Mateu and W. Muller), chap. 9, 207–222. Chichester: Wiley.

Johnson, M., Moore, L. and Ylvisaker, D. (1990) Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Johnson, M. and Nachtsheim, C. (1983) Some guidelines for constructing exact D-optimal designs on convex designs spaces. *Technometrics*, **25**, 271–277.

Kaufman, C., Schervish, M. and Nychka, D. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545–1555.

Kazianka, H. and Pilz, J. (2012) Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, **40**, 304–327.

Kiefer, J. and Wolfowitz, J. (1959) Optimum designs in regression problems. *The Annals of Mathematical Statistics*, **30**, 271–294.

Kitanidis, P. (1986) Parameter uncertainty in estimation of spatial function: Bayesian analysis. *Water Resources Research*, **22**, 499–507.

Kleijnen, J. (2008) *Design and Analysis of Simulation Experiments.* New York: Springer.

Kolmogorov, A. (1941) Interpolation and extrapolation (in Russian). *Izv. Akad. Nauk SSSR, Series Mathematics*, **5**, 3–14.

Krige, D. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical Metallurgical and Mining Society of South Africa*, **52**, 119–139.

Kythe, K. and Schaferkotter, R. (2005) *Handbook of computational methods for integration.* Boca Raton: Chapman & Hall/CRC.

Le, N. and Zidek, J. (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis*, **43**, 351–374.

Leatherman, E., Santner, T. and Dean, A. (2014) Designs for computer experiments that minimize the weighted integrated mean square prediction error. *Submitted.*

Lindley, D. (1956) On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, **27**, 986–1005.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Ye, K. (2006) Variable selection for gaussian process models in computer experiments. *Technometrics*, **48**, 478–490.

Matérn, B. (1960) Spatial variation, stochastic models and their application to some problems in forest surveys and other sampling investigation. *Medd. Statens Skogsforskningsinst*, **5**, 1–144.

Mateu, J. and Muller, W. (2012) *Spatio-Temporal Design: Advances in Efficient Data Acquisition.* Chichester: Wiley.

Matheron, G. (1963) Principles of geostatistcs. *Economic Geology*, **58**, 1246–1266.

McBartney, A., Webster, R. and Burgess, T. (1981) The design of optimal sampling schemes for local estimation and mapping of regionalised variables, I -theory and methods. *Computers and Geosciences*, **7**, 331–334.

McKay, M., Beckaman, R. and Conover, W. (1979) A comparison of three methods for selecting values of inputs variables in the analysis of output from a compute code. *Technometrics*, **21**, 239–245.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meyer, R. and Nachtsheim, C. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, **37**, 60–69.

Morris, M. and Mitchell, T. (1995) Explanatory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.

Morris, M., Mitchell, T. and Ylvisker, D. (1993) Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, **35**, 243–255.

Müller, W. (2007) *Collecting Spatial Data*. Berlin: Springer, third edn.

Müller, W. and Zimmerman, D. (1999) Optimal design for variogram estimation. *Environmetrics*, **10**, 23–37.

Neal, R. (1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Technical Report 9702, University of Toronto*.

O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society: Series B*, **40**, 1–42.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. and Kim, N. (2010) Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, **132**.

Pronzato, L. and Müller, W. (2012) Design of computer experiments: space filling and beyond. *Statistics and Computing*, **22**, 681–701.

Qian, P., Wu, H. and Wu, J. (2008) Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, **50**, 383–396.

Rasmussen, C. and Williams, C. (2006) *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT press.

Ren, C., Sun, D. and He, Z. (2012) Objective Bayesian analysis for a spatial model with nugget effects. *Journal of Statistical Planning and Inference*, **142**, 1933 – 1946.

Ren, C., Sun, D. and Sahu, S. (2013) Objective Bayesian analysis of spatial models with separable correlation functions. *The Canadian Journal of Statistics.*, **41**, 488–507.

Royle, J. (2002) Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, **100**, 121–134.

Sacks, J., Welch, J., Mitchell, J. and Wynn, H. (1989) Design and analysis of computer experiments. *Statistical Science*, **4**, 409–423.

Sang, H. and Huang, J. (2012) A full scale approximation of the covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **74**, 111–132.

Santner, T., Williams, B. and Notz, W. (2003) *The Design and Analysis of Computer Experiments*. New York: Springer.

Sebastiani, P. and Wynn, H. (2000) Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B*, **62**, 145–157.

Shannon, C. (1948) A mathematical theory of communication. *Bell System Bell System Technical Journal*, **27**, 379–423.

Shewry, M. and Wynn, H. (1987) Maximum entropy sampling. *Journal of Applied Statistics*, **14**, 165–170.

Siotani, M. (1967) Some applications of Loewner's ordering on symmetric matrices. *Annals of the Institute of Statistical Mathematics*, **19**, 245–259.

Spöck, G. and Pliz, J. (2010) Spatial sampling design and covariance-robust minimax prediction based on convex design ideas. *Stochastic Environmental Research and Risk Assessment*, **24**, 463–482.

Su, Y. and Cambanis, S. (1993) Sampling designs for estimation of a random process. *Stochastic Processes and their Applications*, **43**, 47–89.

Tang, B. (1993) Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*, **88**, 1392–1397.

Tudose, L. and Jucan, D. (2007) Pareto approach in multi-objective optimal design of helical compression springs. *Annals of the Oradea University, Fascicle of Management and Technological Engineering*, **6**, 991–998.

Uciński, D. and Maciej, P. (2010) Sensor network design for the estimation on spatially distributed processes. *International Journal of Applied Mathematics and Computer Science*, **20**, 459–481.

Wiener, N. (1949) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, Massachusetts: MIT Press.

Wikle, C. (2010) Hierarchical modeling with spatial data. In *Handbook of Spatial Statistics* (eds. A. Gelfan, P. Diggle, M. Fuentes and P. Guttorp), chap. 7, 89–106. Boca Raton: Chapman & Hall/CRC.

Wikle, C. and Royle, J. (1999) Space-time dynamic design of environmental monitoring networks. *Journal of Agriculture, Biological, and Environmental Statistics*, **4**, 489–507.

— (2005) Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Envirmonetrics*, **16**, 507–522.

Woods, D. (2010) Robust designs for binary data: applications of simulated annealing. *Journal of Statistical Computation and Simulation*, **80**, 29–41.

Wu, R. and Kaufman, C. (2014) Comparing non-informative priors for estimation and prediction in spatial models. *Submitted*.

Xia, G., Mirand, M. and Gelfan, A. (2006) Approximately optimal spatial design approaches for environmental data. *Envirometrics*, **17**, 363–385.

Yan, J., Cowles, K., Wang, S. and Armstrong, M. (2007) Parallelizing MCMC for Bayesian spatiotemporal geostatistical models. *Statistics and Computing*, **17**, 323–335.

Yates, F. (1970) *Experimental design: Selected papers of Frank Yates*. London: Griffin.

Zhu, Z. and Stein, M. (2005) Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, **134**, 583–603.

— (2006) Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 24–44.

Zidek, J. and Zimmerman, D. (2010) Hierarchical modeling with spatial data. In *Handbook of Spatial Statistics* (eds. A. Gelfan, P. Diggle, M. Fuentes and P. Guttorp), chap. 10, 131–148. Boca Raton: Chapman & Hall/CRC.

Zimmerman, D. (2006) Optimal network design for spatial prediction, covariance parameter estimation and empirical prediction. *Envirometrics*, **17**, 635–652.

Appendix A

A.1 Proof of Lemma 3.1

Auxiliary facts

Details can be found in Harville (2008).

1. Matrix determinant: Let $\mathbf{A}$ be a non-singular square matrix and $\mathbf{u}$, $\mathbf{v}$ column vectors of an appropriate size. Then

$$|\mathbf{A} + \mathbf{u}\mathbf{v}^\top| = |\mathbf{A}|(1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u}) \tag{A.1}$$

2. Sherman–Morrison formula: Let $\mathbf{A}$ be a non-singular square matrix and $\mathbf{u}$, $\mathbf{v}$ are column vectors. Then $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is nonsingular matrix and

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A^{-1}}\mathbf{u}\mathbf{v}^\top\mathbf{A^{-1}}}{1 + \mathbf{v}^\top\mathbf{A^{-1}}\mathbf{u}} \tag{A.2}$$

3. Sherman Morrison Woodbury formula: Let $\mathbf{A}$, $\mathbf{C}$, $\mathbf{U}, \mathbf{V}$ be $n \times n$, $k \times k, n \times k$ and $k \times n$ matrices respectively, with $\mathbf{A}$ and $\mathbf{C}$ non-singular. Then

$$(\mathbf{A} + \mathbf{UCV})^{\mathbf{-1}} = \mathbf{A}^{-1} - \mathbf{A^{-1}}\mathbf{U}(\mathbf{C^{-1}} + \mathbf{V}\mathbf{A^{-1}}\mathbf{U})^{\mathbf{-1}}\mathbf{V}\mathbf{A^{-1}} \tag{A.3}$$

Proof:

The proof follows similar steps to that found in Ren et al. (2012).

Recall that integrated likelihood $L^I(\phi, \delta^2)$ in (2.31) is

$$\begin{aligned} L^I(\phi, \delta^2) &= \int f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi, \delta^2)\pi(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2 \\ &= \frac{|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{\left[b + \frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)^\top[\boldsymbol{\Sigma} + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}_0)\right]^{a+\frac{n}{2}}}, \end{aligned}$$

which depends on $\rho(\phi; d)$ through $\boldsymbol{\Sigma}$. The continuity of $L^I(\phi, \delta^2)$ is a consequence of continuity of the correlation function $\rho(\phi; d)$.

Part (a): It is sufficient to show the variance-covariance matrix, $\boldsymbol{\Sigma}$ is positive-definite and then $L^I(\phi, \delta^2) > 0$ for all ϕ and δ^2.

- for $(\phi, \delta^2) \to (\phi, 0^+)$,
$$\boldsymbol{\Sigma} = \mathbf{C}(\phi) + \delta^2\mathbf{I} \to \mathbf{C}(\phi)$$
where $\mathbf{C}(\phi)$ is a positive definite correlation matrix for any $\phi > 0$.

- for $(\phi, \delta^2) \to (\infty, \delta^2)$,
$$\mathbf{C}(\phi) \to \mathbf{I} \quad \text{as} \quad \phi \to \infty \quad \text{and} \quad \boldsymbol{\Sigma} = (\delta^2 + 1)\mathbf{I}$$
Hence, $\boldsymbol{\Sigma}$ is positive definite matrix for any δ^2.

- for $(\phi, \delta^2) \to (\infty, 0^+)$, $\boldsymbol{\Sigma} \to \mathbf{I}$

Hence, in each case, $\lim_{\phi\to\infty} L^I(\phi, \delta^2) > 0$.

Part (b): From Assumption 3.3, $\boldsymbol{\Sigma} = \delta^2\mathbf{I} + \mathbf{1_n}\mathbf{1_n}^\top + v(\phi)\mathbf{D} + o(v(\phi))$. Ignoring the $o(\nu(\phi))$ term, and following Kazianka and Pilz (2012) who stated that $\mathbf{1_n}\mathbf{1_n}^\top + v(\phi)\mathbf{D}$ is positive-definite and continuous with respect to its eigenvalues, we have

$$c_1\mathbf{1_n}\mathbf{1_n}^\top + (c_2 v(\phi) + \delta^2)\mathbf{I} \leq \boldsymbol{\Sigma} \leq \mathbf{1_n}\mathbf{1_n}^\top + (c_3 v(\phi) + \delta^2)\mathbf{I} \tag{A.4}$$

where $\leq$ denotes the Lowewner partial ordering and $c_1 < 1$, $c_2 < \min_{\lambda_i>0}\lambda_i$ and $c_3 = \max_i|\lambda_i|$ are positive constants, with λ_i, $i = 1, \ldots, n$, being the eigenvalues of $\mathbf{D}$. Recall the properties of Loewner ordering for determinant and inversion, i.e. $\mathbf{A} \leq \mathbf{B} \Rightarrow \|\mathbf{A}\| \leq \|\mathbf{B}\|$ and $\mathbf{B}^{-1} \leq \mathbf{A}^{-1}$, see Siotani (1967). Then

$$|c_1\mathbf{1_n}\mathbf{1_n}^\top + (c_2 v(\phi) + \delta^2)\mathbf{I}| \leq |\boldsymbol{\Sigma}| \leq |\mathbf{1_n}\mathbf{1_n}^\top + (c_3 v(\phi) + \delta^2)\mathbf{I}|,$$

and using (A.1) we have that

$$(c_2 v(\phi) + \delta^2)^n\left(1 + \frac{c_1 n}{c_2 v(\phi) + \delta^2}\right) \leq |\boldsymbol{\Sigma}| \leq (c_3 v(\phi) + \delta^2)^n\left(1 + \frac{c_1 n}{c_3 v(\phi) + \delta^2}\right),$$

and hence

$$(c_2 v(\phi) + \delta^2)^{n-1}(c_2 v(\phi) + \delta^2 + c_1 n) \leq |\boldsymbol{\Sigma}| \leq (c_3 v(\phi) + \delta^2)^{n-1}(c_3 v(\phi) + \delta^2 + n). \tag{A.5}$$

It follows that, as $\phi \to 0^+$,

$$|\boldsymbol{\Sigma}| = \mathcal{O}\big((\delta^2 + v(\phi))^{n-1}\big). \tag{A.6}$$

The next step is to find the determinant of the matrix $|\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{R}|$. We apply (A.2)

to find the inverse of $\mathbf{\Sigma^{-1}}$.

$$\frac{1}{c_3v(\phi)+\delta^2}\left(\mathbf{I}-\frac{\mathbf{1_n1_n^\top}}{c_3v(\phi)+\delta^2+n}\right)\leq\mathbf{\Sigma^{-1}}\leq\frac{1}{c_2v(\phi)+\delta^2}\left(\mathbf{I}-\frac{c_1\mathbf{1_n1_n^\top}}{c_1n+\delta^2+c_2v(\phi)}\right). \tag{A.7}$$

We make use of

$$|\mathbf{F^T\Sigma^{-1}F+R}|=|\mathbf{F^T\Sigma^{-1}FR^{-1}+I}||\mathbf{R}| \tag{A.8}$$

and Theorem 13.7.3. and Corollary 13.7.4 from Harville (2008), which gives

$$|\mathbf{F^T\Sigma^{-1}FR^{-1}+I}|=\sum_{r=0}^{k}\sum_{\{i_1,\ldots,i_r\}}|\mathbf{F^T\Sigma^{-1}FR^{-1}}^{\{i_1,\ldots,i_r\}}|, \tag{A.9}$$

where $\{i_1,\ldots,i_r\}$ is an $r-$dimensional subset of the first k positive integers, the second summation is over all the $k!/(k-r)!r!$ such subsets and $\mathbf{F^T\Sigma^{-1}FR^{-1}}^{\{i_1,\ldots,i_r\}}$ is the $(k-r)\times(k-r)$ principal submatrix of $\mathbf{F^T\Sigma^{-1}FR^{-1}}$ obtained by striking out the $\{i_1,\ldots,i_r\}th$ rows and columns. We assume $\{i_1,\ldots,i_r\}$ is the empty set for $r=0$.

From equation (A.7) we see that:

$$\mathbf{F^\top\Sigma^{-1}FR^{-1}}\geq\frac{1}{c_3v(\phi)+\delta^2}\left(\mathbf{F^\top FR^{-1}}-\frac{\mathbf{F^\top1_n1_n^\top FR^{-1}}}{c_3v(\phi)+\delta^2+n}\right)$$

$$\mathbf{F^\top\Sigma^{-1}FR^{-1}}\leq\frac{1}{c_2v(\phi)+\delta^2}\left(\mathbf{F^\top FR^{-1}}-\frac{c_1\mathbf{F^\top1_n1_n^\top FR^{-1}}}{c_2v(\phi)+\delta^2+c_1n}\right)$$

and hence

$$\begin{aligned}
|\mathbf{F^\top\Sigma^{-1}FR^{-1}}| &\geq (c_3v(\phi)+\delta^2)^{-k}|\mathbf{F^\top FR^{-1}}||\mathbf{I}-\frac{(\mathbf{F^\top FR^{-1}})^{-1}\mathbf{F^\top1_n1_n^\top FR^{-1}}}{c_3v(\phi)+\delta^2+n}|\\
&\geq (c_3v(\phi)+\delta^2)^{-k}|\mathbf{F^\top FR^{-1}}|\left(1-\frac{\mathbf{1_n^\top F(F^\top F)^{-1}F^\top1_n}}{c_3v(\phi)+\delta^2+n}\right)\\
&\geq |\mathbf{F^\top FR^{-1}}|\left(\frac{c_3v(\phi)+\delta^2+n-\mathbf{1_n^\top F(F^\top F)^{-1}F^\top1_n}}{(c_3v(\phi)+\delta^2+n)(c_3v(\phi)+\delta^2)^k}\right)
\end{aligned} \tag{A.10}$$

$$\begin{aligned}
|\mathbf{F^\top\Sigma^{-1}FR^{-1}}| &\leq (c_2v(\phi)+\delta^2)^{-k}|\mathbf{F^\top FR^{-1}}||\mathbf{I}-\frac{(\mathbf{F^\top FR^{-1}})^{-1}c_1\mathbf{F^\top1_n1_n^\top FR^{-1}}}{c_2v(\phi)+\delta^2+c_1n}|\\
&\leq (c_2v(\phi)+\delta^2)^{-k}|\mathbf{F^\top FR^{-1}}|\left(1-\frac{\mathbf{1_n^\top F(F^\top F)^{-1}F^\top1_n}}{c_2v(\phi)+\delta^2+c_1n}\right)\\
&\leq |\mathbf{F^\top FR^{-1}}|\left(\frac{c_2v(\phi)+\delta^2+c_1n-c_1\mathbf{1_n^\top F(F^\top F)^{-1}F^\top1_n}}{(c_2v(\phi)+\delta^2+c_1n)(c_2v(\phi)+\delta^2)^k}\right)
\end{aligned} \tag{A.11}$$

As (A.10) and (A.11) hold for all submatrices $|\mathbf{F^T\Sigma^{-1}FR^{-1}}^{\{i_1,\ldots,i_r\}}|$, we can use (A.9) to obtain:

$$
\begin{aligned}
|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}\mathbf{R}^{-\mathbf{1}}+\mathbf{I}||\mathbf{R}| &\geq \left(\frac{c_3v(\phi)+\delta^2+n-\mathbf{1_n^\top F(F^\top F)^{-1}F^\top 1_n}}{(c_3v(\phi)+\delta^2+n)(c_3v(\phi)+\delta^2)^k}\right)\sum_{r=0}^{k}\sum_{\{i_1,\ldots,i_r\}}|\mathbf{F}^\top\mathbf{F}\mathbf{R}^{-\mathbf{1}^{\{i_1,\ldots,i_r\}}}||\mathbf{R}| \\
&\geq \left(\frac{c_3v(\phi)+\delta^2+n-\mathbf{1_n^\top F(F^\top F)^{-1}F^\top 1_n}}{(c_3v(\phi)+\delta^2+n)(c_3v(\phi)+\delta^2)^k}\right)|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|
\end{aligned}
$$

$$
\begin{aligned}
|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}\mathbf{R}^{-\mathbf{1}}+\mathbf{I}|\mathbf{R}| &\leq \left(\frac{c_2v(\phi)+\delta^2+c_1n-c_1\mathbf{1_n^\top F(F^\top F)^{-1}F^\top 1_n}}{(c_2v(\phi)+\delta^2+c_1n)(c_2v(\phi)+\delta^2)^k}\right)\sum_{r=0}^{k}\sum_{\{i_1,\ldots,i_r\}}|\mathbf{F}^\top\mathbf{F}\mathbf{R}^{-\mathbf{1}^{\{i_1,\ldots,i_r\}}}||\mathbf{R}| \\
&\leq \left(\frac{c_2v(\phi)+\delta^2+c_1n-c_1\mathbf{1_n^\top F(F^\top F)^{-1}F^\top 1_n}}{(c_2v(\phi)+\delta^2+c_1n)(c_2v(\phi)+\delta^2)^k}\right)|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|
\end{aligned}
$$

Note that $\mathbf{P}_F = \mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{F}^\top$ is a projection matrix. Hence, by Harville (2008) Theorem 12.3.5, if $\mathbf{1}_n \in \mathcal{C}(\mathbf{F})$, $\mathbf{1}_n^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{F}^\top\mathbf{1}_n = \mathbf{1}_n^\top\mathbf{1}_n = n$. Hence,

$$
|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R}| \leq \begin{cases} \frac{|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|(c_2v(\phi)+\delta^2+c_1n-c_1\mathbf{1_n^\top P_x 1_n})}{(c_2v(\phi)+\delta^2+c_1n)(c_2v(\phi)+\delta^2)^k} & \text{if } \mathbf{1}\notin\mathcal{C}(\mathbf{F}) \\ \frac{|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|(c_2v(\phi)+\delta^2)}{(c_2v(\phi)+\delta^2+c_1n)(c_2v(\phi)+\delta^2)} & \text{if } \mathbf{1}\in\mathcal{C}(\mathbf{F}) \end{cases}
$$

and

$$
|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R}| \geq \begin{cases} \frac{|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|(c_3v(\phi)+\delta^2+n-\mathbf{1_n^\top P_x 1_n})}{(c_3v(\phi)+\delta^2+n)(c_3v(\phi)+\delta^2)^k} & \text{if } \mathbf{1}\notin\mathcal{C}(\mathbf{F}), \\ \frac{|\mathbf{F}^\top\mathbf{F}+\mathbf{R}|(c_3v(\phi)+\delta^2)}{(c_3v(\phi)+\delta^2)(c_3v(\phi)+\delta^2)} & \text{if } \mathbf{1}\in\mathcal{C}(\mathbf{F}). \end{cases}
$$

So we have:

$$
|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R}| = \begin{cases} \mathcal{O}\big((\delta^2+v(\phi))^{-k}\big) & \text{if } \mathbf{1}\notin\mathcal{C}(\mathbf{F}), \\ \mathcal{O}\big((\delta^2+v(\phi))^{-k+1}\big) & \text{if } \mathbf{1}\in\mathcal{C}(\mathbf{F}). \end{cases} \tag{A.12}
$$

The final step is to find the limits for $(\mathbf{\Sigma}+\mathbf{F}\mathbf{R}\mathbf{F}^\top)^{-\mathbf{1}}$. Using the Sherman Morrison Woodbury formula (A.3), we have

$$
(\mathbf{\Sigma}+\mathbf{F}\mathbf{R}\mathbf{F}^\top)^{-\mathbf{1}} = \mathbf{\Sigma}^{-\mathbf{1}} - \mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}(\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}} = \mathcal{R}_G \tag{A.13}
$$

From (A.7) $\mathbf{\Sigma}^{-\mathbf{1}} \geq \frac{1}{\delta^2+c_3v(\phi)}\left(\mathbf{I}-\frac{\mathbf{1_n1_n^\top}}{\delta^2+c_3v(\phi)+n}\right)$. Setting $\mathbf{\Sigma}$ equal to this lower-bound and using (A.2) we have

$$
\begin{aligned}
(\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-\mathbf{1}} &= \left(\mathbf{F}^\top\mathbf{F}+\mathbf{R}-\frac{\mathbf{F}^\top\mathbf{1_n}\mathbf{1_n}^\top\mathbf{F}}{n+\delta^2+c_3v(\phi)}\right)^{-1} \\
&= (\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-1}+\frac{(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top\mathbf{1_n}\mathbf{1_n}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}}{n+\delta^2+c_3v(\phi)-\mathbf{1_n}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top\mathbf{1_n}}.
\end{aligned}
$$

Now we denote by $\mathbf{P}_F^*=\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top$ and $c_3^*=n+\delta^2+c_3v(\phi)$ and

$$
\mathbf{F}(\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top = (c_3v(\phi)+\delta^2)\left(\mathbf{P}_F^*+\frac{\mathbf{P}_F^*\mathbf{1}_n\mathbf{1}_n^\top\mathbf{P}_F^*}{c_3^*-\mathbf{1}_n^\top\mathbf{P}_F^*\mathbf{1}_n}\right).
$$

Substituting all terms in (A.13) we have

$$
(\delta^2+c_3v\phi))\mathcal{R}_G = \left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{c_3^*}\right)-\left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{c_3^*}\right)\left(\mathbf{P}_F^*+\frac{\mathbf{P}_F^*\mathbf{1}_n\mathbf{1}_n^\top\mathbf{P}_F^*}{c_3^*-\mathbf{1}_n^\top\mathbf{P}_F^*\mathbf{1}_n}\right)\left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{c_3^*}\right). \tag{A.14}
$$

When $(\phi,\delta^2)\to(0^+,0^+)$, $c_3^*\to n$ and from (A.14) we have that

$$
(\delta^2+c_3v\phi))\mathcal{R}_G \to \left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)-\left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)\left(\mathbf{P}_F^*+\frac{\mathbf{P}_F^*\mathbf{1}_n\mathbf{1}_n^\top\mathbf{P}_F^*}{n-\mathbf{1}_n^\top\mathbf{P}_F^*\mathbf{1}_n}\right)\left(\mathbf{I}-\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right). \tag{A.15}
$$

Now, we need to show that $n-\mathbf{1}_n^\top\mathbf{P}_F^*\mathbf{1}_n\neq 0$

$$
\begin{aligned}
\mathbf{P}_F^* &= \mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top \\
&= \mathbf{F}[(\mathbf{F}^\top\mathbf{F})(\mathbf{I}+(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{R})]^{-\mathbf{1}}\mathbf{F}^\top \\
&= \mathbf{F}[\mathbf{I}+(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{R}]^{-\mathbf{1}}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{F}^\top
\end{aligned}
$$

Applying the Sherman-Morrison-Wodbury formula (A.3) to $[\mathbf{I}+(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{R}]^{-1}$, we obtain

$$
\mathbf{P}_F^*=\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{F}^\top-\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{R}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{F}^\top.
$$

If $\mathbf{1}\in\mathcal{C}(\mathbf{F})$ then $n-\mathbf{1_n}^\top\mathbf{P_x^*}\mathbf{1_n}=n-n+\mathbf{1_n}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{R}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{1_n}>0$ and if $\mathbf{1}\notin\mathcal{C}(\mathbf{F})$ then $n-\mathbf{1_n}^\top\mathbf{P_x^*}\mathbf{1_n}=n-\mathbf{1_n}^\top\mathbf{P_x}\mathbf{1_n}+\mathbf{1_n}^\top\mathbf{F}(\mathbf{F}^\top\mathbf{F}+\mathbf{R})^{-\mathbf{1}}\mathbf{R}(\mathbf{F}^\top\mathbf{F})^{-\mathbf{1}}\mathbf{1_n}>0$.

Therefore $(\delta^2+c_3v(\phi))R_G$ is a bounded, non-zero matrix, and hence

$$
S^2=(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)^\top(\mathbf{\Sigma}+\mathbf{F}\mathbf{R}\mathbf{F}^\top)^{-\mathbf{1}}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}_0)\propto(\delta^2+c_3v(\phi))^{-1},
$$

and

$$
S^2=\mathcal{O}\big((\delta^2+v(\phi))^{-1}\big). \tag{A.16}
$$

Combining statements (A.6), (A.12) and (A.16) completes the proof for (b).

Part (c): for $(\phi, \delta^2) \to (\infty, \infty)$, $(\phi, \delta^2) \to (\phi, \infty)$ or $(\phi, \delta^2) \to (0^+, \infty)$, it is sufficient to show that $\mathbf{\Sigma}$ depends only on δ^2. As $\delta^2 \to \infty$, the correlation matrix $\mathbf{C}(\phi)$ is a function of the correlation parameter ϕ:

$$\mathbf{\Sigma} = \delta^2(\mathbf{I} + \frac{\mathbf{C}(\phi)}{\delta^2}) \to \delta^2\mathbf{I}(1 + o(1)).$$

Hence,$|\mathbf{\Sigma}| = |\delta^2\mathbf{I}(1+o(1))| = (\delta^2)^n$ and $|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F} + \mathbf{R}| \simeq (\delta^2)^{-k}|\mathbf{F}^\top\mathbf{F} + \mathbf{R}|$. Therefore, their respective order are given by the following equations:

$$|\mathbf{\Sigma}| = \mathcal{O}\big((\delta^{2n})\big), \tag{A.17}$$

$$|\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F} + \mathbf{R}| = \mathcal{O}\big((\delta^2)^{-k}\big). \tag{A.18}$$

Finally, we use Sherman-Morrison-Woodbury formula (A.3) to expand the matrix :

$$[\mathbf{\Sigma}+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^\top]^{-1} = \mathbf{\Sigma}^{-\mathbf{1}} - \mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F}(\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{F} + \mathbf{R})^{-\mathbf{1}}\mathbf{F}^\top\mathbf{\Sigma}^{-\mathbf{1}} = (\delta^2)^{-1}(\mathbf{I} - \mathbf{F}(\mathbf{F}^\top\mathbf{F} + \mathbf{R})^{-\mathbf{1}}\mathbf{F}).$$

Therefore,

$$S^2 = \mathcal{O}\big((\delta^2)^{-1}\big). \tag{A.19}$$

Equations (A.17), (A.18) and (A.19) complete the proof for (c).

A.2 Sensitivity Study for $n = 30$

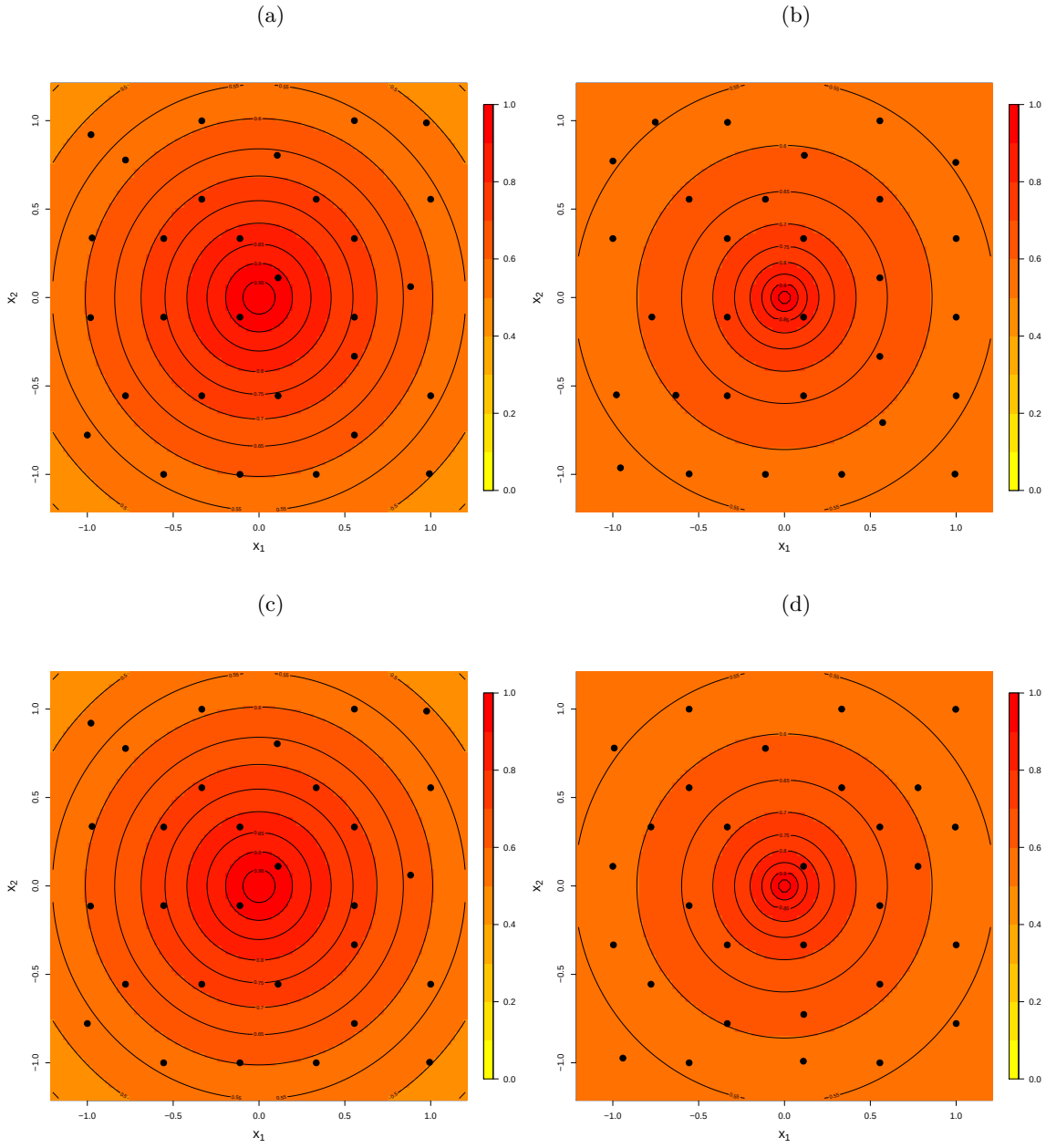


Figure A.1: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), constant mean ($F_2 = 0$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

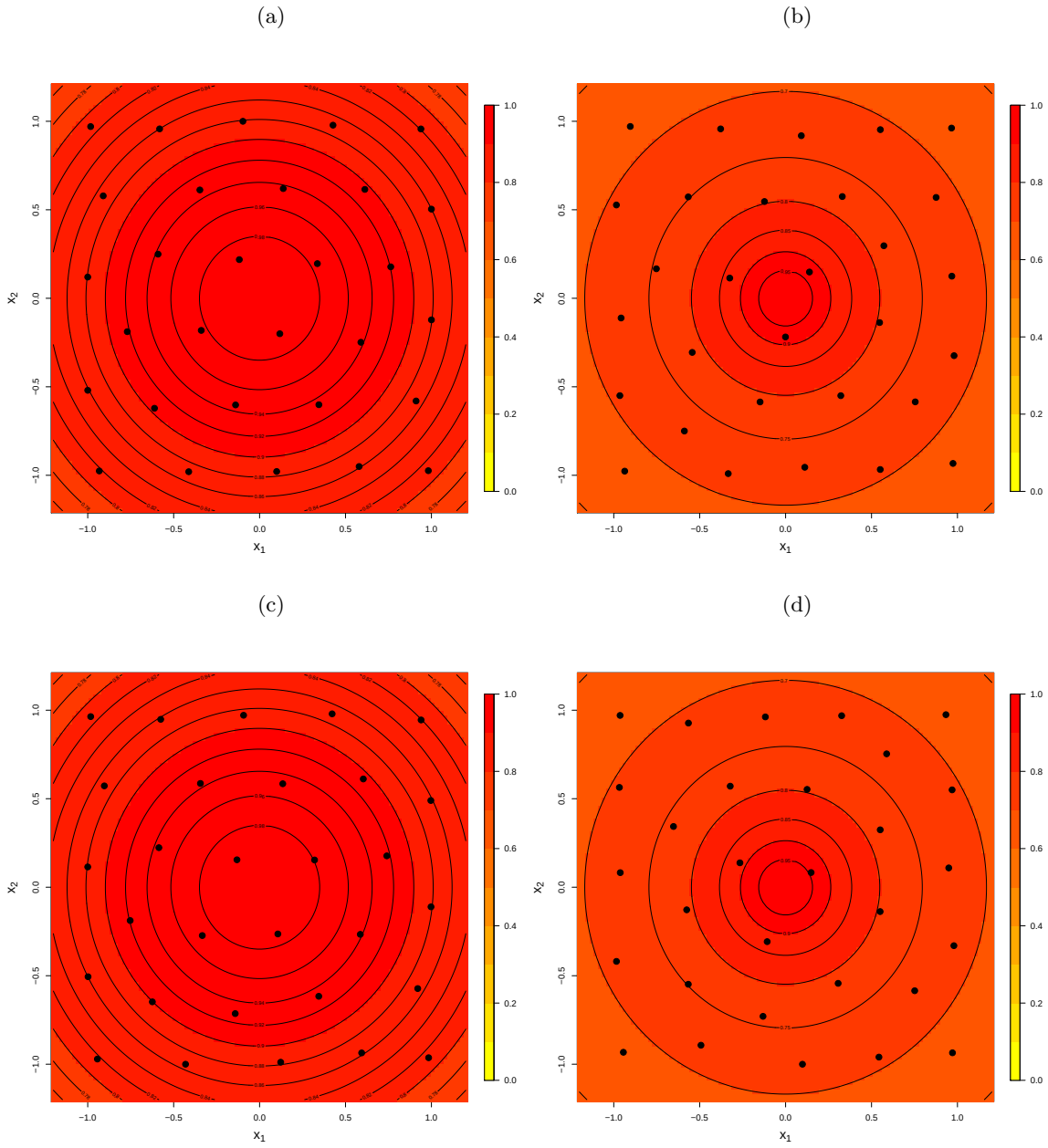


Figure A.2: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), constant mean ($F_2 = 0$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

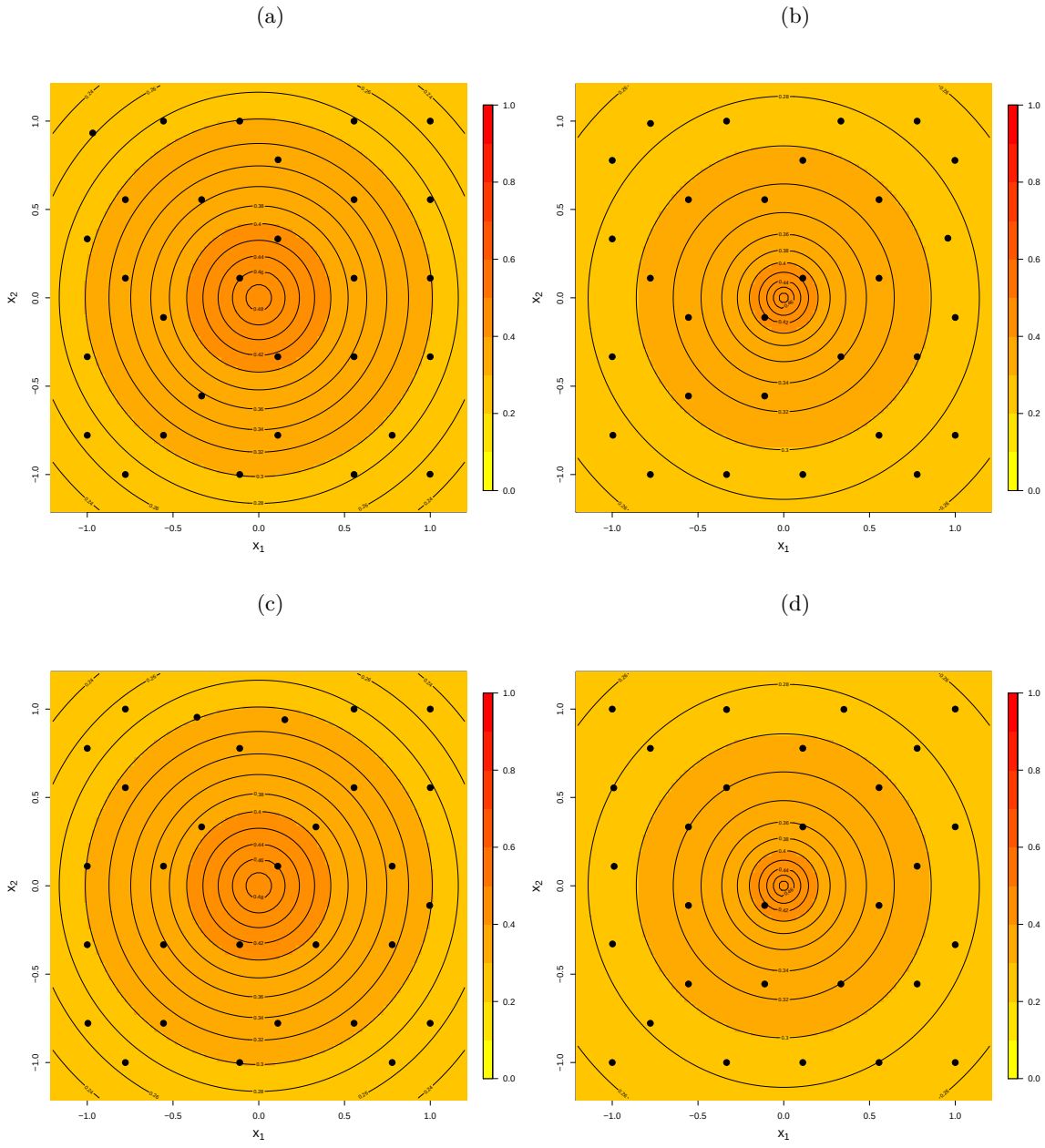


Figure A.3: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), constant mean ($F_2 = 0$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

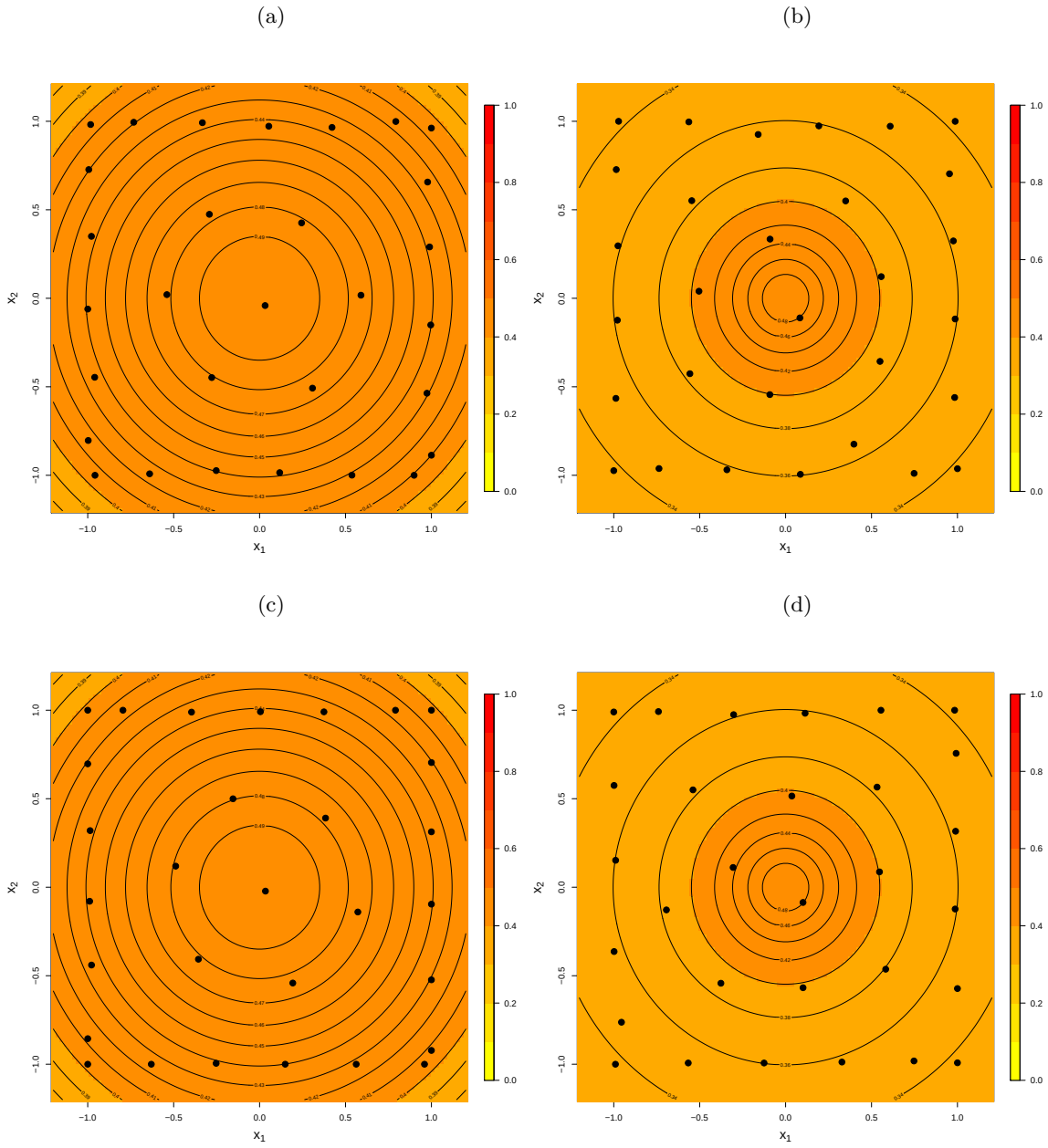


Figure A.4: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), constant mean ($F_2 = 0$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $R^{-1} = 0.25$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $R^{-1} = 0.25$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $R^{-1} = 4$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $R^{-1} = 4$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

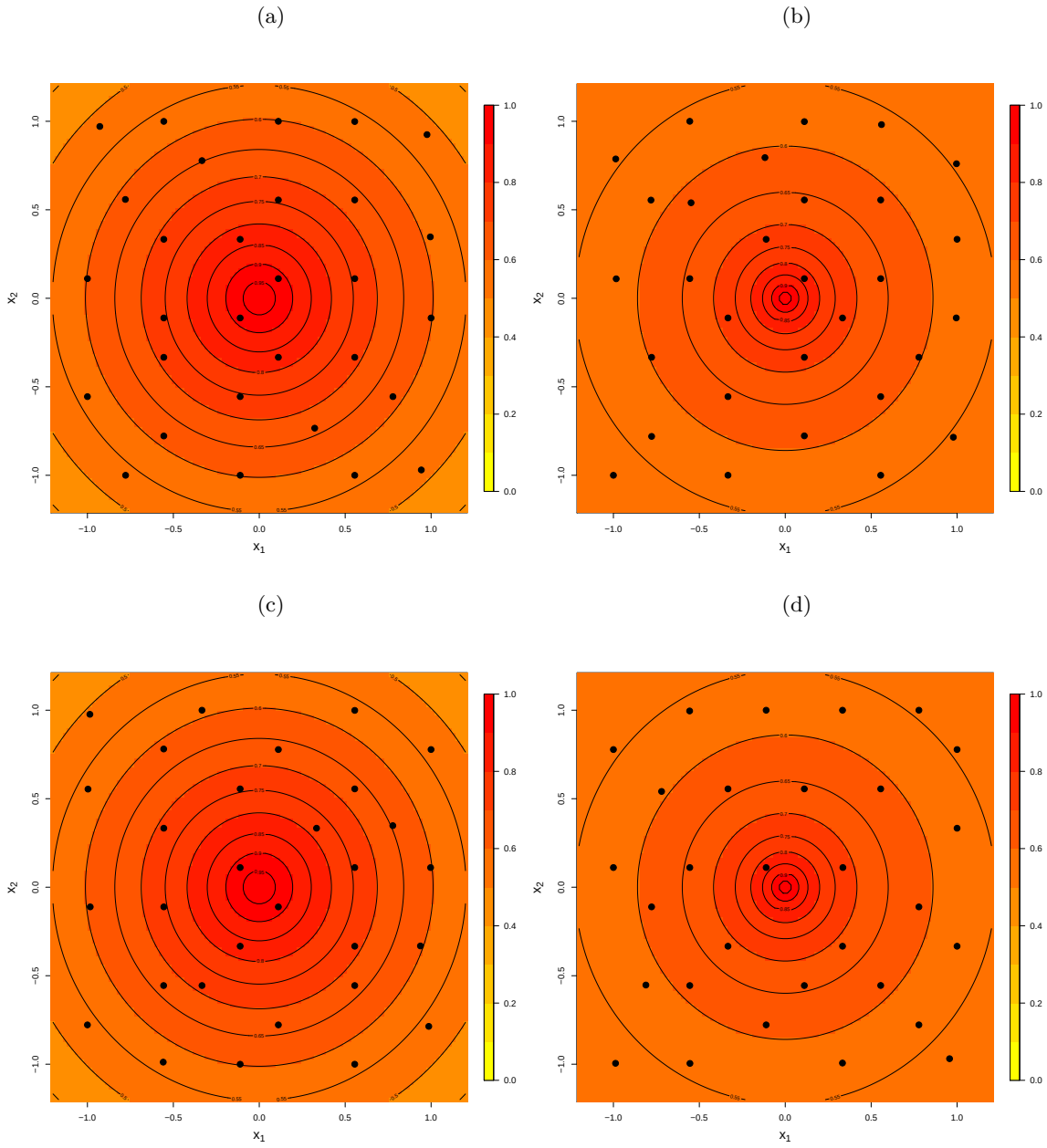


Figure A.5: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), linear mean ($F_2 = 1$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

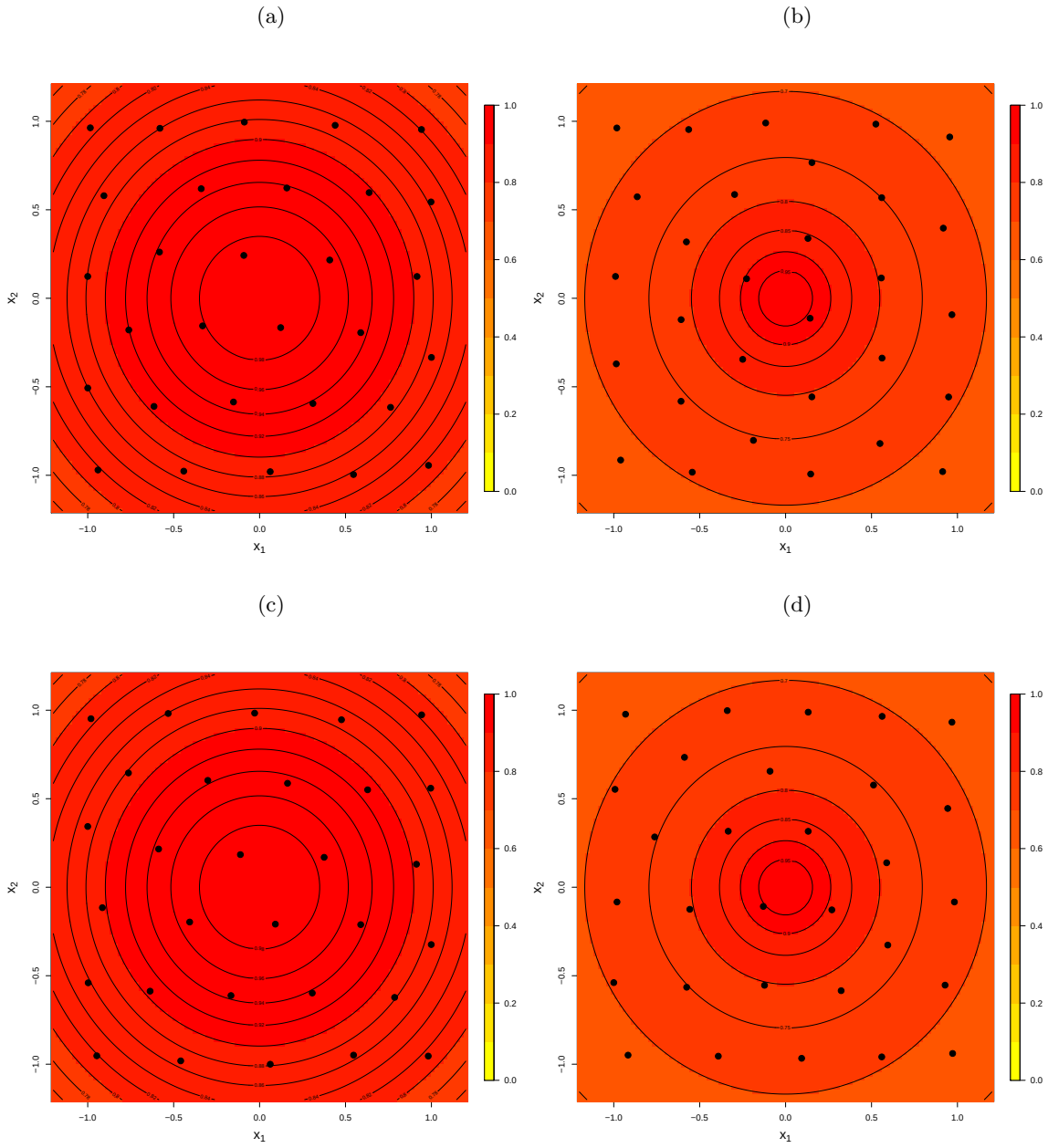


Figure A.6: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), linear mean ($F_2 = 1$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 0$ ($F_4 = 0$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

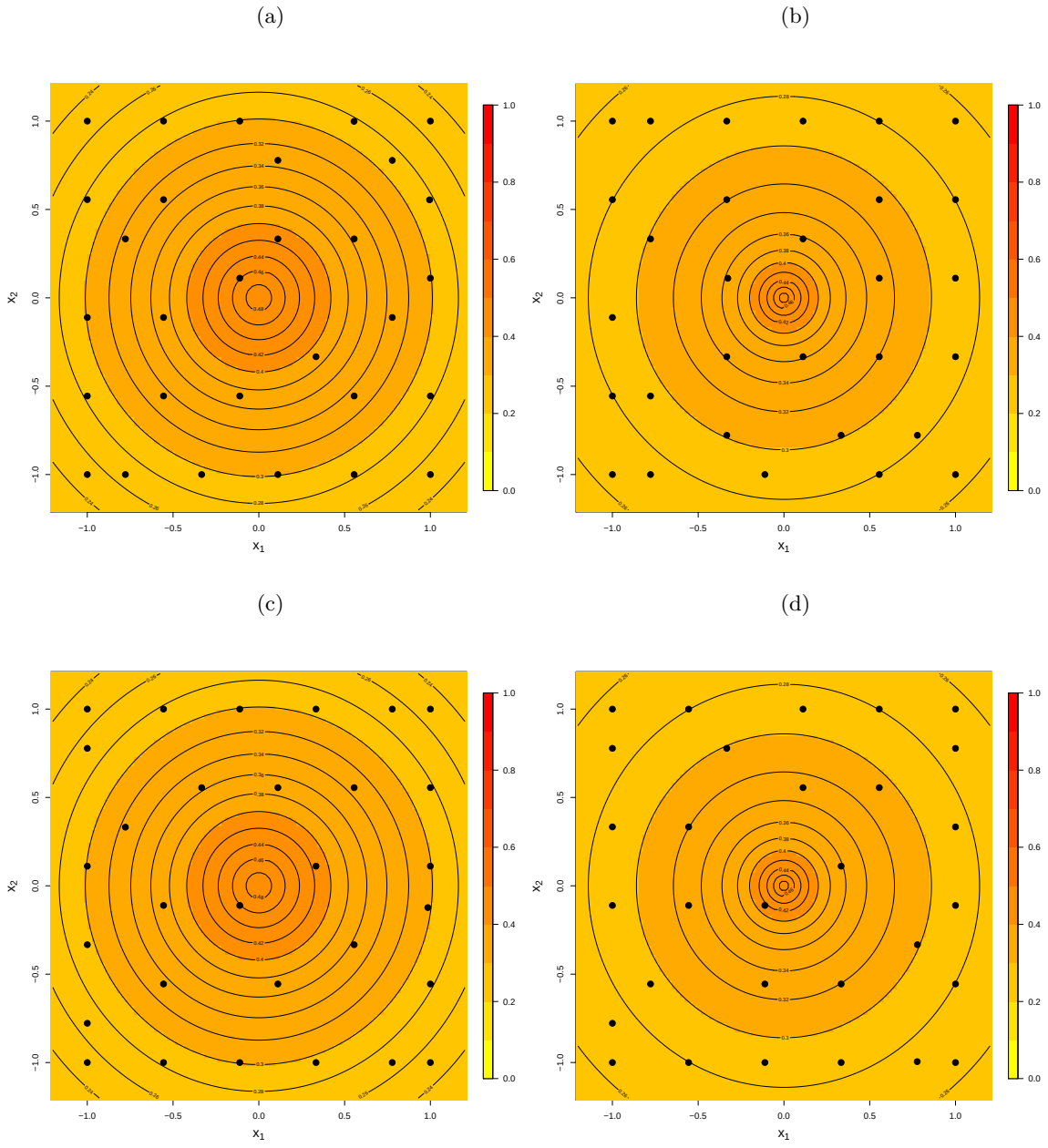


Figure A.7: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), linear mean ($F_2 = 1$), $\nu = 0.5$ ($F_3 = 0$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ.

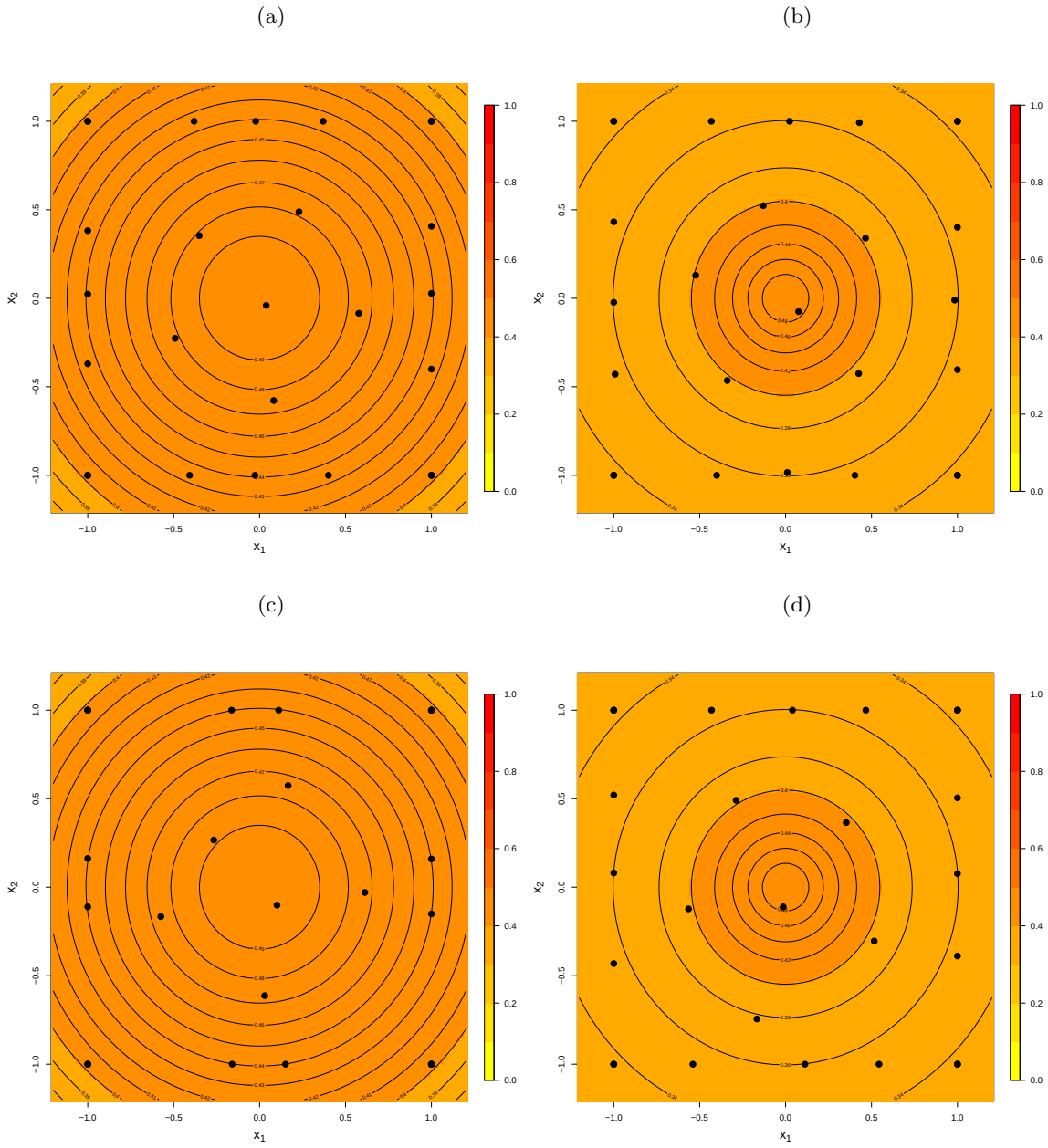


Figure A.8: Ψ-optimal designs for $n = 30$ runs ($F_1 = 1$), linear mean ($F_2 = 1$), $\nu = 1.5$ ($F_3 = 1$) and $\delta^2 = 1$ ($F_4 = 1$) (a) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 0$), (b) $\mathbf{R^{-1}} = 0.25\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 0$) (c) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a uniform prior distribution ($F_5 = 0, F_6 = 1$) and (d) $\mathbf{R^{-1}} = 4\mathbf{I}_3$ and ϕ having a log-normal prior distribution ($F_5 = 1, F_6 = 1$). Contours display correlation between each point in $\mathcal{X}$ and the centre of the design region, averaged across the prior for ϕ. In plot (a) ten points are repeated, (b) eight points are repeated, (c) twelve points are repeated and (d) eight points are repeated.

$\xi_{i_1}^{j_1k_1}$ $i_1 = 1, \cdots 16$	Settings for factors $F_4F_3F_5F_6$ $\xi_{i_2}^{j_2k_2}, i_2 = 1, \cdots 16$																IQR
	0000	0010	0001	0011	0100	0110	0101	0111	1000	1010	1001	1011	1100	1110	1101	1111	
0000	1	1.01	1.00	1.00	0.92	0.95	0.92	0.95	1.02	1.02	1.01	1.02	0.85	0.93	0.85	0.95	0.08
0010	0.99	1	0.99	1.00	0.91	0.94	0.91	0.93	1.02	1.02	1.00	1.02	0.84	0.92	0.85	0.94	0.08
0001	1.00	1.01	1	1.00	0.92	0.95	0.92	0.95	1.02	1.02	1.01	1.02	0.85	0.93	0.85	0.95	0.07
0011	1.00	1.00	1.00	1	0.91	0.94	0.91	0.94	1.02	1.02	1.00	1.02	0.85	0.93	0.85	0.94	0.08
0100	0.84	0.86	0.84	0.81	1	0.98	1.00	0.97	0.87	0.84	0.82	0.86	0.68	0.88	0.66	0.87	0.07
0110	0.92	0.94	0.92	0.90	0.99	1	0.99	0.99	0.92	0.92	0.90	0.92	0.78	0.92	0.76	0.91	0.04
0101	0.84	0.86	0.84	0.81	1.00	0.98	1	0.97	0.87	0.84	0.82	0.86	0.69	0.88	0.66	0.87	0.07
0111	0.92	0.95	0.92	0.91	1.00	1.01	1.00	1	0.93	0.93	0.91	0.93	0.79	0.92	0.77	0.92	0.04
1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.001
1010	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.99	1.00	0.99	1.00	0.002
1001	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.99	1.00	0.001
1011	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	0.99	1.00	0.99	1.00	0.002
1100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.001
1110	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.001
1101	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	0.001
1111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	0.001
IQR	0.02	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.03	0.02	0.17	0.07	0.17	0.07	

Table A.1: Relative efficiencies for $F_1 = 1, j_1, j_2 = 1$ and $F_2 = 0, k_1, k_2 = 0$ together with interquartile range (IQR).

$\xi_{i_1}^{j_1k_1}$ $i_1 = 1, \cdots 16$	Settings for factors $F_4F_3F_5F_6$ $\xi_{i_2}^{j_2k_2}, i_2 = 1, \cdots 16$																IQR
	0000	0010	0001	0011	0100	0110	0101	0111	1000	1010	1001	1011	1100	1110	1101	1111	
0000	1	1.00	0.99	1.00	0.92	0.94	0.92	0.95	1.01	1.01	1.00	1.00	0.72	0.73	0.61	0.74	0.12
0010	1.00	1	1.00	1.01	0.92	0.94	0.91	0.95	1.02	1.01	1.00	1.00	0.73	0.74	0.63	0.75	0.13
0001	1.01	1.00	1	1.01	0.93	0.95	0.93	0.96	1.02	1.02	1.01	1.01	0.73	0.74	0.62	0.75	0.12
0011	1.00	0.99	0.99	1	0.92	0.93	0.91	0.94	1.01	1.01	1.00	1.00	0.73	0.73	0.63	0.75	0.13
0100	0.83	0.76	0.78	0.82	1	0.97	1.00	0.97	0.82	0.82	0.79	0.79	0	0	0	0	0.30
0110	0.86	0.78	0.80	0.84	1.03	1	1.03	1.00	0.85	0.85	0.82	0.81	0	0	0	0	0.31
0101	0.84	0.76	0.78	0.82	1.00	0.97	1	0.97	0.82	0.83	0.79	0.79	0	0	0	0	0.30
0111	0.86	0.78	0.80	0.84	1.03	1.00	1.03	1	0.84	0.85	0.81	0.81	0	0	0	0	0.31
1000	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.99	0.99	0.98	0.99	0.005
1010	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.99	0.99	0.98	0.99	0.005
1001	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1	1.00	0.99	0.99	0.98	0.99	0.01
1011	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1	0.99	0.99	0.98	0.99	0.004
1100	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	0.01
1110	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	0.01
1101	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1	1.00	0.01
1111	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1	0.01
IQR	0.04	0.06	0.06	0.05	0.02	0.03	0.02	0.02	0.05	0.04	0.05	0.05	0.45	0.44	0.53	0.44	

Table A.2: Relative efficiencies for $F_1 = 1, j_1, j_2 = 1$ and $F_2 = 1, k_1, k_2 = 1$together with interquartile range (IQR).

A.3 Examples of Spatial Optimal Designs for $n = 10$

Constant mean function

Figures A.9–A.14 displays Ψ-optimal designs and plots (a)-(d) give the design and the correlation contours for $\delta^2 = 0, 0.5, 1, 2.5$ respectively.

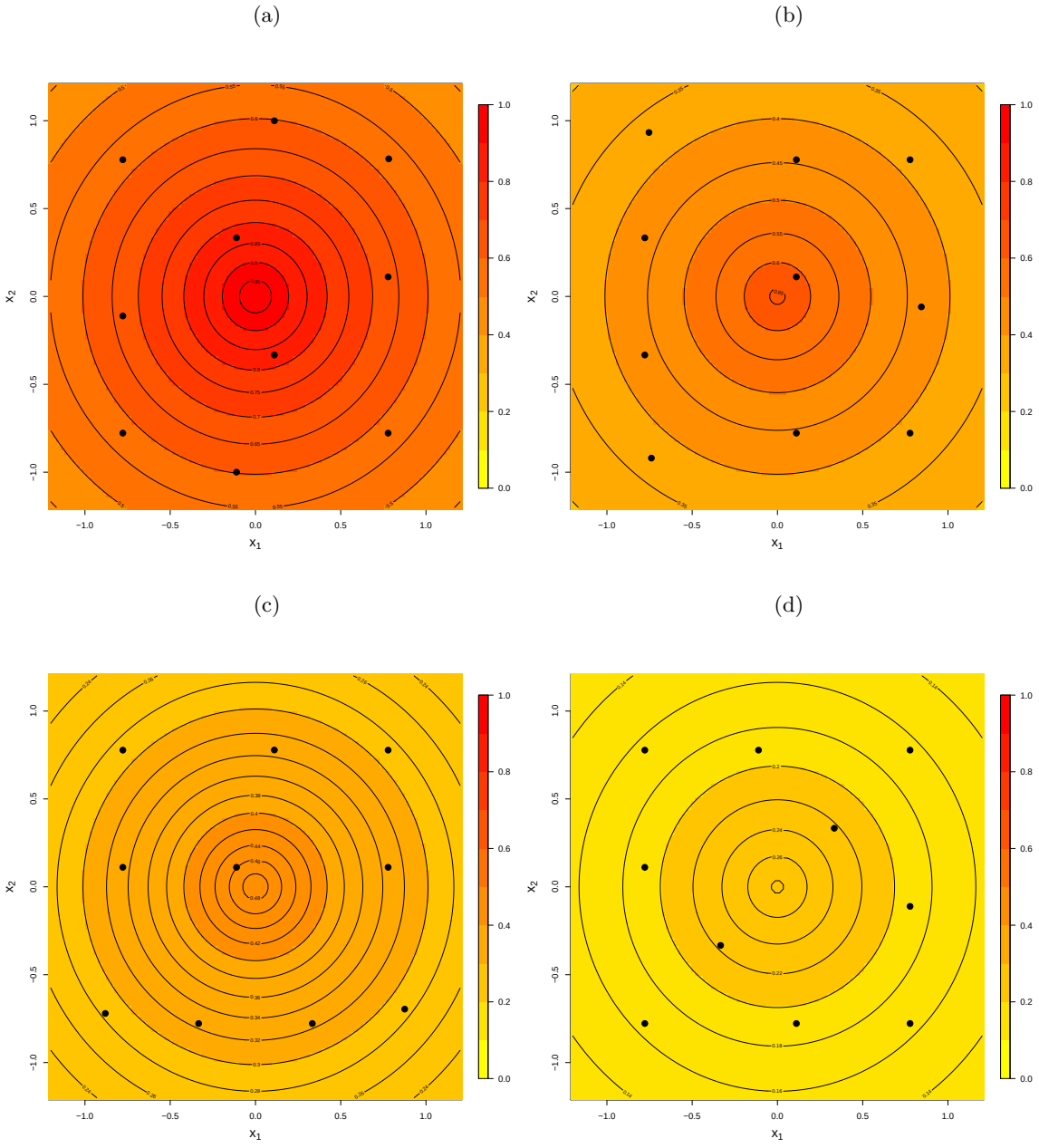


Figure A.9: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 0.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

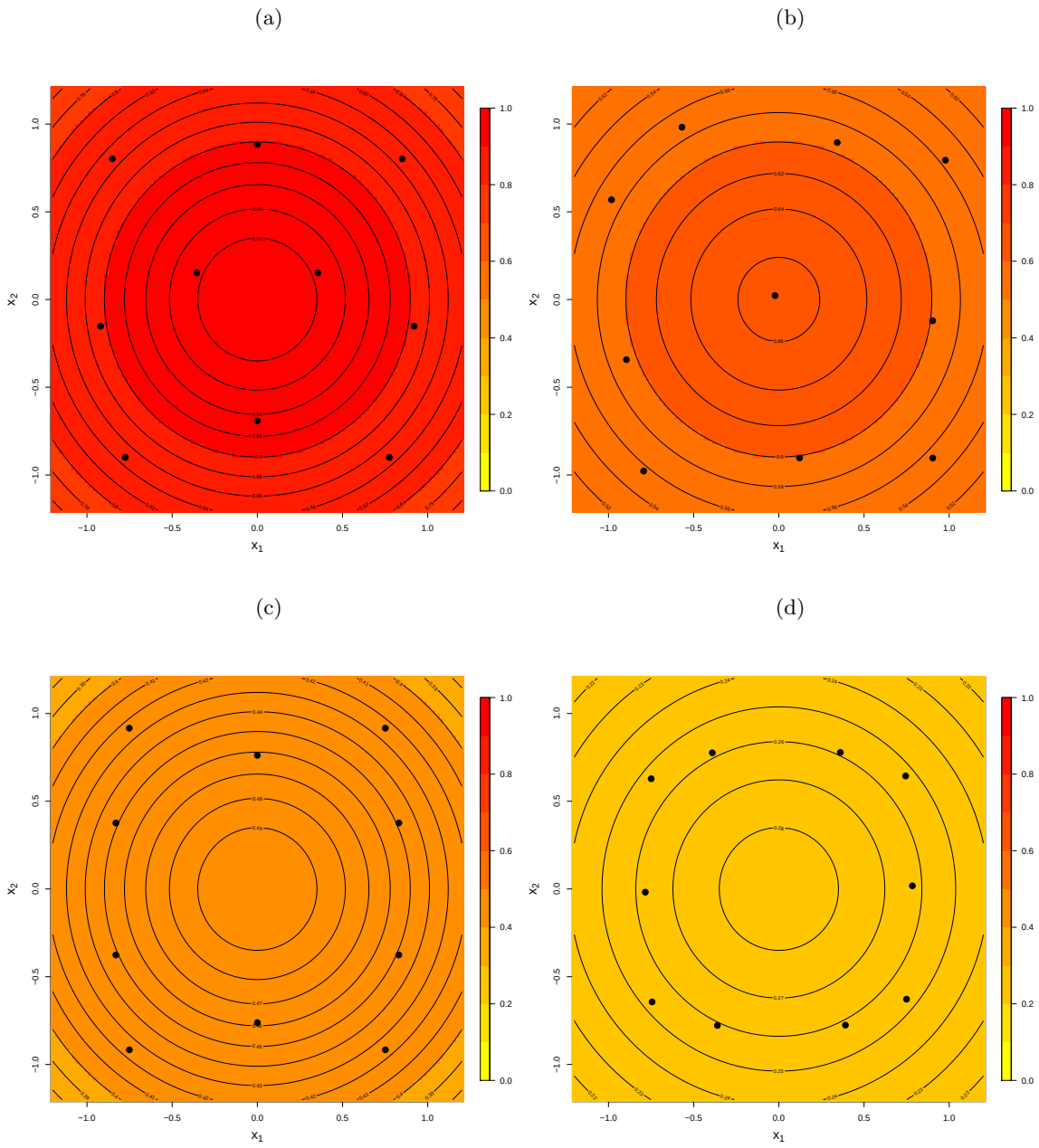


Figure A.10: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 1.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

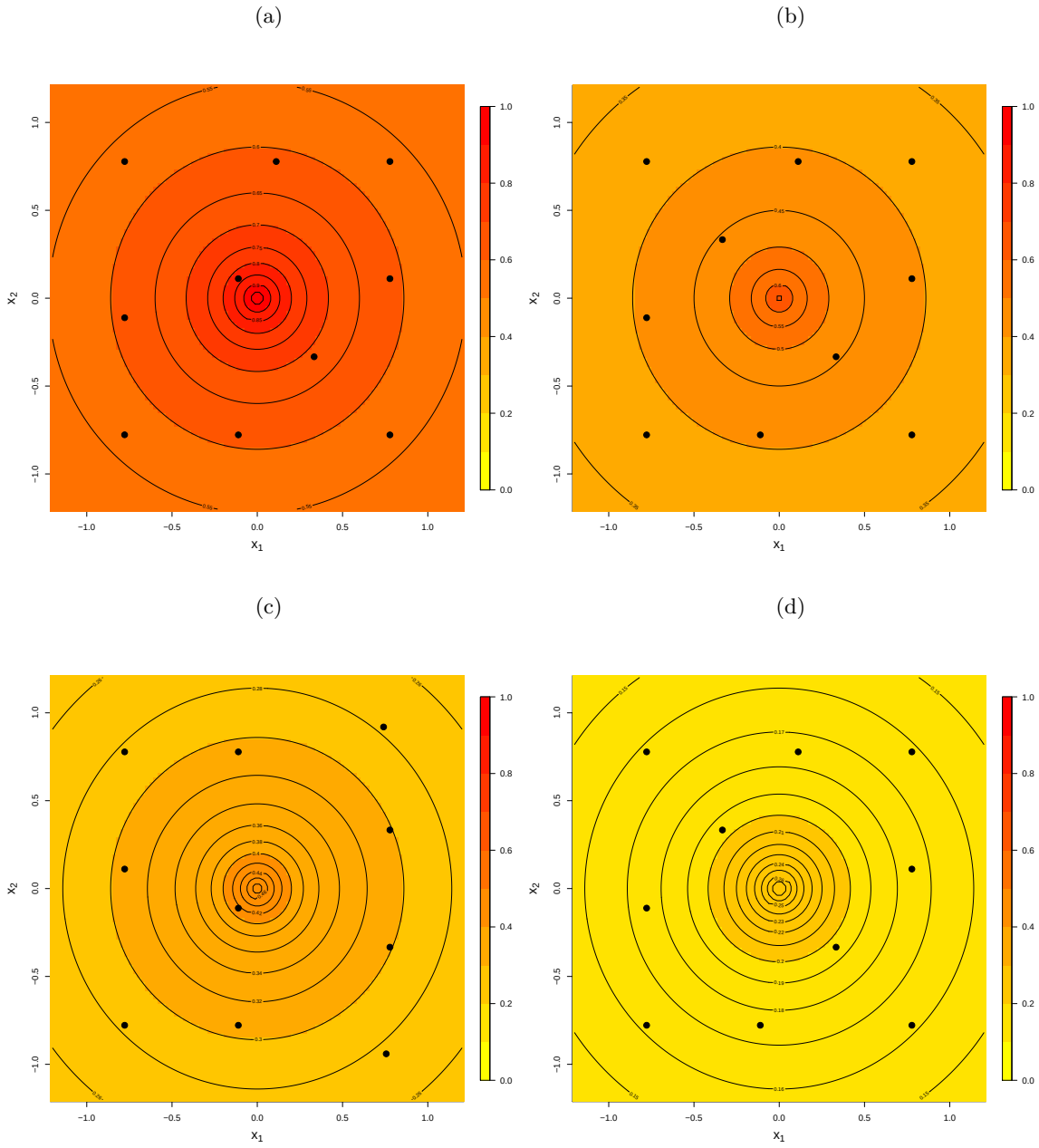


Figure A.11: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 0.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

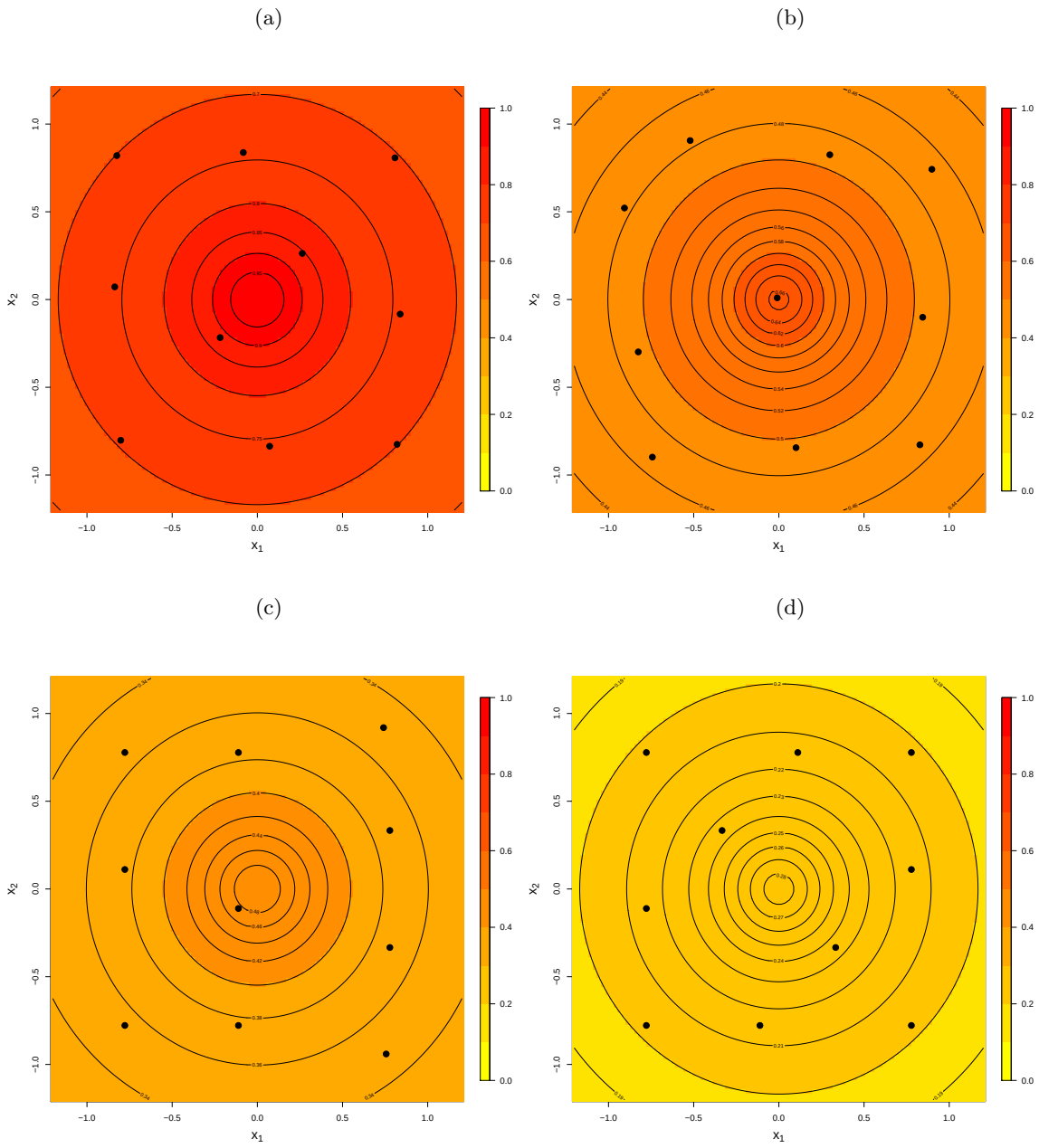


Figure A.12: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 1.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

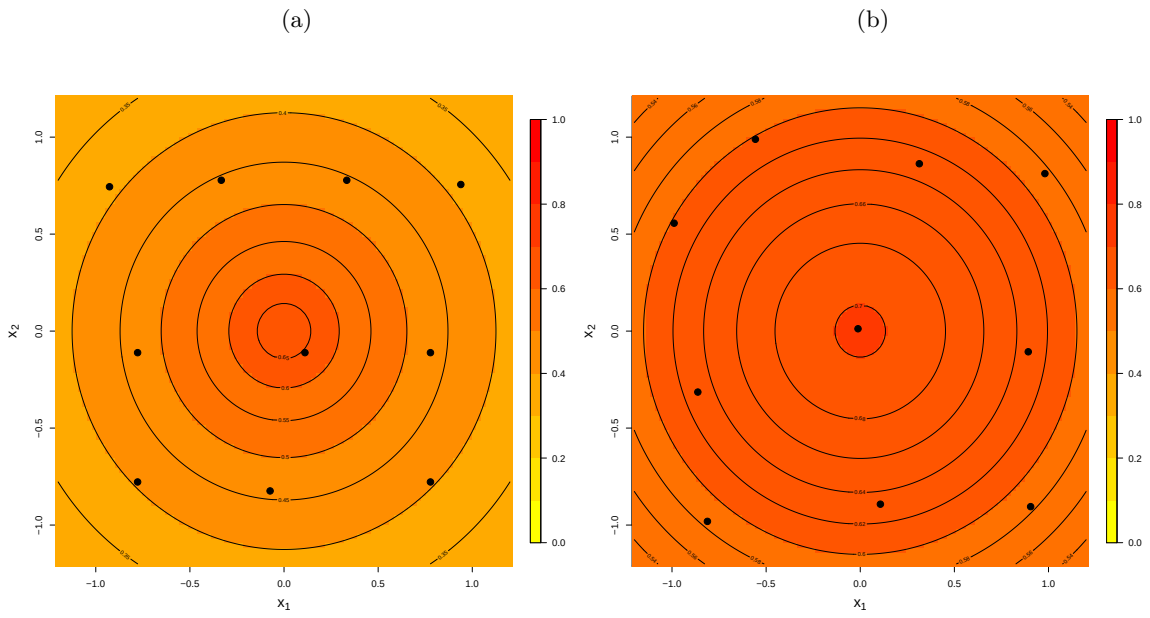


Figure A.13: Ψ-optimal designs for a constant mean function, uniform prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

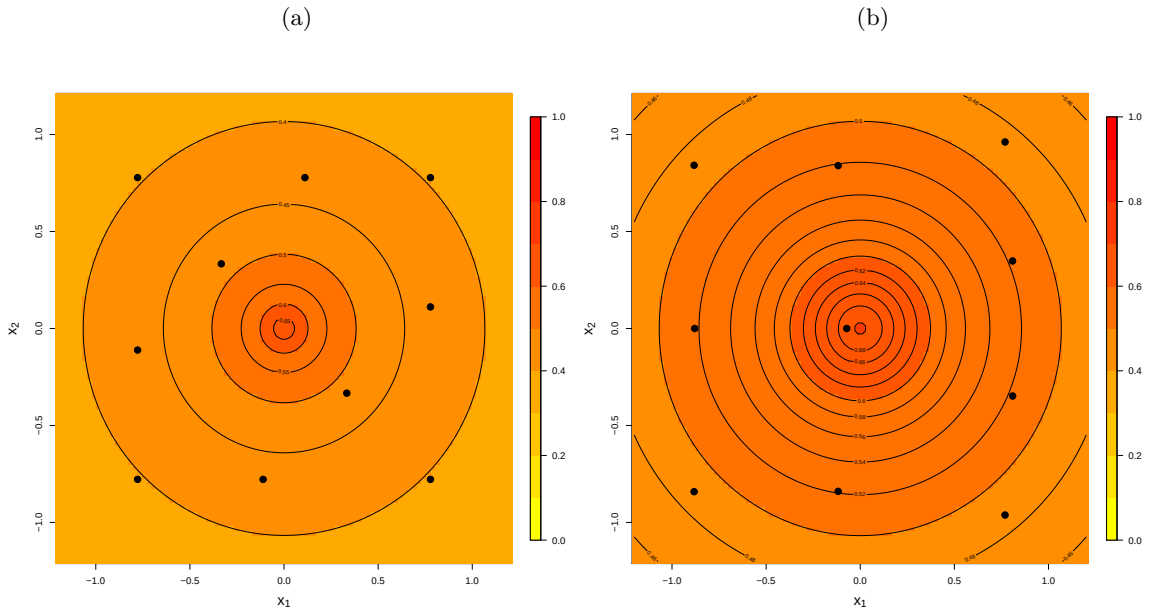


Figure A.14: Ψ-optimal designs for a constant mean function, log-normal prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

A.4 Examples of Spatial Optimal Designs for $n = 20$

Constant mean function

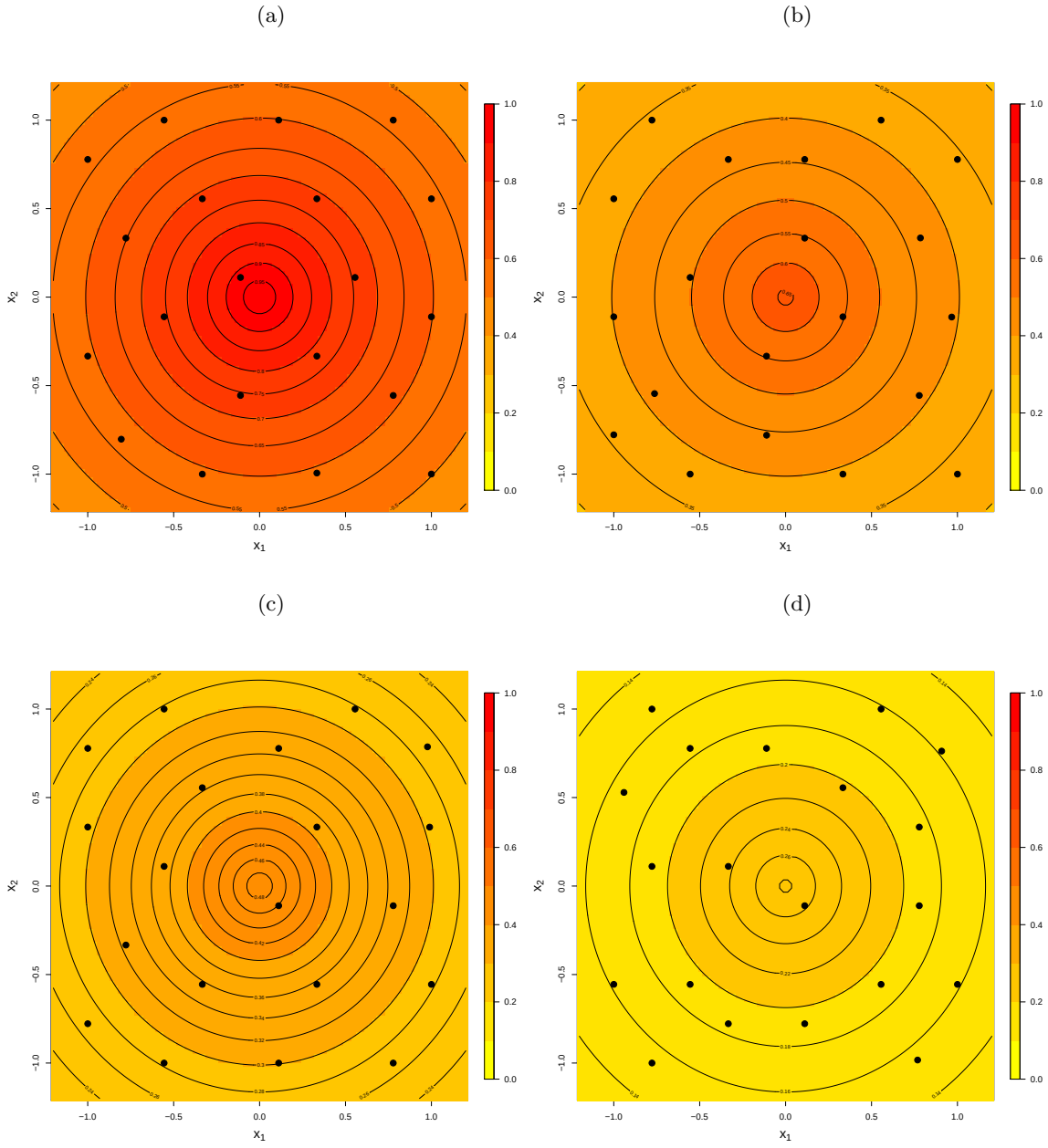


Figure A.15: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 0.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

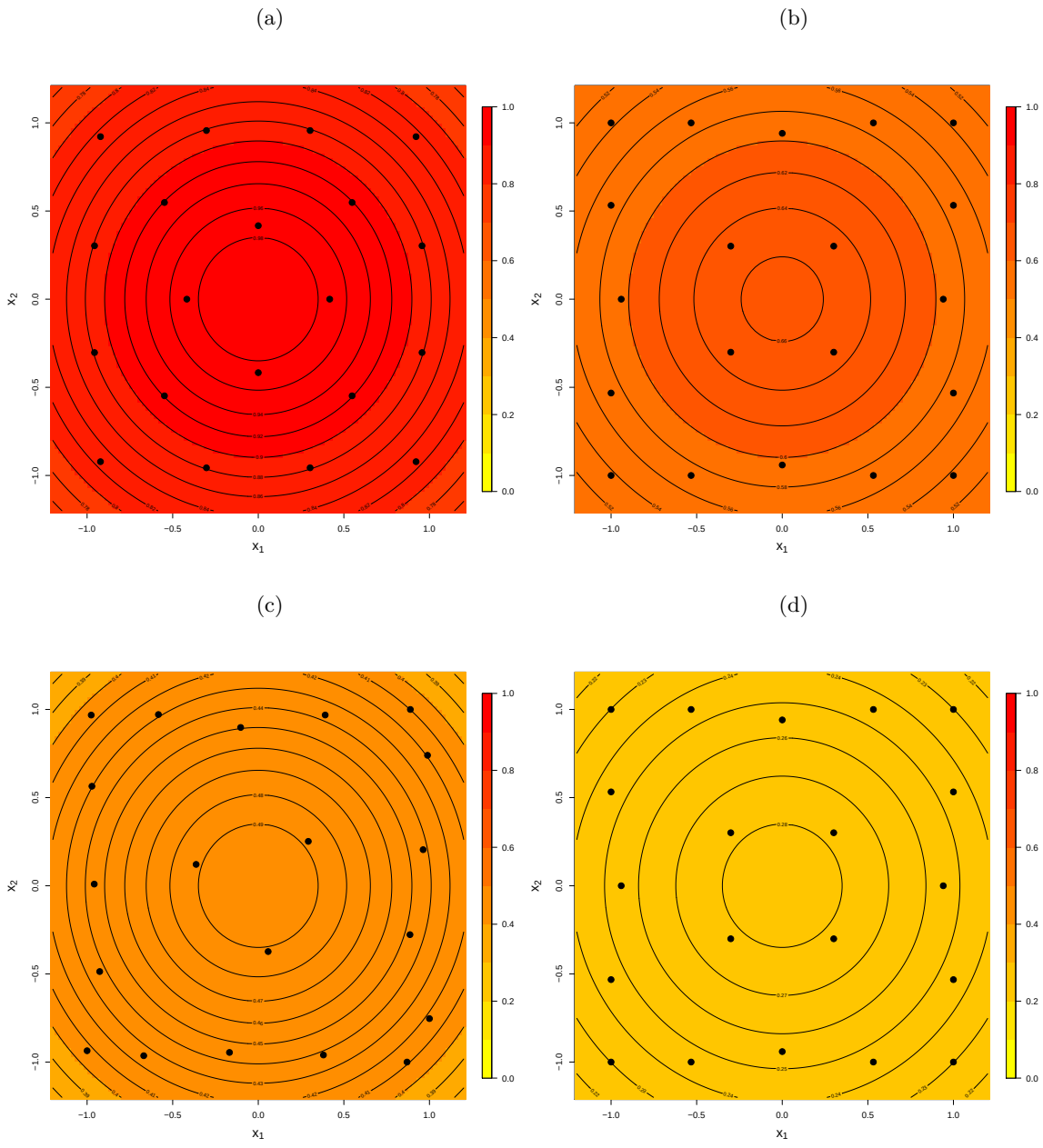


Figure A.16: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 1.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

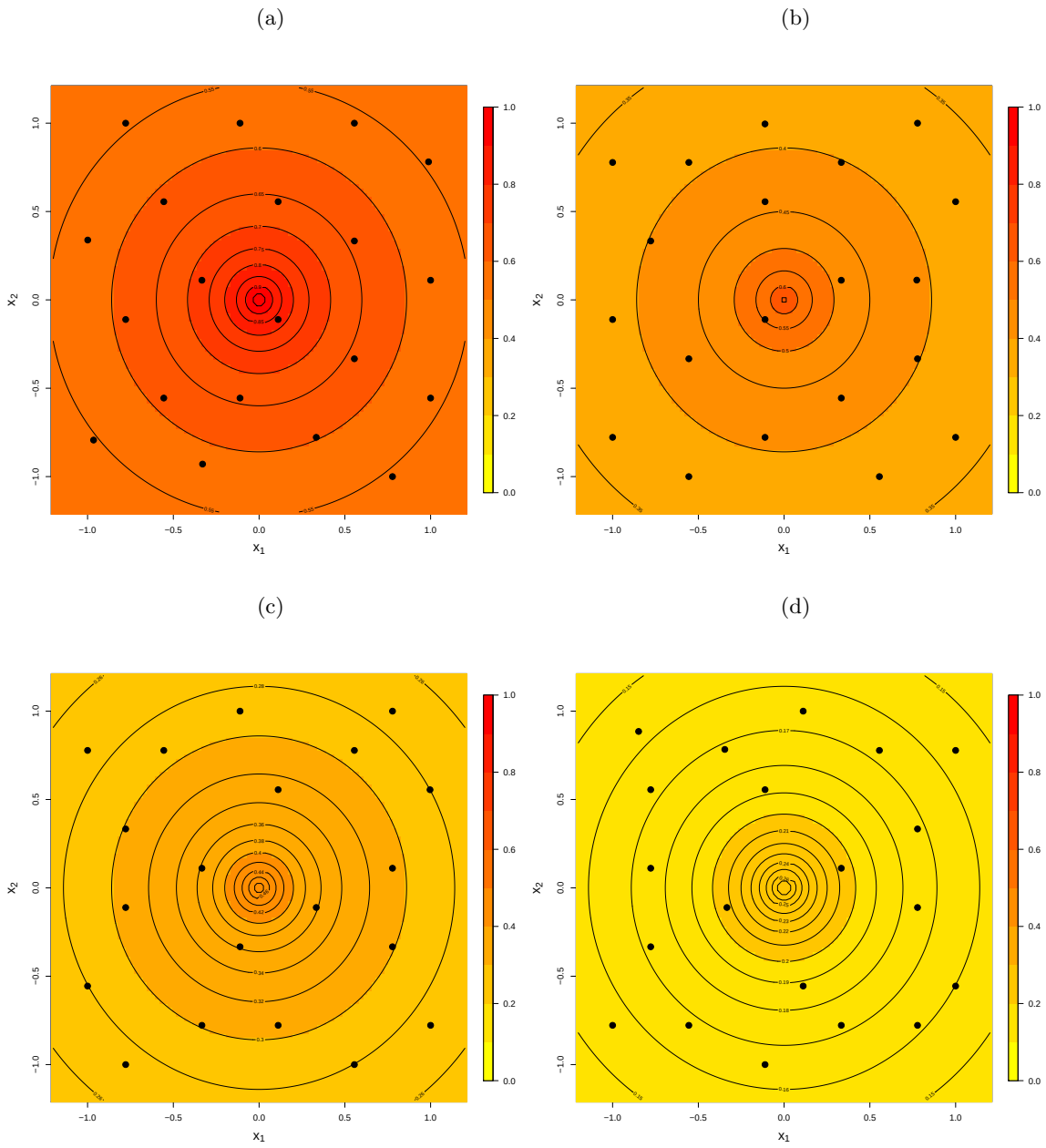


Figure A.17: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 0.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

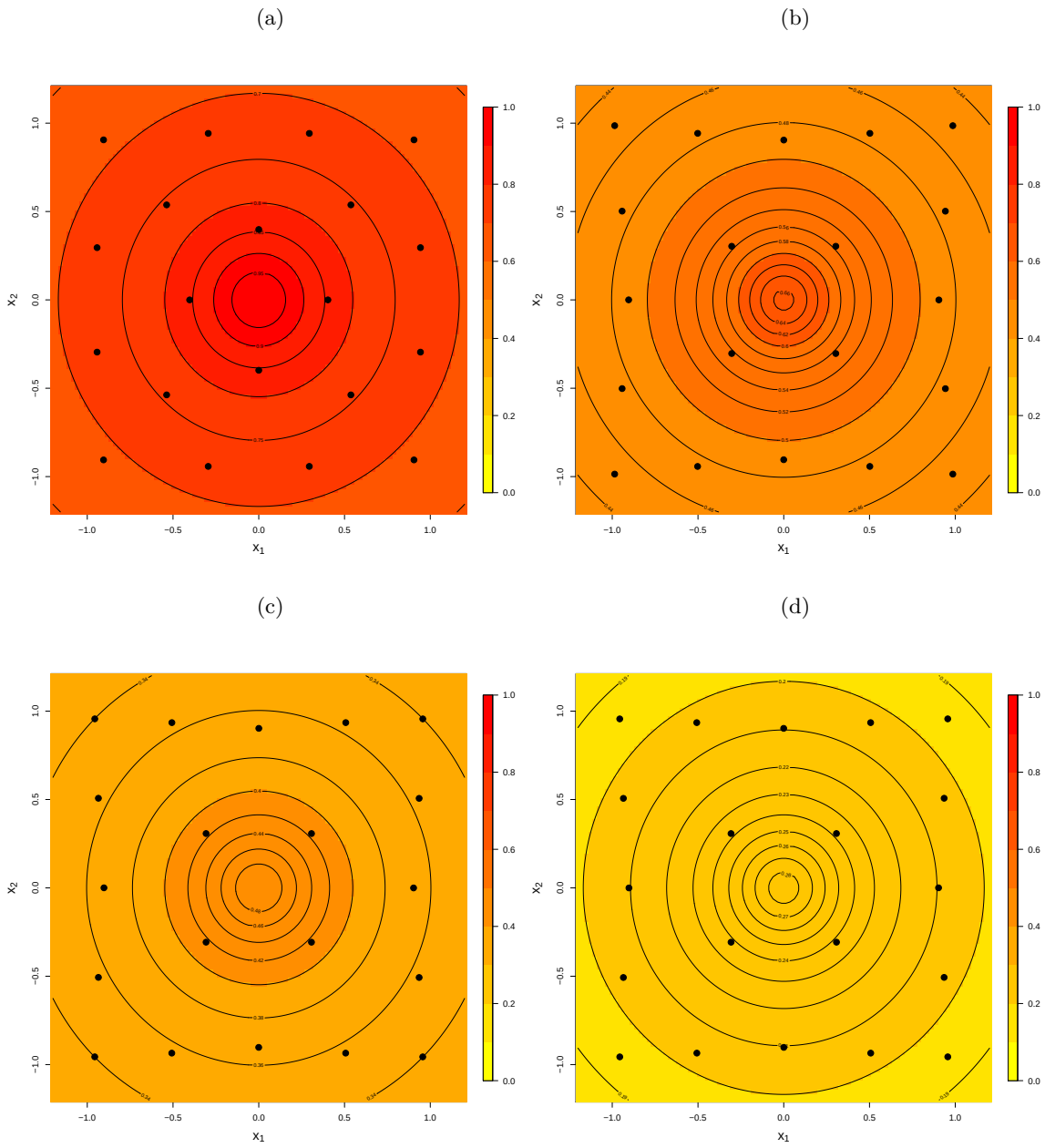


Figure A.18: Ψ-optimal designs for a constant mean function, Matérn correlation function with $\nu = 1.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

(a) (b)

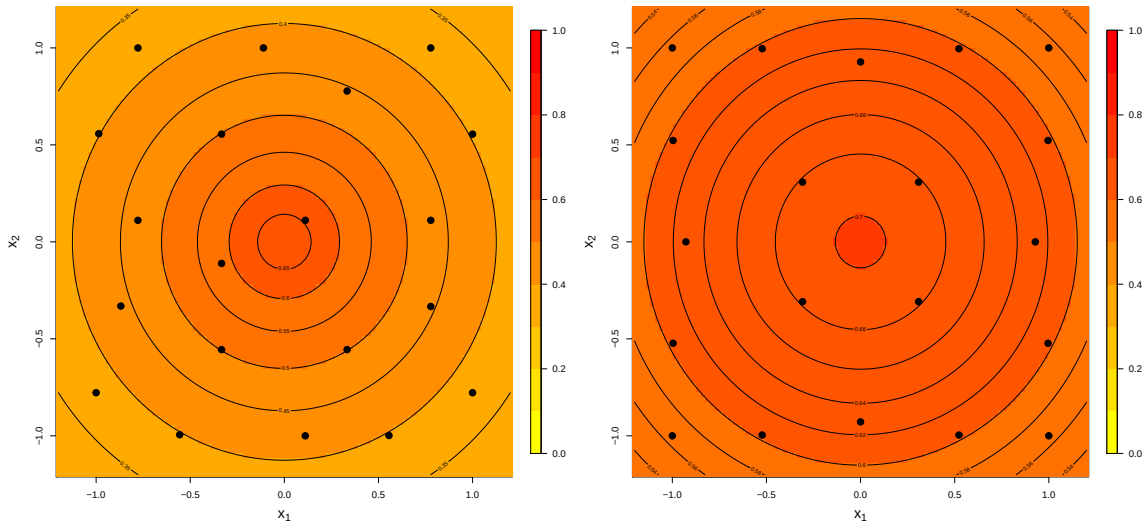


Figure A.19: Ψ-optimal designs for a constant mean function, uniform prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

(a) (b)

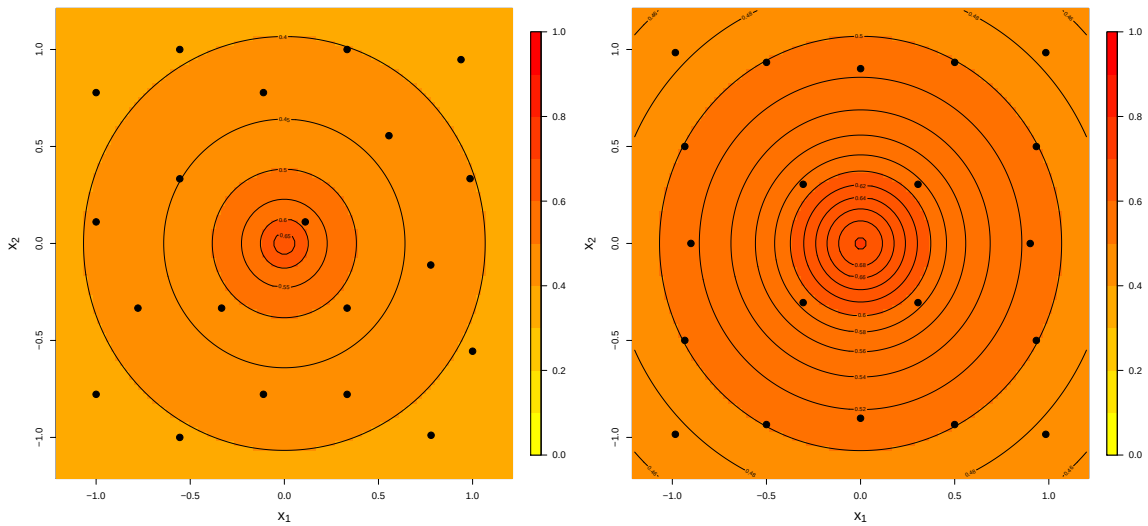


Figure A.20: Ψ-optimal designs for a constant mean function, log-normal prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$.

Linear mean function

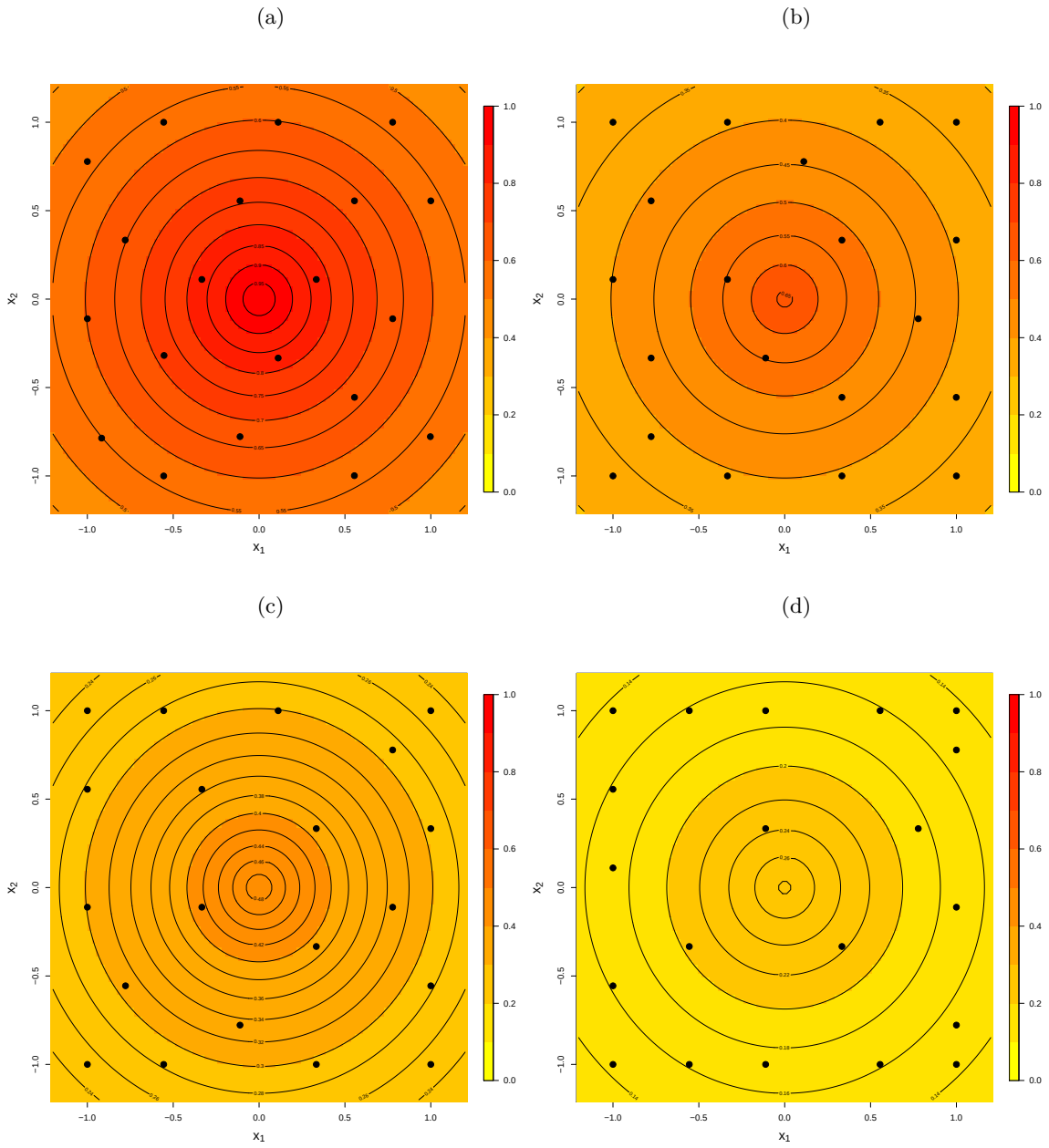


Figure A.21: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 0.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

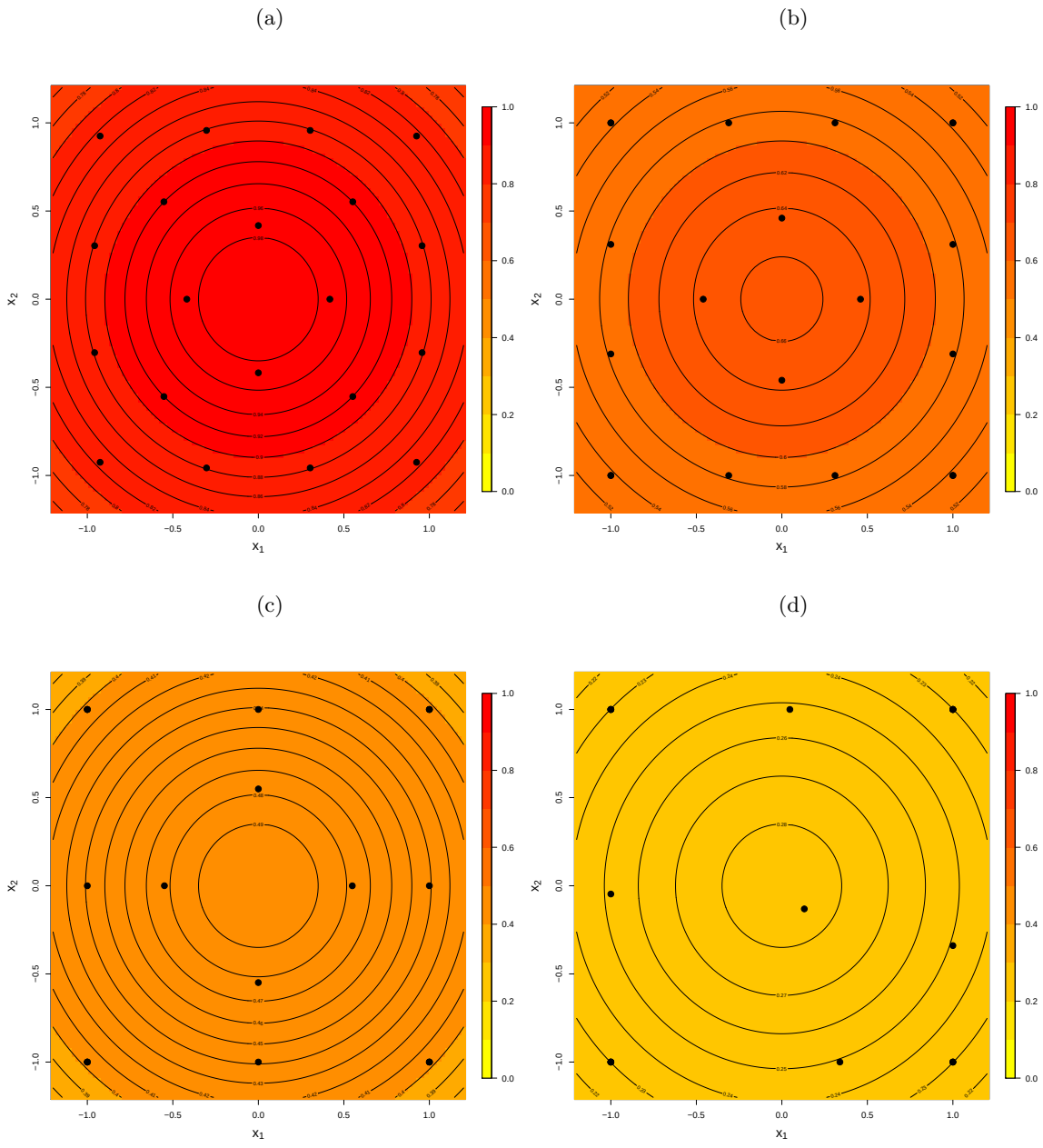


Figure A.22: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 1.5$, uniform prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ. In plot (c) eight points are repeated and plot (d) eleven points are repeated.

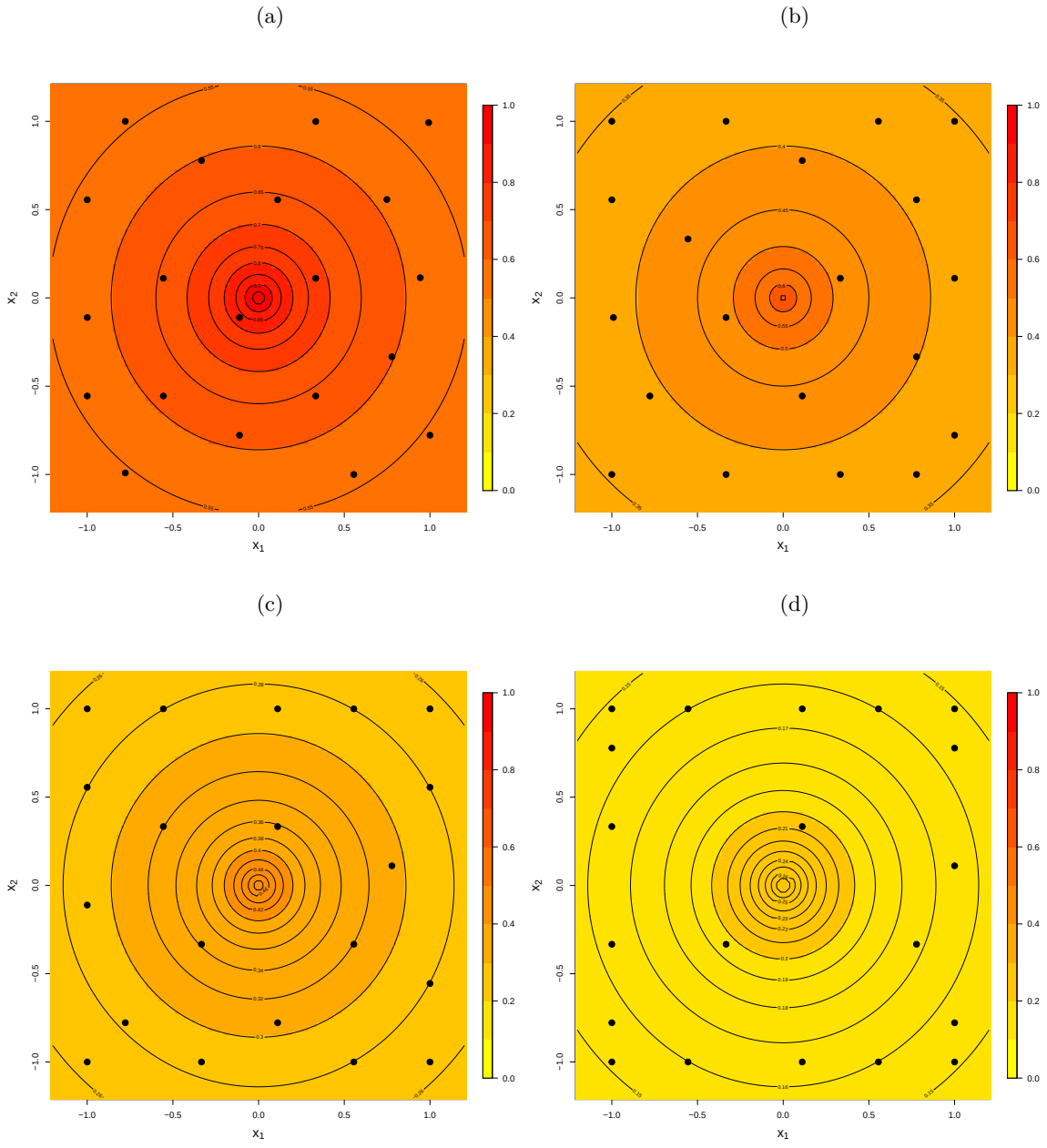


Figure A.23: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 0.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ.

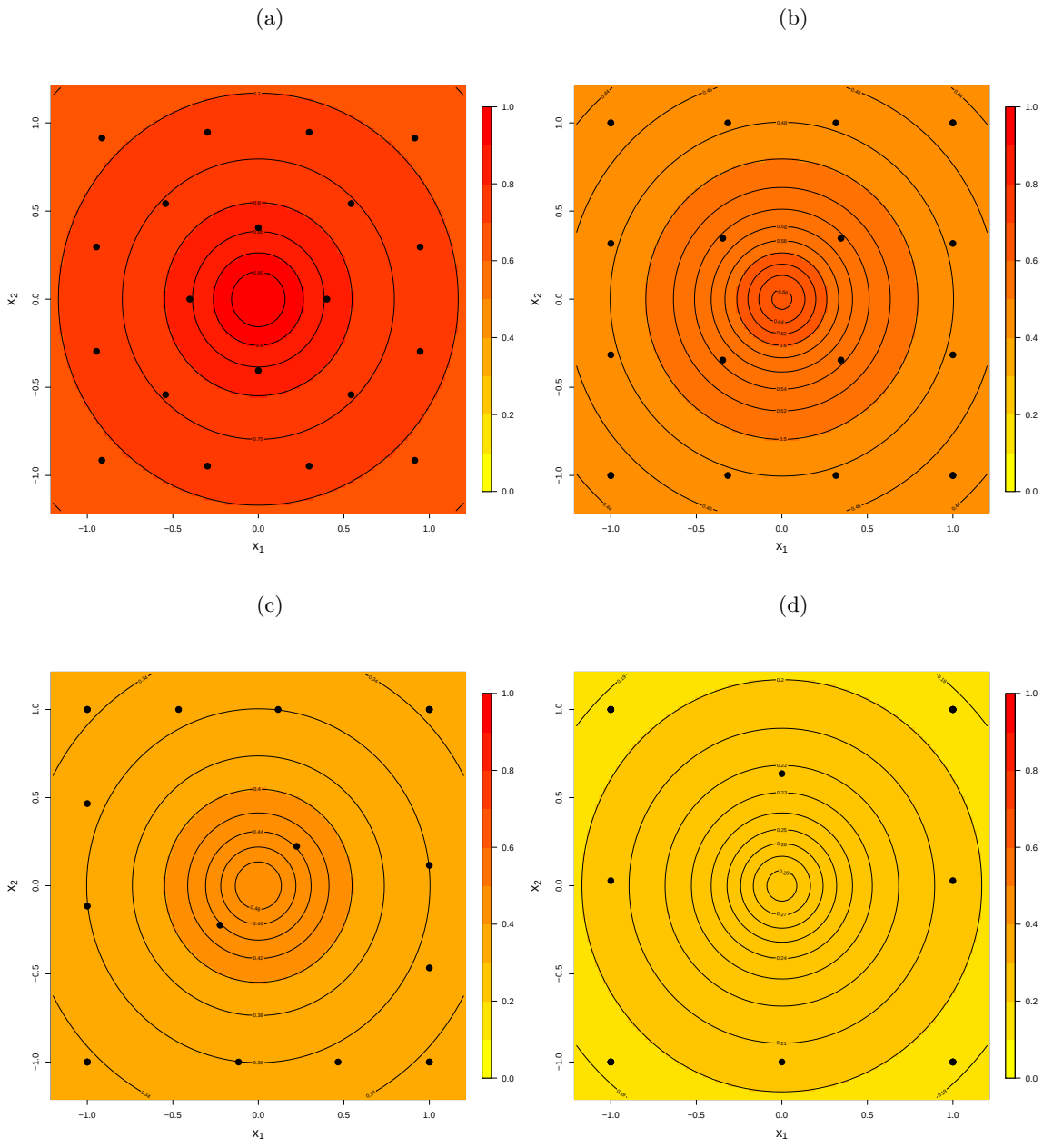


Figure A.24: Ψ-optimal designs for a linear mean function, Matérn correlation function with $\nu = 1.5$, log-normal prior distribution on ϕ and known δ^2: (a) $\delta^2 = 0$, (b) $\delta^2 = 0.5$ (c) $\delta^2 = 1$ and (d) $\delta^2 = 2.5$. Contours display the average correlation between each point and the centre of the design region, averaged across the prior distribution for ϕ. In plot (c) six points are repeated and plot (d) twelve points are repeated.

(a) (b)

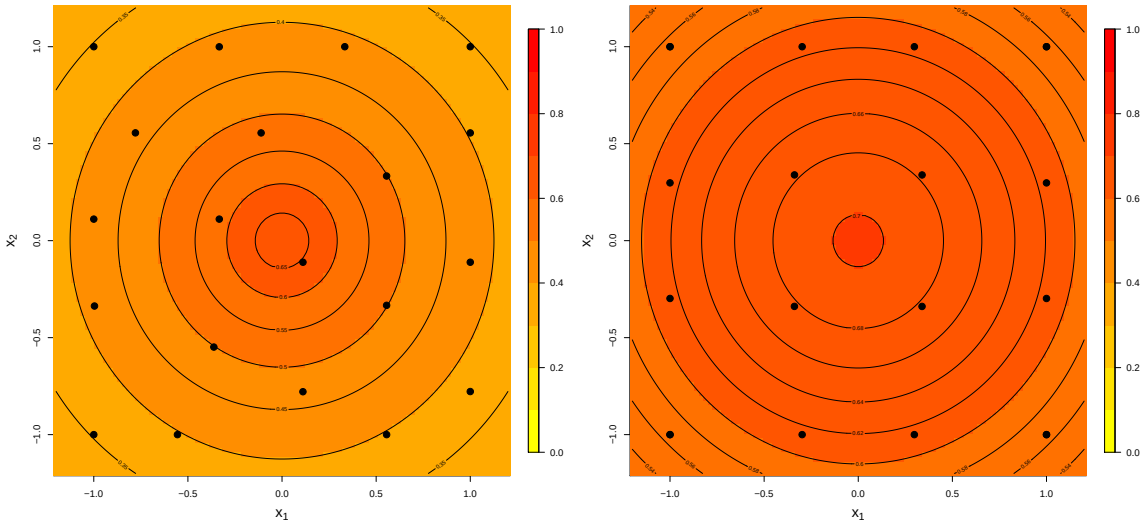


Figure A.25: Ψ-optimal designs for a linear mean function, uniform prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$. In plot (b) four points are repeated.

(a) (b)

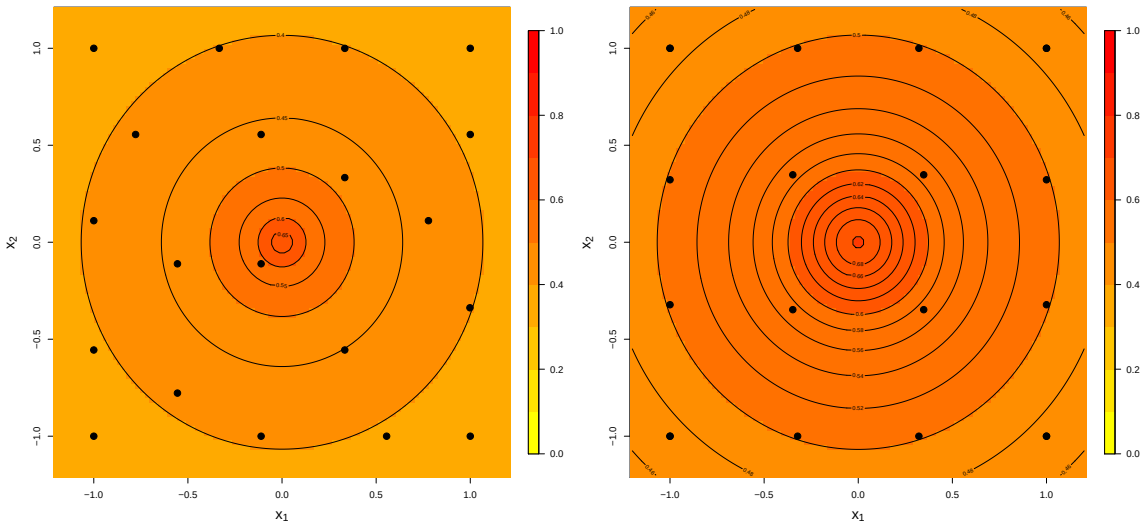


Figure A.26: Ψ-optimal designs for a linear mean function, log-normal prior distribution on ϕ and uniform prior distribution on δ^2 with Matérn correlation function (a) $\nu = 0.5$ and (b) $\nu = 1.5$. In plot (b) four points are repeated.

A.5 Examples of Designs for Computer Experiments $d = 3$ and $n = 5$

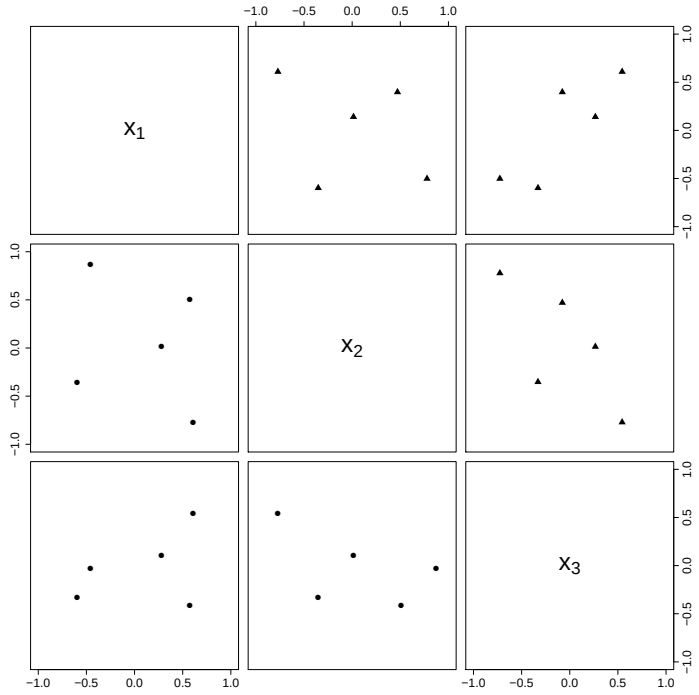


Figure A.27: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 000000 (●) and 000100 (▲).

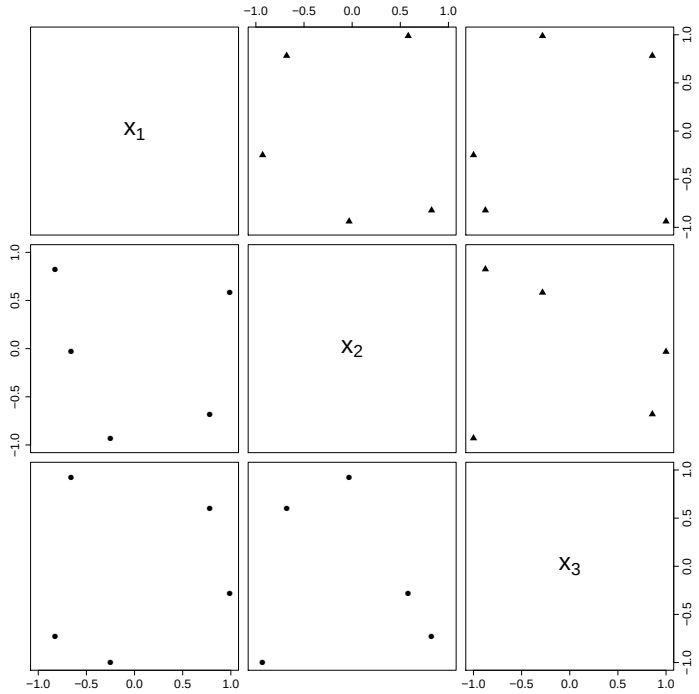


Figure A.28: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 010000 (●) and 010100 (▲).

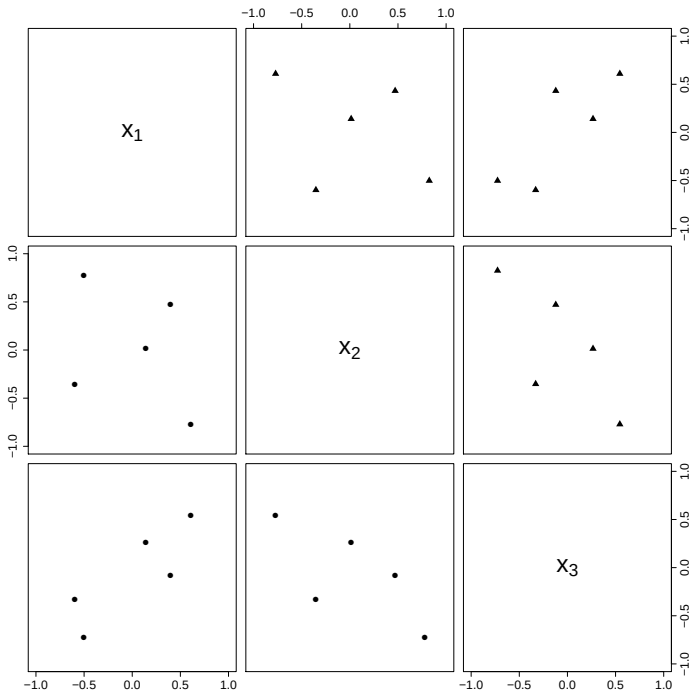


Figure A.29: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 000001 (●) and 000101 (▲).

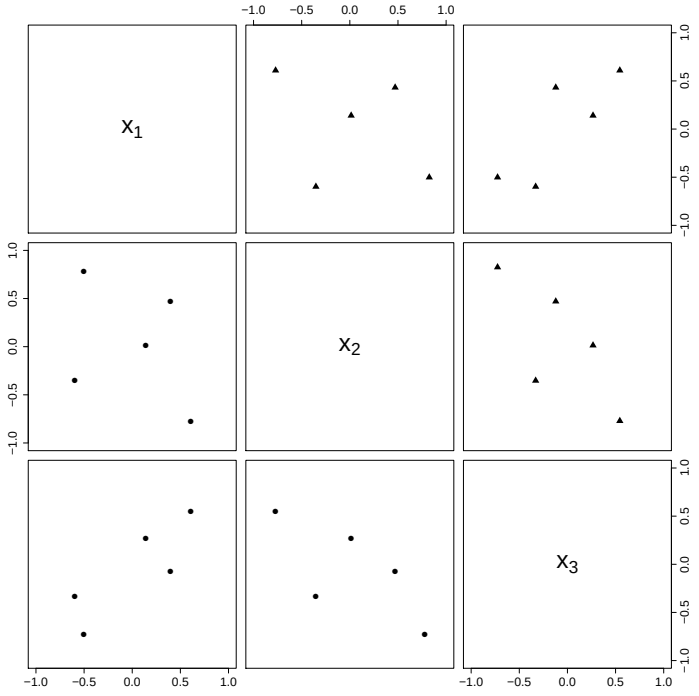


Figure A.30: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 000001 (●) and 000101 (▲).

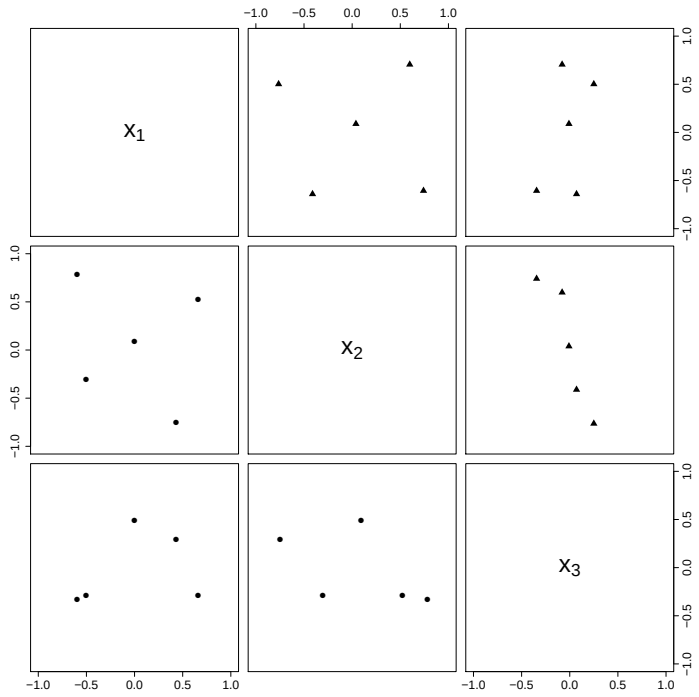


Figure A.31: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 001010 (●) and 001110 (▲).

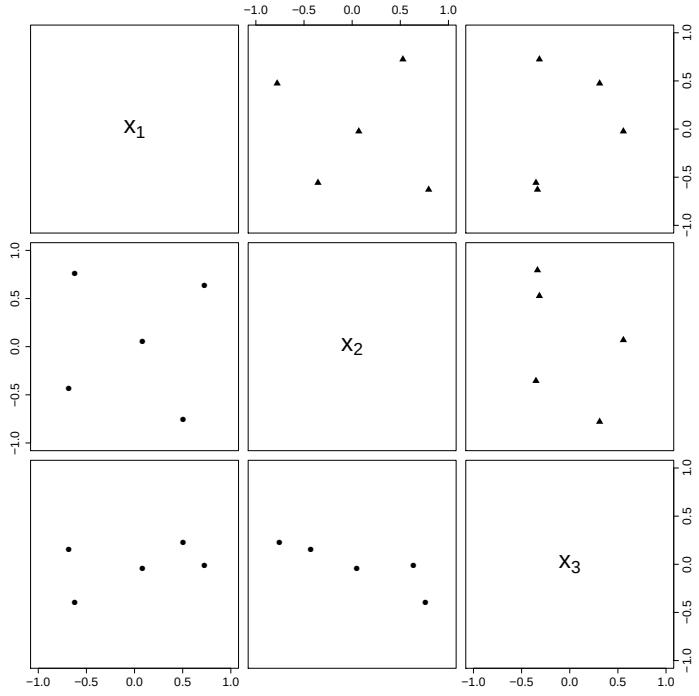


Figure A.32: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 001011 (●) and 001111 (▲).

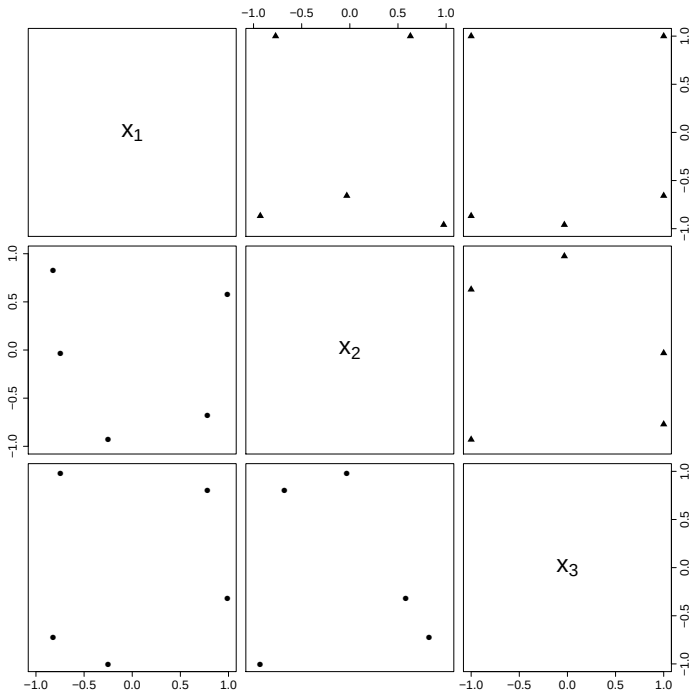


Figure A.33: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 010011 (●) and 010111 (▲).

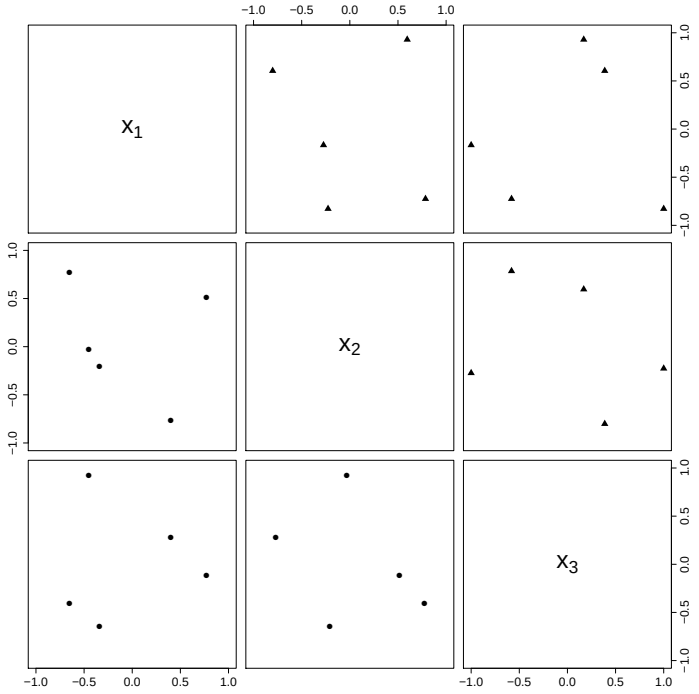


Figure A.34: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 011010 (●) and 011110 (▲).

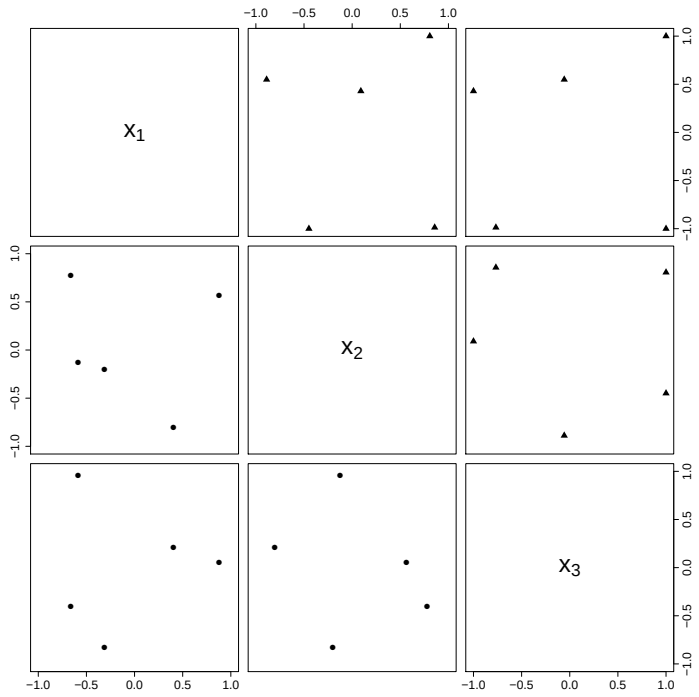


Figure A.35: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 011011 (●) and 011111 (▲).

A.6 Examples of Designs for Computer Experiments $d = 3$ and $n = 10$

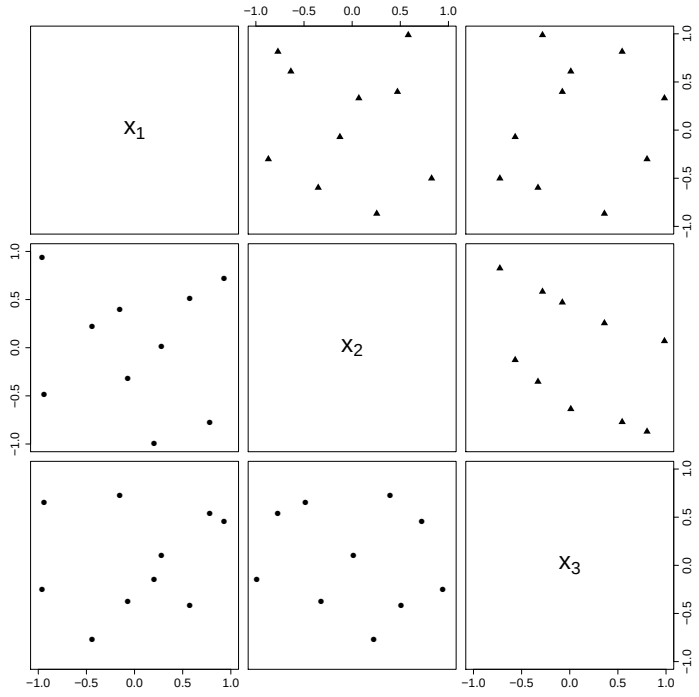


Figure A.36: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 100001 (●) and 100101 (▲).

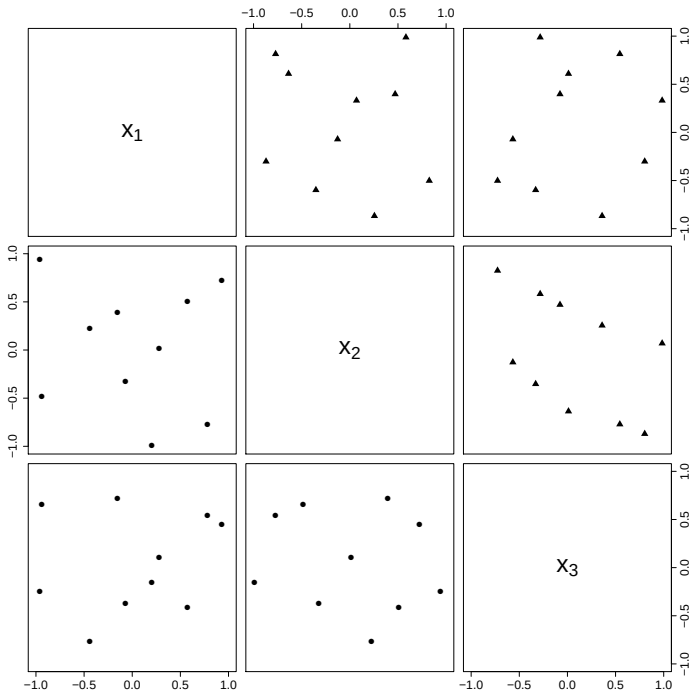


Figure A.37: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 100001 (●) and 100101 (▲).

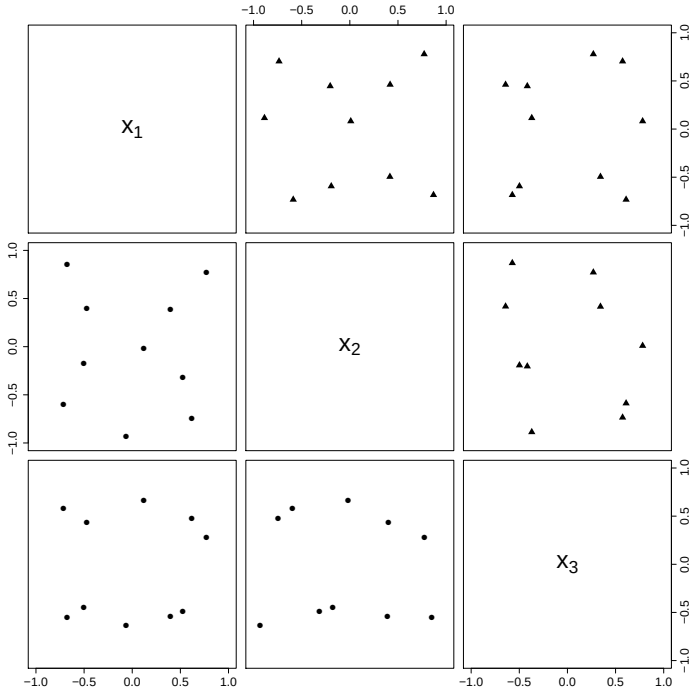


Figure A.38: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 101010 (●) and 101110 (▲).

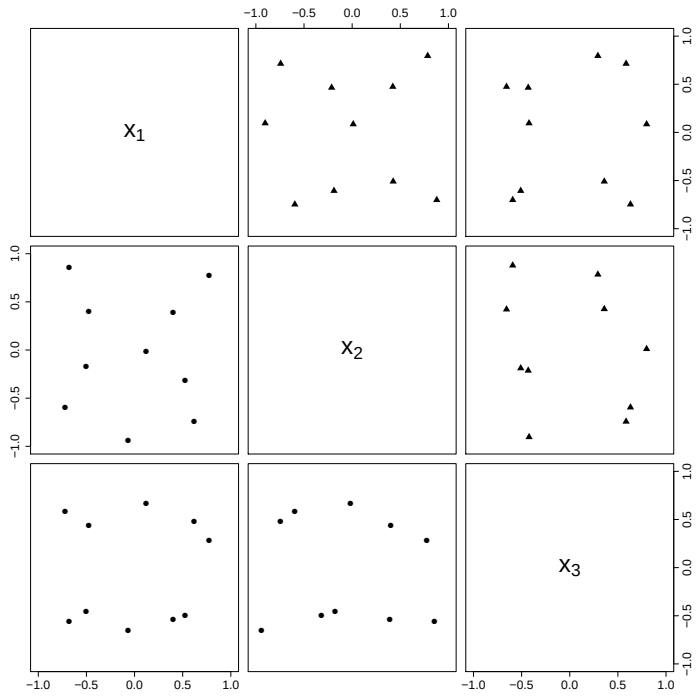


Figure A.39: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 101011 (●) and 101111 (▲).

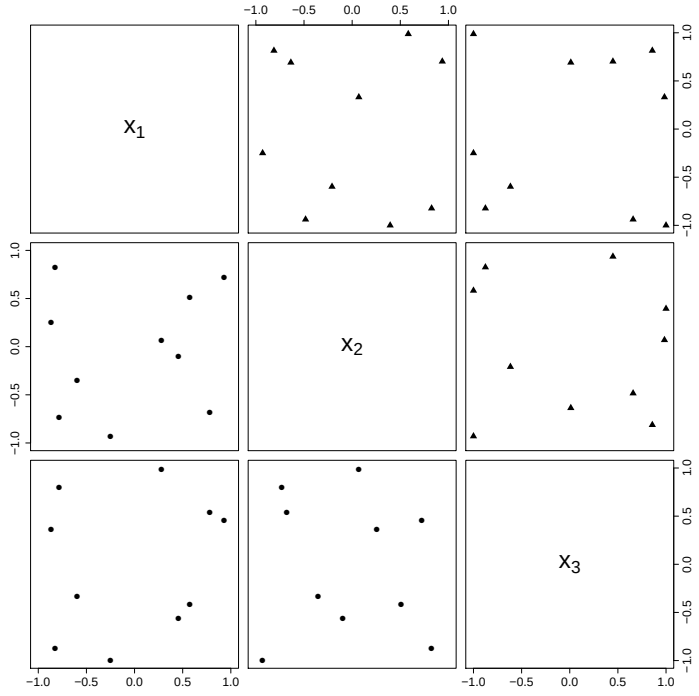


Figure A.40: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 110011 (●) and 110111 (▲).

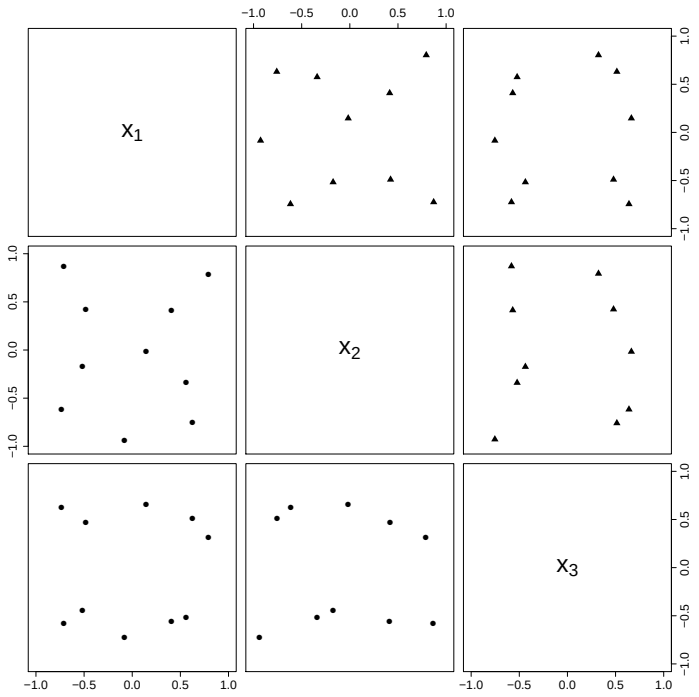


Figure A.41: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 111010 (●) and 111110 (▲).

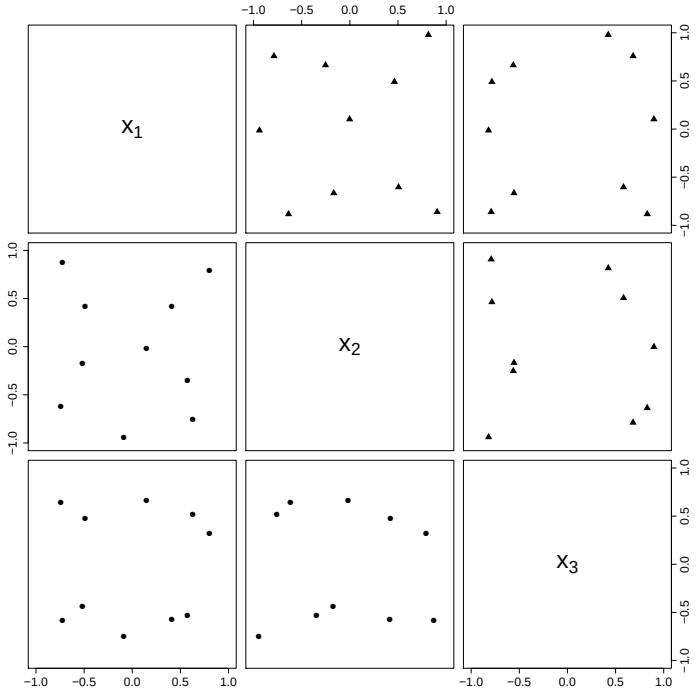


Figure A.42: Two dimensional projections of Ψ-optimal designs for $d = 3$ and for 111011 (●) and 111111 (▲).

A.7 Spatio-temporal Designs

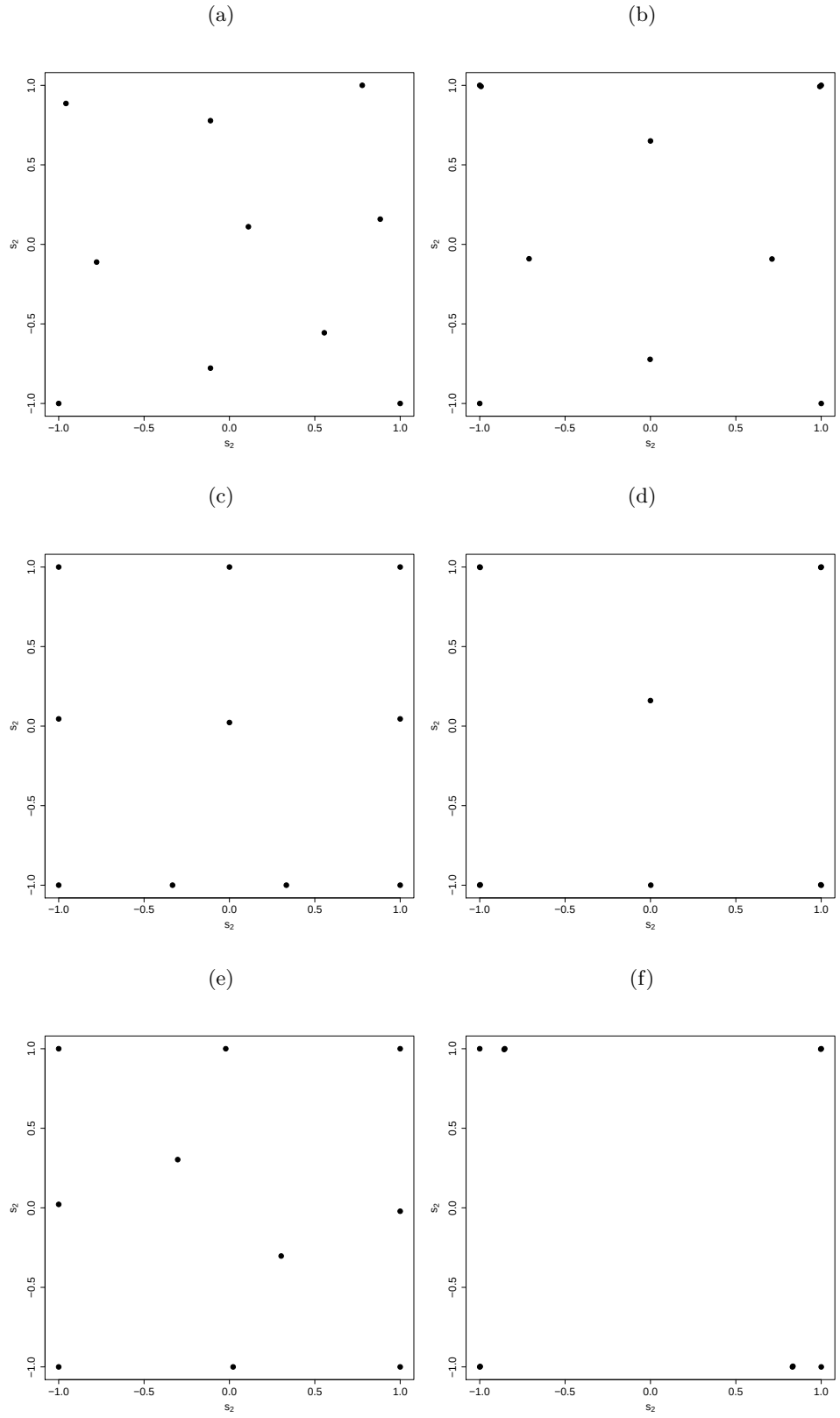


Figure A.43: Spatial Ψ-optimal designs for fixed times (set *Times 2*) (a) $\nu = 0.5$ and $\phi_2 = 0.01$; (b) $\nu = 1.5$ and $\phi_2 = 0.01$; (c) $\nu = 0.5$ and $\phi_2 = 0.5$; (d) $\nu = 1.5$ and $\phi_2 = 0.5$; (e) $\nu = 0.5$ and $\phi_2 = 10$; (f) $\nu = 1.5$ and $\phi_2 = 10$. In plots (d) and (f) four points are repeated.