Robust Designs for Model Discrimination and Prediction of a Threshold Probability

by

Rui Hu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

In the first of the two projects comprising this thesis we consider the construction of experimental designs aimed at the elucidation of a functional relationship between a response variable and various covariates, under various forms of model uncertainty. In such a scenario a central goal may be to design – i.e., to choose those covariates in a sample space at which to observe the response – in order to distinguish between rival models. Specifically we assume that two rival models are available. Usually, one can cast the model discrimination problem as one of hypothesis testing. If the parameters are known (there are adjustments possible otherwise) the Neyman-Pearson test rejects the null hypothesis for large values of test statistic.

The classical design of experiments for discriminating between two rival models are based on the classical assumption that the true model exactly coincides with one of the two rival models. From a point view of robustness it is more realistic to suppose that the correct model lies in one of those only approximately known classes centred around those rival models.

Under the assumption that the true model is in a Hellinger neighbourhood of one of the nominal models, we propose methods of construction for experimental designs by maximizing the worst power of the Neyman-Pearson test over the Hellinger neighbourhoods. The asymptotic properties of the Neyman-Pearson test statistic is derived. The optimal designs are "maximin" designs, which maximize (through the design) the minimum (among the neighbourhoods) asymptotic power function.

To motivate the second project, we note that stochastic processes are widely used in the study of phenomena nowadays. In particular, the question of whether a stochastic process is larger than a critical level is of ecological interest.

Usually, a model used to describe the observed data of the stochastic process includes a deterministic mean perturbed by stochastic errors and uncorrelated, addi-

tive measurement error. People are interested in estimating the threshold probability that the deterministic mean perturbed by stochastic errors at each location is above a fixed threshold. Under certain conditions, the threshold probability is a simple function depending on the variance/covariance structure of the stochastic process, and the regression response function. In practice, however, the variance/covariance structures of the stochastic process, and therefore those of the threshold probability, are only approximately known. As well, the regression response functions might be misspecified.

We consider methods for the construction of robust sampling designs for the estimation of the threshold probability, with particular attention being paid to the effect of spatial correlation between adjoining locations and the regression response functions. A minimax approach is adopted. The optimal design minimizes the maximum over the neighbourhood of the working model of the loss function. The natural loss function is the ratio of mean squared error of the prediction and the true value. We approximate the loss function by the sum of the constant and terms with smaller orders under the increasing domain framework. The approximated loss function is then maximized over realistic neighbourhoods of the fitted linear relationship, and of the assumed variance and correlation structures. This maximized loss function is then minimized over the class of possible samples, yielding an optimally robust design. As an example, the methodology to find the optimal design is applied for a data set previously analyzied in the spatial design literature.

To my parents, Zhichun and Amy

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Douglas Wiens, who has supported me throughout my Ph.D study and research with his excellent guidance, patience and understanding. Without him this thesis would not have been completed or written.

I would like to thank the other members of my committee, Dr. Rohana Karunamuni, Dr. Linglong Kong and Dr. Ying Tsui for the assistance they provided. I would like to thank Dr. William Notz from Ohio State University for taking time to serve as my external examiner.

Special thanks go to Dr. Byron Schmuland, Dr. Giseon Heo, Dr. Thomas Hillen and all the professors who ever taught me for their help and encouragement during my study at the University of Alberta.

Thanks also go to office staffs of the Department of Mathematical and Statistical Sciences for their help with administrative matters.

Last but not the least, I would like to thank all my friends for their support in various forms.

Contents

List of Figures

Chapter 1

Introduction

In most statistical procedures for mathematical convenience there are kinds of explicit or implicit assumptions about models, data independence, randomness and so on. However, some of the most common statistical procedures are excessively sensitive to such assumptions. Minor derivations from the assumptions can cause a disastrous error in the final conclusions. Most time we only have fuzzy knowledge or belief about the underlying situation. A robust procedure should be proposed in these cases which ensures the relative insensitivity to small deviations from the prior assumptions (Box 1953, Huber 1981, Wiens 2015a).

The statistical design of experiments is usually referred to as the process of planning an experiment to collect data appropriately so that they may be analysed by statistical methods resulting in valid and objective conclusions (Montgomery 1984). In classical optimal design theory, one usually has in mind a particular model believed to be correct. An optimal design will be chosen through optimizing a "loss function" which is derived based on the assumed, fitted model. A misspecified form of regression model however can be dangerous as described by researchers (Box and Draper 1959, Ford, Titterington and Kitsos 1989). In robust design theory, one anticipates that the model fitted by the experimenter is not necessarily the true one and the design criterion must work well over classes to which the true model might belong. In this dissertation two design problems are studied which fall into this category: robustness of design for model discrimination problem, and robust design problem for the prediction of a threshold probability.

This work contains five chapters. Chapter 2 and Chapter 4 are independent papers which have been prepared for publication. Chapter 2 studies the model discrimination problem and Chapter 3 provides detailed deviation and proofs to support theoretical results in Chapter 2. Chapter 4 focuses on the robust design problem for the prediction of a threshold probability. This work aims to construct a spatial (sampling)

design which gives accurate and robust estimates of the probabilities that a stochastic process exceeds a particular threshold, possibly as a function of various covariates besides time or location.

This chapter is an introductory chapter presenting a literature review of previous work about designs for model discrimination and designs in spacial statistics. Section 1.1 presents the problem in model discrimination. A brief review of the hypothesis tests will be included in Section 1.1.1 and the properties of test powers underlying different assumptions will be investigated. Especially, test powers are found out to be an increasing function of the Kullback-Leibler divergence which is also briefly introduced. Various optimality criteria based on test power in the literature will be discussed in Section 1.1.2. In this section classical optimal design problems, and also an extension of the regression design problems to possibly misspecified models, are described. Section 1.2 introduces the designs for spatial models: spatial experimental design and spatial sampling design. Section 1.2.1 introduces the classical and robust (experimental and sampling) designs in literature. Section 1.2.2 gives the brief introduction of the preliminary knowledge required for the construction of spatial designs.

1.1 Designs for model discrimination

The primary aim of theoretical studies in scientific disciplines (e.g. physics, chemistry, engineering, etc.) is to discover the actual physical mechanism which governs the process under investigation. Investigators are concerned with a mathematical model relating the response y and the variables $\mathbf{x}$. The mathematical model is often constructed based directly upon an understanding of the mechanism. In practice it is often the case that the investigator will have several plausible models in mind. The central goal of the investigator is therefore to design an experiment distinguishing among these rival models. However, a design that is not well-chosen may result in failure of detecting serious inadequacies of a tentatively entertained model.

Example 1 *Michaelis-Menten kinetics and the exponential response models can be used as mathematical descriptions of the relationship between oxygen consumption and oxygen delivery (Lubarsky, Smith, Sladen, Mault, and Reed 1995). The Michaelis-Menten model is*

$$E(Y|x) = \frac{V_0 x}{K_0 + x}$$

and the exponential response model is

$$E(Y|x) = V_1(1 - \exp\{-K_1 x\})$$

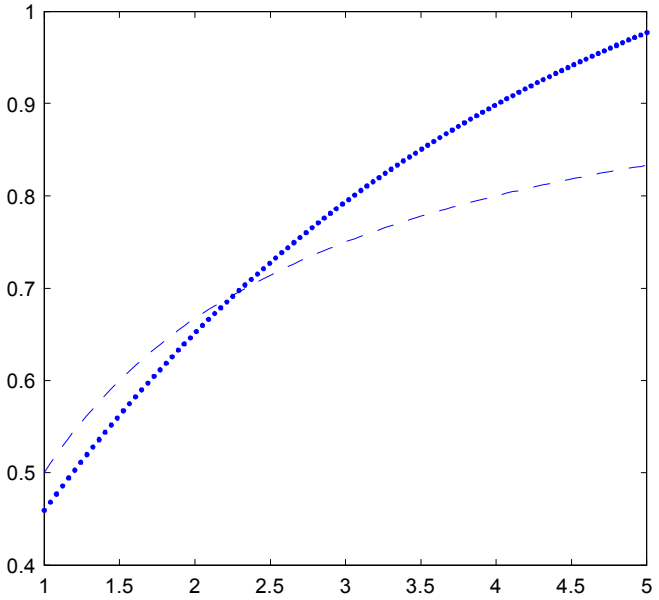

Figure 1.1: Comparison of Michaelis-Menten model and exponential response model. Dashed line represents Michaelis-Menten model. The dotted line represents exponential response model. The horizontal axis denotes variable x and the vertical axis denotes the mean of y at x.

where $E(Y|x)$ is the expected value of the response y with the variable x. When $V_0 = K_0 = 1$, $V_1 = 1.2$, $K_1 = 0.3$, and x varies from 1 to 5, the two curves defined by the above two models respectively intersect at $x = 2.485$ as shown in Fig 1.1. In an extreme situation, an experimenter made most observations at $x = 2.485$. If the true model is one of the two models, the observations would not permit the inadequacy of the other model to be detected.

This undesirable situation illustrated in Example 1 can occur in other cases, and in general is difficult to detect by a mere inspection of the response functions of the rival models, due to the fact that the regression parameters are usually unknown.

The design of experiments for discriminating between models has been investigated in literature under the following possible conditions:

(1) (Hunter and Reiner 1965, Atkinson and Fedorov 1975a) two rival models are available:

$$\begin{aligned} \text{Model I} &: \text{the distribution of the observations is } f_0(y|\mathbf{x}, \mu_0(\mathbf{x}), \boldsymbol{\varphi}_0) \\ \text{Model II} &: \text{the distribution of the observations is } f_1(y|\mathbf{x}, \mu_1(\mathbf{x}), \boldsymbol{\varphi}_1) \end{aligned} \quad (1.1)$$

where f_0 and f_1 are normal densities and

$$\mu_j(\mathbf{x}) = \int y f_j(y|\mathbf{x}, \boldsymbol{\varphi}_j) dy.$$

In particular, there exist functions $\eta_0(x, \theta_0), \eta_1(x, \theta_1)$ such that

$$\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}, \boldsymbol{\theta}_j), \quad j = 0, 1 \tag{1.2}$$

where θ_0, θ_1 are unknown regression parameters. Then (1.1) can be rewritten as

$$\begin{aligned} \text{Model I:} \qquad & y(\mathbf{x}) = \eta_0(\mathbf{x}, \boldsymbol{\theta}_0) + \varepsilon \\ \text{Model II:} \qquad & y(\mathbf{x}) = \eta_1(\mathbf{x}, \boldsymbol{\theta}_1) + \varepsilon \end{aligned} \tag{1.3}$$

where the error terms ε's are normally and independently distributed with constant variance σ^2, i.e. $\boldsymbol{\varphi}_0 = \boldsymbol{\varphi}_1 = \sigma^2$.

(2) (Fedorov and Pazman 1968) There are two rival models as in (1.3) while the error terms are normally and independently distributed with non-constant variances.

(3) (López-Fidalgo, Tommasi and Trandafir 2007) The two rival models, $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}), \boldsymbol{\varphi}_0)$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}), \boldsymbol{\varphi}_1)$, are non-normal models with (1.2) holding.

(4) (Atkinson and Fedorov 1975b) There are several rival models:

$$y(\mathbf{x}) = \eta_j(\mathbf{x}, \boldsymbol{\theta}_j) + \varepsilon, \ j = 1, ...v$$

and the error terms are normally and independently distributed with constant variance σ^2, the same for all models.

(5) (Uciński and Bogacka, 2005) There are several rival models:

$$y(\mathbf{x}) = \eta_j(\mathbf{x}, \boldsymbol{\theta}_j) + \varepsilon, \ j = 1, ...v$$

and the error terms are normally and independently distributed with nonconstant variance.

1.1.1 Hypothesis test for discriminating between two rival models

We assume that two rival models are available as in (1.1). Usually, the model discrimination problem is cast as a problem of hypothesis testing:

$$H_0: \text{ Model I} \quad \text{versus} \quad H_a: \text{ Model II} \tag{1.4}$$

and we suppose that n observations $\mathbf{y}_n = (y_1, y_2, ..., y_n)^T$ at, not necessarily distinct, design points $\mathbf{s} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ have been made. Here $\mathbf{s}$ is chosen from a finite set $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N \subset \boldsymbol{R}^q, q \geq 1$. In the following, for convenience, we denote

$$\boldsymbol{\eta}_j(\boldsymbol{\theta}_j) = (\eta_j(\mathbf{x}_1, \boldsymbol{\theta}_j), ..., \eta_j(\mathbf{x}_n, \boldsymbol{\theta}_j))^T, \quad j = 0, 1.$$

Homoscedastic normal linear regression models: F-test

When condition (1.1) is true and $\eta_j(\mathbf{x}, \boldsymbol{\theta}_j) = g_j^T(\mathbf{x})\boldsymbol{\theta}_j$,

$$\begin{aligned} \boldsymbol{\eta}_0(\boldsymbol{\theta}_0) &= \mathbf{X}_0\boldsymbol{\theta}_0 \text{ for } \mathbf{X}_0 : n \times p_0 \quad \text{and } \boldsymbol{\theta}_0 : p_0 \times 1 \\ \boldsymbol{\eta}_1(\boldsymbol{\theta}_1) &= \mathbf{X}_1\boldsymbol{\theta}_1 \text{ for } \mathbf{X}_1 : n \times p_1 \quad \text{and } \boldsymbol{\theta}_1 : p_1 \times 1 \end{aligned}$$

where $\mathbf{X}_0 = (g_0(\mathbf{x}_1), ..., g_0(\mathbf{x}_n))^T$, $\mathbf{X}_1 = (g_1(\mathbf{x}_1), ..., g_1(\mathbf{x}_n))^T, \mathbf{X}_1 = (\mathbf{X}_0 \colon \mathbf{X}_2), \mathbf{X}_2$ is $n \times (p_1 - p_0)$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ are unknown regression parameters. One can apply an F-test for the model discrimination hypothesis test. The unknown parameters can be estimated by least square estimates:

$$\widehat{\boldsymbol{\theta}}_0 = (\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y} \text{ and } \widehat{\boldsymbol{\theta}}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}.$$

Then the sums of squares of the residuals for these two models are

$$\begin{aligned} SSE_0 &= (\mathbf{y} - \mathbf{X}_0\widehat{\boldsymbol{\theta}}_0)^T(\mathbf{y} - \mathbf{X}_0\widehat{\boldsymbol{\theta}}_0) = \mathbf{y}^T(\mathbf{I} - \mathbf{H}_0)\mathbf{y} \\ SSE_1 &= (\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\theta}}_1)^T(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\theta}}_1) = \mathbf{y}^T(\mathbf{I} - \mathbf{H}_1)\mathbf{y}, \end{aligned}$$

respectively, where $\mathbf{H}_i = \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T$, $i = 0, 1$, are idempotent matrices with $\dim(\mathbf{H}_i) = p_i$. By applying the Gram-Schmidt Theorem, we can obtain QR decompositions for $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\mathbf{X}_0 = (\mathbf{Q}_{01} \colon \mathbf{Q}_{02})\begin{pmatrix}\mathbf{R}_0 \\ \mathbf{0}\end{pmatrix} \text{ and } \mathbf{X}_1 = (\mathbf{Q}_{11} \colon \mathbf{Q}_{12})\begin{pmatrix}\mathbf{R}_1 \\ \mathbf{0}\end{pmatrix}$$

where $\mathbf{Q}_i = (\mathbf{Q}_{i1} \colon \mathbf{Q}_{i2}), i = 0, 1$, are orthogonal matrices and $(\mathbf{Q}_{i1})_{n \times p_i}, (\mathbf{Q}_{i2})_{n \times (n-p_i)}$. Then $\mathbf{H}_i = \mathbf{Q}_{i1}\mathbf{Q}_{i1}^T, \mathbf{I}_{n \times n} - \mathbf{H}_i = \mathbf{Q}_{i2}\mathbf{Q}_{i2}^T$ and

$$\mathbf{y}^T(\mathbf{I} - \mathbf{H}_i)\mathbf{y} = (\mathbf{Q}_{i2}^T\mathbf{y})^T(\mathbf{Q}_{i2}^T\mathbf{y}).$$

If the null hypothesis is true, i.e., the nested model is true, we have

$$\mathbf{Q}_{i2}^T\mathbf{y} \sim MN_{(n-p_i)\times 1}(\mathbf{0}, \sigma^2\mathbf{I}_{(n-p_i)\times(n-p_i)}), i = 0, 1,$$

and

$$SSE_i \sim \sigma^2\boldsymbol{\chi}^2_{n-p_i}, \; SSE_0 - SSE_1 \sim \sigma^2\boldsymbol{\chi}^2_{p_1-p_0} \text{ and } \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n - p_1)} \sim F_{p_1}^{p_1-p_0}.$$

Then F-test reject H_0 for large value of

$$\frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n - p_1)}.$$

Under the alternative hypothesis, i.e., the reduced model is different from the full model (which is assumed to be the true model), $\sigma^{-2}SSE_1$ follows a noncentral chi-square distribution with noncentrality parameter

$$\begin{aligned}\Delta(\boldsymbol{\theta}_1) &= \boldsymbol{\theta}_1^T\mathbf{X}_1^T\mathbf{Q}_{02}\mathbf{Q}_{02}^T\mathbf{X}_1\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^T\mathbf{X}_1^T(\mathbf{I}_{n\times n}-\mathbf{Q}_{01}\mathbf{Q}_{01}^T)\mathbf{X}_1\boldsymbol{\theta}_1 = ||(\mathbf{I}_{n\times n}-\mathbf{Q}_{01}\mathbf{Q}_{01}^T)\mathbf{X}_1\boldsymbol{\theta}_1||^2 \\ &= ||\mathbf{X}_1\boldsymbol{\theta}_1-\mathbf{H}_0\mathbf{X}_1\boldsymbol{\theta}_1||^2 = \inf_{\boldsymbol{\theta}_0}||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2.\end{aligned}$$

The power of the test is an increasing function of $\Delta(\boldsymbol{\theta}_1)$. Notice that $||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2$ is the sum of the Kullback-Leibler divergence between two normal densities with means $\eta_j(\mathbf{x},\boldsymbol{\theta}_j), j=0,1$, and the same standard deviation.

Nonlinear regression models: Neymann-Pearson test

Nonlinear regression models are much more complicated than linear models. Assume that the regression parameters $\boldsymbol{\theta}_0,\boldsymbol{\theta}_1$ and the nuisance parameters $\boldsymbol{\varphi}_0,\boldsymbol{\varphi}_1$ are known. Then Neymann-Pearson test is the uniformly most powerful test. The null hypothesis test is rejected if the value of the test statistic $\mathcal{R}$ is large:

$$\begin{aligned}\mathcal{R} &= 2\log\left\{\frac{\prod_{i=1}^n f_1(y_i|\mu_1(\mathbf{x}_i),\boldsymbol{\varphi}_1)}{\prod_{i=1}^n f_0(y_i|\mu_0(\mathbf{x}_i),\boldsymbol{\varphi}_0)}\right\} \\ &= 2\sum_{i=1}^n\log\frac{f_1(y_i|\mu_1(\mathbf{x}_i),\boldsymbol{\varphi}_1)}{f_0(y_i|\mu_0(\mathbf{x}_i),\boldsymbol{\varphi}_0)}.\end{aligned}$$

If Model II is true, the expectation of the test statistic is

$$E(\mathcal{R}|\text{ Model II}) = 2\sum_{i=1}^n\mathcal{I}\{f_0,f_1|\mathbf{x}_i\}.$$

where

$$\mathcal{I}\{f_0,f_1|\mathbf{x}\} = \int\log\frac{f_1(y|\mu_1(\mathbf{x}),\boldsymbol{\varphi}_1)}{f_0(y|\mu_0(\mathbf{x}),\boldsymbol{\varphi}_0)}f_1(y|\mu_1(\mathbf{x}),\boldsymbol{\varphi}_1)dy$$

is the Kullback-Leibler divergence of f_0 to f_1.

Intuitively, the power of the test should depend on the expectation of the test statistic: the larger $E(\mathcal{R}|$ Model II), the larger the power (López-Fidalgo, et al. 2007). When Model I and Model II are both normal, the claim is provable.

Homoscedastic models When Model I and Model II are both normal with the same nuisance parameter, we can calculate the power of the test explicitly. In this

case,

$$\begin{aligned}\mathcal{R} &= 2\log\left\{\frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\eta_1(\mathbf{x},\boldsymbol{\theta}_1))^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\eta_0(\mathbf{x},\boldsymbol{\theta}_0))^2\right\}}\right\} \\ &= \frac{1}{\sigma^2}\left[2\left(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\mathbf{y}\right)^T\left(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)\right)-||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2\right].\end{aligned}$$

When Model I is true, the test statistic follows a normal distribution with mean

$$-\frac{1}{\sigma^2}||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2$$

and variance

$$\sigma_R^2=\frac{4}{\sigma^2}||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2.$$

When Model II is true, the test statistic follows a normal distribution with mean

$$E(\mathcal{R}|\text{ Model II})=\frac{1}{\sigma^2}||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)-\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2$$

and variance σ_R^2. Given the significance level α, the critical value $c=\sigma_R u_\alpha$ with u_α being the $100(1-\alpha)$ percentile of the standard normal distribution. The power of the test is

$$\begin{aligned}P\left(\mathcal{R}>c|\text{ Model II is true}\right) &= \Phi\left(\frac{E(\mathcal{R}|\text{ Model II})-c}{\sigma_R}\right) \\ &= \Phi\left(\sqrt{E(\mathcal{R}|\text{ Model II})}-u_\alpha\right).\end{aligned}$$

It is clear that the power is an increasing function of $E(\mathcal{R}|$ Model II).

When Model I and Model II are not normal, it is usually hard to investigate the distribution of the test statistic. Wiens (2009b) derived the asymptotic properties of the Neymann-Pearson test statistic.

Theorem 2 *(Theorem 1.1 in Wiens (2009b)) Suppose that Model I and Model II satisfy the following conditions:*

(a) $f_j(y|\mathbf{x},\mu_j,\boldsymbol{\varphi}_j)=f(y|\mathbf{x},\mu_j,\boldsymbol{\varphi}_j)$ *for a density* f*;*

(b) $\mu_1(x_i)=\mu_0(x_i)+\Delta_i/n^{1/2}, i=1,...,N.$

Define

$$\mathcal{D}=\frac{1}{2n}E(\mathcal{R}|\text{ Model II}). \tag{1.5}$$

Then under the regularity conditions $\mathcal{D}$ *is* $O(n^{-1})$ *and the Neymann-Pearson test statistic* $\mathcal{R}$ *is asymptotically normally distributed under each hypothesis:*

$$\text{under } H_0,\ \frac{\mathcal{R}+2n\mathcal{D}}{\sqrt{8n\mathcal{D}}}\xrightarrow{L}N(0,1);$$

$$\text{under } H_1,\ \frac{\mathcal{R}-2n\mathcal{D}}{\sqrt{8n\mathcal{D}}}\xrightarrow{L}N(0,1).$$

Based on this theorem, a Neymann-Pearson test with asymptotic size α has the asymptotic power

$$\beta = \Phi\left(\sqrt{2n\mathcal{D}} - u_\alpha\right) \tag{1.6}$$

which is an increasing function of $\mathcal{D}$.

In particular, when Model I and Model II are normal with the same variance, the test power is exactly the asymptotic power and

$$\mathcal{D} = \frac{1}{2n\sigma^2}||\boldsymbol{\eta}_0(\theta_0) - \boldsymbol{\eta}_1(\theta_1)||^2.$$

Heteroscedastic models When Model I and Model II are both normal with non-constant variances, we can also obtain the power of the test explicitly. In this case,

$$\begin{aligned}
\mathcal{R} &= 2\log\left\{\frac{\exp\left\{-\sum_{i=1}^n \frac{1}{2\sigma_i^2}(y_i - \eta_1(\mathbf{x},\boldsymbol{\theta}_1))^2\right\}}{\exp\left\{-\sum_{i=1}^n \frac{1}{2\sigma_i^2}(y_i - \eta_0(\mathbf{x},\boldsymbol{\theta}_0))^2\right\}}\right\} \\
&= 2\left(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \mathbf{y}\right)^T \Sigma^{-1}(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1)) \\
&\quad -(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))^T\Sigma^{-1}(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))
\end{aligned}$$

where

$$\boldsymbol{\Sigma}^{-1} = \text{diag}(\sigma_1^{-2}, ..., \sigma_n^{-2}).$$

When Model I is true, the test statistic follows a normal distribution with mean

$$-(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))$$

and variance

$$\sigma_R^2 = 4(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1)).$$

When Model II is true, the test statistic follows a normal distribution with mean

$$E(\mathcal{R}|\text{ Model II}) = (\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1))$$

and variance σ_R^2. Given the significant level α, the critical vale $c = \sigma_R u_\alpha$ with u_α being the $100(1-\alpha)$ percentile of the standard normal distribution. The power of the test is

$$P\left(\mathcal{R} > c|\text{ Model II is true}\right) = \Phi\left(\sqrt{E(\mathcal{R}|\text{ Model II})} - u_\alpha\right).$$

When Model I and Model II are not normal, in Chapter 2, asymptotic properties of the test statistic are derived where assumption (a) in Theorem 2 is no longer necessary.

Proposition 3 *Given a design space* $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$, *assume that the experiment has* n_i *replicates at each covariate* $\mathbf{x}_i$, *with* $\sum_{i=1}^N n_i = n$. *Define* $\mathcal{D}$ *as in (1.5) and for any two densities* f_0, f_1 *define*

$$r\left(y|\mathbf{x}_i; f_0, f_1\right) = \frac{f_1(y|\mathbf{x}_i, \mu_1\left(\mathbf{x}_i\right))}{f_0(y|\mathbf{x}_i, \mu_0\left(\mathbf{x}_i\right))}.$$

Assume that two densities f_0, f_1 *satisfy*
(a) for the KL-divergence,

$$n\mathcal{D} = O(1), \tag{1.7}$$

(b) for all $\delta > 0$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i \int_{\{|\log r(y|\mathbf{x}_i; f_0, f_1)| \geq \delta\}} f_0(y|\mathbf{x}_i, \mu_0\left(\mathbf{x}\right)) \left(\sqrt{r\left(y|\mathbf{x}_i; f_0, f_1\right)} - 1\right)^2 dy = 0, \tag{1.8}$$

(c) there is a $\tau > 0$ *such that*

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i \int_{\{\log r(y|\mathbf{x}_i; f_0, f_1) \geq \tau\}} f_1(y|\mathbf{x}_i, \mu_1\left(\mathbf{x}\right)) \log r\left(y|\mathbf{x}_i; f_0, f_1\right) dy = 0. \tag{1.9}$$

Then $\mathcal{R} = 2\sum_{i=1}^N \sum_{l=1}^{n_i} \log r(y|\mathbf{x}_i; f_0, f_1)$ *and:*
(i) under the null hypothesis,

$$\frac{\mathcal{R} + 2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1)$$

(ii) under the alternative hypothesis,

$$\frac{\mathcal{R} - 2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1).$$

Based on this result, the Neymann-Pearson test with asymptotic size α has the same asymptotic power as shown in (1.6) which is an increasing function of the sum of the Kullback-Leibler divergence $\mathcal{D}$ of f_0 from f_1.

Kullback-Leibler divergence

Kullback-Leibler divergence, also called information divergence, is widely used in information theory. It is a non-symmetric measure of the difference between two probability distributions F_0 and F_1. Informally, it quantifies the lost information when F_0 is used to approximate F_1.

Definition 4 *Given two probability distributions, $F_0 << F_1$, the Kullback-Leibler divergence of F_0 from F_1 is*

$$E_{F_1}\left[\log \frac{dF_1}{dF_0}\right].$$

If F_0 is not absolutely continuous with respect to F_1, then the KL-divergence is ∞.

Remark 5 *Here $F_0 << F_1$ denotes that F_0 is absolutely continuous with respect to F_1. That is, there exists a Borel-measurable function g such that $g \geq 0$, $\int_{-\infty}^{+\infty} g(t)dF_1 = 1$ and*

$$F_0(x) = \int_{-\infty}^{x} g(t)dF_1.$$

Remark 6 *If F_0 and F_1 are both absolutely continuous and their densities are f_0 and f_1, respectively, the Kullback-Leibler divergence is*

$$\mathcal{I}\{f_0, f_1\} := \int_{-\infty}^{+\infty} f_1(t) \log \frac{f_1(t)}{f_0(t)} dt.$$

Remark 7 *Kullback-Leibler divergence has the following properties.*

1. *The KL-divergence is nonnegative and is zero iff $F_0 = F_1$ almost everywhere.*
2. *The KL-divergence is not symmetric (therefore, it is not a true metric).*

1.1.2 The design criteria and designs

To discriminate the two models effectively, one should try to choose a design which truly places the hypothesized models in jeopardy, in other words, the optimal design should maximize the power of the discrimination test. Denote $\boldsymbol{\xi}$ as the design measure placing mass ξ_i at $\mathbf{x}_i \in \mathcal{S}$ where $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$. That is, $\boldsymbol{\xi}$ is a vector with elements ξ_i such that $\xi_i \geq 0$ and $\sum_{i=1}^{N} \xi_i = 1$.

Classical optimal designs

The design of experiments for discriminating between rival models has been investigated by numerous authors. A common assumption is that one of the two proposed models is the true one. That is, that model does in fact represent the true physical mechanism in such a way that either Model I or Model II will hold. Under this assumption the powers of the discrimination tests depend on the Kullback-Leibler divergence of the two rival models. In particular, for homoscedastic normal linear

regression models, one should choose a design based on the criterion:

$$\begin{aligned}\boldsymbol{\xi}^* &= \arg\max_{\boldsymbol{\xi}} \inf_{\boldsymbol{\theta}_0} ||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2 \\ &= \arg\max_{\boldsymbol{\xi}} ||\mathbf{X}_1\boldsymbol{\theta}_1 - \mathbf{H}_0\mathbf{X}_1\boldsymbol{\theta}_1||^2 \\ &= \arg\max_{\boldsymbol{\xi}} \boldsymbol{\theta}_1^T \left(\mathbf{X}_1^T\mathbf{X}_1 - \mathbf{X}_1^T\mathbf{X}_0 \left(\mathbf{X}_0^T\mathbf{X}_0\right)^{-1} \mathbf{X}_0^T\mathbf{X}_1\right) \boldsymbol{\theta}_1\end{aligned}$$

where

$$\mathbf{X}_k^T\mathbf{X}_l = \sum_{i=1}^{N} \xi_i g_k(\mathbf{x}_i) g_l^T(\mathbf{x}_i), \quad k, l = 0, 1;$$

for nonlinear heteroscedastic normal regression models, the optimal design is chosen such that

$$\begin{aligned}\boldsymbol{\xi}^* &= \arg\max_{\boldsymbol{\xi}} \Phi\left(\sqrt{E(\mathcal{R}|\text{ Model II})} - u_\alpha\right) \\ &= \arg\max_{\boldsymbol{\xi}} E(\mathcal{R}|\text{ Model II}) \\ &= \arg\max_{\boldsymbol{\xi}} ||\boldsymbol{\eta}_0(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_1(\boldsymbol{\theta}_1)||^2 \\ &= \arg\max_{\boldsymbol{\xi}} \sum_{i=1}^{N} \xi_i (\eta_0(\mathbf{x}_i, \boldsymbol{\theta}_0) - \eta_1(\mathbf{x}_i, \boldsymbol{\theta}_1))^2;\end{aligned}$$

and for non-normal models, the optimal design is

$$\begin{aligned}\boldsymbol{\xi}^* &= \arg\max_{\boldsymbol{\xi}} \Phi\left(\sqrt{E(R|\text{ Model II})} - u_\alpha\right) \\ &= \arg\max_{\boldsymbol{\xi}} \mathcal{D}(\boldsymbol{\xi}).\end{aligned}$$

The optimization criteria above depend on unknown parameters $\boldsymbol{\theta}_j$ $(\in \mathbb{R}^{p_j}), j = 0, 1$. It is a common problem one needs to deal with in the construction of model discrimination designs. There are several methods to address this issue. Two main methods are discussed in literature. The first method assumes that n_0 observations have been made and the unknown parameters can be estimated and updated when new observations are available. This method is used in sequential approaches for constructing designs. Another method is based on the assumption that one of the rival models is true and the regression parameters of that model are known. The unknown parameters in another model are estimated by a least squares estimate, that is, the one that minimizes the L^2-distance between the true model and the other rival model. The method is helpful for constructing static designs.

Sequential methods to construct the optimal designs for model discrimination were first proposed by Hunter and Reiner (1965). Two homoscedastic normal models were considered. The authors assumed that n_0 observations have been made already.

Parameters $\boldsymbol{\theta}_j$ can be estimated by least squares (LS) estimates $\hat{\boldsymbol{\theta}}_j^{(1)}$ and the $(n_0+1)th$ observation then will be made at the new point $\mathbf{x}_{(n+1)}$ maximizing

$$\mathbf{x}_{(n_0+1)} = \arg\max_{\mathbf{x}} (\eta_0(\mathbf{x}, \widehat{\boldsymbol{\theta}}_0^{(1)}) - \eta_1(\mathbf{x}, \widehat{\boldsymbol{\theta}}_1^{(1)}))^2.$$

One can repeat this procedure and update the estimates of $\boldsymbol{\theta}_j$ at each step until n $(n > n_0)$ design points are obtained. Fedorov and Pazman (1968) extended the sequential method to heteroscedastic normal models.

An example of static designs for model discrimination is T-optimal design proposed in Atkinson and Fedorov (1975a). Two homoscedastic nonlinear normal models are assumed. Model I is assumed to be the true model and the parameters $\boldsymbol{\theta}_1$ are known. The optimal design should be the one that maximizes the power of the discrimination test. The problem again is that the power function depends on unknown regression parameters $\boldsymbol{\theta}_0$. Instead of updating the estimate of regression parameters in each iteration, an estimate of $\boldsymbol{\theta}_0$ which minimize the L^2-distance between $\boldsymbol{\eta}_0(\boldsymbol{\theta}_0)$ and $\boldsymbol{\eta}_1(\boldsymbol{\theta}_1)$ is employed and a T-optimal design is

$$\boldsymbol{\xi}^* = \arg\sup_{\boldsymbol{\xi}} \inf_{\theta_0 \in \Theta_0} \sum_{i=1}^{N} \xi_i (\eta_0(\mathbf{x}_i, \boldsymbol{\theta}_0) - \eta_1(\mathbf{x}_i, \boldsymbol{\theta}_1))^2.$$

López-Fidalgo et al. (2007) investigated the construction of KL-optimal designs for non-normal models. Two models as in (1.4) are considered with the first model being assumed to be the true model with known parameter $\boldsymbol{\theta}_1$. The asymptotic power of the discrimination test can proved to be an increasing function of the KL distance between Model I and Model II (Wiens 2009b). It is reasonable to choose the design which maximizes the KL divergence between these two models. The authors employed the method in Atkinson and Fedorov (1975a) to address the issue of unknown parameters $\boldsymbol{\theta}_0$, and the design criterion becomes

$$\boldsymbol{\xi}^* = \arg\sup_{\boldsymbol{\xi}} \inf_{\theta_0 \in \Theta_0} \sum_{i=1}^{N} \xi_i \int \log \frac{f_1(y_i|\eta_1(\mathbf{x}_i, \boldsymbol{\theta}_1), \boldsymbol{\varphi}_1)}{f_0(y_i|\eta_0(\mathbf{x}_i, \boldsymbol{\theta}_0), \boldsymbol{\varphi}_0)} f_1(y_i|\eta_1(\mathbf{x}_i, \boldsymbol{\theta}_1), \boldsymbol{\varphi}_1) dy.$$

Robust designs

In the classical designs introduced above, one of the two rival models is assumed to be exactly correct and the following assumption is common:

$$\mu_1(\mathbf{x}) = \eta_1(\mathbf{x}, \boldsymbol{\theta}_1), \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}, \boldsymbol{\theta}_0).$$

However, in most applications, the assumed model is, at best, only a reasonable approximation to the true model. In the pioneering work of Box and Draper (1959), the inherent dangers can be serious of designing an experiment based on the belief that the assumed models are exactly correct. Therefore, in light of modern notions of robustness, it is more common to suppose that the correct model is a member of one of two only approximately known classes $\mathcal{F}_0$, $\mathcal{F}_1$.

Wiens (2009a) proposed methods of constructing discrimination designs which are robust against model misspecification. Usually, the means $\mu_j(\mathbf{x})$ are only partially known and the function $\eta_j(\mathbf{x}, \boldsymbol{\theta}_j)$ is at best an approximation of the mean $\mu_j(\mathbf{x})$. A problem arising immediately due to model misspecification is that the meaning of the parameter $\boldsymbol{\theta}_j$ becomes unclear. To address the problem, a working response $E(Y|\mathbf{x})$ is adopted and the parameter is defined as

$$\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}} \left[\sum_{\mathcal{S}} \left(E[Y|\mathbf{x}_i] - \eta_j(\mathbf{x}_i, \boldsymbol{\theta}_j) \right)^2 \right]. \tag{1.10}$$

According to the definition, the parameter defined in (1.10) provides the closest agreement between the working response and that in model j. There are different choices of working response. In particular, if $E[Y|\mathbf{x}] = \eta_1(\mathbf{x}, \boldsymbol{\theta}_1)$ with a specified $\boldsymbol{\theta}_1$, the method of determining $\boldsymbol{\theta}_0$ is analogous to that in Atkinson and Fedorov (1975a) and Atkinson and Fedorov (1975b). For more details, see the discussion in Wiens (2009a).

Define $\delta_j(\mathbf{x}) = E[Y|\mathbf{x}] - \eta_j(\mathbf{x}, \boldsymbol{\theta}_j)$ and denote

$$\begin{aligned} \boldsymbol{\delta}_j &= (\delta_j(\mathbf{x}_1), ..., \delta_j(\mathbf{x}_N))^T, \\ \mathbf{U}_j &= \frac{\partial \boldsymbol{\eta}_j(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \eta_j(\mathbf{x}_1, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}}, ..., \frac{\partial \eta_j(\mathbf{x}_N, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}} \right)^T. \end{aligned}$$

Then according to definition (1.10), $\boldsymbol{\delta}_j$ and $\mathbf{U}_j$ should satisfy $\mathbf{U}_j^T \boldsymbol{\delta}_j = 0$. The "errors" should also be bounded. Therefore, two neighbourhoods of $\boldsymbol{\eta}_j(\boldsymbol{\theta}_j)$ considered in Wiens (2009a) are

$$\mathcal{F}_j = \left\{ \boldsymbol{\eta}_j(\boldsymbol{\theta}_j) + \boldsymbol{\delta}_j | \mathbf{U}_j^T \boldsymbol{\delta}_j = 0, ||\boldsymbol{\delta}_j|| \leq \tau_j \right\},$$

for some τ_j such that $\mathcal{F}_0 \cap \mathcal{F}_1 = \emptyset$. The true model is no longer one of the hypothesized models. Instead, the true model is assumed fall in one of the two classes $\mathcal{F}_0, \mathcal{F}_1$. The optimal design still needs to maximize the power of the test. According to the result in Theorem 2, the robust optimal design problem is that of determining

$$\boldsymbol{\xi}^* = \arg\sup_{\boldsymbol{\xi}} \inf_{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1} \sum_{i=1}^{N} \xi_i \int \log \frac{f_1(y_i|\mu_1(\mathbf{x}_i), \boldsymbol{\varphi}_1)}{f_0(y_i|\mu_0(\mathbf{x}_i), \boldsymbol{\varphi}_0)} f_1(y_i|\mu_1(\mathbf{x}_i), \boldsymbol{\varphi}_1) dy,$$

with $\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}, \boldsymbol{\theta}_j) + \delta_j(\mathbf{x})$ and $\boldsymbol{\delta}_j \in \mathcal{C}_j$.

When both models f_0, f_1 are normal with common variance σ^2, robust T-optimal design is proposed. The parameters $\boldsymbol{\theta}_j$ are computed from equation (1.10) and a static design is constructed. When f_0, f_1 are non-normal densities, robust Kullback–Leibler optimal designs are considered. Both static and sequential designs are constructed.

As a natural sequel of Wiens (2009a), in Chapter 2 we assume that the true model is in a Hellinger neighbourhood of one of nominal models. Construction for experimental designs is proposed by maximizing the worst power of the Neyman-Pearson test over the Hellinger neighbourhoods. More specifically, the optimal designs in this project are "maximin" designs, which maximize (through the design) the minimum (among the neighbourhoods $\mathcal{F}_j, j = 0, 1$) asymptotic power function.

1.2 Designs for spatial models

In many areas such as epidemiology, economics, climatology, agriculture and geology, spatial data are observed and recorded. A distinctive characteristic of the spatial data is that the analysis of the spatial data depends on the locations of the objects being collected. A possible spatial model is

$$Y(\mathbf{x}, \mathbf{t}) = \eta(\mathbf{x}, \mathbf{t}) + \delta(\mathbf{t}) + \varepsilon(\mathbf{t}) \tag{1.11}$$

where $\mathbf{t} \in \mathcal{T} := \{\mathbf{t}_1, \cdots, \mathbf{t}_N\} \subset \mathbb{R}^d$ is a location where the object might be measured, and $\mathbf{x} \in \mathbb{R}^q$ is a treatment covariate, and $\eta(\mathbf{x}, \mathbf{t})$ is the deterministic mean (large-scale variation) perturbed by stochastic errors $\delta(\mathbf{t})$ (small-scale variation), and $\varepsilon(\mathbf{t})$ is uncorrelated, additive measurement error. Typically, observations obtained from different locations are correlated and a correlation function is used to describe the statistical correlation between the random variables.

1.2.1 Spatial experimental and sampling designs: classical and robust

As discussed in Cressie (1993) (Wiens 2015b), spatial designs for model (1.11) can be distinguished between spatial experimental designs and spatial sampling designs. In spatial experimental designs, locations are fixed and the design allocates treatments to these locations. In spatial sampling design, however, experimenters need to choose locations to make observations and usually only one observation is available at any spatial location.

Spatial experimental design

Spatial experimental design was derived from agricultural experiments and developed later on. In classical experimental designs (based on randomization, blocking, and replication), the spatial positions of the treatments in the design are often ignored. And the efficiency is always measured based on the error variation. However, the effects of spatial correlation between adjoining plots have been well documented (see, Taplin 1999, Wu and Dutilleul 1999, Legendre, Dale, Fortin, Casgrain and Gurevitch 2004). Petraitis (2001) proposed experimental designs where the effect of spatial variation is controlled by grouping observations and treatments into blocks.

In most existing methods, various specific correlation structures, variance structures and regression responses were modelled. Wiens and Zhou (2008) considered the construction of robust spatial experimental designs for test-control field experiments, with a degree of uncertainty on the model structures. In Wiens and Zhou (2008) the locations of field plots are fixed, and each is to be assigned exactly one treatment, a value of $\mathbf{x}$. The design problem is that of assigning the variable $\mathbf{x}$, varying in a set, in some optional manner. The robust optimal design minimizes the trace of the mean squared error matrix which is first maximized over the neighbourhoods quantifying the model uncertainty.

Spatial sampling design

In spatial sampling design, various designs were proposed due to different purposes, see Thompson (1997). Here, we mainly discuss the optimal sampling designs. Classical method of choosing a spatial sampling design is based on some statistical criterion under the assumption that a model adequately approximates the observations. Such criteria include A-optimality, D-optimality, minimizing the integrated mean squared prediction error, and so on.

An A-optimal design minimizes the trace of the covariance matrix of the least-square estimators of the regression parameters. A D-optimal design minimizes the determinant of the covariance matrix of the regression parameters, the volume of a confidence ellipsoid for the regression parameters. An application of D-optimality is in Müller (2005) where designs are proposed to help the prediction of the chloride concentration (y) at location ($\mathbf{t}$) in a region in the Danube river basin in Austria. Under the assumption that only one monitoring station may be placed at a particular location, the D-optimal design proposed in Müller (2005) maximizes the determinant of the information matrix for the regression parameters.

Constructing spatial sampling designs for efficient prediction of the response at

unsampled locations is usually of the main interest. For that purpose, minimizing the integrated mean squared prediction error (IMSE) over locations in the region of interest is a common aim of the designer and experimenters need to choose locations to make observations. We assume that a stochastic process $\mu(\mathbf{t})$ at sampled locations is observed:

$$Y(\mathbf{t}) = \mu(\mathbf{t}) + \varepsilon(\mathbf{t}) \tag{1.12}$$

where $\varepsilon(\mathbf{t})$ with $\mathbf{t} \in \mathcal{T}$ is uncorrelated, additive measurement error.

In Cressie (1993), the random-field model (1.12) is assumed. The purpose is to predict $\mu(\mathbf{t})$ and the kriging predictors are applied. Due to the optimal properties of kriging predictors, the mean squared prediction errors are simply the kriging (prediction) variances. Then an optimal spatial sampling design minimizes the integrated kriging (prediction) variances. The criterion of minimizing the IMSPEs is also applicable in computer experimental designs, where the deterministic output of computer code is modelled as a random process with spatial correlation structures. Assuming that $Y(\mathbf{t})$ is predicted by universal kriging predictors, Santner, Williams and Notz (2003) discussed different algorithms obtaining IMSPE-optimal designs and MMSPE-optimal (MMSPE: maximum mean squared prediction error) designs in computer experiments.

It is noticeable that many criterion functions, such as IMSPE, require knowledge of the correlation/covariance function, which often depends on unknown correlation/covariance parameters, and sometimes the true values of regression parameters. In practice, a two-stage procedure is used. In the first stage, data are collected using some design strategy (for example, Latin Hypercube design) and used to obtain estimates of the unknown parameters; in the second stage, those estimates are treated as the true parameter values and plugged into the criterion function. Then one can determine an optimal design based on the criterion. The estimation of the unknown correlation/covariance parameters results in only approximately known variance/covariance structures of the stochastic process $Y(\mathbf{t})$. When uncertainties occur, classical methods can cause a disastrous error in the final conclusions. In these cases a robust procedure should be proposed which are the relatively insensitive to small deviations from the prior assumptions (see Box 1953, Huber 1981 and Wiens 2015a).

In Wiens (2005a), the variance/covariance structures of process $Y(\mathbf{t})$ and of the measurement errors are assumed to be only approximately known. A robust (minimax) linear predictor is then obtained, where the loss function is first maximized over the neighbourhoods quantifying the model uncertainty, and then minimized over the coefficients of the predictor subject to a constraint of unbiasedness. The loss function chosen is the mean squared error loss. Notice that the predictor also relies

on the assumption that the fitted model is correct. Wiens (2005b) constructed robust sampling designs which are robust against misspecified regression responses. To obtain the robust designs, the loss function is maximized analytically over a neighbourhood quantifying the departures from the fitted linear regression response and the maximum is minimized numerically to obtain the optimal designs.

Knowing how to predict $\mu(\mathbf{t})$ is a prerequisite for optimal spatial designs. In particular, the design criteria for optimal spatial designs depend on kriging and the estimation of the parameter. The asymptotic analysis of the spatial data can play an important role in solving design problems since the asymptotic results can be used to approximate the complex design criteria to a certain order. In the following, the necessary preliminaries are briefly introduced.

1.2.2 Preliminaries

Prediction: universal kriging

One main purpose of spatial statistics study is prediction. Given data $\mathbf{y}_n = (Y(\mathbf{t}_1), ..., Y(\mathbf{t}_n))^T$ and a location $\mathbf{t}_0 \notin \{\mathbf{t}_1, ..., \mathbf{t}_n\}$, one hopes to predict the value of a measurement-error-free version of $Y(\mathbf{t}_0)$, i.e. $\mu(\mathbf{t}_0)$ in (1.12). Kriging is a minimum-mean-squared-error method for prediction. Different kriging methods have been introduced since Matheron (1963, 1967) developed the theoretical basis for the method of optimal spatial linear prediction. Depending on different model assumptions, different types of kriging can be deduced (Cressie 1993). Here we mainly introduce a common kriging method: universal kriging (the best linear unbiased predictor).

Assume for the random-field model (1.12)

$$E(Y(\mathbf{t})) = \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta}, \tag{1.13}$$

where $\mathbf{f}(\mathbf{t}) = (f_1(\mathbf{t}), \cdots, f_p(\mathbf{t}))^T$ a $p-$dimensional vector of functions. We consider an unbiased linear (in data) predictor $\widehat{\mu}(\mathbf{t}_0)$

$$\widehat{\mu}(\mathbf{t}_0) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{t}_i). \tag{1.14}$$

The unbiasedness implies

$$\sum_{i=1}^{n} \lambda_i \mathbf{f}(\mathbf{t}_i) = \mathbf{f}(\mathbf{t}_0). \tag{1.15}$$

The parameters are chosen to minimize the mean-squared prediction error

$$MSPE = E\left(\mu(\mathbf{t}_0) - \widehat{\mu}(\mathbf{t}_0)\right)^2$$

and satisfy the unbiasedness condition (1.15). The universal kriging equations can be solved by applying the method of Lagrangian multipliers.

Under the assumption that the stochastic process $\mu(\mathbf{t})$ is a Gaussian process, $E(\mu(\mathbf{t}_0)|\,(Y(\mathbf{t}_1),...,Y(\mathbf{t}_n)))$ minimizes the MSPE (see the proof of Theorem 3.2.1 in Santner, Williams, and Notz 2003) and

$$E(\mu(\mathbf{t}_0)|\,(Y(\mathbf{t}_1),...,Y(\mathbf{t}_n))) = \mathbf{f}^T(\mathbf{t}_0)\boldsymbol{\theta} + \boldsymbol{\Sigma}_{1n}\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{y}_n - \mathbf{F}_n\boldsymbol{\theta}) \tag{1.16}$$

where $\boldsymbol{\Sigma}_{1n}$ =Cov$\big(\mu(\mathbf{t}_0), \mathbf{y}_n^T\big)$, $\boldsymbol{\Sigma}_{nn}$ =Var$(\mathbf{y}_n)$ and $\mathbf{F}_n = (\mathbf{f}(\mathbf{t}_1),...,\mathbf{f}(\mathbf{t}_n))^T$. Here we always assume that $\mathbf{F}_n$ has full column rank p.

With $\boldsymbol{\theta}$ in (1.16) being replaced by generalized least squares estimate (GLSE) $\widehat{\boldsymbol{\theta}}_{GLS} = (\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}\mathbf{F}_n)^{-1}\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}\mathbf{y}_n$, the universal kriging is $\widehat{\mu}(\mathbf{t}_0) = \mathbf{a}_n^T\mathbf{y}_n$ (see page 154 Cressie 1993) where

$$\mathbf{a}_n = \big(\mathbf{F}_n(\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}\mathbf{F}_n)^{-1}\mathbf{f}(\mathbf{t}_0) + \big(\mathbf{I}_n - \mathbf{F}_n(\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}\mathbf{F}_n)^{-1}\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}\big)\,\boldsymbol{\Sigma}_{n1}\big)^T\,\boldsymbol{\Sigma}_{nn}^{-1}. \tag{1.17}$$

Here, the covariance matrix $\boldsymbol{\Sigma}_{nn}$ of $\mathbf{y}_n$ and the covariance vector $\boldsymbol{\Sigma}_{n1}$ of $\mu(\mathbf{t}_0)$ and $\mathbf{y}_n^T$ depend on the covariance/variance structure of the process $Y(\mathbf{t})$. In most cases, the parameters in covariance/variance structure are unknown. We will introduce the maximum likelihood method to estimate the unknown parameters.

Estimation of the parameters

In general, there are four methods to estimate parameters in the correlation function (Santner, Williams and Notz 2003): maximum likelihood, restricted maximum likelihood, cross-validation and posterior mode. We mainly talk about maximum likelihood method.

Assume that the correlation function has parameters $\boldsymbol{\psi} = (\psi_1,...,\psi_d)^T$. Let Var$(\mu(\mathbf{t})) = \sigma_\mu^2$ and Var$(\varepsilon(\mathbf{t})) = \sigma_\varepsilon^2$. The log-likelihood is

$$l(\boldsymbol{\theta},\sigma_\mu^2,\sigma_\varepsilon^2,\boldsymbol{\psi}) = -\frac{1}{2}\left[\log\det(\boldsymbol{\Sigma}_{nn}(\boldsymbol{\psi})) + (\mathbf{y}_n - \mathbf{F}_n\boldsymbol{\theta})^T\boldsymbol{\Sigma}_{nn}^{-1}(\boldsymbol{\psi})(\mathbf{y}_n - \mathbf{F}_n\boldsymbol{\theta})\right]. \tag{1.18}$$

We obtain the ML estimates for the parameters:

$$(\boldsymbol{\theta}, \widehat{\sigma}_\mu^2, \widehat{\sigma}_\varepsilon^2, \widehat{\boldsymbol{\psi}}) = \max_{\boldsymbol{\theta},\sigma_\mu^2,\sigma_\varepsilon^2,\boldsymbol{\psi}} l(\boldsymbol{\theta},\sigma_\mu^2,\sigma_\varepsilon^2,\boldsymbol{\psi}).$$

Asymptotics

There are two different asymptotic frameworks applicable to spatial data: **increasing domain asymptotics** and **infill asymptotics**. In increasing domain asymptotics,

the minimum distance between sampling points is bounded away from zero. Therefore, the spatial domain of observation is also increasing; while in infill asymptotics, observations are taken in a fixed and bounded domain, more and more densely.

Consider a spatial process (1.12) with $E(Y(\mathbf{t}))$ satisfying (1.13). The distribution of observations then depends on the parameter $\boldsymbol{\theta} \in R^p$, where $p \in Z^+$. As discussed in Zhang and Zimmerman (2005), the asymptotic behaviour of parameter estimators can be different under increasing domain asymptotics and infill asymptotics. An example is the maximum likelihood estimator (MLE) of the parameter $\boldsymbol{\theta}$.

The work in Mardia and Marshall (1984) is fundamental in discussing the asymptotic properties of a MLE of $\boldsymbol{\theta}$. Under an increasing domain asymptotic framework they showed that if $Y(\mathbf{t})$ is Gaussian the maximum likelihood estimator of $\boldsymbol{\theta}$ is approximately normal with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{I}_n^{-1}(\boldsymbol{\theta}, \boldsymbol{\psi})$ under some regularity conditions. Here,

$$\mathbf{I}_n(\boldsymbol{\theta}, \boldsymbol{\psi}) = E\left(-\frac{\partial^2 l_n(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathbf{y}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = \mathbf{F}_n^T \boldsymbol{\Sigma}_{nn}^{-1}(\boldsymbol{\psi}) \mathbf{F}_n$$

with $l_n(\boldsymbol{\theta}; \boldsymbol{\psi}, Y_n)$ being defined in (1.18).

One of the regularity condition is that the information matrix satisfies

$$\mathbf{I}_n^{-1}(\boldsymbol{\theta}, \boldsymbol{\psi}) \to \mathbf{0} \text{ as } n \to \infty. \tag{1.19}$$

Condition (1.19) is satisfied if the following two conditions hold (Mardia and Marshall 1984)

(i) $\lim_{n\to\infty} \lambda_n = C < \infty$ where λ_n is the maximum eigenvalue of $\boldsymbol{\Sigma}_{nn}(\boldsymbol{\psi})$;

(ii) $\lim_{n\to\infty} (\mathbf{F}_n^T \mathbf{F}_n)^{-1} = \mathbf{0}$.

Condition (ii) can hold when the regression function is suitably chosen. Condition (i) is guaranteed if the row sum norm of $\boldsymbol{\Sigma}_{nn}$ is finite for all n. We consider a simple example when $\mathcal{T}$ is an $N_1 \times ... \times N_d$ regular lattice. Moreover, we assume that the spatial process $Y(\mathbf{t})$ is a covariance stationary process in $\mathbb{R}^d$ with covariance function $\sigma(\mathbf{t}, \mathbf{t}+\mathbf{h}) = \sigma(\mathbf{h})$. Let $\sigma_i = \sum_{j=1}^n |\sigma(\mathbf{h}_{ij})|$ where $\sigma(\mathbf{h}_{ij}) = \sigma(\mathbf{t}_i, \mathbf{t}_j)$ with $\mathbf{h}_{ij} = \mathbf{t}_j - \mathbf{t}_i$. Then condition (i) is guaranteed when

$$\lim_{n\to\infty} \max_{i=1,...,n} \sigma_i < \infty. \tag{1.20}$$

Under infill asymptotics, condition (1.20) may not hold.

Example 8 *On an 1-dimensional lattice, consider a correlation function $\sigma(\cdot, \cdot)$ satisfying*

$$\sigma(t, t+h) = (1+h)^{-2}.$$

Under increasing domain asymptotics, assume that on the lattice the minimum distance between any two sampling points is longer than $r > 0$. we have $|t_i - t_j| = h_{ij} \geq |j - i|r > 0$ and

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \sum_{j=1}^{n} \sigma(h_{ij}) \leq \lim_{n\to\infty} \sum_{j=1}^{n} \frac{2}{(1+rj)^2} < \infty.$$

Under infill asymptotics, we sample points from a finite domain $[0,1]$ by $t_i = (i-1)/n$. Then we have

$$\max_{i=1,\ldots,n} \sum_{j=1}^{n} \sigma(h_{ij}) \geq \sum_{j=1}^{n} \sigma(h_{1j}) = \sum_{j=1}^{n} \left(1 + \frac{j-1}{n}\right)^{-2} \to \infty$$

where $\sigma(h_{1j}) = \sigma(t_1, t_j)$ and $t_1 = 0$.

Infill asymptotics of universal kriging Fazekas and Kukush (2005) studied a linear geostatistical model,

$$Y(\mathbf{t}) = \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} + \boldsymbol{\delta}(\mathbf{t}) + \varepsilon(\mathbf{t}) = \mu(\mathbf{t}) + \varepsilon(\mathbf{t}).$$

They investigated the properties of universal kriging when the locations of the observations are not known precisely, i.e. the locations are no longer non-random points but follow a known distribution. Fazekas and Kukush (2005) proposed a modification of the universal kriging, $\widetilde{\mu}(\mathbf{t})$, and proved that the mean squared error, $E\left[\mu(\mathbf{t}) - \widetilde{\mu}(\mathbf{t})\right]^2$, approaches zero as the number of observations increases. Their method, however, can be used in the case where the locations are non-random points. Following the procedure in Fazekas and Kukush (2005), we can derive the asymptotic property of the regular universal kriging.

The universal kriging estimate $\widehat{\mu}(\mathbf{t})$ of $\mu(\mathbf{t})$ is $\mathbf{a}_n^T(\mathbf{t})\mathbf{y}_n$ with $\mathbf{a}_n(\mathbf{t})$ being defined in (1.17). Notice that according to unbiasedness we should have $\mathbf{f}^T(\mathbf{t}) = \mathbf{a}_n^T(\mathbf{t})\mathbf{F}_n$. Moreover, if $f_1(\mathbf{t}) = 1$ in $\mathbf{f}(\mathbf{t}) = (f_1(\mathbf{t}), \cdots, f_p(\mathbf{t}))^T$, we have $\mathbf{a}_n^T(\mathbf{t})\mathbf{1}_n = 1$. Therefore, the mean squared error of $\widehat{\mu}(\mathbf{t})$ is

$$\begin{aligned}
& E\left[\mu(\mathbf{t}) - \widehat{\mu}(\mathbf{t})\right]^2 \\
= \; & E\left[\mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} + \delta(\mathbf{t}) - \mathbf{a}_n^T(\mathbf{t})\mathbf{F}_n\boldsymbol{\theta} - \mathbf{a}_n^T(\mathbf{t})\boldsymbol{\delta}_n - \mathbf{a}_n^T(\mathbf{t})\varepsilon_n\right]^2 \\
= \; & E\left[\delta(\mathbf{t}) - \mathbf{a}_n^T(\mathbf{t})\boldsymbol{\delta}_n\right]^2 + \sigma_\varepsilon^2 \mathbf{a}_n^T(\mathbf{t})\mathbf{a}_n(\mathbf{t}) \\
= \; & E\left[\mathbf{a}_n^T(\mathbf{t})\delta(\mathbf{t})\mathbf{1}_n - \mathbf{a}_n^T(\mathbf{t})\boldsymbol{\delta}_n\right]^2 + \sigma_\varepsilon^2 \mathbf{a}_n^T(\mathbf{t})\mathbf{a}_n(\mathbf{t}) \\
= \; & \mathbf{a}_n^T(\mathbf{t}) E\left[\delta(\mathbf{t})\mathbf{1}_n - \boldsymbol{\delta}_n\right]^2 \mathbf{a}_n(\mathbf{t}) + \sigma_\varepsilon^2 \mathbf{a}_n^T(\mathbf{t})\mathbf{a}_n(\mathbf{t}). \qquad (1.21)
\end{aligned}$$

Assume that there is a sequence of r positive integers with $r = r(n)$ such that a subset in $\mathcal{T}$ are locations $\left\{\mathbf{t}_1^{(n)}, \mathbf{t}_2^{(n)}, ..., \mathbf{t}_r^{(n)}\right\}$ satisfying the following properties:

$$\alpha_j \leq \frac{1}{r}, j = 1, ..., r, \tag{1.22}$$

$$\lim_{n\to\infty} \sum_{j=1}^{r} \alpha_j = O(n^{-1}) \tag{1.23}$$

where

$$\alpha_j = \max_{0\leq i\leq p} \left\| f_i(\mathbf{t}) - f_i\left(\mathbf{t}_j^{(n)}\right)\right\|, j = 1, ..., r.$$

As remarked in Fazekas and Kukush (2005) assumptions (1.21) and (1.22) imply that the sequence $\left\{\mathbf{t}_1^{(n)}, \mathbf{t}_2^{(n)}, ..., \mathbf{t}_r^{(n)}\right\}$ converges to an unsampled location $\mathbf{t}_*$ in some sense. Moreover, assume that in $\mathcal{T}$ there are $p+1$ locations $\{\mathbf{t}_0, \mathbf{t}_1, ..., \mathbf{t}_p\}$ such that

$$\text{the matrix } \widetilde{\Phi} = (f_i(\mathbf{t}_j))_{i,j=0}^{p} \text{ is invertible} \tag{1.24}$$

and

$$\left\|\widetilde{\Phi}^{-1}\right\| \leq K, \tag{1.25}$$

where $\|\cdot\|$ denotes the spectral norm of a matrix and K is a finite constant not depending on n. This condition implies that the locations $\mathbf{t}_0, \mathbf{t}_1, ..., \mathbf{t}_p$ do not converge to $\mathbf{t}_*$ (Fazekas and Kukush, 2005). Let $\widetilde{\mathbf{T}} = \{\mathbf{t}_1^{(n)}, \mathbf{t}_2^{(n)}, ..., \mathbf{t}_r^{(n)}, \mathbf{t}_0, \mathbf{t}_1, ..., \mathbf{t}_p\}$ and $\widetilde{\boldsymbol{\delta}} = \left(\delta(\mathbf{t}_1^{(n)}), ..., \delta(\mathbf{t}_r^{(n)}), \delta(\mathbf{t}_0), ..., \delta(\mathbf{t}_p)\right)$. Assume that each element in $\Gamma(\widetilde{\mathbf{T}}) := E\left[\delta(\mathbf{t}_*)\mathbf{1}_n - \widetilde{\boldsymbol{\delta}}\right]^2$ satisfies

$$\left|\left(\Gamma(\widetilde{\mathbf{T}})\right)_{i,j}\right| \leq M, \quad i, j = 1, ..., r + p + 1 \tag{1.26}$$

where M is a finite constant not depending on n. Moreover,

$$\lim_{n\to\infty} M_1 = \lim_{n\to\infty} \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} \left|\left(\Gamma(\widetilde{\mathbf{T}})\right)_{i,j}\right| = 0. \tag{1.27}$$

Based on all these assumptions, via a similar proof as that of Theorem 3 in Fazekas and Kukush (2005) we can obtain the following result.

Theorem 9 *Assume that (1.22)-(1.27) are satisfied. Then the mean squared error of universal kriging estimate at an unsampled location* $\mathbf{t}_*$ *approches* 0*, i.e.*

$$\lim_{n\to\infty} E\left[\mu(\mathbf{t}\mathbf{t}_*) - \widehat{\mu}(\mathbf{t}_*)\right]^2 = 0. \tag{1.28}$$

REFERENCES

Atkinson, A.C., and Fedorov, V.V. (1975a), "The Design of Experiments for Discriminating Between Two Rival Models," *Biometrika*, 62, 57-70.

Atkinson, A.C., and Fedorov, V.V. (1975b), "Optimal Design: Experiments for Discriminating Between Several Models," *Biometrika*, 62, 289-303.

Box, G. E. P. (1953), Non-Normality and Tests on Variances, *Biometrika*, 40, 318-335.

Box, G.E.P., and Draper, N.R. (1959), "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, 54, 622-654.

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

Fazekas, I., and Kukush, A. G. (2005), "Kriging and measurement errors. Discussiones Mathematicae," *Probability and Statistics*, 25, 139-159.

Fedorov, V. V. and Pazman, A. (1968), "Design of physical experiments," *Fortschritte der Physik*, 16, 325-355.

Ford, I., Titterington, D. M., and Kitsos, C. P. (1989), "Recent Advances in Nonlinear Experimental Design," *Technometrics*, 31, 49-60.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Hunter, W. G., and Reiner, A. M. (1965), "Designs for discriminating between two rival models," *Technometrics*, 7, 307-323.

Legendre, P., Dale, M. R., Fortin, M. J., Casgrain, P., and Gurevitch, J. (2004), "Effects of spatial structures on the results of field experiments," *Ecology*, 85, 3202-3214.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007), "An Optimal Experimental Design Criterion for Discriminating Between Non-Normal Models," *Journal of the Royal Statistical Society* B, 69, 231-242.

Lubarsky, D. A., Smith, L. R., Sladen, R. N., Mault, J. R., and Reed, R. L. (1995), "Defining the relationship of oxygen delivery and consumption: use of biologic system models," *Journal of Surgical Research*, 58, 503-508.

Mardia, K. V., and Marshall, R. J. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135-146.

Matheron, G. (1963). "Principles of geostatistics," *Economic geology*, 58, 1246-1266.

Matheron, G. (1967). "Kriging or polynomial interpolation procedures," *Transactions of the Canadian Institute of Mining and Metallurgy*, 70, 240-244.

Montgomery, D.C. (1984), *Design and analysis of experiments*. New York: Wiley.

Müller, W.G. (2005), "A comparison of spatial design methods for correlated observations," *Environmetrics*, 16, 495-505.

Petraitis, P. (2001), "Designing experiments that control for spatial and temporal variation," *Methodology Paper Series of the 4th International Conference on ILTER in East Asia and Pacific Region*, Ulaan-Hatgal, Mongolia, 12, 273-287.

Santner, T.J., Williams, B.J. and Notz, W.I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer Verlag.

Taplin, R.H. (1999), "Theory & Methods: An Analysis of Two-Dimensional Agricultural Field Trials in the Presence of Outliers and Fertility Jumps," *Australian & New Zealand Journal of Statistics*, 41, 285-297.

Thompson, S. K. (1997), "Effective sampling strategies for spatial studies," *Metron*, 55, 1–21.

Uciński, D. and Bogacka, B. (2005), "T-optimum designs for discrimination between two multiresponse dynamic models," *Journal of the Royal Statistical Society* B, 67, 3–18.

Wiens, D. and Zhou, J. (2008), "Robust estimators and designs for field experiments," *Journal of Statistical Planning and Inference*, 138, 93–104.

Wiens, D. (2009a), "Robust discrimination designs," *Journal of the Royal Statistical Society (Series B)*; 71, 805-829.

Wiens, D. (2009b), "Asymptotic Properties of a Neyman-Pearson Test for Model Discrimination, with an Application to Experimental Design," *Journal of Statistical Theory and Practice*, 3, 419-427.

Wiens, D. (2015a), *Robustness of Design*, Chapter 20 of Handbook of Design and Analysis of Experiments, Chapman & Hall/CRC

Wiens, D. (2015b), *Optimal Designs for Nonlinear and Spatial Models: Introduction and Historical Overview*, Chapter 12 of Handbook of Design and Analysis of Experiments, Chapman & Hall/CRC.

Wu, T., and Dutilleul, P. (1999), "Validity and efficiency of neighbour analyses in comparison with classical complete and incomplete block analyses of field experiments," *Agronomy J.,* 91, 721-731.

Zhang, H., and Zimmerman, D. L. (2005), "Towards reconciling two asymptotic frameworks in spatial statistics," *Biometrika*, 92, 921-936.

Chapter 2

Robust discrimination designs over Hellinger neighbourhoods

Abstract To aid in the discrimination between two, possibly nonlinear, regression models, we study the construction of experimental designs. Considering that each of these two models might be only approximately specified, robust "maximin" designs are proposed. The rough idea is as follows. We impose neighbourhood structures on each regression response, to describe the uncertainty in the specifications of the true underlying models. We determine the least favourable – in terms of Kullback-Leibler divergence – members of these neighbourhoods. Optimal designs are those maximizing this minimum divergence. Two particular cases are investigated and in each case sequential approaches are studied. Asymptotic optimality is established.

Key words and phrases Hellinger distance; Kullback–Leibler; Maximin; Michaelis–Menten model; Model discrimination; Neyman–Pearson test; Non-linear regression; Optimal design; Robustness; Sequential design

2.1 Introduction

Much of the experimental work in scientific disciplines – physics, chemistry, engineering, etc. – is concerned with the elucidation of a functional relationship between a response variable y and various covariates $\mathbf{x}$. However, in practice it is often the case that the investigator will not know the correct functional form, but instead will have several plausible models in mind. A first aim of the investigator is therefore to design an experiment distinguishing among these rival models. Specifically, we assume that two rival models are available. Under the first model, the data arise from a population

with density $f_0(y|\mathbf{x},\boldsymbol{\varphi}_0)$ while under the other model the density is $f_1(y|\mathbf{x},\boldsymbol{\varphi}_1)$; the conditional means are

$$\mu_j(\mathbf{x}) = \int y f_j(y|\mathbf{x},\boldsymbol{\varphi}_j)dy,\ j=0,1. \tag{2.1}$$

Here $\boldsymbol{\varphi}_0$ and $\boldsymbol{\varphi}_1$ represent nuisance parameters and will not be explicitly mentioned if there is no possibility of confusion. Given a design space $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$, assume that $y_{il}, l = 1, ..., n_i \geq 0$, are observations made at the covariate $\mathbf{x}_i$. Usually, the model discrimination problem is cast as a problem of hypothesis testing (Atkinson and Fedorov 1975a,b; Fedorov 1975):

$$H_0:\ f_0(y|\mathbf{x},\mu_0(\mathbf{x}))\ \text{ versus }\ H_1:\ f_1(y|\mathbf{x},\mu_1(\mathbf{x})),\ \mathbf{x}\in\mathcal{S}.$$

The Neyman-Pearson test then can be used to compare these two hypotheses. Define $\mathcal{R} = \sum_{i,l} R(y_{il})$ with

$$R(y_{il}) = 2\log\left\{\frac{f_1(y_{il}|\mathbf{x}_i,\mu_1(\mathbf{x}_i))}{f_0(y_{il}|\mathbf{x}_i,\mu_0(\mathbf{x}_i))}\right\}. \tag{2.2}$$

We reject H_0 for large values of $\mathcal{R}$. We shall assume that the experimenter models $\mu_j(\mathbf{x})$ parametrically as $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$, with the form of $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ specified but the parameters $\boldsymbol{\theta}_j$ unknown.

The design of experiments for discriminating between rival models has been investigated by numerous authors, among them Fedorov (1975) and Hill (1978). Sequential and static designs are the two most well-studied strategies. In particular, Hunter and Reiner (1965) proposed a sequential design assuming that both densities were Gaussian, viz., $f_j(y|\mathbf{x},\mu_j,\boldsymbol{\sigma}) = \sigma^{-1}\phi\left(\left(y-\mu_j(\mathbf{x})\right)/\sigma\right),\ j=0,1$. Fedorov and Pazman (1968) extended the method to heteroscedastic models. Static, i.e. non-sequential, design strategies were constructed under the normality assumption by Atkinson and Fedorov (1975a,b). López-Fidalgo, Tommasi and Trandafir (2007) extended static design to non-normal models. The criteria to be optimized in these works are seen to be equivalent to the integrated Kullback-Leibler (KL) divergence $\mathcal{D}$:

$$\mathcal{D}(f_0,f_1,\xi|\mu_0,\mu_1) = \int_{\mathcal{S}} \mathcal{I}\{f_0,f_1|\mathbf{x},\mu_0(\mathbf{x}),\mu_1(\mathbf{x})\}\xi(d\mathbf{x}), \tag{2.3}$$

with ξ being the design measure placing mass $\xi_i = n_i/n$ at $\mathbf{x}_i$, and

$$\mathcal{I}\{f_0,f_1|\mathbf{x},\mu_0(\mathbf{x}),\mu_1(\mathbf{x})\} = \int_{-\infty}^{\infty} f_1(y|\mathbf{x},\mu_1(\mathbf{x}))\log\left\{\frac{f_1(y|\mathbf{x},\mu_1(\mathbf{x}))}{f_0(y|\mathbf{x},\mu_0(\mathbf{x}))}\right\}dy \tag{2.4}$$

being the Kullback-Leibler divergence measuring the information lost when $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ is used to approximate $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

In the references above, one of $f_j(y|\mathbf{x},\mu_j)$, $j = 0, 1$, is assumed to correctly represent the true physical mechanism. However, it is dangerous to apply a method that is highly dependent on a specific form (Box and Draper 1959; Huber 1981; Ford, Titterington and Kitsos 1989). From a viewpoint of robustness it is more sensible to suppose only that the correct model lies in a neighbourhood of a specified density. Wiens (2009a) allowed for the means (but not the f_j) to be specified erroneously, and imposed the neighbourhood structure

$$\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j) + \psi_j(\mathbf{x}) \tag{2.5}$$

for specified $\eta_j(\mathbf{x}|\cdot)$ The vectors $\boldsymbol{\psi}_j = (\psi_j(\mathbf{x}_1), ..., \psi_j(\mathbf{x}_N))'$ were allowed to range over classes $\boldsymbol{\Psi}_j$, resulting in the neighbourhoods

$$\mathcal{F}_j = \{f_j(\cdot|\mathbf{x}, \mu_j) \,|\mu_j(\mathbf{x}_i) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j) + \psi_j(\mathbf{x}), \boldsymbol{\psi}_j \in \boldsymbol{\Psi}_j\}, \ j = 0, 1.$$

Under this setting robust Kullback-Leibler optimal designs were obtained in Wiens (2009a) by maximizing the minimum asymptotic power of the Neyman-Pearson test statistic $\mathcal{R}$ over $\mathcal{F}_0$ and $\mathcal{F}_1$. The asymptotic properties of $\mathcal{R}$ were derived in Wiens (2009b) for two rival models with common densities $f_j(y|\mathbf{x}, \mu_j(\mathbf{x})) = f(y|\mathbf{x}, \mu_j(\mathbf{x}))$.

Our work is a natural sequel to Wiens (2009a). Model misspecification is still the problem we would like to address but under a more general scenario as follows. The two rival models are $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}))$, $j = 0, 1$, with $\mu_j(\mathbf{x})$ determined by (2.1) and assumed to be of the form $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ for some $\boldsymbol{\theta}_j$. i.e. $\boldsymbol{\psi}_j \equiv 0$. Define $\mathcal{F}_j$ to be neighbourhoods of $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}))$ used to describe inaccuracies in the specifications of the true underlying densities. The true model lies in one of $\mathcal{F}_j$, $j = 0, 1$. It is our purpose in this paper to propose methods of discrimination design which are robust against the possible model misspecification mentioned above.

There are two possible scenarios:

Case I: under the null hypothesis the density function $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ of the response variable is fixed and its mean μ_0 is as defined in (2.1); under the alternative hypothesis the density function varies over a Hellinger neighbourhood of a nominal density

$f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$. Recall that the Hellinger distance $d_h(f, g)$ between densities f, g is defined by

$$d_h^2(f, g) = \frac{1}{2}\int \left(f^{1/2}(y) - g^{1/2}(y)\right)^2 dy = 1 - \int \sqrt{f(y)g(y)}dy.$$

Here the two classes are $\mathcal{F}_0 = \{f_0(y|\mathbf{x},\mu_0(\mathbf{x})\}$ and $\mathcal{F}_1$ is a Hellinger neighbourhood defined as

$$\mathcal{F}_1(\varepsilon_1) = \{f(y|\mathbf{x})|\max_{\mathbf{x}\in\mathcal{S}} d_h(f(y|\mathbf{x}), f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))) \leq \varepsilon_1\} \tag{2.6}$$

for some $\varepsilon_1 > 0$. The members of $\mathcal{F}_1$ may differ from f_1 because of differences in the functional form of the density, or in their mean structures, or both.

Case II: under the null hypothesis, the response variable has density $f(y|\mathbf{x})$ varying over a Hellinger neighbourhood of a nominal density $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$. The members of $\mathcal{F}_0$ may differ from f_0 because of differences in the functional form of the density, or in their mean structures, or both. Under the alternative hypothesis the density function is $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ with the mean μ_1 defined in (2.1). In this case, the two classes $\mathcal{F}_0(\varepsilon_0)$ and $\mathcal{F}_1$ are defined as

$$\mathcal{F}_0(\varepsilon_0) = \left\{ f(y|\mathbf{x}) \left| \max_{\mathbf{x}\in\mathcal{S}} d_h(f(y|\mathbf{x}), f_0(y|\mathbf{x}, \mu_0(\mathbf{x})) \leq \varepsilon_0 \right.\right\}, \tag{2.7}$$

for some $\varepsilon_0 > 0$, and $\mathcal{F}_1 = \{f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))\}$, respectively.

Thus Case I fixes the null model and allows the alternate to vary over a Hellinger class; in Case II these are reversed. In the upcoming proposition and corollary, we will show that the Neyman-Pearson test for discriminating between any pair in $\mathcal{F}_0 \times \mathcal{F}_1$ is related to the KL-divergence defined in (2.3) between the pair of densities.

Asymptotic properties of the test statistic $\mathcal{R}$

In the asymptotics literature one finds numerous results about the asymptotic distribution of $\mathcal{R}$, the test statistic for the discrimination between a pair of models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$, under various conditions. Wiens (2009b) proved the asymptotic normality of the test statistic $\mathcal{R}$ under standard regularity conditions for likelihood estimation. In Oosterhoff and van Zwet (2012) similar results are proved under certain contiguity assumptions. In the appendix we derive the asymptotic distribution of $\mathcal{R}$ under conditions tailored to our problem. The proof follows that in Oosterhoff and van Zwet (2012).

Proposition 10 *Given a design space* $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$ *, assume that the experiment has* n_i *replicates at each covariate* $\mathbf{x}_i$*, with* $\sum_{i=1}^N n_i = n$*. Define* $\mathcal{D}$ *as in (2.3) and for any two densities* f_0, f_1 *define*

$$r\left(y|\mathbf{x}_i; f_0, f_1\right) = \frac{f_1(y|\mathbf{x}_i, \mu_1\left(\mathbf{x}_i\right))}{f_0(y|\mathbf{x}_i, \mu_0\left(\mathbf{x}_i\right))}.$$

Assume that two densities f_0, f_1 *satisfy*

(a) for the KL-divergence,

$$n\mathcal{D} = O(1), \tag{2.8}$$

(b) for all $\delta > 0$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i \int_{\{|\log r(y|\mathbf{x}_i; f_0, f_1)| \geq \delta\}} f_0(y|\mathbf{x}_i, \mu_0\left(\mathbf{x}\right)) \left(\sqrt{r\left(y|\mathbf{x}_i; f_0, f_1\right)} - 1\right)^2 dy = 0, \tag{2.9}$$

(c) there is a $\tau > 0$ *such that*

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i \int_{\{\log r(y|\mathbf{x}_i; f_0, f_1) \geq \tau\}} f_1(y|\mathbf{x}_i, \mu_1\left(\mathbf{x}\right)) \log r\left(y|\mathbf{x}_i; f_0, f_1\right) dy = 0. \tag{2.10}$$

Then $\mathcal{R} = 2\sum_{i=1}^N \sum_{l=1}^{n_i} \log r(y|\mathbf{x}_i; f_0, f_1)$ *and:*

(i) when the null hypothesis is true,

$$\frac{\mathcal{R} + 2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1)$$

(ii) when the alternative hypothesis is true,

$$\frac{\mathcal{R} - 2n\mathcal{D}}{\sqrt{8n\mathcal{D}}} \xrightarrow{L} N(0,1).$$

Remark 1 Conditions (2.8)-(2.10) in Proposition 10 hold in particular when the two densities $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}), \sigma) = \phi\left(\left(y - \mu_j(\mathbf{x})\right)/\sigma\right)/\sigma$, $j = 0, 1$, have means which satisfy $\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + n^{-1/2}\Delta(\mathbf{x})$ for a bounded function Δ. In particular, condition (2.10) holds for every $\tau > 0$. The same conclusion holds for log-normal densities $f_j(y|\mathbf{x}, \mu_j(\mathbf{x}), v_j^2)$ with $\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + n^{-1/2}\Delta(\mathbf{x})$ and homogeneous variances $v_j^2 = v^2$. These statements are verified in §A.2, A.3 of the Appendix.

Remark 2 Denote by F a distribution function whose density is f. To guarantee that the KL divergence between two densities f_0 and f_1 is finite, F_1 should be absolutely continuous with respect to F_0. Moreover, it is natural to assume that the two rival

models are close to each other in some sense. Therefore, in the following we let $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ have the same support set $\Omega_{\mathbf{x}}$, and its complement set is

$$\Omega_{\mathbf{x}}^c = \{y : f_1(y|\mathbf{x}, \mu_1(\mathbf{x})) = f_0(y|\mathbf{x}, \mu_0(\mathbf{x})) = 0\}.$$

Then to make sure that condition (2.8) is reasonable, F_1 should be absolutely continuous with respect to F_0. Therefore, we only consider $f(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$ such that $f(y|\mathbf{x}) = 0$ on $\Omega_{\mathbf{x}}^c$. For simplicity we assume that the densities we consider in this paper are continuous in the interiors of their support sets.

If the radii ε_j of Hellinger neighbourhoods $\mathcal{F}_0(\varepsilon_0)$ of $f_0(\cdot|\mathbf{x}, \mu_0(\mathbf{x}))$ and $\mathcal{F}_1(\varepsilon_1)$ of $f_1(\cdot|\mathbf{x}, \mu_1(\mathbf{x}))$ shrink at a rate $o(n^{-1/2})$ then the results in Proposition 10 also hold for any pair of densities in $\mathcal{F}_0(\varepsilon_0) \times \mathcal{F}_1(\varepsilon_1)$. This is guaranteed by the result in the following corollary.

Corollary 11 *Assume that the central densities $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ satisfy conditions (2.8)-(2.10) in Proposition 10 and that $\varepsilon_0 = o(n^{-1/2})$, $\varepsilon_1 = o(n^{-1/2})$. Then for any pair $(f^{(0)}(y|\mathbf{x}), f^{(1)}(y|\mathbf{x})) \in \mathcal{F}_0(\varepsilon_0) \times \mathcal{F}_1(\varepsilon_1)$ satisfying (2.10) we have that $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ also satisfy conditions (2.8) and (2.9).*

The main results in Proposition 10 and Corollary 11 show the asymptotic normality of the statistic

$$\mathcal{R}\left(f^{(0)}, f^{(1)}\right) = 2\sum_{i=1}^{N}\sum_{l=1}^{n_i} \log r(y|\mathbf{x}_i; f^{(0)}, f^{(1)})$$

under $f^{(0)}(y|\mathbf{x}) \in \mathcal{F}_0(\varepsilon_0)$ or $f^{(1)}(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$, i.e. the density of the observation variable Y is $f^{(0)}(y|\mathbf{x})$ or $f^{(1)}(y|\mathbf{x})$. In practice, we are more interested in the asymptotic normality of the test statistic $\mathcal{R} := \mathcal{R}(f_0, f_1)$ for the discrimination of the two nominal models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$, when, however, the true model is in $\mathcal{F}_0(\varepsilon_0)$ or $\mathcal{F}_1(\varepsilon_1)$. In the following theorem we show that the asymptotic normality of $\mathcal{R}$ still holds under any density $f \in \mathcal{F}_0(\varepsilon_0)$ or $\mathcal{F}_1(\varepsilon_1)$.

Proposition 12 *Assume that two models $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ and $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ satisfy conditions (2.8)-(2.10) in Proposition 10. Then*
(i) under $f \in \mathcal{F}_1(\varepsilon_1)$,

$$\frac{\mathcal{R} - 2n\mathcal{D}(f_0, f)}{\sqrt{8n\mathcal{D}(f_0, f)}} \xrightarrow{L} N(0, 1);$$

(ii) under $f \in \mathcal{F}_0(\varepsilon_0)$,

$$\frac{\mathcal{R} + 2n\mathcal{D}(f, f_1)}{\sqrt{8n\mathcal{D}(f, f_1)}} \xrightarrow{L} N(0, 1).$$

Based on the results in Proposition 12, we obtain the asymptotic power against a density $f \in \mathcal{F}_0$ or $\mathcal{F}_1$ in the following theorem.

Theorem 13 *The asymptotic power against a density* $f \in \mathcal{F}_0$ *or* $\mathcal{F}_1$ *is*

$$\pi(f) := P_f(\mathcal{R}(f_0, f_1) > c) = \Phi\left(\frac{-c + \gamma(f, f_0, f_1)}{2\sqrt{|\gamma(f, f_0, f_1)|}}\right) + o(1),$$

where $P_f(\cdot)$ *means that the calculations are to be made assuming that* f *is the density of* Y*, and where*

$$\gamma(f, f_0, f_1) = \begin{cases} -2n\mathcal{D}(f, f_1), & \text{if } f \in \mathcal{F}_0, \\ 2n\mathcal{D}(f_0, f), & \text{if } f \in \mathcal{F}_1. \end{cases}$$

The critical value is

$$c = -2n\mathcal{D}(f_0, f_1) + u_\alpha\sqrt{8n\mathcal{D}(f_0, f_1)}$$

determined by

$$\alpha = P_{f_0}(\mathcal{R}(f_0, f_1) > c)$$

with u_α *being the* $(1-\alpha)$*-quantile of the standard normal distribution.*

Design criteria for Case I

In Case I, where $\mathcal{F}_1(\varepsilon_1)$ is a neighbourhood of f_1, the design problem is to find a design to maximize the 'worst' power with controlled Type I error. In particular, a *robust maximin* design $\boldsymbol{\xi}^*$ is constructed which maximizes the minimum power, with significance level α, over $\mathcal{F}_1(\varepsilon_1)$, i.e.

$$\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi}} \min_{f_1 \in \mathcal{F}_1(\varepsilon_1)} P_{f_1}(\mathcal{R} > c), \text{ subject to } \alpha = P_{f_0}(\mathcal{R} > c) \tag{2.11}$$

where c is the critical value defining the rejection region $\{\mathcal{R} > c\}$. According to Theorem 13, asymptotically, the robust design is the solution to the following optimality problem

$$\boldsymbol{\xi}^* = \arg\max \min_{f \in \mathcal{F}_1(\varepsilon_1)} \pi(f) \tag{2.12}$$

and the minimum asymptotic power is

$$\min_{f\in\mathcal{F}_1(\varepsilon_1)} \Phi\left(\frac{-c+2n\mathcal{D}(f_0,f)}{2\sqrt{2n\mathcal{D}(f_0,f)}}\right) \tag{2.13}$$

where c is defined in Theorem 13. Under certain condition, we can solve the minimization problem by minimizing the integrated KL-divergence $\mathcal{D}(f_0,f)$ as shown in the following proposition.

Proposition 14 *Define*

$$f_{1*} = \arg\min_{f\in\mathcal{F}_1(\varepsilon_1)} \mathcal{D}(f_0,f). \tag{2.14}$$

If

$$\mathcal{D}(f_0,f_{1*}) \geq -c, \tag{2.15}$$

then also f_{1*} minimizes $\pi(f)$ in $\mathcal{F}_1(\varepsilon_1)$, and so is the desired minimizer in (2.13).

Remark 3: If $c \geq 0$, then (2.15) is automatic. Otherwise, we check it numerically.

The problem now is to find f_{1*} as at (2.14), and then

$$\boldsymbol{\xi}^* = \arg\max \min_{f\in\mathcal{F}_1(\varepsilon_1)} \pi(f) = \arg\max \mathcal{D}(f_0,f_{1*}). \tag{2.16}$$

For Case II, we view the null hypothesis as composite, in the sense that f_0 is the representative of the whole neighbourhood and the nominal size of the test is evaluated at f_0, and is to be α. We should accept the null if it appears that the f generating the data is anything in $\mathcal{F}_0(\varepsilon_0)$. Then, if $f \in \mathcal{F}_0(\varepsilon_0)$ is generating the data we make an error if we reject the null hypothesis, and we would like to minimize the (maximum) probability of this. A *robust maximin* design $\boldsymbol{\xi}^*$ is then constructed which minimizes the maximum probability of such an error, with significance level α, over $\mathcal{F}_0(\varepsilon_0)$. In this paper, we will only focus Case I and Case II will be followed up in a future work.

As an illustration, we consider two models $f_0(y|\mathbf{x},\mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x},\mu_1(\mathbf{x}))$ with $\mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0)$ and $\mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)$. The design obtained from the criterion (2.16) is robust for testing the hypotheses

$$H_0 : f_0(y|\mathbf{x},\mu_0(\mathbf{x})) \; vs. \; H_1 : f_1(y|\mathbf{x},\mu_1(\mathbf{x}))$$

when the true model is in a small neighbourhood of one of the hypothesized models.

Of course the values of the regression parameters are unknown. To address this problem, Atkinson and Fedorov (1975a) (see also López-Fidalgo *et al.* 2007) assumes a range of plausible values for the parameters. The worst possible values of the parameters, i.e. those that minimize $\mathcal{D}$, are obtained within their respective ranges and the maximin optimal design that maximizes this minimum value is constructed. This method leads to static design strategies. In this paper, we proceed sequentially, with the parameters $\boldsymbol{\theta}_j$ replaced by updated least squares (LS) estimates $\hat{\boldsymbol{\theta}}_j$ before proceeding to the next stage. The next observation then will be made at the point $\mathbf{x}_{new}$ optimizing the discrepancy function (e.g. the KL divergence (2.4)) evaluated at the $\hat{\boldsymbol{\theta}}_j$. This is repeated until sufficiently many design points and observations are obtained. See Hunter and Reiner (1965) and Fedorov and Pazman (1968) for background material.

In Section 2, the minimization problem is solved analytically at each fixed $\mathbf{x} \in \mathcal{S}$. The maximization leading to optimal designs is done numerically in Section 3. A sequential discrimination design is proposed, with the unknown parameters in $\eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ updated as described above. To see how the test performs with our robust designs, we simulate the sizes and powers of the model discrimination test to discriminate between two models $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ when the true model may merely be close to the nominal model $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

Derivations and longer mathematical arguments are in the Appendix.

2.2 Minimization of the discrepancy function

Assume that

$$\varepsilon_1 < \min_{\mathbf{x}\in\mathcal{S}} d_h(f_0(y|\mathbf{x}, \mu_0(\mathbf{x})), f_1(y|\mathbf{x}, \mu_1(\mathbf{x}));$$

this ensures that $\mathcal{F}_0$ and $\mathcal{F}_1(\varepsilon_1)$ are disjoint – otherwise, the minimum power of the test is zero. That ε_1 is sufficiently small will be checked numerically in each example.

For the first step, we minimize $\mathcal{D}$ over the neighbourhood $\mathcal{F}_1(\varepsilon_1)$. Equivalently, we consider the optimization problem

$$\min_f \mathcal{D}(f_0, f, \boldsymbol{\xi}|\mu_0) = \min_f \sum_{i=1}^{N} \xi_i \int f(y|\mathbf{x}_i) \log\left(\frac{f(y|\mathbf{x}_i)}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right) dy \tag{2.17}$$

under the constraints (2.6) and $\int f(y|\mathbf{x})dy = 1$. (The requirement that f be non-negative turns out to be satisfied automatically and need not be prescribed.)

It is sufficient to find, for each $\mathbf{x}$, the minimizer $f_{1*}(y|\mathbf{x})$ of

$$\mathcal{I}\{f_0, f|\mathbf{x}, \mu_0(\mathbf{x})\} = \int f(y|\mathbf{x}) \log\left(\frac{f(y|\mathbf{x})}{f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))}\right) dy \tag{2.18}$$

subject to (i′) $\int \sqrt{f(y|\mathbf{x}) f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))} dy \geq 1 - \varepsilon_1^2$ and (ii′) $\int f(y|\mathbf{x}) dy = 1$. That is, we consider the optimality problem at each $\mathbf{x}$.

To solve this minimization problem, we adopt the Lagrange multiplier method and obtain the following result. For each $\mathbf{x}$, this gives a value $f_{1*}(y|\mathbf{x},\mu_{1*}(\mathbf{x}))$ of the least favourable density, with mean $\mu_{1*}(\mathbf{x})$ given by (2.1).

Proposition 15 *For $\mathbf{x} \in \mathcal{S}$ consider the system*

$$\log \frac{f(y|\mathbf{x})}{f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))} + 1 + \frac{1}{2}\lambda_1 f_1^{1/2}(y|\mathbf{x}, \mu_1(\mathbf{x})) f^{-1/2}(y|\mathbf{x}) + \lambda_2 = 0, \tag{2.19}$$

$$\int_{\Omega_{\mathbf{x}}} f(y|\mathbf{x}) dy = 1, \tag{2.20}$$

$$\int_{\Omega_{\mathbf{x}}} \sqrt{f(y|\mathbf{x}) f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))} dy = 1 - \varepsilon_1^2. \tag{2.21}$$

Define $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ as follows: for $y \in \Omega_{\mathbf{x}}^c$, $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x})) \equiv 0$ and for $y \in \Omega_{\mathbf{x}}$, $f_{1*}(y|\mathbf{x}, \mu_{1*}(\mathbf{x}))$ is a solution to (2.19)-(2.21) with $\lambda_1(\mathbf{x}) < 0$ and $\lambda_2(\mathbf{x}) \in \mathbb{R}$. Then $f_{1*}(y|\mathbf{x},\mu_{1*}(\mathbf{x}))$ is the minimizer of (2.18).

As examples, in Figure 2.1 we plot, for fixed values of $\mathbf{x}$, the densities $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ and the least favourable density $f_{1*}(y|\mathbf{x},\mu_{1*}(\mathbf{x}))$. In Figure 2.1(a) both f_0 and f_1 are normal; in Figure 2.1(b) both are log-normal. In both cases, the shape of the least favourable density obtained from Proposition 15 is, as one would expect, close to the nominal density $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$.

Based on results in Proposition 15, the answer to the optimality problem (2.17) is given in following theorem.

Theorem 16 *For a design $\boldsymbol{\xi}$ placing a fraction ξ_i of the observations at $\mathbf{x}_i$, the minimum divergence $\mathcal{D}$ over $\mathcal{F}_1$ is*

$$\sum_{i=1}^{N} \xi_i \mathcal{I}\left\{f_0, f_{1*}|\mathbf{x}_i, \mu_0(\mathbf{x}_i), \mu_1(\mathbf{x}_i)\right\},$$

where f_{1*} is as given in Proposition 15.

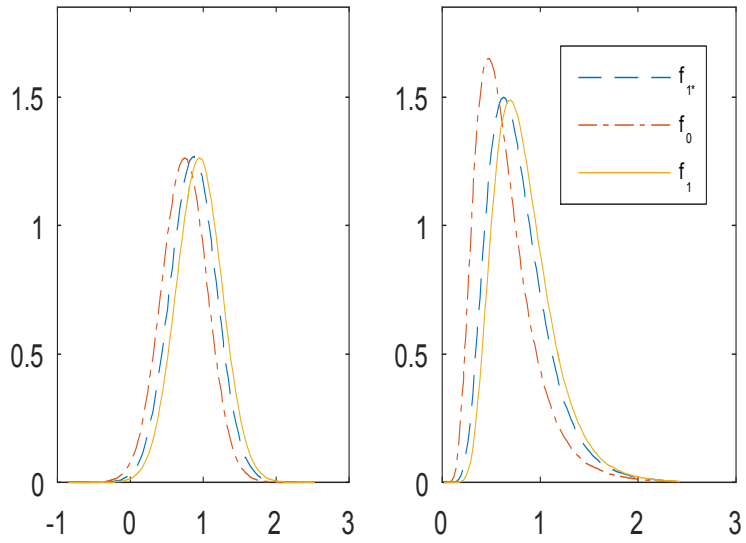

Figure 2.1: Dotted line represents $f_0(y|x,\mu_0)$ and solid line is $f_1(y|x,\mu_1)$ at $x = 2.8947$. (a) f_0 and f_1 are **normal** densities (see Example 1 in Section 2.2.1). (b) f_0 and f_1 are **lognormal** densities (see Example 2 in Section 2.2.2). Here f_0 and f_1 have mean functions $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ being given by (2.23) and (2.24), respectively, with $V = K = 1$. The two nominal densities have the same variance 0.1. Dashed line represents least favourable density in $\mathcal{F}_1(0.045)$.

2.3 Sequential discrimination designs

We employ a sequential approach to obtain the optimal design. To address the problem of unknown parameters, we adopt a 'working model' similar to that in Wiens (2009a).

The designs we propose are robust against the assumption that the hypotheses are correctly specified. To demonstrate this in finite samples, we shall simulate the sizes and minimum powers of the model discrimination tests based on our robust designs and those based on 'classically optimal' designs – those which entertain no neighbourhood structure on the models, i.e. $\varepsilon_0 = \varepsilon_1 = 0$. On this basis we will compare the methods.

Given a working 'null' model $\mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_{0true})$, we simulate the observations from either $f_0(y|\eta_0(\mathbf{x}|\boldsymbol{\theta}_{0true}))$ or from a chosen member $f(y|\mathbf{x})$ of $\mathcal{F}_1(\varepsilon_1)$. We simulate from f_0 when investigating the sizes of the tests associated with the robust and classically optimal designs. To investigate the minimum powers we simulate from $f_{1*}(y|\mathbf{x},\mu_{1*}(\mathbf{x}))$.

We construct sequential designs as follows.

Step 1: Choose an initial design $\boldsymbol{\xi}_0 = \{\xi_{0i}\}_{i=1}^N$ of size $n_{init} = \sum_{i=1}^N n_{i,init}$, where $\xi_{0i} = n_{i,init}/n_{init}$.

Step 2: Simulate n_i observations at each covariate $\mathbf{x}_i$.

Carry out steps 3-5, starting with $m = 0$, until an n-point design is obtained.

Step 3: Estimate both parameter vectors $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. In each case, estimation of $\boldsymbol{\theta}_j$ is done assuming that $\mu_j(\mathbf{x}) = \eta_j(\mathbf{x}|\boldsymbol{\theta}_j)$ exactly. We denote these estimates by $\widehat{\boldsymbol{\theta}}_m = \left(\widehat{\boldsymbol{\theta}}_{0m}, \widehat{\boldsymbol{\theta}}_{1m}\right)$.

Step 4: The next design point in the classical design is

$$\mathbf{x}_{new}^{(c)} = \arg\max_{\mathbf{x}\in\mathcal{S}} \mathcal{I}\left\{f_0, f_1 \left| \mathbf{x}, \mu_0(\mathbf{x}) = \eta_0\left(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{0m}\right), \mu_1(\mathbf{x}) = \eta_1\left(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{1m}\right)\right.\right\}.$$

The next design point in the robust design is

$$\mathbf{x}_{new}^{(r)} = \arg\max_{\mathbf{x}\in\mathcal{S}} \mathcal{I}\left\{f_0, f_{1*} \left| \mathbf{x}, \mu_0(\mathbf{x}) = \eta_0\left(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{0m}\right), \mu_1(\mathbf{x}) = \eta_1\left(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{1m}\right)\right.\right\}.$$

We note, but for ease of presentation do not emphasize, that the estimates $\widehat{\boldsymbol{\theta}}_m$ will depend on the designs used up to this point.

Step 5: Simulate y_{new} as described above.

These steps in the construction of sequential design will be illustrated in detail in the examples which follow. Before doing this, we state the related result that, under appropriate conditions, the sequential designs so obtained are asymptotically optimal. We entertain two sequences of models $f_{0n}(y|\mathbf{x}, \mu_0(\mathbf{x}))$, $f_{1n}(y|\mathbf{x}, \mu_1(\mathbf{x}))$ indexed by the sample size n. We assume that

$$|\eta_0(\mathbf{x}|\boldsymbol{\theta}_{0n}) - \eta_1(\mathbf{x}|\boldsymbol{\theta}_{1n})| = O(n^{-1/2}).$$

This guarantees, as in Remark 1, that if $f_{0n}(y|\mathbf{x}, \mu_0(\mathbf{x}))$ and $f_{1n}(y|\mathbf{x}, \mu_1(\mathbf{x}))$ are both normal or both lognormal densities with $\mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_{0n})$, $\mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_{1n})$ and the same nuisance parameters, then conditions (2.8) - (2.10) in Proposition 10 hold. It is also needed for (i) of Theorem 2, which is based on Theorem 3.1 in Sinha and Wiens (2003). By (i) of Theorem 17, the LSEs $\widehat{\boldsymbol{\theta}}_{jn}$ updated in each iteration are consistent for sequences $\boldsymbol{\theta}_{jn}$ of parameters defined as

$$\boldsymbol{\theta}_{jn} = \arg\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^N \{E[Y_n|\mathbf{x}_i] - \eta_j(\mathbf{x}_i|\boldsymbol{\theta})\}^2\right]. \tag{2.22}$$

In fact, $\widehat{\boldsymbol{\theta}}_{jn} - \boldsymbol{\theta}_{jn} \overset{a.s.}{\to} 0$ as shown in Sinha and Wiens (2003). With this consistency we then can obtain that the sequential designs $\{\boldsymbol{\xi}_n\}$ constructed as in Steps 1-5 above are asymptotically optimal. We require assumptions (B1)-(B5) and A3′ as stated in Sinha and Wiens (2003). Moreover, we have two additional assumptions:

(B6) For each fixed $\mathbf{x}$, the KL-divergence in Proposition 15, given by $\mathcal{I}\{f_0, f_{1*}|\mathbf{x}, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\}$, is Lipschitz continuous with respect to $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$.

(B7) The size of the initial sample n_{init} satisfies

$$\lim_{n_{init}\to\infty} \frac{n_{init}}{n} = 0.$$

Sequential optimality has been treated elsewhere in the literature. In particular Wynn (1970) proposed a sequential method converging to a D-optimal design. Wiens and Li (2014) gave a sequential estimation method yielding both consistent variance estimates and an asymptotically V-optimal design. Our proof of the following optimal design theorem closely parallels those in Wynn (1970) and Wiens and Li (2014).

Theorem 17 *Under assumptions (B1)-(B7) and (A3′), as $n_{init} \to \infty$, there are sequences $\{\boldsymbol{\theta}_{jn}\}$ for which*

(i) the LS estimates $\widehat{\boldsymbol{\theta}}_{jn} - \boldsymbol{\theta}_{jn} \overset{a.s.}{\to} 0$, $j = 0, 1$, and

(ii) $\mathcal{D}\left(\boldsymbol{\xi}_n, \widehat{\boldsymbol{\theta}}_n\right) - \max_{\xi\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) \overset{pr}{\to} 0$ with $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{0n}, \widehat{\boldsymbol{\theta}}_{1n})$ and $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_{0n}, \boldsymbol{\theta}_{1n})$.

Here $\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta})$ is the KL-divergence between $f_0(y|\mathbf{x},\mu_0(\mathbf{x}))$ and the least favourable density $f_{1*}(y|\mathbf{x},\mu_{1*}(\mathbf{x}))$:

$$\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \xi_i \mathcal{I}\{f_0, f_{1*}|\mathbf{x}_i, \mu_0(\mathbf{x}) = \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \mu_1(\mathbf{x}) = \eta_1(\mathbf{x}|\boldsymbol{\theta}_1)\},$$

and $\mathcal{P}$ is the set of all possible n-point designs.

In the following we consider several examples in a 20-point design space $\mathcal{S} = \{1, ..., 5\}$, dividing $[1, 5]$ into 50 equal subintervals.

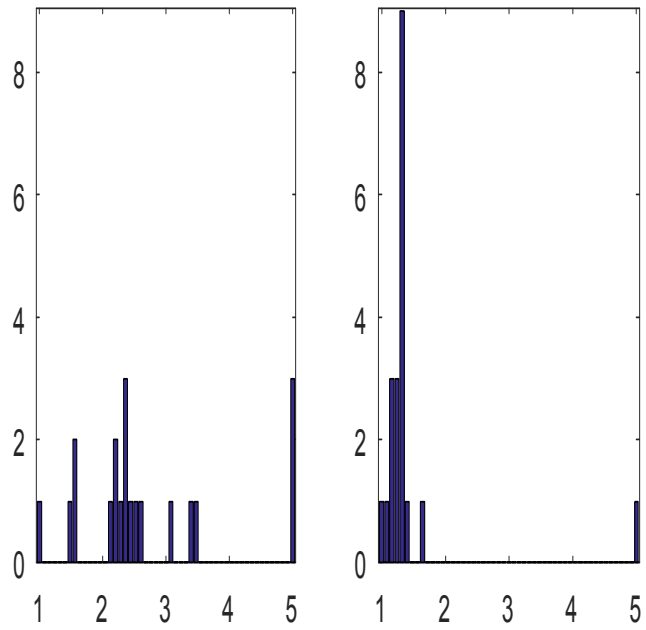

Figure 2.2: In Example 1, (a) classical design with sample size 20 (b) robust design with sample size 20 for $\varepsilon_1 = .17$. In this graph and the following graph, the horizontal axis denotes the value of covariate x and the vertical axis denotes the number of observations made at each covariate value.

2.3.1 Example 1.

Assume that both f_0 and f_1 are normal densities with mean $\eta_j(x|\boldsymbol{\theta}_j)$, $j = 0, 1$, and common variance $\sigma^2 = 0.1$. We consider the Michaelis-Menten and exponential response models

$$\eta_0(x|\boldsymbol{\theta}_0) = \frac{V_0 x}{K_0 + x}, \tag{2.23}$$

$$\eta_1(x|\boldsymbol{\theta}_1) = V_1(1 - \exp\{-K_1 x\}), \tag{2.24}$$

where $\boldsymbol{\theta}_0 = (V_0, K_0)', \boldsymbol{\theta}_1 = (V_1, K_1)'$.

Following steps 1-5 as described above, we obtain robust (or classical) sequential designs with sample size 20. Fig 2.2 shows a robust design and a classical design obtained in this example.

After a robust (or classical) design, i.e., a design measure $\boldsymbol{\xi} = (\xi_1, ..., \xi_N)$, is obtained, observations at the design points will be simulated from the 'null' model or the 'alternative' model. To investigate the sizes of the tests, we simulate from the

Table 1. Simulated sizes and minimum powers (standard errors in parentheses) for Example 1 $\varepsilon_1 = 0.058$

Criterion	Classical design	Robust design
Size	0.049(0.0068)	0.055(0.0072)
minimum power	0.684(0.0147)	0.73(0.0140)

Table 2. Simulated sizes and minimum powers (standard errors in parentheses) for Example 1 $\varepsilon_1 = 0.063$

Criterion	Classical design	Robust design
Size	0.044(.0065)	.043(.0064)
minimum power	0.703(0.0144)	0.711(0.0143)

working 'null' model $f_0\left(y|\eta_0(x|\boldsymbol{\theta}_{0true})\right)$ with $\boldsymbol{\theta}_{0true} = (1,1)^T$. To investigate the minimum powers we simulate from the 'alternative' model, the least favourable density $f_{1*}\left(y|x,\mu_{1*}\left(x\right)\right)$. Then the observations are substituted into the test statistic $\mathcal{R}$ and a model discrimination test is performed for the following hypotheses

$$H_0 : f_0(y|\eta_0(x|\boldsymbol{\theta}_0)) \; vs. \; H_1 : f_1(y|\eta_1(x|\boldsymbol{\theta}_1)).$$

We simulate 1000 robust (classical) designs and do the hypothesis tests of size $\alpha = .05$. The number of rejections is counted and the ratio of number of rejections to number of tests is the estimate of the size (if the working model is the 'null' model) or the minimum power (if the working model is the 'alternative' model). To complete the definition of the neighbourhood $\mathcal{F}_1(\varepsilon_1)$, we take $\varepsilon_1 = 0.058, 0.063, 0.1, 0.17$. The simulated results are recorded in Tables 1-4.

According to the four tables, the sizes of model discrimination tests for both classical and robust designs are close to the test size $\alpha = .05$. The minimum powers for robust designs are higher than those of classical designs. As the neighbourhood $\mathcal{F}_1(\varepsilon_1)$ is enlarged with respect to ε_1, the minimum powers are decreasing because the least favourable densities are found in a bigger neighbourhood.

2.3.2 Example 2.

Suppose that under each model the observations are log-normal, i.e. $\log Y$ is normally distributed. Assume that the logarithm of the observation has mean $\alpha_j(x)$ and

Table 3. Simulated sizes and minimum powers (standard errors in parentheses) for Example 1 $\varepsilon_1 = 0.1$

Criterion	Classical design	Robust design
Size	.054(.0071)	.052(.007)
minimum power	0.681(0.0147)	0.707(0.0144)

Table 4. Simulated sizes and minimum powers (standard errors in parentheses) for Example 1 $\varepsilon_1 = 0.17$

Criterion	Classical design	Robust design
Size	0.054(0.0071)	0.052(0.007)
power	0.67(0.0149)	0.699(0.0145)

variance $\sigma_j^2(x)$. Then

$$\begin{aligned} E_{\text{model j}}[Y|x] &= \mu_j(x) = \exp(\sigma_j^2(x)/2 + \alpha_j(x)), \\ var_{\text{model j}}(Y|x) &= v_j^2(x) = \mu_j^2(x)\{\exp(\sigma_j^2(x)) - 1\}. \end{aligned}$$

The density of Y is

$$f_j(y|x, \mu_j(x)) = \frac{1}{y\sigma_j(x)}\phi\left(\frac{\log y - \alpha_j(x)}{\sigma_j(x)}\right) I(y > 0).$$

In the following we assume homoscedastic models and specify the variance function $v_j^2(x) \equiv v^2 = 0.1$.

Let f_0 and f_1 be log-normal densities with means $\eta_0(x|\boldsymbol{\theta}_0)$ and $y|\eta_1(x|\boldsymbol{\theta}_1)$. Here $\eta_j(x|\boldsymbol{\theta}_j)$, $j = 0, 1,$ are the Michaelis-Menten and exponential response models defined in (2.23) and (2.24), respectively. The robust designs and classical optimal designs can be obtained by following steps 1-5. As an example a robust design and a classical design are illustrated in Fig 2.3.

To investigate the sizes of the tests, we simulate from a working 'null' model $f_0\left(y|\eta_0(x|\boldsymbol{\theta}_{0true})\right)$ with $\boldsymbol{\theta}_{0true} = (1, 1)^T$. To assess the minimum powers we simulate from $f_{1*}\left(y|\mathbf{x}, \mu_{1*}\left(\mathbf{x}\right)\right)$, the least favourable density in the neighbourhood $\mathcal{F}_1(\varepsilon_1)$ for $\varepsilon_1 = 0.01, 0.032, 0.17, 0.2$, respectively.

As described in Example 1, We simulate 1000 robust (classical) designs and perform model discrimination tests with size $\alpha = .05$. The estimates of type-I error and minimum powers are as follows:

In this example, the numerical results show that the robust designs are superior to the classical designs in the sense that the powers for robust designs in the worst

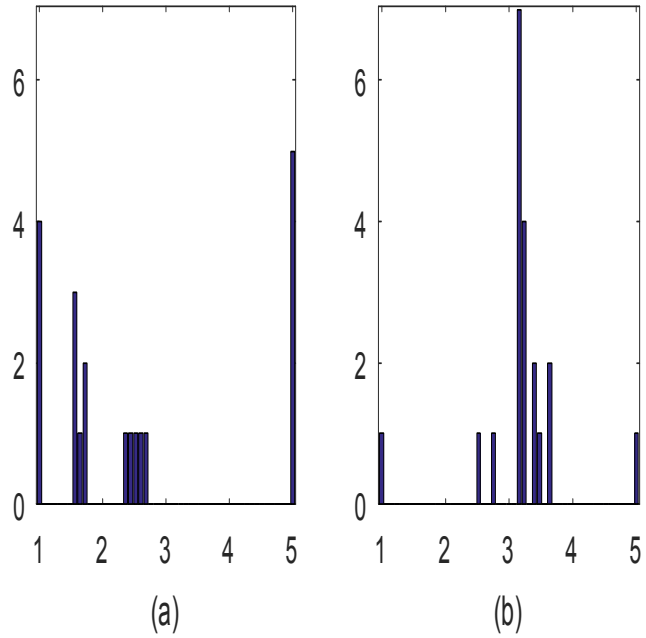

Figure 2.3: In Example 2, (a) classical design with sample size 20 (b) robust design with sample size 20 for $\varepsilon_1 = .17$.

Table 5. Simulated sizes and minimum powers (standard errors in parentheses) for Example 2 $\varepsilon_1 = 0.01$

Criterion	Classical design	Robust design
Size	0.052(0.007)	0.049(0.0068)
minimum power	0.602(0.0155)	0.649(0.0151)

Table 6. Simulated sizes and minimum powers (standard errors in parentheses) for Example 2 $\varepsilon_1 = 0.032$

Criterion	Classical design	Robust design
Size	0.053(0.0071)	0.054(0.0071)
minimum power	0.641(0.0152)	0.675(0.0148)

Table 7. Simulated sizes and minimum powers (standard errors in parentheses) for Example 2 $\varepsilon_1 = 0.17$

Criterion	Classical design	Robust design
Size	0.0610(0.0076)	0.05(0.0069)
minimum power	0.601(0.0155)	0.63(0.0153)

Table 8. Simulated sizes and minimum powers (standard errors in parentheses) for Example 2 $\varepsilon_1 = 0.2$

Criterion	Classical design	Robust design
Size	0.043(0.0064)	0.053(0.0071)
minimum power	0.593(0.0155)	0.623(0.0153)

case are higher than those for classical designs. One can also notice that the sizes are all around the significance size $\alpha = .05$ and the 'worst' powers decreases as the size of neighbourhoods are increasing.

2.4 Summarizing remarks

We have considered the construction of robust model discrimination designs, to aid in the choice of regression models. In the existing literature on discrimination designs, the construction has generally been based on the assumption that the true model is one of the nominal models. Our method instead assumes that the true model is an unknown member of Hellinger neighbourhoods of the nominal models. The model discrimination problem can be cast as a problem of hypothesis testing. In particular, a case is considered: let the neighbourhood be a singleton - a fixed density function, under the null hypothesis; under the alternative hypothesis the density lies in a Hellinger neighbourhood of the (possibly incorrectly) hypothesized density. We aimed at constructing experimental designs by maximizing the worst power of the Neyman-Pearson test, i.e. the minimum power over the Hellinger neighbourhood. We derived the asymptotic properties of the Neyman-Pearson test statistic and proved that the power of the Neyman-Pearson test is a monotonic function of the Kullback-Leibler divergence between the two rival models under certain condition. Therefore, we have proposed designs that maximize the minimum KL divergence in the neighbourhood.

The minimization part of this procedure has been carried out analytically; the optimal designs are obtained by maximizing the minimized discrepancy function sequentially. Different examples are discussed, especially, when nominal densities are both normal or log-normal. In the examples we have obtained sequential designs via simulating observations from a 'true' model which might be slightly away from the two nominal models. We obtained 1000 robust designs and 1000 classical designs for each example. To illustrate that the robust designs is robust against slight deviation from the nominal model, we compared the minimum powers of model discrimination

tests based on robust designs and classical designs. According to the results in Tables 1-8, we can see that for all the examples the minimum test powers based on robust designs are higher than those based on classical designs.

REFERENCES

Atkinson, A. C., and Fedorov, V. V. (1975a), "The Design of Experiments for Discriminating Between Two Rival Models," *Biometrika*, 62, 57-70.

Atkinson, A. C., and Fedorov, V. V. (1975b), "Optimal Design: Experiments for Discriminating Between Several Models," *Biometrika*, 62, 289-303.

Box, G. E. P., and Draper, N. R. (1959), "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, 54, 622-654.

Fedorov, V. V. (1975), "Optimal Experimental Designs for Discriminating Two Rival Regression Models," *A Survey of Statistical Design and Linear Models*, ed. Srivastava, J.N., Amsterdam: North Holland.

Fedorov, V. V. and Pazman, A. (1968), "Design of physical experiments," *Fortschritte der Physik*, 16, 325-355.

Ford, I., Titterington, D. M., and Kitsos, C. P. (1989), "Recent Advances in Nonlinear Experimental Design," *Technometrics*, 31, 49-60.

Hill, P. D. H. (1978), "A Review of Experimental Design Procedures for Regression Model Discrimination," *Technometrics*, 20, 15-21.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Hunter, W. G., and Reiner, A. M. (1965), "Designs for discriminating between two rival models," *Technometrics*, 7, 307-323.

Loeve, M. (1963), *Probability theory* (3rd ed.), Van Nostrand, New York.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007), "An Optimal Experimental Design Criterion for Discriminating Between Non-Normal Models," *Journal of the Royal Statistical Society* B, 69, 231-242.

Oosterhoff, J. and van Zwet, W. R. (2012), *A note on contiguity and Hellinger distance*, Springer New York.

Sinha, S. and Wiens, D. (2003), "Asymptotics for robust sequential designs in misspecified regression models," in *Festschrift for Constance van Eeden* (eds. M. Moore, C. Léger and S. Froda), pp. 233-248. Hayward: Institute of Mathematical Statistics.

Wiens, D. (2009a), "Robust discrimination designs," *Journal of the Royal Statistical Society (Series B)*; 71, 805-829.

Wiens, D. (2009b), "Asymptotic Properties of a Neyman-Pearson Test for Model Discrimination, with an Application to Experimental Design," *Journal of Statistical Theory and Practice*, 3, 419-427.

Wiens, D. and Li, P. (2014), "V-optimal designs for heteroscedastic regression," *Journal of Statistical Planning and Inference*, 145, 125-138.

Wynn, H. P. (1970), "The sequential generation of D-optimum experimental design," *Annals of Mathematical Statistics*, 5, 1655-1664.

Chapter 3

Derivation and proofs for Chapter 2

3.1 Proof of Proposition 10

Denote by F_{1i}, F_{0i} the distribution functions with densities $f_1(y|\mathbf{x}_i)$ and $f_0(y|\mathbf{x}_i)$, respectively. We denote the probability of an event A, under F, as $F(A)$. To prove Proposition 10, we need several preliminary results. In this section we abbreviate $r(y|\mathbf{x}_i; f_0, f_1)$ by $r(y|\mathbf{x}_i)$.

Lemma 18 *Condition (2.9) holds iff each of*

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i F_{1i}\left(\log r(y|\mathbf{x}_i) \geq \delta\right) = 0, \tag{3.1}$$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i F_{0i}\left(\log r(y|\mathbf{x}_i) \leq -\delta\right) = 0, \tag{3.2}$$

holds for all $\delta > 0$.

Proof. This is proven in Theorem 2 of Oosterhoff and van Zwet (2012). ■

Lemma 19 *Define $F_0^{(n)} = \Pi_{i=1}^{N}(F_{0i})^{n_i}$ and $F_1^{(n)} = \Pi_{i=1}^{N}(F_{1i})^{n_i}$. Then (2.8) and (2.9) imply that the sequences $\left\{F_0^{(n)}\right\}$ and $\left\{F_1^{(n)}\right\}$ are contiguous with respect to each other, i.e. for any sequence $\{A_n\}$ of measurable sets, $\lim_{n\to\infty} F_0^{(n)}(A_n) = 0$ iff $\lim_{n\to\infty} F_1^{(n)}(A_n) = 0$. Moreover, $\left\{F_0^{(n)}\right\}$ and $\left\{F_1^{(n)}\right\}$ being contiguous with respect to each other also implies that*

$$\lim_{n\to\infty} \sum_{i=1}^{N} n_i F_{0i}(A_{n_i}) = 0 \Leftrightarrow \lim_{n\to\infty} \sum_{i=1}^{N} n_i F_{1i}(A_{n_i}) = 0 \tag{3.3}$$

for any collection of measurable sets A_{n_i}.

Proof. Corollary 1 in Oosterhoff and van Zwet (2012) shows that $\{F_0^{(n)}\}$ and $\{F_1^{(n)}\}$ are contiguous with respect to each other under (3.1), (3.2) and the condition

$$\limsup_{n\to\infty} \sum_{i=1}^{N} n_i d_h^2(f_1(y|\mathbf{x}_i), f_0(y|\mathbf{x}_i)) < \infty. \tag{3.4}$$

Lemma 18 shows that condition (2.9) is equivalent to (3.1) and (3.2). Therefore, to establish contiguity it suffices to show that (2.8) implies (3.4). For this we note that the KL distance between $f_0(y|\mathbf{x}_i), f_1(y|\mathbf{x}_i)$ is larger than twice their squared Hellinger distance, i.e.

$$\begin{aligned} \int_{-\infty}^{\infty} f_1(y|\mathbf{x}_i) \log(r(y|\mathbf{x}_i))dy &\geq \int_{-\infty}^{\infty} \left(f_1^{1/2}(y|\mathbf{x}_i) - f_0^{1/2}(y|\mathbf{x}_i)\right)^2 dy \\ &= 2d_h^2(f_1(y|\mathbf{x}_i), f_0(y|\mathbf{x}_i)); \end{aligned}$$

now (2.8) gives $\sum_{i=1}^{N} n_i d_h^2(f_1(y|\mathbf{x}_i), f_0(y|\mathbf{x}_i)) = O(1)$. The equivalence in (3.3) follows from (2.1) in Oosterhoff and van Zwet (2012). ■

Lemma 20 *Under conditions (2.8) and (2.9), for all* $\delta > 0$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i F_{1i}(|\log r(y|\mathbf{x}_i)| \geq \delta) = 0, \tag{3.5}$$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i F_{0i}(|\log r^{-1}(y|\mathbf{x}_i)| \geq \delta) = 0. \tag{3.6}$$

Proof. Under (2.8) and (2.9) we have both (3.2) and (3.3), hence

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i F_{1i}(\log r(y|\mathbf{x}_i) \leq -\delta) = 0;$$

this with (3.1) gives (3.5). The proof of (3.6) is completely analogous. ■

The following lemma shows two different expansions of $\log r(y|\mathbf{x}_i)$. Both expansions are used in the proof of the asymptotic distribution of the test statistic $\mathcal{R}$ under the null and alternative hypotheses.

Lemma 21 *Assume that* $|\log r(y|\mathbf{x}_i)| < \delta$ *for some* $\delta \in (0,1)$. *Then* $r(y|\mathbf{x}_i)$ *can be expanded as*

$$\log r(y|\mathbf{x}_i) = -2(r^{-1/2}(y|\mathbf{x}_i) - 1) + (r^{-1/2}(y|\mathbf{x}_i) - 1)^2(1 + \rho_{1i\delta}), \tag{3.7}$$

and

$$\log r(y|\mathbf{x}_i) = 2(r^{1/2}(y|\mathbf{x}_i) - 1) - (1 - r^{1/2}(y|\mathbf{x}_i))^2(1 + \rho_{2i\delta}), \tag{3.8}$$

where $|\rho_{ji\delta}| < 3\delta, j = 1, 2$.

Proof. Expansion (3.8) is (3.11) in Oosterhoff and van Zwet (2012); expansion (3.7) is obtained in a similar manner. ■

The following lemma gives two expansions of $(\log r(y|\mathbf{x}_i))^2$ under the conditions of Lemma 21. It is a rather immediate consequence of Lemma 21 and so is presented without proof. The first expansion (3.9) is used in the determination of the asymptotic distribution of the test statistic $\mathcal{R}$ under the alternative hypothesis H_1; the second expansion (3.10) is used in the proof of the asymptotic distribution of the test statistic $\mathcal{R}$ under the null hypothesis H_0.

Lemma 22 *Assume* $|\log r(y|\mathbf{x}_i)| < \delta$. *Let* $0 < \delta < 1$. *Then*

$$(\log r(y|\mathbf{x}_i))^2 = 4(r^{-1/2}(y|\mathbf{x}_i) - 1)^2 + (1 + \rho_{1i\delta})(r^{-1/2}(y|\mathbf{x}_i) - 1)^2\varpi_{1i\delta} \tag{3.9}$$

and

$$(\log r(y|\mathbf{x}_i))^2 = 4(1 - r^{1/2}(y|\mathbf{x}_i))^2 + (1 + \rho_{2i\delta})(1 - r^{1/2}(y|\mathbf{x}_i))^2\varpi_{2i\delta} \tag{3.10}$$

where $|\varpi_{ji\delta}| < 6\delta$, $j = 1, 2$ *and* $i = 1, ..., N$.

We now give the proof of the main result Proposition 10. Notice that for the test statistic $\mathcal{R} = 2\sum_{i=1}^{N} n_i \log r(y|\mathbf{x}_i)$, the means and variances under the two hypotheses are

$$\begin{aligned} E_{H_0}(\mathcal{R}) &= 2\sum_{i=1}^{N} n_i \int (\log r(y|\mathbf{x}_i)) f_0(y|\mathbf{x}_i)dy, \\ \text{VAR}_{H_0}(\mathcal{R}) &= 4\left[\sum_{i=1}^{N} n_i \int (\log r(y|\mathbf{x}_i))^2 f_0(y|\mathbf{x}_i)dy - (E_{H_0}(\mathcal{R}))^2\right]; \end{aligned}$$

and

$$\begin{aligned} E_{H_1}(\mathcal{R}) &= 2\sum_{i=1}^{N} n_i \int (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy, \\ \text{VAR}_{H_1}(\mathcal{R}) &= 4\left[\sum_{i=1}^{N} n_i \int (\log r(y|\mathbf{x}_i))^2 f_1(y|\mathbf{x}_i)dy - (E_{H_1}(\mathcal{R}))^2\right]. \end{aligned}$$

According to the Normal Convergence Criterion (Loeve 1963, p. 316) and an equivalent form of this result, to prove Proposition 10 it is sufficient to prove the following:

(1) under H_0,

$$\lim_{n\to\infty}\sum_{i=1}^{N} n_i F_{0i}(|\log r(Y|\mathbf{x}_i)| \geq \delta) = 0, \text{ for every } \delta > 0, \tag{3.11}$$

$$\lim_{\delta\to 0^+}\lim_{n\to\infty}\sum_{i=1}^{N} n_i \left(\int_{|\log r(y|\mathbf{x}_i)|\leq\delta} (\log r(y|\mathbf{x}_i)) f_0(y|\mathbf{x}_i)dy + \mathcal{D}\right) = 0, \tag{3.12}$$

$$\lim_{\delta\to 0^+}\lim_{n\to\infty}\sum_{i=1}^{N} n_i \left\{ \begin{array}{c} \int_{|\log r(y|\mathbf{x}_i)|\leq\delta}(\log r(y|\mathbf{x}_i))^2 f_0(y|\mathbf{x}_i)dy - \\ \left(\int_{|\log r(y|\mathbf{x}_i)|\leq\delta}(\log r(y|\mathbf{x}_i)) f_0(y|\mathbf{x}_i)dy\right)^2 - 2\mathcal{D} \end{array} \right\} = 0. \tag{3.13}$$

(2) under H_1,

$$\lim_{n\to\infty}\sum_{i=1}^{N} n_i F_{1i}(|\log r(Y|\mathbf{x}_i)| \geq \delta) = 0, \text{ for every } \delta > 0, \tag{3.14}$$

$$\lim_{\delta\to 0^+}\lim_{n\to\infty}\sum_{i=1}^{N} n_i \left(\int_{|\log r(y|\mathbf{x}_i)|\leq\delta} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy - \mathcal{D}\right) = 0, \tag{3.15}$$

$$\lim_{\delta\to 0^+}\lim_{n\to\infty}\sum_{i=1}^{N} n_i \left\{ \begin{array}{c} \int_{|\log r(y|\mathbf{x}_i)|\leq\delta}(\log r(y|\mathbf{x}_i))^2 f_1(y|\mathbf{x}_i)dy - \\ \left(\int_{|\log r(y|\mathbf{x}_i)|\leq\delta}(\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy\right)^2 - 2\mathcal{D} \end{array} \right\} = 0. \tag{3.16}$$

Conditions (3.11) and (3.14) are direct results of Lemma 20. We prove (3.12); the proof of (3.15) is similar. For $\tau > \delta > 0$ we first split $n\mathcal{D}$ into three terms

$$n\mathcal{D} = \sum_{j=1}^{3}\sum_{i=1}^{N} n_i \int_{E_j} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy,$$

where $E_1 = \{|\log r(y|\mathbf{x}_i)| \leq \delta\}$, $E_2 = \{\delta < |\log r(y|\mathbf{x}_i)| \leq \tau\}$, $E_3 = \{|\log r(y|\mathbf{x}_i)| > \tau\}$. Then in (3.12) we have that

$$\begin{aligned} & \left|\sum_{i=1}^{N} n_i \left(\int_{|\log r(y|\mathbf{x}_i)|\leq\delta} (\log r(y|\mathbf{x}_i))\, f_0(y|\mathbf{x}_i)dy + \mathcal{D}\right)\right| \\ \leq\ & \left|\sum_{i=1}^{N} n_i \int_{E_1} \left[(\log r(y|\mathbf{x}_i))\left\{f_0(y|\mathbf{x}_i) + f_1(y|\mathbf{x}_i)\right\}\right] dy\right| \qquad (3.17) \\ & + \sum_{j=2,3}\left|\sum_{i=1}^{N} n_i \int_{E_j} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy\right|. \qquad (3.18) \end{aligned}$$

We first prove that the terms in (3.18) tend to zero as $n \to \infty$. For $j = 2$, using (3.5) we have

$$\left|\sum_{i=1}^{N} n_i \int_{E_2} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i)dy\right| \leq \tau \sum_{i=1}^{N} n_i F_{1i}(|\log r(y|\mathbf{x}_i)| \geq \delta) \to 0.$$

Now consider $j = 3$. Using (3.5) we have

$$\sum_{i=1}^{N} n_i \int_{\log r(y|\mathbf{x}_i)<-\tau} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy < -\tau \sum_{i=1}^{N} n_i F_{1i}(\log r(Y|\mathbf{x}_i) < -\tau) \to 0$$

and since $-\log u > 1 - u$

$$\begin{aligned}
& \sum_{i=1}^{N} n_i \int_{\log r(y|\mathbf{x}_i)<-\tau} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy \\
= & \sum_{i=1}^{N} n_i \int_{\log r(y|\mathbf{x}_i)<-\tau} (-\log r^{-1}(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy \\
\geq & \sum_{i=1}^{N} n_i \int_{\log r(y|\mathbf{x}_i)<-\tau} (1 - r^{-1}(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy \\
= & \sum_{i=1}^{N} n_i F_{1i}(\log r(y|\mathbf{x}_i) < -\tau) - \sum_{i=1}^{N} n_i F_{0i}(\log r(y|\mathbf{x}_i) < -\tau) \to 0.
\end{aligned}$$

Therefore,

$$\left| \sum_{i=1}^{N} n_i \int_{\log r(y|\mathbf{x}_i)<-\tau} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy \right| \to 0.$$

Combining this with (2.10),

$$\left| \sum_{i=1}^{N} n_i \int_{E_3} (\log r(y|\mathbf{x}_i)) f_1(y|\mathbf{x}_i) dy \right| \to 0.$$

Finally, the sum in (3.17) is, by Lemma 21, equal to

$$\begin{aligned}
& \sum_{i=1}^{N} n_i \int_{E_1} \left[(f_1^{1/2}(y|\mathbf{x}_i) - f_0^{1/2}(y|\mathbf{x}_i))^2 (\rho_{1i\delta} - \rho_{2i\delta}) \right] dy \\
& +2 \sum_{i=1}^{N} n_i \int_{E_1} (f_1(y|\mathbf{x}_i) - f_0(y|\mathbf{x}_i)) dy. \qquad (3.19)
\end{aligned}$$

The second sum in (3.19) tends to 0 as $n \to \infty$ by (3.3) and (3.5). For the first we have, again using Lemma 21, that

$$\begin{aligned}
& \lim_{\delta \to 0^+} \overline{\lim_{n \to \infty}} \left| \sum_{i=1}^{N} n_i \int_{E_1} \left[(f_1^{1/2}(y|\mathbf{x}_i) - f_0^{1/2}(y|\mathbf{x}_i))^2 (\rho_{1i\delta} - \rho_{2i\delta}) \right] dy \right| \\
\leq & \lim_{\delta \to 0^+} \overline{\lim_{n \to \infty}} \left| 6\delta \sum_{i=1}^{N} n_i \int_{E_1} \left[(f_1^{1/2}(y|\mathbf{x}_i) - f_0^{1/2}(y|\mathbf{x}_i))^2 \right] dy \right| \\
\leq & \lim_{\delta \to 0^+} \overline{\lim_{n \to \infty}} |6\delta n \mathcal{D}| = 0.
\end{aligned}$$

This completes the proof of (3.12).

The proofs of (3.13) and (3.16) run along the same lines, but using the bounds of Lemma 22 rather than those of Lemma 21. ■

3.2 Applying (2.8)-(2.10) to normal densities

As in Remark 1, assume that $f_j(y|\mathbf{x}_i, \mu_j(\mathbf{x}_i), \sigma) = \frac{1}{\sigma}\phi\left(\frac{y-\mu_j(\mathbf{x}_i)}{\sigma}\right)$, $j = 0, 1$, $i = 1, ..., N$, satisfy $\mu_1(\mathbf{x}_i) = \mu_0(\mathbf{x}_i) + \Delta(\mathbf{x}_i)/\sqrt{n}$. Without loss of generality we take $\mu_0(\mathbf{x}) = 0$, $\sigma = 1$. Condition (2.8) follows from

$$\int f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)) \log \frac{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} dy = \frac{\Delta^2(\mathbf{x}_i)}{2n}.$$

For condition (2.9) we note that

$$\log \frac{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} = \mu_1(\mathbf{x}_i)\left(y - \mu_1(\mathbf{x}_i)/2\right),$$

so that for any $\delta > 0$

$$|\log r(y|\mathbf{x}_i)| \geq \delta \Leftrightarrow \frac{\Delta^2(\mathbf{x}_i)}{2\sqrt{n}} - \sqrt{n}\delta > y\Delta(\mathbf{x}_i) \text{ or } \frac{\Delta^2(\mathbf{x}_i)}{2\sqrt{n}} + \sqrt{n}\delta < y\Delta(\mathbf{x}_i).$$

For sufficiently large n there exists $\widehat{\delta}$ satisfying

$$\min\left(\left|\frac{\Delta(\mathbf{x}_i)}{2\sqrt{n}} - \sqrt{n}\frac{\delta}{\Delta(\mathbf{x}_i)}\right|, \left|\frac{\Delta(\mathbf{x}_i)}{2\sqrt{n}} + \sqrt{n}\frac{\delta}{\Delta(\mathbf{x}_i)}\right|\right) > \sqrt{n}\widehat{\delta}.$$

Then

$$\begin{aligned}
& \int_{\{|\log r(y|\mathbf{x}_i)| \geq \delta\}} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right)^2 dy \\
\leq & \int_{\{|y| \geq \sqrt{n}\widehat{\delta}\}} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right)^2 dy \\
= & \Phi\left(\mu_1(\mathbf{x}_i) - \sqrt{n}\widehat{\delta}\right) + \Phi\left(-\mu_1(\mathbf{x}_i) - \sqrt{n}\widehat{\delta}\right) + 2\Phi\left(-\sqrt{n}\widehat{\delta}\right) \\
& -2e^{-\mu_1^2(\mathbf{x}_i)/8}\left(\Phi\left(\frac{\mu_1(\mathbf{x}_i)}{2} - \sqrt{n}\widehat{\delta}\right) + \Phi\left(-\frac{\mu_1(\mathbf{x}_i)}{2} - \sqrt{n}\widehat{\delta}\right)\right).
\end{aligned}$$

Now (2.9) follows from the observation that

$$n\Phi\left(\widehat{\Delta}(\mathbf{x}_i)n^{-1/2} - \sqrt{n}\widehat{\delta}\right) \to 0 \text{ as } n \to \infty,$$

for each of $\widehat{\Delta}(\mathbf{x}_i) = \pm\Delta(\mathbf{x}_i), \pm\Delta(\mathbf{x}_i)/2$ or 0.

Finally, we show that condition (2.10) holds for any $\tau > 0$. We take $\Delta(\mathbf{x}_i) > 0$; if $\Delta(\mathbf{x}_i) < 0$ the procedure is similar. For $\tau > 0$ we have, by calculations similar to those above, that

$$\begin{aligned} 0 &\leq n \int_{\log r \geq \tau} f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)) \log \frac{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} dy \\ &\leq n \int_{y \geq \tau\sqrt{n}/\Delta(\mathbf{x}_i)} f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i)) \log \frac{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} dy \\ &= \sqrt{n}\Delta(\mathbf{x}_i)\phi\left(\frac{\Delta(\mathbf{x}_i)}{\sqrt{n}} - \frac{\tau\sqrt{n}}{\Delta(\mathbf{x}_i)}\right) + \frac{\Delta^2(\mathbf{x}_i)}{2n}\Phi\left(\frac{\Delta(\mathbf{x}_i)}{\sqrt{n}} - \frac{\tau\sqrt{n}}{\Delta(\mathbf{x}_i)}\right) \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

For $\Delta(\mathbf{x}_i) < 0$, we reach the same conclusion. ■

3.3 Applying (2.8)-(2.10) to log-normal densities

Under the conditions of Remark 1 the densities are

$$\begin{aligned} f_j(y|\mathbf{x}, \mu_j(\mathbf{x}), v^2) &= y^{-1}\left(2\pi\sigma_j^2(\mathbf{x})\right)^{-1/2} e^{-\frac{(\log y - \vartheta_j(\mathbf{x}))^2}{2\sigma_j^2(\mathbf{x})}} I(y > 0), \qquad (3.20) \\ \text{where } \vartheta_j(\mathbf{x}) &= \log\left[\frac{\mu_j(\mathbf{x})}{\{1 + v_j^2(\mathbf{x})/\mu_j^2(\mathbf{x})\}^{1/2}}\right], \\ \sigma_j^2(\mathbf{x}) &= \log\left[1 + \frac{v^2}{\mu_j^2(\mathbf{x})}\right]. \end{aligned}$$

Using $\mu_1(\mathbf{x}_i) = \mu_0(\mathbf{x}_i) + \Delta(\mathbf{x}_i)/\sqrt{n}$ we have that

$$\vartheta_1(\mathbf{x}_i) = \vartheta_0(\mathbf{x}_i) + \kappa(\mathbf{x}_i)/\sqrt{n} + o(n^{-1/2}), \qquad (3.21)$$

for

$$\kappa(\mathbf{x}_i) = \frac{1}{\mu_0(\mathbf{x}_i)} + \frac{v^2}{\mu_0^3(\mathbf{x}_i) + v^2\mu_0(\mathbf{x}_i)}, \qquad (3.22)$$

and that $\sigma_1^2(\mathbf{x}) = \sigma_0^2(\mathbf{x}) + o(1)$. One can now substitute (3.21) and (3.22) into (3.20) and proceed as in §A.2 to conclude that $f_0(y|\mathbf{x}, \mu_0(\mathbf{x}), v^2)$ and $f_1(y|\mathbf{x}, \mu_1(\mathbf{x}), v^2)$ satisfy all the conditions in Proposition 10. Moreover, condition (2.10) holds for any $\tau > 0$. ■

3.4 Proof of Corollary 11

We first show that (2.9) holds, under the conditions in the statement of the Corollary. For arbitrary $\left(f^{(0)}, f^{(1)}\right)$ and any $\delta > 0$ denote

$$\Xi = \left\{ y : |\log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)| \geq \delta \right\}.$$

Then

$$\begin{aligned}
& \int_{\Xi} f^{(0)}(y|\mathbf{x}_i) \left(\sqrt{r(y|\mathbf{x}_i; f^{(0)}, f^{(1)})} - 1 \right)^2 dy \\
= & \int_{\Xi} \left(\begin{array}{c} \left[\sqrt{f^{(0)}(y|\mathbf{x}_i)} - \sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} \right] + \\ \left[\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right] \\ + \left[\sqrt{f_1(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f^{(1)}(y|\mathbf{x}_i)} \right] \end{array} \right)^2 dy \\
\leq & 3 \sum_{j=0,1} \int_{\Xi} \left(\sqrt{f^{(j)}(y|\mathbf{x}_i)} - \sqrt{f_j(y|\mathbf{x}_i, \mu_j(\mathbf{x}_i))} \right)^2 dy \qquad (3.23) \\
& + 3 \int_{\Xi} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy
\end{aligned}$$

We show that each term in (3.23) is $o(n^{-1})$. Since $\varepsilon_0 = o(n^{-1/2}), \varepsilon_1 = o(n^{-1/2})$, the first term is $o(n^{-1})$ for each j, and

$$\lim_{n \longrightarrow \infty} \sum_{i=1}^{N} n_i \int_{\Xi} \left(\sqrt{f^{(j)}(y|\mathbf{x}_i)} - \sqrt{f_j(y|\mathbf{x}_i, \mu_j(\mathbf{x}_i))} \right)^2 dy = 0, \ j = 0, 1. \qquad (3.24)$$

With

$$\begin{aligned}
\Xi_1 &= \left\{ |\log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)| \geq \delta \ \& \ |\log r\left(y|\mathbf{x}_i; f_0, f_1\right)| \geq \frac{\delta}{n} \right\}, \\
\Xi_2 &= \left\{ |\log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)| \geq \delta \ \& \ |\log r\left(y|\mathbf{x}_i; f_0, f_1\right)| < \frac{\delta}{n} \right\},
\end{aligned}$$

the second term in (3.23) can be divided into two terms

$$\begin{aligned}
& \int_{\Xi} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \\
= & \int_{\Xi_1} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \\
& + \int_{\Xi_2} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy. \qquad (3.25)
\end{aligned}$$

Notice that $\Xi_1 \subseteq \Xi_3 = \left\{ \left| \log r\left(y|\mathbf{x}_i; f_0, f_1\right) \right| \geq \frac{\delta}{n} \right\}$. Then the first term in (3.25) satisfies

$$\begin{aligned} & \int_{\Xi_1} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \\ \leq & \int_{\Xi_3} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy. \end{aligned}$$

Moreover, since $f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))$ and $f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))$ satisfy (2.9), we have

$$\begin{aligned} & \lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\Xi_1} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \\ \leq & \lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\Xi_3} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \qquad (3.26) \\ = & 0. \end{aligned}$$

For the second term in (3.25), noticing that

$$\left| \log r\left(y|\mathbf{x}_i; f_0, f_1\right) \right| < \frac{\delta}{n} \Leftrightarrow e^{-\delta/n} \leq r\left(y|\mathbf{x}_i; f_0, f_1\right) \leq e^{\delta/n},$$

we have

$$\begin{aligned} & \int_{\Xi_2} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy \\ \leq & \max\{(1-e^{-\delta/2n})^2, (1-e^{\delta/2n})^2\} = o(n^{-1}), \end{aligned}$$

and then

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\Xi_2} \left(\sqrt{f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i))} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))} \right)^2 dy = 0. \qquad (3.27)$$

Therefore, combining (3.24), (3.26) and (3.27) we have that

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\Xi} f^{(0)}(y|\mathbf{x}_i) \left(r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right) - 1 \right)^2 dy = 0,$$

i.e., $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.9).

To prove that $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.8), first write

$$\begin{aligned} n\mathcal{D} = & \sum_{i=1}^{N} n_i \int_{\Xi} f^{(1)}(y|\mathbf{x}_i) \log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right) dy \\ & + \sum_{i=1}^{N} n_i \int_{\Xi^c} f^{(1)}(y|\mathbf{x}_i) \log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right) dy. \qquad (3.28) \end{aligned}$$

We will prove that the limit of the first term in (3.28) is 0 as $n \to \infty$ and the limit of the second term is finite.

By the triangle inequality for Hellinger distance we have

$$d_h^2(f^{(0)}(y|\mathbf{x}_i), f^{(1)}(y|\mathbf{x}_i)) \leq \varepsilon_0^2 + \varepsilon_1^2 + d_h^2(f_0(y|\mathbf{x}_i, \mu_0(\mathbf{x}_i)), f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))) \leq O(n^{-1}). \tag{3.29}$$

According to Oosterhoff and van Zwet (2012), (3.29) and condition (2.9) imply that $\left\{\Pi_{i=1}^N \left(F_i^{(0)}\right)^{n_i}\right\}$ and $\left\{\Pi_{i=1}^N \left(F_i^{(1)}\right)^{n_i}\right\}$ are contiguous with respect to each other, where $F_i^{(0)}$ and $F_i^{(1)}$ are the distributions correcponding to $f^{(0)}(y|\mathbf{x}_i)$ and $f^{(1)}(y|\mathbf{x}_i)$, respectively. Therefore, we have conclusions that are similar to (3.5) and (3.6): for all $\delta > 0$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i F_i^{(1)} \left(\left|\log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)\right| \geq \delta\right) = 0, \tag{3.30}$$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i F_i^{(0)} \left(\left|\log r^{-1}\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)\right| \geq \delta\right) = 0. \tag{3.31}$$

Moreover, according to the contiguity of $\left\{\Pi_{i=1}^N \left(F_i^{(0)}\right)^{n_i}\right\}$ and $\left\{\Pi_{i=1}^N \left(F_i^{(1)}\right)^{n_i}\right\}$, we have

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^N n_i F_i^{(0)} \left(\left|\log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right)\right| \geq \delta\right) = 0. \tag{3.32}$$

Then by similar analysis as in the proof of Proposition 10 and using condition (2.10), the limit of the first term in (3.28) is 0. To prove that the second term in (3.28) has a finite limit, we apply the expansion of (3.7) in Lemma 21 to $\log r(y|\mathbf{x}_i)$,

obtaining

$$\begin{aligned}
&\left|\sum_{i=1}^{N} n_i \int_{\Xi^c} f^{(1)}(y|\mathbf{x}_i) \log r\left(y|\mathbf{x}_i; f^{(0)}, f^{(1)}\right) dy\right| \\
\leq &\left\{\begin{array}{c} 2\sum_{i=1}^{N} n_i \int_{\Xi^c} \left(\sqrt{f^{(1)}(y|\mathbf{x}_i)} - \sqrt{f^{(0)}(y|\mathbf{x}_i)}\right)^2 dy \\ + \left|\sum_{i=1}^{N} n_i \int_{\Xi^c} \left(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\right) dy\right| \\ + \left|\sum_{i=1}^{N} n_i \int_{\Xi^c} \rho_{1i\delta} \left(\sqrt{f^{(1)}(y|\mathbf{x}_i)} - \sqrt{f^{(0)}(y|\mathbf{x}_i)}\right)^2 dy\right| \end{array}\right\} \\
\leq &\left\{\begin{array}{c} 4\sum_{i=1}^{N} n_i d_h^2(f^{(0)}(y|\mathbf{x}_i), f^{(1)}(y|\mathbf{x}_i)) \\ + \left|\sum_{i=1}^{N} n_i \int_{\Xi^c} \left(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\right) dy\right| \\ +3\delta \sum_{i=1}^{N} n_i d_h^2(f^{(0)}(y|\mathbf{x}_i)), f^{(1)}(y|\mathbf{x}_i)) \end{array}\right\} \qquad (3.33) \\
= &\ O(1).
\end{aligned}$$

Here the second term in (3.33) satisfies

$$\left|\sum_{i=1}^{N} n_i \int_{\Xi^c} \left(f^{(0)}(y|\mathbf{x}_i) - f^{(1)}(y|\mathbf{x}_i)\right) dy\right| = \left|\sum_{i=1}^{N} n_i \int_{\Xi} \left(f^{(1)}(y|\mathbf{x}_i) - f^{(0)}(y|\mathbf{x}_i)\right) dy\right| \to 0 \qquad (3.34)$$

due to (3.30) and (3.32). Combining (3.33) and (3.34), we can conclude that $f^{(0)}(y|\mathbf{x})$ and $f^{(1)}(y|\mathbf{x})$ satisfy (2.8). ■

3.5 Proof of Proposition 12

we first prove (i). According to Proposition 10, if $f(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$ is the true model, the test statistic

$$\mathcal{R}(f_0, f) = 2\sum_{i=1}^{N}\sum_{l=1}^{n_i} \log\left\{\frac{f(y_{il}|\mathbf{x}_i)}{f_0(y_{il}|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right\}$$

is normally distributed with mean $-2n\mathcal{D}(f_0, f)$ and standard deviation $\sqrt{8n\mathcal{D}(f_0, f)}$. Recall that

$$\mathcal{R} := \mathcal{R}(f_0, f_1) = 2\sum_{i=1}^{N}\sum_{l=1}^{n_i} \log\left\{\frac{f_1(y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}{f_0(y_{il}|\mathbf{x}_i, \mu_0(\mathbf{x}_i))}\right\}.$$

Therefore, if we can prove that when $f(y|\mathbf{x}) \in \mathcal{F}_1(\varepsilon_1)$ is the true model

$$z_n = \sum_{i,l} \log\left\{\frac{f(y_{il}|\mathbf{x}_i)}{f_1(y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\} = o_p(1)$$

then the asymptotic normality of $\mathcal{R}$ in (i) is proved since

$$\mathcal{R} = \mathcal{R}(f_0, f) - 2z_n$$

and $n\mathcal{D}(f_0, f) = O(1)$ according to Corollary 11. With the notation as in Proposition 10, under $f(y|\mathbf{x})$ to prove $z_n = o_p(1)$, we need to prove that for any $\epsilon > 0$

$$\lim_{n\to\infty} F^{(n)}\left(\left|\sum_{i,l} \log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \epsilon\right) = 0. \tag{3.35}$$

Notice that

$$\begin{aligned} & F^{(n)}\left(\left|\sum_{i,l} \log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \epsilon\right) \\ \leq\ & F^{(n)}\left(\sum_{i,l} \left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \epsilon\right) \\ \leq\ & F^{(n)}\left(\max_{i,l} \left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \frac{\epsilon}{n}\right). \end{aligned} \tag{3.36}$$

If we can prove that

$$\sum_{i,l} F^{(n)}\left(\left|\log\left\{\frac{f(Y_{il}|\mathbf{x}_i)}{f_1(Y_{il}|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right\}\right| > \frac{\epsilon}{n}\right) \to 0, \tag{3.37}$$

according to (3.36), (3.35) holds.

Notice that the Hellinger distance between $f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))$ and $f(y|\mathbf{x}_i)$ is at most $\varepsilon_1 = o(n^{-1/2})$, and then

$$\sum_{i=1}^{N} n_i \int \left(\sqrt{f(y|\mathbf{x}_i)} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right)^2 dy = o(1).$$

Based on (1.5) in Oosterhoff and van Zwet (2012), we have $F_1^{(n)}$ and $F^{(n)}$ are contiguous with respect to each other.

Then according to the proof of Theorem 2 in Oosterhoff and van Zwet (2012), to prove (3.37), it is equivalent to prove that for any $\epsilon > 0$

$$\lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\left\{|\log r(y|\mathbf{x}_i; f, f_1)| \geq \frac{\epsilon}{n}\right\}} \left(\sqrt{f(y|\mathbf{x}_i)} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right)^2 dy = 0. \tag{3.38}$$

It is true since

$$\begin{aligned} & \lim_{n\longrightarrow\infty} \sum_{i=1}^{N} n_i \int_{\left\{|\log r(y|\mathbf{x}_i; f, f_1)| \geq \frac{\epsilon}{n}\right\}} \left(\sqrt{f(y|\mathbf{x}_i)} - \sqrt{f_1(y|\mathbf{x}_i, \mu_1(\mathbf{x}_i))}\right)^2 dy \\ \leq\ & \lim_{n\longrightarrow\infty} 2n\varepsilon_1^2 = 0. \end{aligned}$$

We then finished the proof that the test statistic $\mathcal{R}$ is normally distributed with means being $-2n\mathcal{D}(f_0, f)$ and standard deviation $\sqrt{8n\mathcal{D}(f_0, f)}$ under $f(y|\mathbf{x})$. Similarly, we can prove (ii). ■

3.6 Proof of Proposition 14

For an arbitrary $f \in \mathcal{F}_1(\varepsilon_1)$ set $t = \mathcal{D}(f_0, f)$, and define $t_0 = \mathcal{D}(f_0, f_{1*})$, so that, by definition and assumption,

$$t \geq t_0 \geq -c. \tag{3.39}$$

In this notation we are to show that $\Phi\left(\frac{-c+t_0}{2\sqrt{t_0}}\right) \leq \Phi\left(\frac{-c+t}{2\sqrt{t}}\right)$, i.e. that

$$\frac{-c+t_0}{\sqrt{t_0}} \leq \frac{-c+t}{\sqrt{t}} \text{ for } t \geq t_0.$$

After a re-arrangement this condition becomes

$$-c \leq \sqrt{t_0}\sqrt{t}.$$

This is obvious if $c \geq 0$, otherwise it follows from (3.39). ■

3.7 Proof of Proposition 15

In the following, we denote $f(y|\mathbf{x}), f_0(y|\mathbf{x}, \mu_0(\mathbf{x})), f_1(y|\mathbf{x}, \mu_1(\mathbf{x}))$ by $f(y), f_0(y|\mu_0)$, $f_1(y|\mu_1)$. Define

$$\mathcal{L}(f(y), \lambda_1, \lambda_2) = f(y) \log \frac{f(y)}{f_0(y|\mu_0)} + \lambda_1 \sqrt{f(y) f_1(y|\mu_1)} + \lambda_2 f(y) \text{ for } y \in \Omega_{\mathbf{x}}.$$

For each fixed $y \in \Omega_{\mathbf{x}}$, the function $\mathcal{L}(f(y), \lambda_1, \lambda_2)$ is convex with respect to $f(y) > 0$. It follows that the critical point which is a solution to (2.19) is a minimizer of $\mathcal{L}(f(y), \lambda_1, \lambda_2)$. Then the solution to (2.19)-(2.21) is also the solution to the optimality problem (2.18), as we now show. Assume that $(f_{1*}(y|\mu_{1*}), \lambda_1, \lambda_2)$ is a solution to the equation system. For any $f(y)$ such that $f(y)$ vanishes on $\Omega_{\mathbf{x}}^c$ and satisfies the constraints of the optimization problem (2.18), it is clear that

$$\mathcal{L}(f(y), \lambda_1, \lambda_2) \geq \mathcal{L}(f_{1*}(y), \lambda_1, \lambda_2),$$

i.e.

$$\begin{aligned} & f(y)\log\left(\frac{f(y)}{f_0(y|\mu_0)}\right)+\lambda_1\sqrt{f(y)f_1(y|\mu_1)}+\lambda_2 f(y) \\ \geq\ & f_{1*}(y|\mu_{1*})\log\left(\frac{f_{1*}(y|\mu_{1*})}{f_0(y|\mu_0)}\right)+\lambda_1\sqrt{f_{1*}(y)f_1(y|\mu_1)}+\lambda_2 f_{1*}(y|\mu_{1*}), \end{aligned}$$

and then

$$\begin{aligned} \mathcal{I}\{f_0,f|\mu_0\} &\geq \int_{\Omega_{\mathbf{x}}} f_{1*}(y|\mu_{1*})\log\left(\frac{f_{1*}(y|\mu_{1*})}{f_0(y|\mu_0)}\right)dy \\ &\quad +\lambda_1\int_{\Omega_{\mathbf{x}}}\left(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)}-\sqrt{f(y)f_1(y|\mu_1)}\right)dy \\ &= \mathcal{I}\{f_0,f_{1*}|\mu_0,\mu_{1*}\} \\ &\quad +\lambda_1\int_{\Omega_{\mathbf{x}}}\left(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)}-\sqrt{f(y)f_1(y|\mu_1)}\right)dy \\ &\geq \mathcal{I}\{f_0,f_{1*}|\mu_0,\mu_{1*}\}, \end{aligned}$$

since

$$\lambda_1\int_{\Omega_{\mathbf{x}}}\left(\sqrt{f_{1*}(y|\mu_{1*})f_1(y|\mu_1)}-\sqrt{f(y)f_1(y|\mu_1)}\right)dy\geq 0.$$

Therefore, the solution to equations (2.19),(2.20) and (2.21) is the minimizer of the optimality problem (2.18).

The multiplier λ_1 is strictly negative. For, if $\lambda_1=0$ then according to (2.19) we have $f_{1*}(y|\mu_{1*})=f_0(y|\mu_0)\exp\{-1-\lambda_2\}$. However, constraint (2.20) then implies $\lambda_2=-1$ and $f_{1*}(y|\mu_{1*})=f_0(y|\mu_0)$. But then constraint (2.21) cannot be satisfied, since $\min_{\mathbf{x}\in\mathcal{S}} d_h(f_0,f_1|\mathbf{x})>\varepsilon_1$. Therefore, $\lambda_1<0$.

Finally, by using the constraints (2.20) and (2.21) the minimum of the optimality problem (2.18) can be simplified:

$$\mathcal{I}\{f_0,f_{1*}|\mu_0,\mu_{1*}\}=-1-\frac{1}{2}\lambda_1(1-\varepsilon_1^2)-\lambda_2.$$

■

3.8 Proof of Theorem 17

(i) The consistency of the LS estimates is a direct result of Theorem 3.1 in Sinha and Wiens (2003).

(ii) We prove that $\mathcal{D}(\boldsymbol{\xi}^{(n)},\widehat{\boldsymbol{\theta}}_n)-\max_{\boldsymbol{\xi}\in\mathcal{P}}\mathcal{D}(\boldsymbol{\xi},\boldsymbol{\theta}_n)\xrightarrow{pr}0$ by verifying that

(E1) $\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) \xrightarrow{pr} 0$ as $n_{init} \to \infty$,

(E2) $\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) \xrightarrow{pr} 0$ as $n_{init} \to \infty$.

We first prove (E1). Let $\boldsymbol{\xi}_n^*$ be a design such that $\mathcal{D}(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n) = \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n)$ and let $\boldsymbol{\xi}_{n0}$ be the design such that $\mathcal{D}(\boldsymbol{\xi}_{n0}, \boldsymbol{\theta}_n) = \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$. Then

$$L_n := \mathcal{D}(\boldsymbol{\xi}_{0n}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}_{0n}, \boldsymbol{\theta}_n) \le \mathcal{D}(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}_{0n}, \boldsymbol{\theta}_n) \le \mathcal{D}(\boldsymbol{\xi}_n^*, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}_n^*, \boldsymbol{\theta}_n) =: U_n.$$

Recall that

$$\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \xi_i \mathcal{I}\left\{f_0, f_{1*} | \mathbf{x}_i, \eta_0(\mathbf{x}_i|\boldsymbol{\theta}_0), \eta_1\left(\mathbf{x}_i|\boldsymbol{\theta}_1\right)\right\}.$$

According to condition (B6), the integrand $\mathcal{I}\left\{f_0, f_{1*} | \mathbf{x}, \eta_0(\mathbf{x}|\boldsymbol{\theta}_0), \eta_1\left(\mathbf{x}|\boldsymbol{\theta}_1\right)\right\}$ is Lipschitz continuous with respect to $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. Via the consistency of $\widehat{\boldsymbol{\theta}}_n$, and the linearity of $\mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta})$ with respect to ξ we have that for any design ξ,

$$\left|\mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)\right| \le \max_{i=1,\ldots,N} \left| \begin{array}{c} \mathcal{I}\left\{f_0, f_{1*} | \mathbf{x}_i, \eta_0(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{0n}), \eta_1\left(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{1n}\right)\right\} \\ -\mathcal{I}\left\{f_0, f_1 | \mathbf{x}_i, \eta_0(\mathbf{x}_i|\boldsymbol{\theta}_{0n}), \eta_1(\mathbf{x}_i|\boldsymbol{\theta}_{1n})\right\} \end{array} \right| \xrightarrow{a.s.} 0. \tag{3.40}$$

Therefore, $L_n, U_n \xrightarrow{a.s.} 0$ and (E1) follows.

To prove (E2) we first write

$$\begin{aligned} & \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n) \\ = & \left(\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}})\right) + \left(\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{init}})\right) \\ & + \left(\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_n)\right). \end{aligned} \tag{3.41}$$

The first and last terms in (3.41) converge to 0 in probability as $n_{init} \to \infty$, due to (3.40) and (E1). Then it suffices to prove that, for any $\varepsilon > 0$,

$$\Pr\left(\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \ge -\varepsilon\right) \to 1. \tag{3.42}$$

Recall that given the $(n-1)^{th}$ design $\boldsymbol{\xi}^{(n-1)}$ and the estimates $\widehat{\boldsymbol{\theta}}_n$, the next design point is

$$\mathbf{x}_{new} = \arg\max_{i=1,\ldots,N} \mathcal{I}\left\{f_0, f_{1*} | \mathbf{x}_i, \eta_0(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{0n}), \eta_1\left(\mathbf{x}_i|\widehat{\boldsymbol{\theta}}_{1n}\right)\right\}.$$

Then the n^{th} design is

$$\boldsymbol{\xi}^{(n)} = \frac{n-1}{n}\boldsymbol{\xi}^{(n-1)} + \frac{1}{n}\delta_n(\mathbf{x}),$$

where $\delta_n(\mathbf{x}) = I(\mathbf{x} = \mathbf{x}_{new})$. Therefore, the KL-divergence for the n^{th} design is

$$\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) = \frac{n-1}{n}\mathcal{D}(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{1}{n}\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}})$$

and the difference between the KL-divergence with the n^{th} design and that with $(n-1)^{th}$ design is

$$\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) = \frac{\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}})}{n-1}. \tag{3.43}$$

Now to establish (3.42), denote $\boldsymbol{\xi}^*_{init} = \arg\max_{\boldsymbol{\xi}\in\mathcal{P}} \mathcal{D}(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}_{n_{init}})$. For any $\varepsilon > 0$, divide the sequence $\left\{\boldsymbol{\xi}^{(n)}\right\}$ into two disjoint subsequences $S_1(\varepsilon)$ and $S_2(\varepsilon)$ such that

$$\begin{aligned} S_1(\varepsilon) &: = \left\{\boldsymbol{\xi}^{(n)} : \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \geq \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \varepsilon/2\right\}, \\ S_2(\varepsilon) &: = \left\{\boldsymbol{\xi}^{(n)} : \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) < \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \varepsilon/2\right\}. \end{aligned}$$

We first show that $S_1(\varepsilon)$ is non-empty for each $\varepsilon > 0$. If not, there must exist an ε such that for any n we have

$$\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) < \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \varepsilon/2.$$

Then according to (3.43), and with

$$z_{1n} = \left(\mathcal{D}(\delta_n(\mathbf{x}); \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\delta_n(\mathbf{x}); \widehat{\boldsymbol{\theta}}_n)\right) + \left(\mathcal{D}(\boldsymbol{\xi}^*_{init}; \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}^*_{init}; \widehat{\boldsymbol{\theta}}_{n_{init}})\right),$$

we have that

$$\begin{aligned} & \mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^{(n-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \\ > & \frac{\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \varepsilon/2}{n-1} \\ = & \frac{\varepsilon}{2(n-1)} + \frac{z_{1n}}{n-1} + \frac{\mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}})}{n-1} \\ \geq & \frac{\varepsilon}{2(n-1)} + \frac{z_{1n}}{n-1}, \end{aligned} \tag{3.44}$$

since $\mathcal{D}(\delta_n(\mathbf{x}), \widehat{\boldsymbol{\theta}}_n) \geq \mathcal{D}(\xi^*_{Init}, \widehat{\boldsymbol{\theta}}_n)$ by the definition of $\delta_n(\mathbf{x})$.

We can prove $z_{1n} \overset{a.s.}{\to} 0$ by applying (3.40). According to the proof of Theorem 3.1(i) in Sinha and Wiens (2003), there exists small enough q (> 0) such that $\lim_{n\to\infty} n^q(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) < \infty$ almost surely. Then because of condition (B6) (or (B6′)),

we also have $\lim_{n\to\infty} n^q z_{1n} < \infty$ almost surely. Moreover, since z_{1n} is bounded by the maximum KL-divergence, we have

$$\sum_{m=1}^{\infty} \frac{z_{1m}}{m} < \infty \ a.s. \tag{3.45}$$

Since $n_{init}/n \to 0$ as $n_{init} \to \infty$ we have

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\xi}^{(n)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) &= \mathcal{D}(\boldsymbol{\xi}^{(n_{init})}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \sum_{m=n_{Init}+1}^{n} \Big(\mathcal{D}(\boldsymbol{\xi}^{(m)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^{(m-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \Big) \\
&> \mathcal{D}(\boldsymbol{\xi}^{(n_{init})}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \sum_{m=n_{Init}+1}^{n} \left(\frac{\varepsilon}{2(m-1)} + \frac{z_{1m}}{m} \right) \overset{a.s.}{\to} \infty
\end{aligned}$$

as $n_{init} \to \infty$, a contradiction to the assumption that the maximum KL-divergence is finite. Therefore, for any $\varepsilon > 0$, $S_1(\varepsilon)$ is nonempty and we can find a sequence $\left\{ \boldsymbol{\xi}^{(n_l)} \right\}_{l=1}^{\infty} \subset S_1(\varepsilon)$, i.e. $\mathcal{D}(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{init}})$ arbitrarily close to $\mathcal{D}(\boldsymbol{\xi}^{*}_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}})$. By (3.43) we have

$$\begin{aligned}
& \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \\
= \ & \mathcal{D}(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}})}{n_l} \\
= \ & \mathcal{D}(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1})}{n_l} \\
& + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1})}{n_l} + \frac{\mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}})}{n_l} \\
= \ & \mathcal{D}(\boldsymbol{\xi}^{(n_l)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1})}{n_l} + \frac{z_{2n_l}}{n_l} \\
\geq \ & \mathcal{D}(\boldsymbol{\xi}^{*}_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \frac{\varepsilon}{2} + \frac{z_{2n_l}}{n_l},
\end{aligned}$$

for $\boldsymbol{\xi}^{(n_l+1)} \in S_1(\varepsilon)$ or $S_2(\varepsilon)$, where

$$z_{2n} = \Big(\mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\delta_{n_l+1}(\mathbf{x}), \widehat{\boldsymbol{\theta}}_{n_l+1}) \Big) + \Big(\mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_l+1}) - \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \Big)$$

and, similar to z_{1n}, we also have $z_{2n} \to 0$ a.s.

In summary, as in (3.44), for all $\boldsymbol{\xi}^{(n_k)} \in S_2(\varepsilon)$ we have

$$\mathcal{D}(\boldsymbol{\xi}^{(n_k)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) > \mathcal{D}(\boldsymbol{\xi}^{(n_k-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{\varepsilon}{2(n_k-1)} + \frac{z_{1n_k}}{n_k} > \mathcal{D}(\boldsymbol{\xi}^{(n_k-1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \frac{z_{1n_k}}{n_k},$$

iterating this gives

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\xi}^{(n_k)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) &> \mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) + \sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m} \\
&> \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \frac{\varepsilon}{2} + \frac{z_{2n_l}}{n_l} + \sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m} \\
&= \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \frac{\varepsilon}{2} + Y_{n_k},
\end{aligned}$$

with

$$Y_n = \begin{cases} \frac{z_{2n_l}}{n_l}, \text{ if } n = n_l + 1 \\ \frac{z_{2n_l}}{n_l} + \sum_{m=n_l+2}^{n_k} \frac{z_{1m}}{m}, \text{if } n \neq n_l + 1 \end{cases}.$$

Notice that $Y_n \xrightarrow{pr} 0$ by (3.45). Therefore, for any $\varepsilon > 0$,

$$\Pr\left(\mathcal{D}(\boldsymbol{\xi}^{(n_l+1)}, \widehat{\boldsymbol{\theta}}_{n_{init}}) - \mathcal{D}(\boldsymbol{\xi}^*_{init}, \widehat{\boldsymbol{\theta}}_{n_{init}}) \geq -\varepsilon\right) \to 1$$

as $n_{init} \to \infty$. Then we have proved (3.42) holds which implies (E2). ■

Chapter 4

Robust design for the estimation of a threshold probability

Abstract We consider the construction of robust sampling designs for the estimation of threshold probabilities in spatial studies. A threshold probability is a probability that the value of a stochastic process at a particular location exceeds a given threshold. We propose designs which estimate a threshold probability efficiently, and also deal with two possible model uncertainties: misspecified regression responses and misspecified variance/covariance structures. The designs minimize a loss function based on the relative mean squared error of the predicted values (i.e., relative to the true values). To this end an asymptotic approximation of the loss function is derived. To address the uncertainty of the variance/covariance structures of this process, we average this loss over all such structures in a neighbourhood of the experimenter's nominal choice. We then maximize this averaged loss over a neighbourhood of the experimenter's fitted model. Finally the maximum is minimized, to obtain a minimax design.

Key words and phrases Environmental monitoring; Increasing domain asymptotics; Robust design; Spatial sampling design; Threshold probability; Universal kriging

4.1 Introduction

Stochastic processes are widely used in the study of, in particular, environmental phenomena. For instance, whether or not a stochastic process is larger than a certain critical level is of ecological interest. Consider aerosol optical depth (AOD) as an

example. AOD is an indirect measure of radiative forcing and air quality (Oleson, Kumar and Smith 2012). High AOD (> 0.6) has been shown to have detrimental effects on people's health, and to adversely affect cloud formation (Prasad, Singh, Singh and Kafatos 2005).

Let $\{Y(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ be a stochastic process, for instance representing realized AOD. Here $\mathbf{t} = (t_1, ..., t_d)^T$, $d \geq 1$, is a "location" – we use the term loosely, to include spatial locations and possibly non-spatial covariates as well – from a set $\mathcal{T} = \{\mathbf{t}_1, \cdots, \mathbf{t}_N\} \subset \mathbb{R}^d$ at which the process might be measured. Observations $\mathbf{y}_n = (y(\mathbf{t}_{s1}), \cdots, y(\mathbf{t}_{sn}))^T$ will be obtained once sample locations $\mathcal{S} = \{\mathbf{t}_{s1}, \cdots, \mathbf{t}_{sn}\} \subset \mathcal{T}$ are chosen. A model used to describe the observed data is

$$Y(\mathbf{t}) = \mu(\mathbf{t}) + \varepsilon(\mathbf{t}) \tag{4.1}$$

where $\mu(\mathbf{t}) = \eta(\mathbf{t}) + \delta(\mathbf{t})$, $\eta(\mathbf{t})$ is the deterministic mean perturbed by stochastic errors $\delta(\mathbf{t})$, and $\varepsilon(\mathbf{t})$ is uncorrelated, additive measurement error. Moreover, $\{\delta(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ and $\{\varepsilon(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ are independent. Our interest in this article is to construct a design which is robust against some uncertainties for the estimation of the probability that the value of the μ process, at a given location $\mathbf{t}$, is above a fixed threshold u_*, i.e.

$$z(\mathbf{t}) = P\left(\mu(\mathbf{t}) > u_*\right).$$

We refer to $z(\mathbf{t})$ as the "threshold probability".

The problem of design related to a probability has received considerable attention. In Bect, Ginsbourger, Li, Picheny and Vazquez (2012) sequential designs were proposed for the estimation of the volume of the excursion set of a Gaussian process ς above a fixed threshold u_*, $\alpha = P\{x : \varsigma(x) > u_*\}$. This probability corresponds to a probability of a system failure in the industrial world.

Gaussian processes are among the most widely used stochastic processes for modelling dependent data (Fahrmeir and Tutz 1994; Rosenblatt 2000). Thus, in the following context, both $\delta(\mathbf{t})$ and $\varepsilon(\mathbf{t})$ are Gaussian. Moreover, assume that $\{\delta(\mathbf{t})|\mathbf{t} \in \mathcal{T}\}$ has covariance matrix $\mathbf{G}_{N\times N} = (g(\mathbf{t}_i, \mathbf{t}_j))_{i,j=1}^N$ and $\{\varepsilon(\mathbf{t})|\mathbf{t} \in \mathcal{T}\}$ has known, uniformly bounded variance matrix $\mathbf{H}_{N\times N}$ =diag$(h(\mathbf{t}_i))_{i=1}^N$, with $h(\mathbf{t}) \in [h_1, h_N]$ for all $\mathbf{t} \in \mathcal{T}$ and $0 < h_1 \leq h_N < \infty$. A natural and optimal (Cressie 1993) estimator of $z(\mathbf{t})$ is

$$\hat{z}_n(\mathbf{t}) := E\left[\mathbf{1}(\mu(\mathbf{t}) > u_*)|\mathbf{y}_n\right].$$

Under the Gaussian assumption, and a linear structure for $E(Y(\mathbf{t}))$, we shall obtain an explicit form of $\hat{z}_n(\mathbf{t})$. This is based on the result in Lemma 23 below. Before stating this lemma, we introduce some notation.

When $\mathbf{t} = \mathbf{t}_i$, let $\mathbf{e_t}$ denote the i^{th} column of $\mathbf{I}_N$. Let $\boldsymbol{\xi}$ be the $N \times 1$ "design" vector, with elements $\xi_i = I(\mathbf{t}_i \in \mathcal{S})$. Here $I(\cdot)$ is an indicator function. Let $(\mathbf{M}_\xi)_{n\times N}$ be those rows of $\mathbf{I}_N$ for which $\xi_i = 1$, and $(\mathbf{L}_\xi)_{(N-n)\times N}$ those for which $\xi_i = 0$. Then $\left(\begin{array}{cc} \mathbf{M}_\xi^T & \mathbf{L}_\xi^T \end{array}\right)^T$ is a permutation of the rows of $\mathbf{I}_N$ and for $\mathbf{t} \in \mathcal{T}$,

$$\begin{aligned} \mathrm{Var}(\mu(\mathbf{t})) &= \mathrm{Var}(\delta(\mathbf{t})) = \mathbf{e}_\mathbf{t}^T \mathbf{G} \mathbf{e}_\mathbf{t} =: \sigma_\mathbf{t}^2(\mathbf{G}), & (4.2)\\ \mathrm{Var}(\mathbf{y}_n) &= \mathbf{M}_\xi(\mathbf{G}+\mathbf{H})\mathbf{M}_\xi^T =: \boldsymbol{\Sigma}_{nn}(\mathbf{G}), & (4.3)\\ \mathrm{Cov}(\mu(\mathbf{t}), \mathbf{y}_n^T) &= \mathbf{e}_\mathbf{t}^T \mathbf{G} \mathbf{M}_\xi^T =: \boldsymbol{\Sigma}_{n1\mathbf{t}}^T(\mathbf{G}). & (4.4) \end{aligned}$$

We write $\mu(\mathbf{t})|\mathbf{y}_n \sim GP(\hat{\mu}(\mathbf{t}), \sigma_{n\mathbf{t}}^2(\mathbf{G}))$ if the conditional mean and covariance functions of the Gaussian process $\mu(\mathbf{t})|\mathbf{y}_n$ are $\hat{\mu}(\mathbf{t})$ and $\sigma_{n\mathbf{t}}^2(\mathbf{G})$, respectively.

Lemma 23 *Assume that $E(Y(\mathbf{t})) = \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta}$ for a $p-$dimensional vector of functions $\mathbf{f}(\mathbf{t}) = (f_1(\mathbf{t}), \cdots, f_p(\mathbf{t}))^T$, and that $\boldsymbol{\theta} \sim \mathcal{U}_{\mathbb{R}^p}$, the improper uniform distribution over $\mathbb{R}^p$. Define $\mathbf{F}_n = (\mathbf{f}(\mathbf{t}_{s1}), \cdots, \mathbf{f}(\mathbf{t}_{sn}))^T$. Then*

$$\mu(\mathbf{t})|\mathbf{y}_n \sim GP(\hat{\mu}(\mathbf{t}), \sigma_{n\mathbf{t}}^2(\mathbf{G})),$$

where the conditional mean of $\mu(\mathbf{t})$ (given data $\mathbf{y}_n$) is the best linear unbiased predictor (BLUP)

$$\hat{\mu}(\mathbf{t}) = \mathbf{a}_{n\mathbf{t}}^T(\mathbf{G})\mathbf{y}_n,$$

with

$$\mathbf{a}_{n\mathbf{t}}^T(\mathbf{G}) = \left\{ \begin{array}{c} \mathbf{f}^T(\mathbf{t})(\mathbf{F}_n^T \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{G})\mathbf{F}_n)^{-1}\mathbf{F}_n^T \\ +\boldsymbol{\Sigma}_{n1\mathbf{t}}^T(\mathbf{G})(\mathbf{I} - \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{G})\mathbf{F}_n(\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{G})\mathbf{F}_n)^{-1}\mathbf{F}_n^T) \end{array} \right\} \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{G}). \quad (4.5)$$

The conditional variance $\sigma_{n\mathbf{t}}^2(\mathbf{G})$ of $\mu(\mathbf{t})$ given data $\mathbf{y}_n$ is the mean squared prediction error of the BLUP,

$$\begin{aligned} \sigma_{n\mathbf{t}}^2(G) &= Var(\mu(t) - \hat{\mu}(t)) \\ &= \sigma_\mathbf{t}^2(\mathbf{G}) - (\mathbf{f}^T(\mathbf{t}), \boldsymbol{\Sigma}_{n1\mathbf{t}}^T(\mathbf{G})) \left(\begin{array}{cc} \mathbf{0} & \mathbf{F}_n^T \\ \mathbf{F}_n & \boldsymbol{\Sigma}_{nn}(\mathbf{G}) \end{array}\right)^{-1} \left(\begin{array}{c} \mathbf{f}(\mathbf{t}) \\ \boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{G}) \end{array}\right). \quad (4.6) \end{aligned}$$

The proof of this lemma is similar to that of Theorem 4.1.1 in Santner, Williams and Notz (2003) and is omitted. Based on Lemma 23, we have

$$\hat{z}_n(\mathbf{t}) = \Phi\left(\frac{\hat{\mu}(\mathbf{t}) - u_*}{\sigma_{n\mathbf{t}}(\mathbf{G})}\right).$$

Note that the estimator $\hat{z}_n(\mathbf{t})$ is derived under quite restrictive assumptions:
(U1) the variance/covariance structure of $Y(\mathbf{t})$ is known.
These kinds of structures are only approximately known in applications. For instance, aerosol concentrations (and therefore the value of AOD) at some sites are significantly affected by those of adjoining locations. But the correlations can be influenced by meteorological conditions whose effect on the correlation is difficult to measure.
(U2) The mean response $E(Y(\mathbf{t}))$ is linear in these regressors $\mathbf{f}(\mathbf{t})$.
In real life, the mean structure is generally only partially known and unlikely to be exactly linear.

We anticipate that the investigator will not address these challenges at the estimation/prediction stage, but hopes to do so through the design. That is, the estimate $\hat{z}_n(\mathbf{t})$ is still computed based on the possibly incorrect response and a nominal covariance matrix $\mathbf{G_0}$. Our interest is to develop a design such that the estimate of $z(\mathbf{t})$ is robust against uncertainties (**U1**) and (**U2**).

The necessity of robust design in spatial studies has been documented in the literature. Wiens (2005a) obtained a robust predictor of $Y(\mathbf{t})$ in the face of uncertainty (**U1**). As a natural sequel to Wiens (2005a), Wiens (2005b) considered designs robust against misspecified models. Wiens and Zhou (2008) considered the construction of robust designs for test-control field experiment with particular attention paid to the effects of both (**U1**) and (**U2**).

To address the first uncertainty – (**U1**) – we assume that the covariance matrix $\mathbf{G}$ varies over a neighbourhood $\mathcal{G}$. We entertain a scenario under which the true covariance matrix $\mathbf{G}$ of $\{\delta(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ is an element of the random set

$$\mathcal{G} = \left\{ \begin{array}{c} \mathbf{G_d}\text{: } \mathbf{G_d} = \mathbf{U}\text{diag}(\lambda_i e^{\mathbf{d}/\sqrt{n}})_{i=1}^N \mathbf{U}^T, \\ \mathbf{d} \text{ is a bounded random variable with} \\ \text{mean 0 and standard deviation } \omega_{\mathbf{d}}^2\ , \\ \text{and } -\infty < d_1 \leq \mathbf{d} \leq d_2 < \infty \end{array} \right\}.$$

Here $\lambda_1 \leq ... \leq \lambda_N$ are the eigenvalues of a nominal, positive definite covariance matrix $\mathbf{G_0}$ specified by the experimenter, and $\mathbf{U}$ is the orthogonal matrix whose columns are the corresponding eigenvectors.

Where necessary, for $\mathbf{G_d} \in \mathcal{G}$, we simply denote the functions of $\mathbf{G_d}$ by functions of $\mathbf{d}$. For example, $\mathbf{\Sigma}_{nn}(\mathbf{G_d})$ and $\mathbf{\Sigma}_{n1\mathbf{t}}(\mathbf{G_d})$ as defined in (4.3) and (4.4) will be denoted as $\mathbf{\Sigma}_{nn}(\mathbf{d}), \mathbf{\Sigma}_{n1\mathbf{t}}(\mathbf{d})$, respectively. Similarly, $\sigma^2_{\mathbf{t}}(\mathbf{G_d})$ in (4.2), $\mathbf{a}_{n\mathbf{t}}(\mathbf{G_d})$ and $\sigma^2_{n\mathbf{t}}(\mathbf{G_d})$ in Lemma 23 will be denoted as $\sigma^2_{\mathbf{t}}(\mathbf{d})$, $\mathbf{a}_{n\mathbf{t}}(\mathbf{d})$, $\sigma^2_{n\mathbf{t}}(\mathbf{d})$. The distribution of $Y(\mathbf{t})$ is then determined by the random variable $\mathbf{d}$. Given the regression parameter $\boldsymbol{\theta}$, and given $\mathbf{d}$, the conditional distribution of $Y(\mathbf{t})$ is that $Y(\mathbf{t})|(\boldsymbol{\theta}, \mathbf{d}) \sim N(\mathbf{f}^T(\mathbf{t})\boldsymbol{\theta}, \sigma^2_{\mathbf{t}}(\mathbf{d}) + h(\mathbf{t}))$.

To address the second uncertainty – (**U2**) – we suppose that $E(Y(\mathbf{t})) \approx \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta}$. Without further restrictions the parameter vector $\boldsymbol{\theta}$ is unidentifiable; we shall define it by

$$\boldsymbol{\theta} = \arg\min_{\mathbf{v}} \sum_{\mathbf{t}\in\mathcal{T}} \left(E(Y(\mathbf{t})) - \mathbf{f}^T(\mathbf{t})\mathbf{v}\right)^2.$$

Define $\psi(\mathbf{t})$ by

$$E(Y(\mathbf{t})) = \eta(\mathbf{t}) = \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} + \frac{\psi(\mathbf{t})}{\sqrt{n}}. \tag{4.7}$$

The factor $n^{-1/2}$ is for the asymptotics, and is analogous to the requirement of contiguity in the asymptotic theory of hypothesis testing. These definitions imply the orthogonality condition

$$\mathbf{F}_N^T \mathbf{\Psi}_N := \sum_{\mathbf{t}\in\mathcal{T}} \mathbf{f}(\mathbf{t})\psi(\mathbf{t}) = \mathbf{0}$$

where $\mathbf{F}_N = (\mathbf{f}(\mathbf{t}_1), \cdots, \mathbf{f}(\mathbf{t}_N))^T$ and $\mathbf{\Psi}_N = (\psi_1, \cdots, \psi_N)^T$ with $\psi_i = \psi(\mathbf{t}_i)$. We then let $\mathbf{\Psi}_N$ vary over a set quantifying the model uncertainty:

$$\mathbf{\Psi} = \{\mathbf{\Psi}_N : \mathbf{F}_N^T\mathbf{\Psi}_N = 0,\ ||\mathbf{\Psi}_N|| \le \tau^2\ \},$$

where $\|\cdot\|$ is the Euclidean norm. With the above notations, the true model should be

$$\begin{aligned} Y(\mathbf{t}) &= \eta(\mathbf{t}) + \varepsilon(\mathbf{t}) + \delta(\mathbf{t}) \\ &= \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} + \frac{\psi(\mathbf{t})}{\sqrt{n}} + \varepsilon(\mathbf{t}) + \delta(\mathbf{t}) \\ &= \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} + \frac{\mathbf{e}_{\mathbf{t}}^T\mathbf{\Psi}_N}{\sqrt{n}} + \varepsilon(\mathbf{t}) + \delta(\mathbf{t}). \end{aligned}$$

Let $\boldsymbol{\eta}_n = (\eta(\mathbf{t}_{s1}), ..., \eta(\mathbf{t}_{sn}))^T$, $\boldsymbol{\delta}_n = (\delta(\mathbf{t}_{s1}), ..., \delta(\mathbf{t}_{sn}))^T$, $\boldsymbol{\varepsilon}_n = (\varepsilon(\mathbf{t}_{s1}), ..., \varepsilon(\mathbf{t}_{sn}))^T$ and $\mathbf{\Psi}_n = (\psi(\mathbf{t}_{s1}), ..., \psi(\mathbf{t}_{sn}))^T$. Notice that

$$\boldsymbol{\eta}_n = \mathbf{F}_n\boldsymbol{\theta} + \frac{\mathbf{\Psi}_n}{\sqrt{n}},$$

$\boldsymbol{\delta}_n + \boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\mathbf{z}_n$ with $\mathbf{z}_n \sim N_n(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Psi}_n = \mathbf{M}_\xi \boldsymbol{\Psi}_N$. With these notations,

$$\begin{aligned} \mathbf{y}_n &= \boldsymbol{\eta}_n + \boldsymbol{\delta}_n + \boldsymbol{\varepsilon}_n \\ &= \mathbf{F}_n\boldsymbol{\theta} + \frac{\boldsymbol{\Psi}_n}{\sqrt{n}} + \boldsymbol{\delta}_n + \boldsymbol{\varepsilon}_n \\ &= \mathbf{F}_n\boldsymbol{\theta} + \frac{\mathbf{M}_\xi \boldsymbol{\Psi}_N}{\sqrt{n}} + \boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\mathbf{z}_n. \end{aligned} \tag{4.8}$$

An optimally robust design $\boldsymbol{\xi}^*$ is a design that optimizes a chosen loss function $\mathcal{L}_0$ in the face of uncertainties. This loss will be averaged, with respect to a "prior" distribution on $\mathbf{d}$, as a means of relaxing (**U1**). The "averaged" loss is then maximized over $\boldsymbol{\Psi}$ to handle the non-robustness source (**U2**). An optimal design is the one that minimizes this maximized loss. Here, we choose the loss function $\mathcal{L}_0$ to be the relative conditional mean squared prediction error (MSPE), averaged over locations in $\mathcal{T}\backslash\mathcal{S}$ at which observations are not obtained:

$$\mathcal{L}_0\left(\boldsymbol{\xi}|\boldsymbol{\Psi}_\mathbf{N}, \boldsymbol{\theta}, \mathbf{d}\right) = \frac{1}{N-n}\sum_{\mathbf{t}\in\mathcal{T}\backslash\mathcal{S}} \frac{E_{\mathbf{y}_n|\boldsymbol{\theta},\mathbf{d}}\left(z(\mathbf{t}) - \hat{z}_n(\mathbf{t})\right)^2}{z^2(\mathbf{t})}.$$

Upon taking an expectation with respect to $\mathbf{d}$, the loss becomes

$$\mathcal{L}_0\left(\boldsymbol{\xi}|\boldsymbol{\Psi}_\mathbf{N}, \boldsymbol{\theta}\right) = \frac{1}{N-n}\sum_{\mathbf{t}\in\mathcal{T}\backslash\mathcal{S}} E_{\mathbf{d}|\boldsymbol{\theta}}\left(\frac{E_{\mathbf{y}_n|\boldsymbol{\theta},\mathbf{d}}\left(z(\mathbf{t}) - \hat{z}_n(\mathbf{t})\right)^2}{z^2(\mathbf{t})}\right). \tag{4.9}$$

A problem that arises immediately is that this loss depends on the unknown values of the parameters $\boldsymbol{\theta}$. There are various methods of handling this problem. One is by constructing a "locally optimal" design – one that is optimal only at a particular value $\boldsymbol{\theta}_0$ of the parameter. This local parameter value might arise from the experimenter's prior knowledge, or perhaps as an estimate from an earlier experiment. To allow for uncertainty about the parameter values, one might first maximize the loss function over a neighbourhood of a local parameter $\boldsymbol{\theta}_0$ and then minimize the maximized loss function over the class of designs. For details of this second approach see Atkinson and Fedorov (1975a,b), King and Wong (2000), Dette and Biedermann (2003), López-Fidalgo, Tommasi and Trandafir (2007), Dette and Pepelyshev (2008). Bayesian methods are also applicable in eliminating the parameters from the loss function – the loss function is averaged, with respect to a prior distribution on the parameters before being minimized (Dette and Neugebauer 1997, Karami and Wiens 2014). These three methods allow for static, i.e. non-sequential, design construction.

A method which is arguably more practical is sequential design. In this approach (Hunter and Reiner 1965; Fedorov and Pazman 1968; Sinha and Wiens 2002), estimates are computed using the available data and subsequent observations are made at new design points minimizing the loss function, evaluated at the current estimates. Here we propose such a strategy.

In Section 2, we give an expansion of the loss function (4.9) up to and including terms that are $O(n^{-1})$ under the increasing domain asymptotic framework – this derivation is in the appendix. We then approximate the loss by the terms of this expansion, ignoring the remainder. We maximize, over $\mathbf{\Psi}$, the average, over $\mathbf{d}$, value of this asymptotic loss. This maximized loss is exhibited in Proposition 26 in Section 3. In Section 4, we describe and apply the sequential algorithm to choose designs minimizing the maximum loss. In order to illustrate an application of the construction of robust designs, we find optimal designs for the coal-ash data (Gomez and Hazen 1970, Cressie 1993), where the fitted model is

$$E(Y(\mathbf{t})) = \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta}$$

with $\mathbf{f}^T(\mathbf{t}) = (1, t_1, t_2), \mathbf{t} = (t_1, t_2)$. The nominal correlation structure of the stochastic error δ, from which $\mathbf{G_0}$ is computed, is determined by a certain isotropic correlation/covariance function.

4.2 Increasing domain asymptotics and expansion of the loss

Our 'large n and N' approach is one of increasing domain, rather than infill, asymptotics (Cressie 1993). Mardia and Marshall (1984) discussed asymptotic normality of the MLE under an increasing domain framework. Evangelou and Zhu (2012) assumed increasing domain asymptotics and proposed optimal prediction designs for spatial generalized linear mixed models. In this section, we expand the loss function $\mathcal{L}_0\left(\boldsymbol{\xi}|\mathbf{\Psi}_N, \boldsymbol{\theta}\right)$ under an increasing domain asymptotic framework in the spirit of Mardia and Marshall (1984).

To discuss the asymptotic properties, we require the following definition.

Definition 24 *A matrix* $\mathbf{A}$ *is* $\mathbf{O}(n^{-q}), q \in \mathbb{Z}$, *if* $||\mathbf{A}|| = O(n^{-q})$, *where* $||\mathbf{A}||^2$ *is the maximum eigenvalue of* $\mathbf{A}^T\mathbf{A}$.

Assumptions (A2)-(A4), listed below, are similar to those in Mardia and Marshall (1984); they together with (A1) are critical for the discussion of asymptotic expansion of the loss.

(A1) The eigenvalues $\lambda_1 \leq ... \leq \lambda_N$ of $\mathbf{G_0}$ are bounded, and bounded away from 0, as $N \to \infty$.

(A2) $(\mathbf{F}_n^T \mathbf{F}_n)^{-1} = \mathbf{O}(n^{-1})$ and $\mathbf{F}_n^T \mathbf{F}_n = \mathbf{O}(n)$.

(A3) $\mathbf{\Sigma}_{n1\mathbf{t}}(0) = \mathbf{O}(1)$, where $\mathbf{\Sigma}_{n1\mathbf{t}}^T(0) = \mathbf{e_t^T G_0 M}_\xi^T$.

(A4) $n/N \to r$ with $r \in (0, 1]$.

At each location $\mathbf{t} \in \mathcal{T} \backslash \mathcal{S}$, the expected relative mean squared prediction error (MSPE) $E_{\mathbf{d}|\boldsymbol{\theta}} \left(\frac{E_{\mathbf{y}_n|\mathbf{d},\mathbf{\Psi}_N,\boldsymbol{\theta}}(z(\mathbf{t}) - \hat{z}_n(\mathbf{t}))^2}{z^2(\mathbf{t})} \right)$ satisfies

$$E_{\mathbf{d}|\boldsymbol{\theta}} \left(\frac{E_{\mathbf{y}_n|\boldsymbol{\theta},\mathbf{d}} \left(z(\mathbf{t}) - \hat{z}_n(\mathbf{t}) \right)^2}{z^2(\mathbf{t})} \right) = E_{\mathbf{d},\mathbf{y_n}|\boldsymbol{\theta}} \left(1 - F_{\mathbf{t}} \left(\mathbf{\Psi_N}, \mathbf{d} \right) \right)^2, \tag{4.10}$$

where $F_{\mathbf{t}} \left(\mathbf{\Psi_N}, \mathbf{d} \right)$ is the ratio between the threshold probability $z(\mathbf{t})$ at location $\mathbf{t}$ and its estimate $\hat{z}_n(\mathbf{t})$

$$F_{\mathbf{t}} \left(\mathbf{\Psi_N}, \mathbf{d} \right) = \frac{\hat{z}_n(\mathbf{t})}{z(\mathbf{t})}. \tag{4.11}$$

Recall that according to Lemma 23

$$\hat{z}_n(\mathbf{t}) = \Phi \left(\frac{\hat{\mu}(\mathbf{t}) - u_*}{\sigma_{n\mathbf{t}}(\mathbf{d})} \right), \tag{4.12}$$

where $\hat{\mu}(\mathbf{t}) = \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{y}_n$ is the BLUP of $\mu(\mathbf{t})$ and $\sigma_{n\mathbf{t}}(\mathbf{d})$ is the mean squared prediction error of the BLUP as defined in Lemma 23.

Notice that

$$\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{F}_n = \mathbf{f}^T(\mathbf{t}),$$

because the BLUP is unbiased. Substituting (4.8) into $\hat{\mu}(\mathbf{t}) = \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{y}_n$, then $\hat{z}_n(\mathbf{t})$ becomes

$$\hat{z}_n(\mathbf{t}) = \Phi \left(\widetilde{y}_{n\mathbf{dt}} \sigma_{n\mathbf{t}}(0) \sigma_{n\mathbf{t}}^{-1}(\mathbf{d}) + \frac{\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{M}_\xi}{\sigma_{n\mathbf{t}}(\mathbf{d})} \frac{\mathbf{\Psi}_N}{\sqrt{n}} \right), \tag{4.13}$$

where

$$\widetilde{y}_{n\mathbf{dt}} = \frac{\mathbf{b}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{z}_n + \mathbf{f}^T(\mathbf{t}) \boldsymbol{\theta} - u_*}{\sigma_{n\mathbf{t}}(0)},$$

with $\mathbf{b}_{n\mathbf{t}}^T(\mathbf{d}) := \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{\Sigma}_{nn}^{1/2}(\mathbf{d})$.

Recall that the threshold probability $z(\mathbf{t})$ at location $\mathbf{t}$ is defined as

$$z(\mathbf{t}) = P \left(\mu(\mathbf{t}) > u_* \right),$$

where $\mu(\mathbf{t}) = \eta(\mathbf{t}) + \delta(\mathbf{t})$. Since $\delta(\mathbf{t}) \sim N(0, \sigma_{\mathbf{t}}(\mathbf{d}))$ where $\sigma_{\mathbf{t}}(\mathbf{d})$ is the variance of $\delta(\mathbf{t})$ as defined in (4.2), the threshold probability $z(\mathbf{t})$ is

$$z(\mathbf{t}) = \Phi\left(\frac{\eta(\mathbf{t}) - u_*}{\sigma_{\mathbf{t}}(\mathbf{d})}\right).$$

Substituting (4.7), the true model of $\eta(\mathbf{t})$, into $z(\mathbf{t})$, we have

$$z(\mathbf{t}) = \Phi\left(x_{0\mathbf{t}}\sigma_{\mathbf{t}}(0)\sigma_{\mathbf{t}}^{-1}(\mathbf{d}) + \frac{\mathbf{e}_{\mathbf{t}}^T}{\sigma_{\mathbf{t}}(\mathbf{d})}\frac{\mathbf{\Psi}_N}{\sqrt{n}}\right), \tag{4.14}$$

with

$$x_{0\mathbf{t}} = \frac{\mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} - u_*}{\sigma_{\mathbf{t}}(0)}.$$

Combining (4.13) and (4.14), $F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})$ becomes

$$F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d}) = \frac{\Phi\left(\widetilde{y}_{n\mathbf{dt}}\sigma_{n\mathbf{t}}(0)\sigma_{n\mathbf{t}}^{-1}(\mathbf{d}) + \frac{\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\mathbf{M}_\xi}{\sigma_{n\mathbf{t}}(\mathbf{d})}\frac{\mathbf{\Psi}_N}{\sqrt{n}}\right)}{\Phi\left(x_{0\mathbf{t}}\sigma_{\mathbf{t}}(0)\sigma_{\mathbf{t}}^{-1}(\mathbf{d}) + \frac{\mathbf{e}_{\mathbf{t}}^T}{\sigma_{\mathbf{t}}(\mathbf{d})}\frac{\mathbf{\Psi}_N}{\sqrt{n}}\right)}.$$

In the following, when there is no confusion, $E_{\mathbf{d},\mathbf{y_n}|\boldsymbol{\theta}}$ is simplified as E.

We will expand the MSPE in (4.10) up to order $O(n^{-1})$. That is, we seek an expansion of the form

$$\begin{aligned} & E\left(1 - F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})\right)^2 \\ = \;& \varphi_0(C_{0\mathbf{t}}) + 2E(\varphi_1(C_{0\mathbf{t}}, C_{1\mathbf{t}}))\frac{1}{\sqrt{n}} + E(\varphi_2(C_{0\mathbf{t}}, C_{1\mathbf{t}}, C_{2\mathbf{t}}))\frac{1}{n} + R_{n\mathbf{t}}, \end{aligned} \tag{4.15}$$

where $\varphi_j(\cdot)$, $j = 0, 1, 2$, are polynomial functions, and $C_{j\mathbf{t}}$ is the coefficient of $n^{j/2}$ in the expansion of $F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})$.

To obtain (4.15), we first expand $F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})$ as

$$F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d}) = C_{0\mathbf{t}} + C_{1\mathbf{t}}\frac{1}{\sqrt{n}} + C_{2\mathbf{t}}\frac{1}{n} + r_{n\mathbf{t}}, \tag{4.16}$$

such that $C_{0\mathbf{t}}, E(C_{1\mathbf{t}}^2), E(C_{2\mathbf{t}}^2) < \infty$, the remainder term $r_{n\mathbf{t}}$ is $o_p(n^{-1})$ with $E(r_{n\mathbf{t}}) = O(n^{-3/2})$ and $E(r_{n\mathbf{t}}^2) = O(n^{-3})$. We substitute the expansion (4.16) into $(1 - F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d}))^2$, and take expectation with respect to random variables ($\mathbf{d}$ and $\mathbf{y}_n$) in the terms conditional on the unknown regression parameter $\boldsymbol{\theta}$. Then expansion, (4.15), is obtained, in which the constant term $\varphi_0(C_{0\mathbf{t}})$ and the expectations are bounded, and the remain term $R_{n\mathbf{t}}$ is $O(n^{-3/2})$. Explicit forms of $C_{j\mathbf{t}}$, the calculation of expectations in (4.15) and their properties are discussed in Chapter 5.

Taking the average of (4.15) over $\mathbf{t} \in \mathcal{T} \backslash \mathcal{S}$ we approximate the loss (4.9) by

$$\frac{1}{N-n} \sum_{t \in \mathcal{T} \backslash \mathcal{S}} \left[\varphi_0(C_{0\mathbf{t}}) + 2E(\varphi_1(C_{0\mathbf{t}}, C_{1\mathbf{t}})) \frac{1}{\sqrt{n}} + E(\varphi_2(C_{0\mathbf{t}}, C_{1\mathbf{t}}, C_{2\mathbf{t}})) \frac{1}{n} \right].$$

Denote

$$\begin{aligned}
D^1_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d}) &= \frac{\partial F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})}{\partial \mathbf{d}/\sqrt{n}}, \quad D^2_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d}) = \frac{\partial^2 F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})}{\partial (\mathbf{d}/\sqrt{n})^2}, \\
D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{\Psi}_N, \mathbf{d}) &= \mathbf{J}_{F_{\mathbf{t}}}(\mathbf{\Psi}_N, \mathbf{d}) = \left(\frac{\partial F_{\mathbf{t}}(\mathbf{\Psi}_N, \mathbf{d})}{\partial \psi_1/\sqrt{n}}, ..., \frac{\partial F_{\mathbf{t}}(\mathbf{\Psi}_N, \mathbf{d})}{\partial \psi_N/\sqrt{n}} \right) : 1 \times N, \\
D^2_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{\Psi}_N, \mathbf{d}) &= \mathbf{H}_{F_{\mathbf{t}}}(\mathbf{\Psi}_N, \mathbf{d}) = \left(\frac{\partial^2 F_{\mathbf{t}}(\mathbf{\Psi}_N, \mathbf{d})}{\partial (\psi_i/\sqrt{n}) \, \partial (\psi_j/\sqrt{n})} \right)_{i,j=1}^{N} : N \times N.
\end{aligned}$$

Then, the above results are summarized in the following theorem.

Theorem 25 *Apart from terms which are $o(n^{-1})$, the loss (4.9) is*

$$\mathcal{L}_0(\xi | \mathbf{\Psi}_N, \boldsymbol{\theta}) = \frac{1}{N-n} \sum_{t \in \mathcal{T} \backslash \mathcal{S}} \left[\mathbf{\Psi}_N^T \mathbf{A}_{\mathbf{t}\xi\boldsymbol{\theta}} \mathbf{\Psi}_N \frac{1}{n} + 2\mathbf{b}^T_{\mathbf{t}\xi\boldsymbol{\theta}} \mathbf{\Psi}_N \frac{1}{\sqrt{n}} + \left(c_{1\mathbf{t}\xi\boldsymbol{\theta}} + c_{2\mathbf{t}\xi\boldsymbol{\theta}} \frac{\omega^2_{\mathbf{d}}}{n} \right) \right], \tag{4.17}$$

where

$$\begin{aligned}
c_{1\mathbf{t}\xi\boldsymbol{\theta}} &= E_{\mathbf{y}_n|\boldsymbol{\theta}} \left(1 - F_{\mathbf{t}}(\mathbf{0},0)\right)^2, \\
c_{2\mathbf{t}\xi\boldsymbol{\theta}} &= E_{\mathbf{y}_n|\boldsymbol{\theta}} \left[\left(D^1_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0) \right)^2 - D^2_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0) + F_{\mathbf{t}}(\mathbf{0},0) D^2_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0) \right], \\
\mathbf{b}^T_{\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}} &= E_{\mathbf{y}_n|\boldsymbol{\theta}} \left[\left(F_{\mathbf{t}}(\mathbf{0},0) - 1 \right) D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0) \right], \\
\mathbf{A}_{\mathbf{t}\xi\boldsymbol{\theta}} &= E_{\mathbf{y}_n|\boldsymbol{\theta}} \left[\begin{array}{c} -D^2_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0) + \left(D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0) \right)^T D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0) \\ + \left(F_{\mathbf{t}}(\mathbf{0},0) D^2_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0) \right) \end{array} \right].
\end{aligned}$$

In the following, we will use (4.17), the approximation of the original loss function, as the new loss function. To deal with the second non-robustness source (**U2**) – the possible model misspecification – in the following section we maximize the loss function (4.17) over the set $\mathbf{\Psi}$ which quantifies the model misspecification.

4.3 Maximization of the loss $\mathcal{L}_0(\xi | \mathbf{\Psi}_N, \boldsymbol{\theta})$ over $\mathbf{\Psi}_N \in \mathbf{\Psi}$

In this section we maximize $\mathcal{L}_0(\xi | \mathbf{\Psi}_N, \boldsymbol{\theta})$ over the set $\mathbf{\Psi}$. The orthogonality requirement $\mathbf{F}_N^T \mathbf{\Psi}_N = 0$ is equivalent to the statement that $\mathbf{\Psi}_N$ lies in the orthogonal

complement to the column space of $\mathbf{F}_N$. Let $\mathbf{K} : N \times (N-p)$ be a matrix whose columns form an orthonormal basis for this orthogonal complement. Then $\mathbf{\Psi}_N = \mathbf{K}\mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^{N-p}$ with $||\mathbf{v}|| = ||\mathbf{\Psi}_N||$. Thus, maximizing $\mathcal{L}_0(\xi|\mathbf{\Psi}_N, \boldsymbol{\theta})$ over $\mathbf{\Psi}$ is equivalent to solving the problem

$$\max_{\mathbf{v}\in\mathbb{R}^{N-p}:||\mathbf{v}||\leq\tau} \mathcal{L}_0(\xi|\mathbf{v}, \boldsymbol{\theta}) = \frac{1}{N-n} \max_{\mathbf{v}\in\mathbb{R}^{N-p}:||\mathbf{v}||_2\leq\tau} \left(\mathbf{v}^T\mathbf{A}_{\boldsymbol{\xi},\boldsymbol{\theta}}\mathbf{v} + 2\mathbf{b}_{\boldsymbol{\xi},\boldsymbol{\theta}}^T\mathbf{v} + c_{\boldsymbol{\xi},\boldsymbol{\theta}}\right), \quad (4.18)$$

where

$$\begin{aligned}
\mathbf{A}_{\boldsymbol{\xi},\boldsymbol{\theta}} &= \frac{1}{n}\sum_{t\in\mathcal{T}\setminus\mathcal{S}} \mathbf{K}^T\mathbf{A}_{\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}}\mathbf{K}, \\
\mathbf{b}_{\boldsymbol{\xi},\boldsymbol{\theta}}^T &= \frac{1}{\sqrt{n}}\sum_{t\in\mathcal{T}\setminus\mathcal{S}} \mathbf{b}_{\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}}^T\mathbf{K}, \\
c_{\boldsymbol{\xi},\boldsymbol{\theta}} &= \sum_{t\in\mathcal{T}\setminus\mathcal{S}} \left(c_{1\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}} + c_{2\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}}\frac{\omega_{\mathbf{d}}^2}{n}\right).
\end{aligned}$$

Based on Lemmas 2.4 and 2.8 in Sorensen (1982), we can give a characterization of the solution to problem (4.18) .

Proposition 26 *The solution $\mathbf{v}_{\xi,\boldsymbol{\theta}}^*$ to problem (4.18) is the solution of*

$$(\lambda_{\xi,\boldsymbol{\theta}}\mathbf{I}_{N\times N} - \mathbf{A}_{\boldsymbol{\xi},\boldsymbol{\theta}})\mathbf{v}_{\xi,\boldsymbol{\theta}}^* = \mathbf{b}_{\boldsymbol{\xi},\boldsymbol{\theta}},$$

and the maximum loss is

$$\mathcal{L}_0\left(\xi|\mathbf{v}_{\xi,\boldsymbol{\theta}}^*, \boldsymbol{\theta}\right) = \frac{1}{N-n}\left(\mathbf{v}_{\xi,\boldsymbol{\theta}}^{*T}\mathbf{A}_{\boldsymbol{\xi},\boldsymbol{\theta}}\mathbf{v}_{\xi,\boldsymbol{\theta}}^* + \mathbf{2}\mathbf{b}_{\boldsymbol{\xi},\boldsymbol{\theta}}^T\mathbf{v}_{\xi,\boldsymbol{\theta}}^* + c_{\boldsymbol{\xi},\boldsymbol{\theta}}\right), \quad (4.19)$$

where $\lambda_{\xi,\boldsymbol{\theta}}$ is chosen such that $\lambda_{\xi,\boldsymbol{\theta}}(\left\|\mathbf{v}_{\xi,\boldsymbol{\theta}}^*\right\| - \tau) = 0$ and $\lambda_{\xi,\boldsymbol{\theta}}\mathbf{I}_{N\times N} - \mathbf{A}_{\boldsymbol{\xi},\boldsymbol{\theta}}$ is positive semi-definite.

In Sections 2 and 3, for each design $\boldsymbol{\xi}$, the loss function (4.9) is first approximated by (4.17). Within a neighbourhood of the parametric model thought to be a reasonable approximation to the true response, the approximated loss function is maximized as shown in Proposition 26. Let $\mathcal{L}_{\max}(\boldsymbol{\xi}|\boldsymbol{\theta}) = \mathcal{L}_0\left(\boldsymbol{\xi}|\mathbf{v}_{\boldsymbol{\xi},\boldsymbol{\theta}}^*, \boldsymbol{\theta}\right)$, with $\mathbf{v}_\xi^*$ as defined in Proposition 26. In the following section, we will find by numerical methods a design ξ_0 which minimizes $\mathcal{L}_{\max}(\boldsymbol{\xi}|\boldsymbol{\theta})$, among all designs of the same size.

4.4 Robust optimal designs and case study

To find the optimal design which minimizes the maximized loss $\mathcal{L}_{\max}(\boldsymbol{\xi}|\boldsymbol{\theta})$, we proceed sequentially. The MATLAB code for the computations in this section is available from the authors.

4.4.1 Sequential algorithm

The design criterion $\mathcal{L}_{\max}(\boldsymbol{\xi}|\boldsymbol{\theta})$ to be optimized depends on unknown regression parameters $\boldsymbol{\theta}$ and other nuisance parameters, including the variances of the measurement error $\varepsilon(\mathbf{t})$, and the parameters of the variance/covariance function of the δ process. For convenience, we denote the nuisance parameters as $\boldsymbol{\varphi}$ and the design criterion as $\mathcal{L}_{\max}(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{\varphi})$. To address the issue of dependence, we propose a sequential design – one design point at a time – replacing $\boldsymbol{\theta}$ by the GLS estimates at each iteration, and $\boldsymbol{\varphi}$ by the maximum likelihood estimates.

Step 1: choose an initial design $\boldsymbol{\xi}_{n_0}$.

For $m = 0, 1, ...$ until an n-point design $\boldsymbol{\xi}_n$ is obtained carry out steps 2-5.

Step 2: make observation at the sampled locations of the current design $\boldsymbol{\xi}_m = \{\mathbf{t}_{s1}, ..., \mathbf{t}_{sm}\}$.

Step 3: the regression parameters that are required in the evaluation of the loss are replaced by GLS estimation $\hat{\boldsymbol{\theta}}_m$, and nuisance parameters $\boldsymbol{\varphi}$ by MLE $\widehat{\boldsymbol{\varphi}}_m$.

Step 4: substitute $\widehat{\boldsymbol{\theta}}_m, \widehat{\boldsymbol{\varphi}}_m$ into the loss function and obtain $\mathcal{L}_{\max}(\boldsymbol{\xi}_m|\widehat{\boldsymbol{\theta}}_m, \widehat{\boldsymbol{\varphi}}_m)$ by maximizing the loss function over the set $\boldsymbol{\Psi}$.

Step 5: make the next observation at

$$\mathbf{t}_{new} = \arg\min_{\mathbf{t}\in\mathcal{T}} \mathcal{L}_{\max}(\{\mathbf{t}_{s1}, ..., \mathbf{t}_{sm}, \mathbf{t}\}\,|\hat{\boldsymbol{\theta}},\widehat{\boldsymbol{\varphi}}).$$

4.4.2 Coal-ash data example

We study the "coal-ash" data to illustrate our methods. These data are given by Gomez and Hazen (1970) and described in Cressie (1993). There are 208 coal-ash core measurements obtained from a Pittsburgh coal seam, at sites throughout a grid. In the data set each location is determined by its relative east-west and relative north-south positions. The percent coal ash at each location is recorded.

Table 1. Minimax losses for varying τ and n_0

	τ^2			
	0.3	0.5	1	1.5
$n_0 = 20$	489.1	326.7	806.4	901.6
$n_0 = 41$	107.7	108.2	109.3	110.2
$n_0 = 52$	42.9	43.2	43.8	44.3

Chemicals, such as arsenic and selenium, in coal ash above a certain threshold are key facts to determine that coal ash is hazardous according to the U.S. Environmental Protection Agency. Burning coal with high ash and sulfur content will affect human health. Therefore, our study aim to obtain a robust design which helps to estimate the probability that the percent coal ash at each site of a Pttsburgh coal seam exceeds a certain value u_* based on the coal-ash data.

The proposed random-process model is

$$y(\mathbf{t}) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \delta(\mathbf{t}) + \varepsilon(\mathbf{t}),\ \mathbf{t} = (t_1, t_2). \tag{4.20}$$

Denote $\mathrm{Var}(\delta(\mathbf{t})) = \sigma_1^2$ and $\mathrm{Var}(\varepsilon(\mathbf{t})) = \sigma_2^2$. The nominal correlation structure of δ process is determined by an isotropic correlation function (Wiens 2005a)

$$\rho_\kappa(h) = (1 + \kappa h)^{-2},$$

where $h = ||\mathbf{t} - \mathbf{t}'||, \mathbf{t}, \mathbf{t}' \in \mathcal{T}$.

The values of σ_1^2, σ_2^2 and κ are unknown, and in practice are replaced by estimates $\widehat{\sigma}_1^2, \widehat{\sigma}_2^2$ and $\widehat{\kappa}$, thus

$$\mathbf{G}_0 = \widehat{\sigma}_1^2 \left((1 + \widehat{\kappa}||t_i - t_j||)^{-2}\right)_{i,j=1}^{N}.$$

To obtain the estimates, we assume that an initial sample at n_0 locations has been collected. With these data, initial maximum likelihood estimates $\widehat{\sigma}_1^2, \widehat{\sigma}_2^2$ and $\widehat{\kappa}$ were obtained; these are updated as more data are collected. From the remaining $208 - n_0$ locations we then sequentially determine the robust design.

We take $n_0 = 20, 41, 52$, and the initial sample locations are spread out across the whole domain. We assume that the true covariance/variance structure varies in the neighbourhood $\mathcal{G}$ of $\mathbf{G}_0$ with $\mathbf{d} \sim U(-1, 1)$, $\omega_{\mathbf{d}}^2 = 1/3$. As an example, we take the third quartile of all the y values as the threshold value and $u_* = 10.575$.

The robust sequential designs with size $n_0 + n = n_0 + 30$ for the prediction of the threshold probability $P(y(\mathbf{t}) > 10.575)$ are shown in Figs 4.1- 4.4. To find the

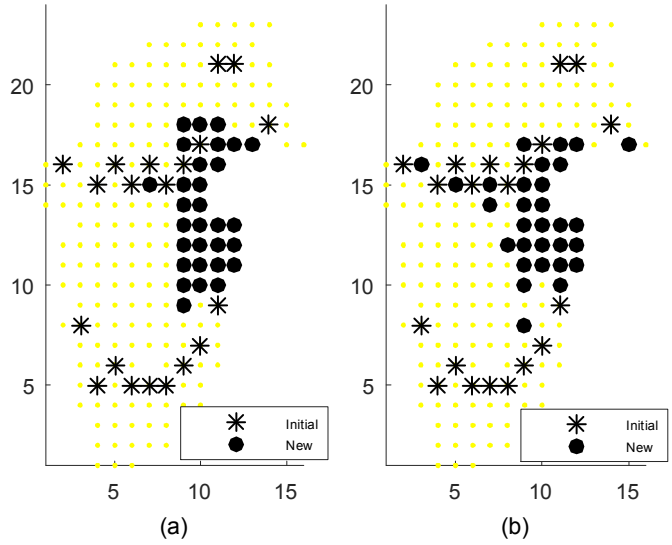

Figure 4.1: Optimal design for Coal-ash data example with $N = 208$ locations. Observations are obtained at $n_0 = 20$ initial sample locations (denoted by asterisks in the graph). A minimax design (denoted by filled circles) is obtained with $n = 30$ sites chosen among the remaining 188 locations. (a) $\tau = 0.3$; (b) $\tau = 0.5$.

maximum of the loss, we choose $\tau^2 = 0.3, 0.5, 1, 1.5$, which defines the neighbourhood quantifying the misspecified model. The minimax losses obtained for different combinations of the parameters (n_0 and τ^2) are shown in Table 1. From there, we notice that for a given initial sample the minimax losses generally increase with respect to the value of τ^2 determining the range of the set $\mathbf{\Psi}$. When the sample size of the initial sample is small, the prediction error is large. And when we increase the sample size of the initial sample (and therefore, the size of the design) the prediction error decreases dramatically.

Figure 4.1 shows the $n_0 = 20$ initial sample locations denoted by asterisks in the graph and the robust designs consisting of a further 30 points (denoted by filled circles) when $\tau^2 = 0.3$ and 0.5, respectively. Fig 4.2 shows the robust designs when $\tau^2 = 1$ and 1.5, respectively. In Figures 4.3, initial 41-point designs were chosen and the minimax designs with design points 71 for $\tau^2 = 0.3, 0.5, 1, 1.5$ are the same as shown.

Fig 4.4 illustrate the $n_0 = 52$ initial sample locations and the robust designs of size 82 for $\tau^2 = 0.3, 0.5, 1, 1.5$. The robust designs are the same for all these τ^2.

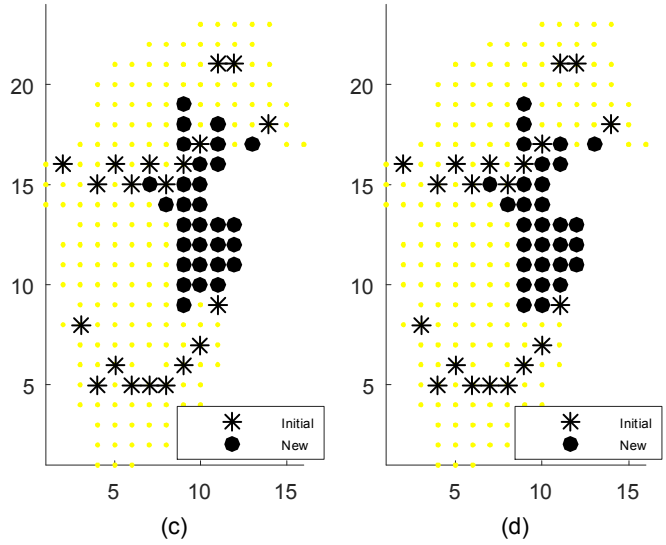

Figure 4.2: Optimal design for Coal-ash data example with $N = 208$ locations. Observations are obtained at $n_0 = 20$ initial sample locations (denoted by asterisks in the graph) for parameter estimation. A minimax design (denoted by filled circles) is obtained with $n = 30$ sites chosen among the remaining 188 locations. (a) $\tau = 1$; (b) $\tau = 1.5$.

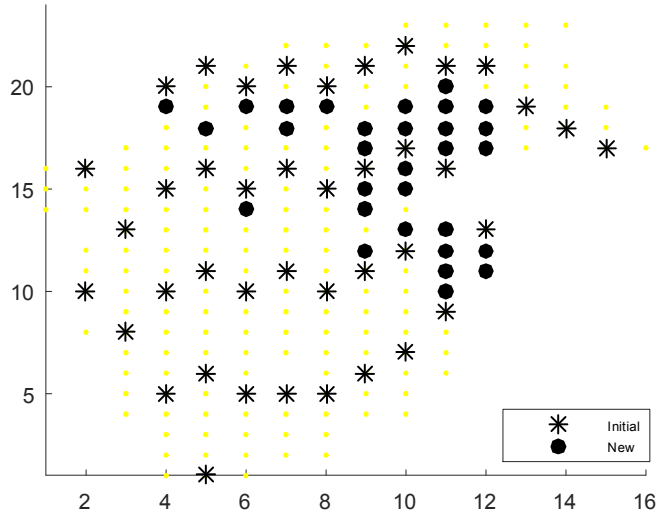

Figure 4.3: Optimal design for Coal-ash data example with $N = 208$ locations. Observations are obtained at $n_0 = 41$ initial sample locations (denoted by asterisks in the graph) . A minimax design (denoted by filled circles) is obtained with $n = 30$ sites chosen among the remaining 167 locations. $\tau = 0.3, 0.5, 1, 1.5$.

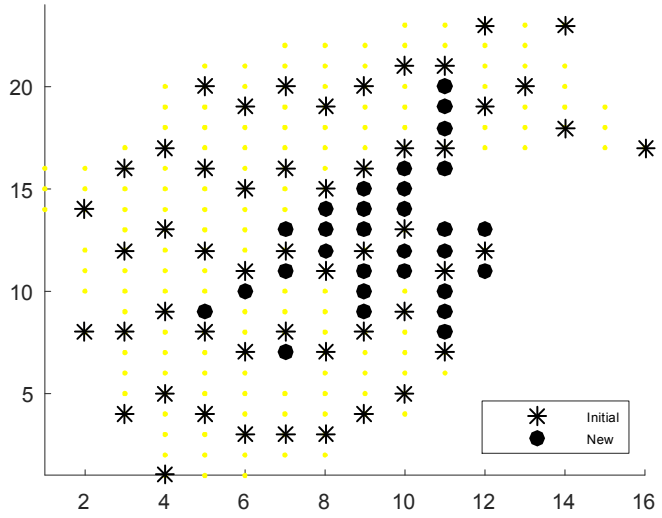

Figure 4.4: Optimal design for Coal-ash data example with $N = 208$ locations. Observations are obtained at $n_0 = 52$ initial sample locations (denoted by asterisks in the graph). A minimax design (denoted by filled circles) is obtained with $n = 30$ sites chosen among the remaining 156 locations. Here the best designs for $\tau = 0.3, 0.5, 1$ and $\tau = 1.5$ are the same.

It is noticeable that the design points vary in a non-subtle way for different combinations of parameters (n_0 and τ^2). However, the x-coordinate values of most design points range from 9 to 11. This might be because the effect of the east-west direction is stronger than that of the the north-south direction.

REFERENCES

Atkinson, A. C. and Fedorov, V. V. (1975a), "The Design of Experiments for Discriminating Between Two Rival Models," *Biometrika*, 62, 57-70.

Atkinson, A. C. and Fedorov, V. V. (1975b), "Optimal Design: Experiments for Discriminating Between Several Models," *Biometrika*, 62, 289-303.

Bect, J., Ginsbourger, D., Li, L., Picheny, V. and Vazquez, E. (2012), "Sequential design of computer experiments for the estimation of a probability of failure," *Statistics and Computing*, 22, 773-793.

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

Dette, H., and Biedermann, S. (2003), "Robust and Efficient Designs for the Michaelis–Menten Model," *Journal of the American Statistical Association*, 98, 679-686.

Dette, H., and Pepelyshev, A. (2008), "Efficient Experimental Designs for Sigmoidal Growth Models," *Journal of Statistical Planning and Inference*, 138, 2-17.

Dette, H., and Neugebauer, H. M. (1997), "Bayesian D-optimal Designs for Exponential Regression Models," *Journal of Statistical Planning and Inference*, 60, 331-349.

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.

Fahrmeir, L. and Tutz, G. (1994), *Multivariate statistical modelling based on generalized linear models*. New York: Springer.

Fang, Y., Loparo, K. A., and Feng, X. (1994), "Inequalities for the trace of matrix product," *Automatic Control, IEEE Transactions on*, 39, 2489-2490.

Fedorov, V. V. and Pazman, A. (1968), "Design of physical experiments," *Fortschritte der Physik*, 16, 325-355.

Gomez, M. and Hazen, K. (1970), "Evaluation of sulphur and ash distribution in coal seams by statistical response surface regression analysis," *US Bureau of Mines Report of Investigation 7377*.

Hunter, W. G., and Reiner, A. M. (1965), "Designs for discriminating between two rival models," *Technometrics*, 7, 307-323.

Karami, J. H., and Wiens, D. P. (2014), "Robust static designs for approximately specified nonlinear regression models," *Journal of Statistical Planning and Inference*, 144, 55-62.

King, J., and Wong, W. K. (2000), "Minimax D-Optimal Designs for the Logistic Model," *Biometrics*, 56, 1263-1267.

Kleinman, L., and Michael A. (1968), "The design of suboptimal linear time-varying systems," *Automatic Control, IEEE Transactions,* 13, 150-159.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007), "An Optimal Experimental Design Criterion for Discriminating Between Non-Normal Models," *Journal of the Royal Statistical Society B*, 69, 231-242.

Mardia, K. V., and Marshall, R. J. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135-146.

Oleson, J., Kumar, N., Smith, B. (2013), "Spatiotemporal modeling of irregularly spaced aerosol optical depth data", Environmental and ecological statistics, 20, 297-314.

Prasad, A. K., Singh, R. P., Singh, A. and Kafatos, M. (2005), "Seasonal Variability of Aerosol Optical Depth over Indian Subcontinent," in *Analysis of Multi-Temporal Remote Sensing Images*, pp. 35-38.

Rosenblatt, M. (2000), *Gaussian and non-Gaussian linear time series and random fields*. Springer.

Sinha, S. and Wiens, D. P. (2002), "Robust Sequential Designs for Nonlinear Regression," *The Canadian Journal of Statistics*, 30, 601-618.

Sorensen, D. C. (1982), "Newton's Method with a Model Trust Region Modification," *SIAM Journal on Numerical Analysis*, 19, 409-426.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer Verlag: New York.

Wiens, D. (2005a), "Robustness in spatial studies I: minimax prediction," *Environmetrics*, 16, 191-203.

Wiens, D. (2005b), "Robustness in spatial studies II: minimax design," *Environmetrics*, 16, 205-217.

Wiens, D. and Zhou, J. (2008), "Robust estimators and designs for field experiments," *Journal of Statistical Planning and Inference*, 138, 93-104.

Chapter 5

Derivation and proofs for Chapter 4

To prove Theorem 25, we require some preliminary results. Recall that $\sigma^2_{\mathbf{t}}(\mathbf{d}) = \mathbf{e}_{\mathbf{t}}^T \mathbf{G}_{\mathbf{d}} \mathbf{e}_{\mathbf{t}}$, $\mathbf{a}_{n\mathbf{t}}(\mathbf{d})$ is defined in (4.5), $\sigma^2_{n\mathbf{t}}(\mathbf{d})$ is the variance of the prediction error defined in (4.6). For convenience, in the following context for any vector function $\mathbf{g}(\mathbf{d}) = (g_1(\mathbf{d}), ..., g_m(\mathbf{d}))^T$ we denote

$$D^0_{\mathbf{d}}\mathbf{g}(\mathbf{d}) = \mathbf{g}(\mathbf{d}) \text{ and } D^k_{\mathbf{d}}\mathbf{g}(\mathbf{d}) = \left(\frac{d^k g_1(\mathbf{d})}{d(\mathbf{d}/\sqrt{n})^k}, ..., \frac{d^k g_m(\mathbf{d})}{d(\mathbf{d}/\sqrt{n})^k} \right)^T, \text{ for } k = 1, 2, ...$$

The derivatives of $\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})$, $D^k_{\mathbf{d}}\sigma^2_{\mathbf{t}}(\mathbf{d})$ and $D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(\mathbf{d})$ for $0 \leq k \leq 3$ play important roles in the expansion of the loss function in Theorem 25. To prove Theorem 25, we need to show that the expectation of the squared norms (or squared norms when $\mathbf{d} = \mathbf{0}$) of the derivatives are all $O(1)$. In particular, that the expectations of powers, up to 16th, of $\left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|$ are $O(1)$ is needed in the proof. This is the result of the following proposition.

Proposition 27 *Under assumptions (A1)-(A3), we have for $k, l \in \mathbb{Z}$ and $0 \leq k \leq 3$, $2 \leq l \leq 16$,*

$$\begin{aligned} E\left(\left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^l\right) &= O(1), \text{ and } \left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{0})\right\|^l = O(1), \\ E\left(D^k_{\mathbf{d}}\sigma^2_{\mathbf{t}}(\mathbf{d})\right)^2 &= O(1), \text{ and } D^k_{\mathbf{d}}\sigma^2_{\mathbf{t}}(\mathbf{0}) = O(1), \\ E\left(D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(\mathbf{d})\right)^2 &= O(1), \text{ and } D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(0) = O(1). \end{aligned}$$

The proof of Proposition 27 is in Appendix B.

5.1 Proof of Theorem 25

5.1.1 Taylor expansions

We first expand $F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)$ around $\mathbf{\Psi}_N/\sqrt{n} = \mathbf{0}$, obtaining

$$F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right) = F_{\mathbf{t}}\left(\mathbf{0}, \mathbf{d}\right) + D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}\left(\mathbf{0}, \mathbf{d}\right) \frac{\mathbf{\Psi}_N}{\sqrt{n}} + \frac{\mathbf{\Psi}_N^T D^2_{\mathbf{\Psi}_N} F_{\mathbf{t}}\left(\mathbf{0}, \mathbf{d}\right) \mathbf{\Psi}_N}{2n} \tag{5.1}$$
$$+\frac{1}{6}\sum_{i,j,l=1}^{N} \frac{\partial^3 F_{\mathbf{t}}(\widehat{\mathbf{\Psi}}_N, \mathbf{d})}{\partial\psi_i \partial\psi_j \partial\psi_l} \frac{\psi_i\psi_j\psi_l}{n^{3/2}}.$$

Here ψ_i, ψ_j, ψ_l are elements of $\mathbf{\Psi}_N$ and $\widehat{\mathbf{\Psi}}_N \in \mathbf{R}^N$ is between $\mathbf{0}$ and $\mathbf{\Psi}_N/\sqrt{n}$ (i.e., all elements of $\widehat{\mathbf{\Psi}}_N$ are between 0 and the corresponding elements of $\mathbf{\Psi}_N/\sqrt{n}$).

We then expand the terms in (5.1) with order higher than $O(n^{-3/2})$ at $\mathbf{d}/\sqrt{n} = 0$. With the same notation as in (4.16)

$$\begin{aligned} F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right) &= C_{0\mathbf{t}} + C_{1\mathbf{t}}\frac{1}{\sqrt{n}} + C_{2\mathbf{t}}\frac{1}{n} + C_{3\mathbf{t}}\frac{1}{n^{3/2}} \\ &= C_{0\mathbf{t}} + C_{1\mathbf{t}}\frac{1}{\sqrt{n}} + C_{2\mathbf{t}}\frac{1}{n} + r_{n\mathbf{t}}, \end{aligned} \tag{5.2}$$

where

$$C_{0\mathbf{t}} = F_{\mathbf{t}}(\mathbf{0},0), \tag{5.3a}$$

$$C_{1\mathbf{t}} = D^1_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0)\mathbf{d} + D^1_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0)\mathbf{\Psi}_N, \tag{5.3b}$$

$$\begin{aligned} C_{2\mathbf{t}} &= D^{1,1}_{\mathbf{\Psi}_N\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0)\mathbf{\Psi}_N\mathbf{d} + \frac{1}{2}\mathbf{\Psi}_N^T D^2_{\mathbf{\Psi}_N} F_{\mathbf{t}}(\mathbf{0},0)\mathbf{\Psi}_N \\ &\quad + \frac{1}{2} D^2_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},0)\mathbf{d}^2, \end{aligned} \tag{5.3c}$$

$$\begin{aligned} C_{3\mathbf{t}} &= \frac{1}{6} D^3_{\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\mathbf{d}^3 + \frac{1}{2} D^{1,2}_{\mathbf{\Psi}_N\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\mathbf{\Psi}_N\mathbf{d}^2 \\ &\quad + \frac{1}{2}\mathbf{\Psi}_N^T D^{2,1}_{\mathbf{\Psi}_N\mathbf{d}} F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\mathbf{\Psi}_N\mathbf{d} + \frac{1}{6}\sum_{i,j,l} \frac{\partial^3 F_{\mathbf{t}}(\widehat{\mathbf{\Psi}}_N, \mathbf{d})}{\partial\psi_i \partial\psi_j \partial\psi_l}\psi_i\psi_j\psi_l, \end{aligned} \tag{5.3d}$$

with $\widehat{\mathbf{d}}$ between $\mathbf{0}$ and $\mathbf{d}/\sqrt{n}$. As shown in the Section 5.1.2, $E(C^2_{i\mathbf{t}}) < \infty, i = 1, 2, 3$. Therefore, $r_{n\mathbf{t}} = C_{3\mathbf{t}}/n^{3/2}$ is $o_p(n^{-1})$ with $E(r_{n\mathbf{t}}) = o(n^{-1})$ and $E(r^2_{n\mathbf{t}}) = o(n^{-2})$.

Substituting (5.2) into $\left(1 - F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)\right)^2$, with the same notations as in (4.15) we have

$$\left(1 - F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)\right)^2 = \varphi_0(C_{0\mathbf{t}}) + 2\varphi_1(C_{0\mathbf{t}}, C_{1\mathbf{t}})\frac{1}{\sqrt{n}} + \varphi_2(C_{0\mathbf{t}}, C_{1\mathbf{t}}, C_{2\mathbf{t}})\frac{1}{n} + R_{n\mathbf{t}}, \tag{5.4}$$

where

$$\begin{aligned}\varphi_0(C_{0\mathbf{t}}) &= (1 - F_{\mathbf{t}}(\mathbf{0},0))^2\,,\\ \varphi_1(C_{0\mathbf{t}}, C_{1\mathbf{t}}) &= F_{\mathbf{t}}(\mathbf{0},0)C_{1\mathbf{t}} - C_{1\mathbf{t}},\\ \varphi_2(C_{0\mathbf{t}}, C_{1\mathbf{t}}, C_{2\mathbf{t}}) &= C_{1\mathbf{t}}^2 - 2C_{2\mathbf{t}} + 2F_{\mathbf{t}}(\mathbf{0},0)C_{2\mathbf{t}},\end{aligned}$$

and the remainder

$$\begin{aligned}R_{n\mathbf{t}} &= 2\left(F_{\mathbf{t}}(\mathbf{0},0)C_{3\mathbf{t}} - C_{3\mathbf{t}} + C_{1\mathbf{t}}C_{2\mathbf{t}}\right)\frac{1}{n^{3/2}} + \left(C_{2\mathbf{t}}^2 + 2C_{1\mathbf{t}}C_{3\mathbf{t}}\right)\frac{1}{n^2}\\ &\quad + C_{3\mathbf{t}}^2\frac{1}{n^3} + 2C_{2\mathbf{t}}C_{3\mathbf{t}}\frac{1}{n^{5/2}}\end{aligned}$$

satisfies $E(R_{n\mathbf{t}}) = O(n^{-3/2})$, since $E(C_{i\mathbf{t}}^2) < \infty, i = 1, 2, 3$.

Recall that

$$E(\mathbf{d}) = 0 \text{ and } E(\mathbf{d}^2) = \omega_{\mathbf{d}}^2.$$

Substituting $C_{1\mathbf{t}}, C_{2\mathbf{t}}$ into (5.4) and taking expectation with respect to $\mathbf{d}$ and $\mathbf{y}_n$ gives, after a calculation, that the expectation of (5.4) is

$$E\left(1 - F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)\right)^2 = \mathbf{\Psi}_N^T\mathbf{A}_{\mathbf{t}\xi\boldsymbol{\theta}}\mathbf{\Psi}_N\frac{1}{n} + 2\mathbf{b}_{\mathbf{t}\xi\boldsymbol{\theta}}^T\frac{\mathbf{\Psi}_N}{\sqrt{n}} + \left(c_{1\mathbf{t}\xi\boldsymbol{\theta}} + c_{2\mathbf{t}\xi\boldsymbol{\theta}}\frac{\omega_{\mathbf{d}}^2}{n}\right) + E(R_n)$$

with $\mathbf{A}_{\mathbf{t}\xi\boldsymbol{\theta}}, \mathbf{b}_{\mathbf{t}\xi\boldsymbol{\theta}}^T, c_{1\mathbf{t}\xi\boldsymbol{\theta}}$ and $c_{2\mathbf{t}\xi\boldsymbol{\theta}}$ as defined in Theorem 25.

Then an expansion of the loss (4.9) is

$$\begin{aligned}\mathcal{L}_0\left(\xi|\mathbf{\Psi}_N, \boldsymbol{\theta}\right) &= \frac{1}{N-n}\sum_{t\in\mathcal{T}\setminus\mathcal{S}}\left[\mathbf{\Psi}_N^T\mathbf{A}_{\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}}\mathbf{\Psi}_N\frac{1}{n} + 2\mathbf{b}_{\mathbf{t}\boldsymbol{\xi}\boldsymbol{\theta}}^T\mathbf{\Psi}_N\frac{1}{\sqrt{n}} + \left(c_{1\mathbf{t}\xi\boldsymbol{\theta}} + c_{2\mathbf{t}\xi\boldsymbol{\theta}}\frac{\omega_{\mathbf{d}}^2}{n}\right)\right]\\ &\quad + \frac{1}{N-n}\sum_{t\in\mathcal{T}\setminus\mathcal{S}}E(R_n),\end{aligned}$$

where, according to Assumption (A4), the remainder, $\frac{1}{N-n}\sum_{t\in\mathcal{T}\setminus\mathcal{S}}E(R_n)$, is $o(n^{-1})$.

5.1.2 Proof that $E(C_{i\mathbf{t}}^2) < \infty, i = 1, 2, 3$

Recall (5.3b) - (5.3d). The derivatives of $F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)$ with respect to $\mathbf{d}/\sqrt{n}$ and $\mathbf{\Psi}_N/\sqrt{n}$ are shown in Appendix C. The second moments of $C_{1\mathbf{t}}, C_{2\mathbf{t}}$ and $C_{3\mathbf{t}}$ include the moments of $\widetilde{y}_{n\mathbf{0t}}$ and $D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(0)\mathbf{z}_n$. In fact, these two random variables are normally distributed with finite mean and variance. Since $\sigma_{n\mathbf{t}}(0)$ is bounded as shown in Lemma 31 (in Section 5.2.1), $E\left(\|\mathbf{a}_{n\mathbf{t}}(0)\|^2\right), E\left(\|\mathbf{a}_{n\mathbf{t}}(0)\|^4\right) < \infty$ as shown in Proposition 27, and $\mathbf{z}_n \sim N_n(\mathbf{0}, \mathbf{I})$, the random variable

$$\widetilde{y}_{n\mathbf{0t}} = \frac{\mathbf{b}_{n\mathbf{t}}^T(0)\mathbf{z}_n + \mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} - u_*}{\sigma_{n\mathbf{t}}(0)},$$

with $\mathbf{b}_{n\mathbf{t}}^T(0) := \mathbf{a}_{n\mathbf{t}}^T(0)\boldsymbol{\Sigma}_{nn,\mathbf{0}}^{1/2}$, has mean

$$E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\widetilde{y}_{n\mathbf{0t}}\right) = \frac{\mathbf{f}^T(\mathbf{t})\boldsymbol{\theta} - u_*}{\sigma_{n\mathbf{t}}(0)} < \infty, \tag{5.5}$$

and its variance is

$$\begin{aligned}
\mathrm{Var}\left(\widetilde{y}_{n\mathbf{0t}}\right) &= \frac{\left\|\mathbf{b}_{n\mathbf{t}}(0)\right\|^2}{\sigma_{n\mathbf{t}}^2(0)} \\
&\leq \lambda_{\max}\left(\boldsymbol{\Sigma}_{nn,\mathbf{0}}\right)\frac{\left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2}{\sigma_{n\mathbf{t}}^2(0)} \\
&\leq (\lambda_N + h_N)\frac{\left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2}{\sigma_{n\mathbf{t}}^2(0)} \\
&< \infty.
\end{aligned}$$

Similarly, since $\left\|\mathbf{D}_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{0}}(\mathbf{t})\right\| < \infty$, $D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(0)\mathbf{z}_n$ is also normally distributed with finite mean and variance. Therefore, all the moments of $\widetilde{y}_{n\mathbf{0t}}$ and $D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(0)\mathbf{z}_n$ are finite.

Next, we show that the second moments of each terms in $C_{1\mathbf{t}}, C_{2\mathbf{t}}, C_{3\mathbf{t}}$ are finite. Since

$$C_{1\mathbf{t}}^2 \leq 2\left(D_{\mathbf{d}}^1 F_{\mathbf{t}}(\mathbf{0},0)\right)^2\mathbf{d}^2 + 2\left(D_{\boldsymbol{\Psi}_N}^1 F_{\mathbf{t}}(\mathbf{0},0)\boldsymbol{\Psi}_N\right)^2, \tag{5.6}$$

we need to prove the expectation of each term is finite. Substituting (B.18) in Section 5.3 into $E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\left(D_{\mathbf{d}}^1 F_{\mathbf{t}}(\mathbf{0},0)\right)^2\right)$ gives

$$\begin{aligned}
E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\left(D_{\mathbf{d}}^1 F_{\mathbf{t}}(\mathbf{0},0)\right)^2\right) &\leq \frac{3}{2\pi\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}^2(0)}\left[\left\|D_{\mathbf{d}}^1\mathbf{b}_{n\mathbf{t}}^T(0)\right\| + E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\widetilde{y}_{n\mathbf{0t}}^2\right)\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2\right] \\
&\quad + 3\frac{\phi^2\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}^2}{\sigma_{\mathbf{t}}^2(0)\Phi^4\left(x_{0\mathbf{t}}\right)}\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2 \\
&\leq \frac{3}{2\pi\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}^2(0)}\left(\begin{array}{c}\left\|D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}^T(0)\boldsymbol{\Sigma}_{nn,\mathbf{0}}\right\|^2 + \lambda_N\left\|\mathbf{a}_{n\mathbf{t}}^T(0)\right\|^2 \\ + E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\widetilde{y}_{n\mathbf{0t}}^2\right)\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2\end{array}\right) \\
&\quad + \frac{3}{2\pi}\frac{x_{0\mathbf{t}}^2\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2}{\Phi^4\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}^2(0)}.
\end{aligned}$$

For some positive constant k_1 we have

$$\begin{aligned}
k_1^{-1}E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\left(D_{\mathbf{d}}^1 F_{\mathbf{t}}(\mathbf{0},0)\right)^2\right) &\leq \frac{\left\|D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(0)\right\|^2 + \lambda_N\left\|\mathbf{a}_{n\mathbf{t}}^T(0)\right\|^2}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}^2(0)} \\
&+ \frac{E\left(\widetilde{y}_{n\mathbf{0t}}^2\right)\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}^2(0)} + \frac{x_{0\mathbf{t}}^2\left(D_{\mathbf{d}}^1\sigma_{n\mathbf{t}}(0)\right)^2}{\Phi^4\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}^2(0)}.
\end{aligned}$$

Recall that $\left\|D^1_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(0)\right\|^2, \left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2, D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0) < \infty$ as shown in Proposition 27, $\sigma_{n\mathbf{t}}(0)$ is bounded as shown in Lemma 31 (in Appendix B.1), and $\sigma_{\mathbf{t}}(0)$ is bounded by λ_1 and λ_N. Then we have

$$k_1^{-1}E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(\left(D^1_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0)\right)^2\right) < \infty. \tag{5.7}$$

Similarly, for the second term in (5.6) there exists some positive constant k_2 such that

$$k_2^{-1}E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(D^1_{\boldsymbol{\Psi}_N}F_{\mathbf{t}}(\mathbf{0},0)\boldsymbol{\Psi}_N\right)^2 \leq \frac{\left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2}{\sigma^2_{n\mathbf{t}}(0)\Phi^2\left(x_{0\mathbf{t}}\right)} + \frac{1}{\sigma^2_{\mathbf{t}}(0)\Phi^4\left(x_{0\mathbf{t}}\right)} < \infty. \tag{5.8}$$

Therefore, combining (5.7) and (5.8), we have

$$\begin{aligned} E_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}(C^2_{1\mathbf{t}}) &\leq 2E_{\mathbf{z}_n|\boldsymbol{\theta}}(D^1_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0))^2E(\mathbf{d}^2) + 2E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(D^1_{\boldsymbol{\Psi}_N}F_{\mathbf{t}}(\mathbf{0},0)\boldsymbol{\Psi}_N\right)^2 \\ &= 2E_{\mathbf{z}_n|\boldsymbol{\theta}}(D^1_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0))^2\omega^2_{\mathbf{d}} + 2E_{\mathbf{z}_n|\boldsymbol{\theta}}\left(D^1_{\boldsymbol{\Psi}_N}F_{\mathbf{t}}(\mathbf{0},0)\boldsymbol{\Psi}_N\right)^2 \\ &< \infty. \end{aligned}$$

Via a similar process, we can prove that $E_{\mathbf{z}_n|\boldsymbol{\theta}}(C^2_{2\mathbf{t}}) < \infty$. The proof is omitted.

Next, we find the upper bound for the expectation of $\left(\sum_{i,j,l}\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}\psi_i\psi_j\psi_l\right)^2$. For convenience, we denote $\frac{\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\mathbf{M}_\xi}{\sigma_{n\mathbf{t}}(\mathbf{d})}$, $\frac{\mathbf{e}^T_{\mathbf{t}}(\mathbf{t})}{\sigma_{\mathbf{t}}(\mathbf{d})}$ as $\widetilde{\mathbf{c}}^T_1 = (c_{11},...,c_{1N})$ and $\widetilde{\mathbf{c}}^T_2 = (c_{21},...,c_{2N})$, respectively. Then $F_{\mathbf{t}}\left(\boldsymbol{\Psi}_{\mathbf{N}},\mathbf{d}\right)$ becomes

$$F_{\mathbf{t}}\left(\boldsymbol{\Psi}_{\mathbf{N}},\mathbf{d}\right) = \frac{\Phi\left(\widetilde{y}_{n\mathbf{dt}}\sigma_{n\mathbf{t}}(0)\sigma^{-1}_{n\mathbf{t}}(\mathbf{d}) + \frac{\widetilde{\mathbf{c}}^T_1\boldsymbol{\Psi}_N}{\sqrt{n}}\right)}{\Phi\left(x_{0\mathbf{t}}\sigma_{\mathbf{t}}(0)\sigma^{-1}_{\mathbf{t}}(\mathbf{d}) + \frac{\widetilde{\mathbf{c}}^T_2\boldsymbol{\Psi}_N}{\sqrt{n}}\right)}.$$

In this notation, the third order derivative $\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}$ is shown in (B.19) in Section 5.3. Note that all the denominators in (B.19) are $\Phi\left(\widehat{x}_{\mathbf{0t}}\right)$ with powers up to 4. Here

$$\widehat{x}_{\mathbf{0t}} = x_{\mathbf{0t}}\sigma_{\mathbf{t}}(0)\sigma^{-1}_{\mathbf{t}}(\mathbf{d}) + \widetilde{\mathbf{c}}^T_2\widehat{\boldsymbol{\Psi}}_{\mathbf{N}}.$$

We first prove that $\Phi^{-1}\left(\widehat{x}_{\mathbf{0t}}\right)$ is bounded. In Lemma 31, we showed that $\sigma_{n\mathbf{t}}(\mathbf{d}) \geq \left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|\sqrt{h_1}$. Moreover, since $\left\|\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\mathbf{M}_\xi\right\| \leq \left\|\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\right\|$ we have $\|\widetilde{\mathbf{c}}_1\| \leq 1/\sqrt{h_1}$. Since $d_1 \leq \mathbf{d} \leq d_2$, for sufficiently n,

$$\|\widetilde{\mathbf{c}}_2\| \leq \frac{1}{\sigma_{\mathbf{t}}(0)}e^{-d_1/(2\sqrt{n})} \leq \frac{\sqrt{2}}{\sigma_{\mathbf{t}}(0)},$$

and

$$\begin{aligned} (c_{1i}c_{1j}c_{1l})^2 &\leq \frac{1}{h_1^3}, (c_{1i}c_{1j}c_{2l})^2 \leq \frac{2}{h_1^2\sigma^2_{\mathbf{t}}(0)}, \\ (c_{1i}c_{2j}c_{2l})^2 &\leq \frac{4}{h_1\sigma^4_{\mathbf{t}}(0)}, (c_{2i}c_{2j}c_{2l})^2 \leq \frac{2^{3/2}}{\sigma^6_{\mathbf{t}}(0)}. \end{aligned}$$

Therefore,

$$\begin{aligned} E_{\mathbf{d}|\boldsymbol{\theta}}\left(c_{ki}c_{sj}c_{tl}\right)^2 &= O(1) \text{ and } E_{\mathbf{d}|\boldsymbol{\theta}}\left(c_{ki}c_{sj}c_{tl}\right)^4 = O(1), \\ \text{for } k,s,t &= 1,2, \text{ and } i,j,l = 1,...,N. \end{aligned}$$

Since $\left|x_{\mathbf{0t}}\sigma_{\mathbf{t}}(0)\sigma_{\mathbf{t}}^{-1}(\mathbf{d}) + \widetilde{\mathbf{c}}_2^T\widehat{\boldsymbol{\Psi}}_{\mathbf{N}}\right| \leq \left(|x_{\mathbf{0t}}| + \sigma_{\mathbf{0}}^{-1}(\mathbf{t})\tau n^{-1/2}\right)\sqrt{2}$ when n is large enough, we have

$$\begin{aligned} \Phi^{-1}\left(\widehat{x}_{\mathbf{0t}}\right) &= \Phi^{-1}\left(x_{\mathbf{0t}}\sigma_{\mathbf{t}}(0)\sigma_{\mathbf{t}}^{-1}(\mathbf{d}) + \widetilde{\mathbf{c}}_2^T\widehat{\boldsymbol{\Psi}}_{\mathbf{N}}\right) \\ &\leq \Phi^{-1}\left(-\left(|x_{\mathbf{0t}}| + \sigma_{\mathbf{0}}^{-1}(\mathbf{t})\tau n^{-1/2}\right)\sqrt{2}\right). \end{aligned}$$

Based on (B.19), we have for some positive constant k_3

$$\begin{aligned} & k_3^{-1}E^2_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left[\left(\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}\right)^2\right] \\ \leq\ & \frac{1}{\Phi^8\left(-\left(|x_{\mathbf{0t}}| + \sigma_{\mathbf{0}}^{-1}(\mathbf{t})\tau n^{-1/2}\right)\sqrt{2}\right)} \\ & \times\left\{\begin{array}{c} \frac{E_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left(\widetilde{y}^8_{n\mathbf{dt}}\right)+1}{h^8_{\mathbf{1}}} + \frac{E^{1/2}_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left(\widetilde{y}^4_{n\mathbf{dt}}\right)}{h^2_{\mathbf{1}}\sigma^2_{\mathbf{t}}(0)} \\ + \frac{\left(|x_{\mathbf{0t}}|+\sigma_{\mathbf{t}}^{-1}(0)\tau n^{-1/2}\right)^2+1}{h_{\mathbf{1}}\sigma^4_{\mathbf{t}}(0)} + \frac{\left(|x_{\mathbf{0t}}|+\sigma_{\mathbf{t}}^{-1}(0)\tau n^{-1/2}\right)^2+1}{\sigma^6_{\mathbf{t}}(0)} \end{array}\right\} \\ =\ & : k_4. \end{aligned}$$

Noticing that $E\left(\left\|\mathbf{D}^1_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^l\right) = O(1)$ for $2 \leq l \leq 16$ as shown in Proposition 27 and $\sigma_{n\mathbf{t}}^{-l}(0)$ are bounded, via a direct proof the moments of $\widetilde{y}_{n\mathbf{dt}}$ in k_4 are finite. Therefore, $k_4 = O(1)$ and $E_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left[\left(\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}\right)^2\right]$ is bounded by k_4. Moreover, we have

$$\begin{aligned} & E_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left(\sum_{i,j,l}\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}\psi_i\psi_j\psi_l\right)^2 \\ =\ & E_{\mathbf{z}_n,\mathbf{d}|\boldsymbol{\theta}}\left(\sum_{i,j,l,k,s,t}\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_i\partial\psi_j\partial\psi_l}\frac{\partial^3 F_{\mathbf{t}}(\widehat{\boldsymbol{\Psi}}_N,\mathbf{d})}{\partial\psi_k\partial\psi_s\partial\psi_t}\psi_i\psi_j\psi_l\psi_k\psi_s\psi_t\right) \\ \leq\ & k_3k_4||\boldsymbol{\Psi}_N||_2^6 < \infty. \end{aligned}$$

Via direct calculation similar as before,

$$E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\left(D^3_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\right)^2\right),$$

$$E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\left(D^{1,2}_{\boldsymbol{\Psi}_N\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\boldsymbol{\Psi}_N\right)^2\right)$$

and

$$E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(D^{2,1}_{\boldsymbol{\Psi}_N\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},\widehat{\mathbf{d}})\boldsymbol{\Psi}_N\right)^2$$

can be bounded by

$$g\left(\begin{array}{c} E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\widetilde{y}^{k_1}_{n\widehat{\mathbf{d}}\mathbf{t}}\right), E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\left\|D^{k_2}_{\mathbf{d}}\mathbf{a}^T_{n\mathbf{t}}\left(\widehat{\mathbf{d}}\right)\right\|^2\right), E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(D^{k_2}_{\mathbf{d}}\sigma_{n\mathbf{t}}\left(\widehat{\mathbf{d}}\right)\right)^2, \\ k_1=2,4, k_2=0,1,2,3 \end{array}\right)$$
$$\times\Phi^{-8}\left(-\left(|x_{\mathbf{0t}}|+\sigma^{-1}_{\mathbf{t}}(0)\tau n^{-1/2}\right)\sqrt{2}\right).$$

Here $g(\cdot)$ is a polynomial function of $E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\widetilde{y}^{k_1}_{n\widehat{\mathbf{d}}\mathbf{t}}\right)$, $E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(\left\|D^{k_2}_{\mathbf{d}}\mathbf{a}^T_{n\mathbf{t}}\left(\widehat{\mathbf{d}}\right)\right\|^2\right)$ and $E_{\mathbf{z}_n,\widehat{\mathbf{d}}|\boldsymbol{\theta}}\left(D^{k_2}_{\mathbf{d}}\sigma_{n\mathbf{t}}\left(\widehat{\mathbf{d}}\right)\right)^2$ which are all $O(1)$ according to Proposition 27. Therefore, $E\left(C^2_{3\mathbf{t}}\right)<\infty$.

5.2 Proof of Proposition 27

5.2.1 Preliminary results for the proof of Proposition 27

Before proving Proposition 27, we establish the following results.

Lemma 28 *With assumptions (A1),* $\lambda_{\min}(\mathbf{H})\geq h_1>0, \lambda_{\max}(\mathbf{H})\leq h_N<\infty$ *and* $d_1\leq\mathbf{d}\leq d_2$*, there exist* $0<c_1<c_2<\infty$ *such that*

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_{nn}(\mathbf{d})) &\geq c_1\\ \lambda_{\max}(\boldsymbol{\Sigma}_{nn}(\mathbf{d})) &\leq c_2.\end{aligned}$$

In particular, the eigenvalues of $\boldsymbol{\Sigma}_{nn,\mathbf{0}}$ *are all bounded from below and above by* c_1 *and* c_2*, respectively.*

Proof. Recall that $\boldsymbol{\Sigma}_{nn}(\mathbf{d})=\mathbf{M}_\xi(\mathbf{G}_\mathbf{d}+\mathbf{H})\mathbf{M}^T_\xi$ and $\mathbf{G}_\mathbf{d}=\mathbf{U}\text{diag}(\lambda_i e^{\mathbf{d}/\sqrt{n}})^N_{i=1}\mathbf{U}^T$. According to Weyl's inequality, we have

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_{nn}(\mathbf{d})) &\geq \lambda_{\min}(\mathbf{M}_\xi\mathbf{G}_\mathbf{d}\mathbf{M}^T_\xi)+\lambda_{\min}(\mathbf{M}_\xi\mathbf{H}\mathbf{M}^T_\xi),\\ \lambda_{\max}(\boldsymbol{\Sigma}_{nn}(\mathbf{d})) &\leq \lambda_{\max}(\mathbf{M}_\xi\mathbf{G}_\mathbf{d}\mathbf{M}^T_\xi)+\lambda_{\max}(\mathbf{M}_\xi\mathbf{H}\mathbf{M}^T_\xi).\end{aligned}$$

Notice that $\mathbf{M}_\xi\mathbf{M}^T_\xi=\mathbf{I}_{n\times n}$, $\lambda_{\min}(\mathbf{H})\geq h_1>0$ and $\lambda_{\max}(\mathbf{H})\leq h_N<\infty$. Then we have

$$\lambda_{\min}(\mathbf{M}_\xi\mathbf{H}\mathbf{M}^T_\xi)\geq h_1>0$$

and

$$\lambda_{\max}(\mathbf{M}_\xi \mathbf{H} \mathbf{M}_\xi^T) \leq h_N < \infty.$$

According to Assumption (A1), the following holds

$$\lambda_{\min}(\mathbf{M}_\xi \mathbf{G}_\mathbf{d} \mathbf{M}_\xi^T) \geq \lambda_1 e^{\mathbf{d}/\sqrt{n}} > \lambda_1 e^{d_1/\sqrt{n}} > 0,$$

and

$$\lambda_{\max}(\mathbf{M}_\xi \mathbf{G}_\mathbf{d} \mathbf{M}_\xi^T) \leq \lambda_N e^{\mathbf{d}/\sqrt{n}} < \lambda_N e^{d_2/\sqrt{n}}.$$

Therefore,

$$\begin{aligned} \lambda_{\min}(\mathbf{\Sigma}_{nn}(\mathbf{d})) &\geq \lambda_1 e^{d_1/\sqrt{n}} + h_1 =: c_1, \\ \lambda_{\max}(\mathbf{\Sigma}_{nn}(\mathbf{d})) &\leq \lambda_N e^{d_2/\sqrt{n}} + h_N =: c_2. \end{aligned}$$

Let $\mathbf{d} = 0$. Via the same procedure, we can show that the eigenvalues of $\mathbf{\Sigma}_{nn,\mathbf{0}}$ are bounded by c_1 and c_2. ■

Lemma 29 *Denote*

$$\mathbf{M}_{nn}(\mathbf{d}) := \left(\mathbf{F}_n^T \mathbf{\Sigma}_{nn}^{-1}(\mathbf{d}) \mathbf{F}_n\right)^{-1}. \tag{B.1}$$

Under assumptions (A1)-(A3), we have

$$E\left(\mathbf{M}_{nn}(\mathbf{d})\right) = \mathbf{O}(n^{-1}).$$

Especially,

$$\mathbf{M}_{nn}(0) = \mathbf{O}(n^{-1}).$$

Proof. We first prove that $\mathbf{M}_{nn}(0) = \mathbf{O}(n^{-1})$. Since $\mathbf{M}_{nn}^{-1}(0) = \mathbf{F}_n^T \mathbf{\Sigma}_{nn,\mathbf{0}}^{-1} \mathbf{F}_n$ is positive definite, we can decompose this matrix as $\mathbf{M}_{nn}^{-1}(0) = \mathbf{Q}\mathbf{K}\mathbf{Q}^T$ where $\mathbf{Q}$ is a matrix with orthogonal eigenvectors of $\mathbf{M}_{nn}^{-1}(0)$ as its columns and $\mathbf{K}$ is the diagonal matrix with eigenvalues on the diagonal. It is straightforward that

$$\mathbf{M}_{nn}(0) = \mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T,$$

and that

$$\|\mathbf{M}_{nn}(0)\| = \max\left(\operatorname{diag}\left(\mathbf{K}^{-1}\right)\right) = \frac{1}{\min\left(\operatorname{diag}\left(\mathbf{K}\right)\right)}.$$

Therefore, to prove that $\|\mathbf{M}_{nn}(0)\| = O(n^{-1})$, we need to prove that $\min(\mathrm{diag}(\mathbf{K})) \geq O(n)$. This holds since

$$\begin{aligned}
\min(\mathrm{diag}(\mathbf{K})) &= \min_{||x||=1, x\in\mathbb{R}^p} x^T \mathbf{M}_{nn}^{-1}(0) x \\
&= \min_{||x||=1} x^T \mathbf{F}_n^T \boldsymbol{\Sigma}_{nn,\mathbf{0}}^{-1} \mathbf{F}_n x \\
&\geq c_2^{-1} \min_{||x||=1} x^T \mathbf{F}_n^T \mathbf{F}_n x \\
&= c_2^{-1} ||(\mathbf{F}_n^T \mathbf{F}_n)^{-1}||^{-1} = O(n).
\end{aligned}$$

Now since

$$\|E(\mathbf{M}_{nn}(\mathbf{d}))\| \leq E(\|\mathbf{M}_{nn}(\mathbf{d})\|),$$

we can follow the same strategy as above and show that the norm is $O(n^{-1})$, i.e., $E(\mathbf{M}_{nn}(\mathbf{d})) = O(n^{-1})$. ■

Lemma 30 *Under Assumption (3), we have for each* $\mathbf{t} \in \mathcal{T}\backslash\mathcal{S}$

$$E\left(||\boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})||^2\right) = O(1)$$

Proof. Since the elements of $\mathbf{G}_\mathbf{d}$ are bounded, the proof is straightforward. ■

Lemma 31 *Under assumptions (A1)-(A3) and* $d_1 \leq \mathbf{d} \leq d_2$, *for each* $\mathbf{t} \in \mathcal{T}\backslash\mathcal{S}$ *the prediction error variance* $\sigma_{n\mathbf{t}}^2(\mathbf{d})$ *is bounded by* $\sigma_\mathbf{t}^2(\mathbf{d}) + \varpi_\mathbf{t}(\mathbf{d})$ *and* $h_1 \|\mathbf{f}(\mathbf{t})\|^2$,

$$\sigma_\mathbf{t}^2(\mathbf{d}) + \varpi_\mathbf{t}(\mathbf{d}) \geq \sigma_{n\mathbf{t}}^2(\mathbf{d}) \geq h_1 \|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\|^2 \geq h_1 \|\mathbf{f}(\mathbf{t})\|^2,$$

where with $\mathbf{M}_{nn}(\mathbf{d})$ *being defined in (B.1)*

$$\varpi_\mathbf{t}(\mathbf{d}) = (\mathbf{f}(\mathbf{t}) - \mathbf{F}_n^T \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d}) \boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{d}))^T \mathbf{M}_{nn}(\mathbf{d}) (\mathbf{f}(\mathbf{t}) - \mathbf{F}_n^T \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d}) \boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{d}))$$

is positive.

Proof. Recall that

$$\sigma_{n\mathbf{t}}^2(\mathbf{d}) = \mathrm{Var}(\mu(\mathbf{t}) - \hat{\mu}(\mathbf{t})),$$

and that

$$\begin{aligned}
\mu(\mathbf{t}) &= \eta(\mathbf{t}) + \delta(\mathbf{t}), \\
\hat{\mu}(\mathbf{t}) &= \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \mathbf{y}_n \\
&= \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \boldsymbol{\eta}_n + \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \boldsymbol{\delta}_n + \mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) \boldsymbol{\varepsilon}_n,
\end{aligned}$$

where $\boldsymbol{\eta}_n = (\eta(\mathbf{t}_{s1}), ..., \eta(\mathbf{t}_{sn}))^T$ are the deterministic means, $\boldsymbol{\delta}_n = (\delta(\mathbf{t}_{s1}), ..., \delta(\mathbf{t}_{sn}))^T$ and $\boldsymbol{\varepsilon}_n = (\varepsilon(\mathbf{t}_{s1}), ..., \varepsilon(\mathbf{t}_{sn}))^T$. Notice that $\delta(\mathbf{t}_l)$ and $\varepsilon(\mathbf{t}_k)$ are independent. Then

$$\begin{aligned}
\sigma^2_{n\mathbf{t}}(\mathbf{d}) &= \mathrm{Var}\left(\mu(\mathbf{t}) - \hat{\mu}(\mathbf{t})\right) \\
&= \mathrm{Var}\left[(\eta(\mathbf{t}) + \delta(\mathbf{t})) - \left(\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\eta}_n + \mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\delta}_n\right) - \mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\varepsilon}_n\right] \\
&= \mathrm{Var}\left[(\eta(\mathbf{t}) + \delta(\mathbf{t})) - \left(\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\eta}_n + \mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\delta}_n\right)\right] + \mathrm{Var}\left(\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\varepsilon}_n\right) \\
&\geq \mathrm{Var}\left(\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\boldsymbol{\varepsilon}_n\right) = \mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\mathbf{M}_\xi\mathbf{H}\mathbf{M}^T_\xi\mathbf{a}_{n\mathbf{t}}(\mathbf{d}) \geq h_1 \left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2.
\end{aligned}$$

Notice that $\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\mathbf{y}_n$ is an estimate of $\mu(\mathbf{t})$ with $\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})\mathbf{F}_n = \mathbf{f}^T(\mathbf{t})$. Another estimate of $\mu(\mathbf{t})$ is $\mathbf{a}^T_{*\mathbf{t}}(\mathbf{d}) = \mathbf{f}^T(\mathbf{t})\left(\mathbf{F}^T_n\mathbf{F}_n\right)^{-1}\mathbf{F}^T_n$ which also satisfies $\mathbf{a}^T_{*\mathbf{t}}(\mathbf{d})\mathbf{F}_n = \mathbf{f}^T(\mathbf{t})$. We can show that $\left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2 \geq \left\|\mathbf{a}^T_{*\mathbf{t}}(\mathbf{d})\right\|^2 = ||\mathbf{f}(\mathbf{t})||^2$. And then, $\sigma^2_{n\mathbf{t}}(\mathbf{d})$ is further bounded below by $h_1||\mathbf{f}(\mathbf{t})||^2$. The proof is straightforward and we omit it here.

On the other hand, we have that

$$\begin{aligned}
\sigma^2_{n\mathbf{t}}(\mathbf{d}) &= \sigma^2_{\mathbf{t}}(\mathbf{d}) - (\mathbf{f}^T(\mathbf{t}), \boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d}))\begin{pmatrix} \mathbf{0} & \mathbf{F}^T_n \\ \mathbf{F}_n & \boldsymbol{\Sigma}_{nn}(\mathbf{d}) \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{f}(\mathbf{t}) \\ \boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{d}) \end{pmatrix} \\
&\leq \sigma^2_{\mathbf{t}}(\mathbf{d}) + \varpi_{\mathbf{t}}(\mathbf{d})
\end{aligned}$$

since

$$\begin{aligned}
& (\mathbf{f}^T(\mathbf{t}), \boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d}))\begin{pmatrix} \mathbf{0} & \mathbf{F}^T_n \\ \mathbf{F}_n & \boldsymbol{\Sigma}_{nn}(\mathbf{d}) \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{f}(\mathbf{t}) \\ \boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{d}) \end{pmatrix} \\
= \; & \boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\boldsymbol{\Sigma}^{-1}_{nn}(\mathbf{d})\boldsymbol{\Sigma}_{n1\mathbf{t}}(\mathbf{d}) - \varpi_{\mathbf{t}}(\mathbf{d}).
\end{aligned}$$

■

5.2.2 Proof of Proposition 27

In this Proposition, we investigate the expectations of $\mathbf{a}^T_{n\mathbf{t}}(\mathbf{d})$ and $\sigma_{n\mathbf{t}}(\mathbf{d})$ at each location $\mathbf{t} \in \mathcal{T}\backslash\mathcal{S}$. The following inequality in Kleinman and Athans (1968) (see also in Fang, Y., Loparo, K. A., and Feng, X. 1994) is very useful in our proof:

$$\lambda_{\min}(\mathbf{A})tr(\mathbf{B}) \leq tr(\mathbf{AB}) \leq \lambda_{\max}(\mathbf{A})tr(\mathbf{B}) \tag{B.2}$$

for both $\mathbf{A}$ and $\mathbf{B}$ being positive semi-definite matrices.

First, we prove $E\left(\left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(1)$ and $\left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2 = O(1)$. To prove

$$E\left(\left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(1), \tag{B.3}$$

it is sufficient to prove that

$$E\left(\left\|\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\right\|^2\right) = O(1), \tag{B.4}$$

since, according to Lemma 28 and (B.2),

$$E\left(\left\|\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\right\|^2\right) \geq c_1 E\left(\left\|\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2\right).$$

We prove (B.4) instead of (B.3) because the form of $\left\|\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\right\|^2$ is easier than $\left\|\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\right\|^2$.

To establish (B.4), first let

$$\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d}) = \mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d}) + \mathbf{a}_{2n\mathbf{t}}^T(\mathbf{d}), \tag{B.5}$$

where with $\mathbf{M}_{nn}(\mathbf{d})$ being defined in (B.1),

$$\mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d}) = \mathbf{f}^T(\mathbf{t})\mathbf{M}_{nn}(\mathbf{d})\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d}), \tag{B.6}$$

and

$$\begin{aligned}\mathbf{a}_{2n\mathbf{d}}^T &= \boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})(\mathbf{I} - \boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{F}_n\mathbf{M}_{nn}(\mathbf{d})\mathbf{F}_n^T)\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d}) \\ &= \boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d})(\mathbf{I} - \mathbf{Q}_{nn}(\mathbf{d}))\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d}),\end{aligned} \tag{B.7}$$

with

$$\mathbf{Q}_{nn}(\mathbf{d}) = \boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{F}_n\mathbf{M}_{nn}(\mathbf{d})\mathbf{F}_n^T\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d}) \tag{B.8}$$

being an idempotent matrix.

Notice that

$$\mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}) = 0. \tag{B.9}$$

Then

$$\left\|\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{1/2}(\mathbf{d})\right\|^2 = \mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{1n\mathbf{t}}(\mathbf{d}) + \mathbf{a}_{2n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}). \tag{B.10}$$

Substituting (B.6) into the first term of (B.10), we have

$$E\left(\mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})\right) = \mathbf{f}^T(\mathbf{t})E\left(\mathbf{M}_{nn}(\mathbf{d})\right)\mathbf{f}(\mathbf{t}).$$

Moreover, according to the result in Lemma 29, we have

$$E\left(\mathbf{a}_{1n\mathbf{t}}^T(\mathbf{d})\boldsymbol{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})\right) \leq \left\|E\left(\mathbf{M}_{nn}(\mathbf{d})\right)\right\|\mathbf{f}^T(\mathbf{t})\mathbf{f}(\mathbf{t}) = O(n^{-1}). \tag{B.11}$$

As for the second term in (B.10), let

$$\mathbf{a}_{2n\mathbf{t}}^T(\mathbf{d})\mathbf{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}) = tr[\mathbf{AB}]$$

where

$$\mathbf{A} = \mathbf{I} - \mathbf{Q}_{nn}(\mathbf{d}),$$

$$\mathbf{B} = \mathbf{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{\Sigma}_{n1\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}^{-1/2}(\mathbf{d})$$

with $\mathbf{Q}_{nn}(\mathbf{d})$ being defined in (B.8). Since $\mathbf{Q}_{nn}(\mathbf{d})$ is an idempotent matrix, we have that the matrices $\mathbf{A}, \mathbf{B}$ are both positive semi-definite and that $\lambda_{\max}(\mathbf{A}) = 1$. Applying the inequality (B.2) we have $E\left(tr(\mathbf{AB})\right) \geq 0$ and according to Lemma 30

$$\begin{aligned} E\left(tr(\mathbf{AB})\right) &\leq E\left(tr(\mathbf{B})\lambda_{\max}(\mathbf{A})\right) \\ &= E\left(\mathbf{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{\Sigma}_{n1\mathbf{t}}(\mathbf{d})\right) \\ &\leq \frac{E\left(\left\|\mathbf{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\right\|^2\right)}{c_1} = O(1). \end{aligned} \tag{B.12}$$

Combining (B.11) and (B.12), we have that (B.4), and therefore (B.3), holds. Following the same procedure, we can prove that $\left\|\mathbf{a}_{n\mathbf{t}}(0)\right\|^2 = O(1)$.

Next, we prove $E\left(\left\|D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(1)$ and $\left\|D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(0)\right\|^2 = O(1)$. Similar as before, we split $\mathbf{a}_{n\mathbf{t}}(\mathbf{d})$ into two parts $\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})$ and $\mathbf{a}_{2n\mathbf{t}}(\mathbf{d})$. Then

$$D_{\mathbf{d}}^1\mathbf{a}_{n\mathbf{t}}(\mathbf{d}) = D_{\mathbf{d}}^1\mathbf{a}_{1n\mathbf{t}}(\mathbf{d}) + D_{\mathbf{d}}^1\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}).$$

We can prove that

$$E\left(\left\|D_{\mathbf{d}}^1\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(n^{-1}) \text{ and } E\left(||D_{\mathbf{d}}^1\mathbf{a}_{2n\mathbf{t}}(\mathbf{d})||^2\right) = O(1), \tag{B.13}$$

and at $\mathbf{d} = \mathbf{0}$,

$$\left\|D_{\mathbf{d}}^1\mathbf{a}_{1n\mathbf{t}}(0)\right\|^2 = O(n^{-1}) \text{ and } ||D_{\mathbf{d}}^1\mathbf{a}_{2n\mathbf{0}}||^2 = O(1).$$

Via direct calculation, we have

$$D_{\mathbf{d}}^1\mathbf{a}_{1n\mathbf{t}}(\mathbf{d}) = -\mathbf{f}^T(\mathbf{t})\mathbf{M}_{nn}(\mathbf{d})\mathbf{F}_n^T\mathbf{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{M}_\xi\widetilde{\mathbf{G}}_{\mathbf{d}},$$

and

$$D_{\mathbf{d}}^1\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}) = \begin{bmatrix} \mathbf{e}_{\mathbf{t}}^T - \mathbf{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{M}_\xi \\ +\mathbf{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{Q}_{nn,\mathbf{d}}\mathbf{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{M}_\xi \end{bmatrix}\widetilde{\mathbf{G}}_{\mathbf{d}}, \tag{B.14}$$

where

$$\widetilde{\mathbf{G}}_{\mathbf{d}} = \mathbf{G}_{\mathbf{d}}\mathbf{M}_{\xi}^{T}\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d})\left(\mathbf{I}-\mathbf{Q}_{nn}(\mathbf{d})\right)\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d}).$$

To prove (B.13), it is enough to prove that the mean square of each term in $D^1_{\mathbf{d}}\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})$ and $D^1_{\mathbf{d}}\mathbf{a}_{2n\mathbf{t}}(\mathbf{d})$ is $O(1)$. First, notice that $\mathbf{I}-\mathbf{Q}_{nn}(\mathbf{d})$ in $\widetilde{\mathbf{G}}_{\mathbf{d}}$ is an idempotent matrix and its maximum eigenvalue is 1. We have

$$\widetilde{\mathbf{G}}_{\mathbf{d}}\widetilde{\mathbf{G}}_{\mathbf{d}}^{T} \preceq c_1^{-1}\mathbf{G}_{\mathbf{d}}\mathbf{M}_{\xi}^{T}\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{M}_{\xi}\mathbf{G}_{\mathbf{d}}^{T} \preceq c_1^{-2}\lambda_N^2\mathbf{I}_{N\times N}. \tag{B.15}$$

Here, for any two symmetric matrix $\mathbf{A}$ and $\mathbf{B}$ we say $\mathbf{A}\preceq\mathbf{B}$ if $\mathbf{B}-\mathbf{A}$ is a positive semi-definite matrix.

Applying the inequality (B.15) on $D^1_{\mathbf{d}}\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})$ we have

$$E\left(\left\|D^1_{\mathbf{d}}\mathbf{a}_{1n\mathbf{t}}(\mathbf{d})\right\|^2\right) \le c_1^{-3}\lambda_N^2\mathbf{f}^T(\mathbf{t})E\left(\mathbf{M}_{nn}(\mathbf{d})\right)\mathbf{f}(\mathbf{t}) = O(n^{-1}).$$

Similarly, for the first and second terms in (B.14), we have

$$E\left(\left\|\mathbf{e}_{\mathbf{t}}^T\widetilde{\mathbf{G}}_{\mathbf{d}}\right\|^2\right) \le c_1^{-2}\lambda_N^2 = O(1),$$

$$E\left(\left\|\boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{-1}(\mathbf{d})\mathbf{M}_{\xi}\widetilde{\mathbf{G}}_{\mathbf{d}}\right\|^2\right) \le c_1^{-4}\lambda_N^2\left\|\boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\right\|^2 = O(1).$$

Moreover, since $\mathbf{Q}_{nn,\mathbf{d}}$ is also an idempotent matrix, for the third term in (B.14)

$$E\left(\left\|\boldsymbol{\Sigma}_{1n\mathbf{t}}(\mathbf{d})\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{Q}_{nn,\mathbf{d}}\boldsymbol{\Sigma}_{nn}^{-1/2}(\mathbf{d})\mathbf{M}_{\xi}\widetilde{\mathbf{G}}_{\mathbf{d}}\right\|^2\right) = O(1).$$

Therefore, $E\left(\left\|D^1_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(1)$. Moreover, $\left\|D^1_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(0)\right\|^2 = O(1)$.

Via similar calculation, it is straightforward to prove that $E\left(\left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(\mathbf{d})\right\|^l\right) = O(1)$ and $\left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{n\mathbf{t}}(0)\right\|^l = O(1)$ for $0\le k\le 3, 2\le l\le 16$, $k,l\in\mathbb{Z}$.

For the variance of the stochastic error δ at location $\mathbf{t}$, since $\sigma^2_{\mathbf{t}}(\mathbf{d}) = \mathbf{e}_{\mathbf{t}}^T\mathbf{G}_{\mathbf{d}}\mathbf{e}_{\mathbf{t}}$ and the eigenvalues of $\mathbf{G}_{\mathbf{d}}$ are bounded by some constants, it is straightforward that

$$E\left(\left(D^k_{\mathbf{d}}\sigma^2_{\mathbf{t}}(\mathbf{d})\right)^2\right) = O(1) \text{ and } D^k_{\mathbf{d}}\sigma^2_{\mathbf{t}}(0) = O(1). \tag{B.16}$$

In the following, we prove that

$$E\left(\left(D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(\mathbf{d})\right)^2\right) = O(1) \text{ and } \left(D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right)^2 = O(1). \tag{B.17}$$

We first investigate $E\left(\sigma^2_{n\mathbf{t}}(\mathbf{d})\right)$ and $E\left(\left(D^1_{\mathbf{d}}\sigma^2_{n\mathbf{t}}(\mathbf{d})\right)^2\right)$. According to Lemma 31, the prediction error variance $\sigma^2_{n\mathbf{t}}(\mathbf{d})$ is dominated by $\sigma^2_{\mathbf{t}}(\mathbf{d})+\varpi_{\mathbf{t}}(\mathbf{d})$ from above and

bounded below by $h_1||\mathbf{f}(\mathbf{t})||^2$. We can show that $E(\varpi_{\mathbf{t}}(\mathbf{d})) = O(1)$. Actually, according to Lemma 28 and Assumption (A3), applying inequality (B.2) we can directly prove that each term in $E(\varpi_{\mathbf{t}}(\mathbf{d}))$ is at most $O(1)$. The proof is omitted here. Therefore

$$h_1||\mathbf{f}(\mathbf{t})||^2 \leq E\left(\sigma^2_{n\mathbf{t}}(0)\right) \leq E\left(\sigma^2_{\mathbf{t}}(\mathbf{d})\right) + E\left(\varpi_{\mathbf{t}}(\mathbf{d})\right) = O(1).$$

Moreover, we have

$$|E(\sigma_{n\mathbf{t}}(\mathbf{d}))| \leq \sqrt{E\left(\sigma^2_{n\mathbf{t}}(0)\right)} = O(1).$$

Since

$$\begin{aligned}\sigma^2_{n\mathbf{t}}(\mathbf{d}) &= \mathrm{Var}(\mu(\mathbf{t}) - \hat{\mu}(\mathbf{t})) \\ &= \sigma^2_{\mathbf{t}}(\mathbf{d}) - \mathbf{a}^T_{2n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{2n\mathbf{t}}(\mathbf{d}) + \mathbf{a}^T_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{nn}(\mathbf{d})\mathbf{a}_{1n\mathbf{t}}(\mathbf{d}) - 2\mathbf{a}^T_{1n\mathbf{t}}(\mathbf{d})\mathbf{\Sigma}_{n1\mathbf{t}}(\mathbf{d}),\end{aligned}$$

and $E\left(\left\|\mathbf{D}^k_{\mathbf{d}}\mathbf{a}_{jn\mathbf{t}}(\mathbf{d})\right\|^2\right) = O(1)$, it is a direct proof that

$$E\left(D^k_{\mathbf{d}}\sigma^2_{n\mathbf{t}}(\mathbf{d})\right)^2 = O(1) \text{ and } \left(D^k_{\mathbf{d}}\sigma^2_{n\mathbf{t}}(0)\right)^2 = O(1).$$

According to Lemma 31,

$$\left|E\left(D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(\mathbf{d})\right)^2\right| = \left|E\left(\frac{D^1_{\mathbf{d}}\sigma^2_{n\mathbf{t}}(\mathbf{d})}{2\sigma_{n\mathbf{t}}(\mathbf{d})}\right)^2\right| \leq \frac{\left|E\left(D^1_{\mathbf{d}}\sigma^2_{n\mathbf{t}}(\mathbf{d})\right)^2\right|}{4h_1||\mathbf{f}(\mathbf{t})||^2} = O(1).$$

By deduction, we can further conclude that $E\left(D^k_{\mathbf{d}}\sigma_{n\mathbf{t}}(\mathbf{d})\right)^2 = O(1)$.

5.3 Derivatives in Theorem 25

Theorem 25 includes derivatives of the function $F_{\mathbf{t}}(\mathbf{\Psi_N}, \mathbf{d})$ with respect to its arguments. In this subsection, we calculate them term by term. First,

$$\begin{aligned}D^1_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0) &= \frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi\left(x_{0\mathbf{t}}\right)}\left(\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)} - \frac{\widetilde{y}_{n\mathbf{0t}}}{\sigma_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right) \\ &\quad + \frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}}{\sigma_{\mathbf{t}}(0)\Phi^2\left(x_{0\mathbf{t}}\right)}D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0),\end{aligned} \tag{B.18}$$

$$D^1_{\psi_N}F_{\mathbf{t}}(\mathbf{0},0) = \frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi}{\Phi\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}(0)} - \frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)\mathbf{e}^T_{\mathbf{t}}}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}(0)},$$

$$
\begin{aligned}
D^2_{\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0) &= \frac{-\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\widetilde{y}_{n\mathbf{0t}}}{\Phi\left(x_{0\mathbf{t}}\right)}\left(\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)}-\frac{\widetilde{y}_{n\mathbf{0t}}}{\sigma_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right)^2 \\
&+\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi\left(x_{0\mathbf{t}}\right)}\begin{pmatrix} \frac{D^2_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)}+\frac{2\widetilde{y}_{n\mathbf{0t}}\left(D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right)^2}{\sigma^2_{n\mathbf{t}}(0)} \\ -\frac{\widetilde{y}_{n\mathbf{0t}}D^2_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)}{\sigma_{n\mathbf{t}}(0)}-\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma^2_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0) \end{pmatrix} \\
&+\frac{2\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}(0)}\left(\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)}-\frac{\widetilde{y}_{n\mathbf{0t}}}{\sigma_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right) \\
&+\frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)\left(D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\right)^2\left(x^2_{0\mathbf{t}}-2x_{0\mathbf{t}}\right)}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)}+\frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}D^2_{\mathbf{d}}\sigma_{\mathbf{t}}(0)}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}(0)} \\
&+\frac{2\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi^2\left(x_{0\mathbf{t}}\right)x^2_{0\mathbf{t}}\left(D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\right)^2}{\Phi^3\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)},
\end{aligned}
$$

$$
\begin{aligned}
D^{1,1}_{\boldsymbol{\psi}_N\mathbf{d}}F_{\mathbf{t}}(\mathbf{0},0) &= -\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\widetilde{y}_{n\mathbf{0t}}}{\Phi\left(x_{0\mathbf{t}}\right)}\frac{\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi}{\sigma_{\mathbf{t}}(0)}\left(\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)}-\frac{\widetilde{y}_{n\mathbf{0t}}}{\sigma_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right) \\
&-\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi\left(x_{0\mathbf{t}}\right)}\frac{\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi}{\sigma^2_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)+\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi\left(x_{0\mathbf{t}}\right)}\frac{D^1_{\mathbf{d}}\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi}{\sigma_{n\mathbf{t}}(0)} \\
&+\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi}{2\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{n\mathbf{t}}(0)\sigma_{\mathbf{t}}(0)} \\
&-\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)}{\Phi^2\left(x_{0\mathbf{t}}\right)}\left(\frac{D^1_{\mathbf{d}}\mathbf{b}^T_{n\mathbf{t}}(0)\mathbf{z}_n}{\sigma_{n\mathbf{t}}(0)}-\frac{\widetilde{y}_{n\mathbf{0t}}}{\sigma_{n\mathbf{t}}(0)}D^1_{\mathbf{d}}\sigma_{n\mathbf{t}}(0)\right)\frac{\mathbf{e}^T_{\mathbf{t}}}{\sigma_{\mathbf{t}}(0)} \\
&+\frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\mathbf{e}^T_{\mathbf{t}}}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)} \\
&-\frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x^2_{0\mathbf{t}}D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\mathbf{e}^T_{\mathbf{t}}}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)}-\frac{2\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi^2\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}D^1_{\mathbf{d}}\sigma_{\mathbf{t}}(0)\mathbf{e}^T_{\mathbf{t}}}{\Phi^3\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)},
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\boldsymbol{\psi}_N}F_{\mathbf{t}}(\mathbf{0},0) \\
&= -\frac{\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\widetilde{y}_{n\mathbf{0t}}}{\Phi\left(x_{0\mathbf{t}}\right)\sigma^2_{n\mathbf{t}}(0)}\mathbf{M}^T_\xi\mathbf{a}_{n\mathbf{t}}(0)\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi+2\frac{\phi^2\left(x_{0\mathbf{t}}\right)\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi^3\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)}\mathbf{e}_{\mathbf{t}}\mathbf{e}^T_{\mathbf{t}} \\
&-\frac{\phi\left(x_{0\mathbf{t}}\right)\phi\left(\widetilde{y}_{n\mathbf{0t}}\right)}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma_{\mathbf{t}}(0)\sigma_{n\mathbf{t}}(0)}\left(\mathbf{e}_{\mathbf{t}}\mathbf{a}^T_{n\mathbf{t}}(0)\mathbf{M}_\xi+\mathbf{M}^T_\xi\mathbf{a}_{n\mathbf{t}}(0)\mathbf{e}^T_{\mathbf{t}}\right) \\
&+\frac{\Phi\left(\widetilde{y}_{n\mathbf{0t}}\right)\phi\left(x_{0\mathbf{t}}\right)x_{0\mathbf{t}}}{\Phi^2\left(x_{0\mathbf{t}}\right)\sigma^2_{\mathbf{t}}(0)}\mathbf{e}_{\mathbf{t}}\mathbf{e}^T_{\mathbf{t}}
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial^3 F_{\mathbf{t}}\left(\mathbf{\Psi_N}, \mathbf{d}\right)}{\partial\psi_i \partial\psi_j \partial\psi_l} \\
= \; & -\frac{\phi(\widehat{y}_{n\mathbf{dt}})\widehat{y}_{n\mathbf{dt}}^2 c_{1i}c_{1j}c_{1l}}{\Phi\left(\widehat{x}_{\mathbf{0t}}\right)} - \frac{\phi^2\left(\widehat{y}_{n\mathbf{dt}}\right)c_{1i}c_{1j}c_{1l}}{\Phi^3\left(\widehat{x}_{\mathbf{0t}}\right)} + \frac{\phi\left(\widehat{y}_{n\mathbf{dt}}\right)\phi\left(\widehat{x}_{\mathbf{0t}}\right)\widehat{y}_{n\mathbf{dt}}c_{1i}c_{1j}c_{2l}}{\Phi^2\left(\widehat{x}_{\mathbf{0t}}\right)} \\
& +2\frac{\phi\left(\widehat{y}_{n\mathbf{dt}}\right)\phi\left(\widehat{x}_{\mathbf{0t}}\right)\widehat{y}_{n\mathbf{dt}}c_{1i}c_{2j}c_{1l}}{\Phi^2\left(\widehat{x}_{\mathbf{0t}}\right)} + 2\frac{\phi\left(\widehat{y}_{n\mathbf{dt}}\right)\phi\left(\widehat{x}_{\mathbf{0t}}\right)\widehat{x}_{\mathbf{0t}}c_{1i}c_{2j}c_{2l}}{\Phi^2\left(\widehat{x}_{\mathbf{0t}}\right)} \\
& +4\frac{\phi\left(\widehat{y}_{n\mathbf{dt}}\right)\phi^2\left(\widehat{x}_{\mathbf{0t}}\right)c_{1i}c_{2j}c_{2l}}{\Phi^3\left(\widehat{x}_{\mathbf{0t}}\right)} \\
& +\begin{pmatrix} \frac{\phi(\widehat{y}_{n\mathbf{dt}})\phi^2(\widehat{x}_{\mathbf{0t}})c_{2i}c_{2j}c_{1l}}{\Phi^2(\widehat{x}_{\mathbf{0t}})} - 2\frac{\Phi(\widehat{y}_{n\mathbf{dt}})\phi^2(\widehat{x}_{\mathbf{0t}})\widehat{x}_{\mathbf{0t}}c_{2i}c_{2j}c_{2l}}{\Phi^2(\widehat{x}_{\mathbf{0t}})} \\ -2\frac{\Phi(\widehat{y}_{n\mathbf{dt}})\phi^2(\widehat{x}_{\mathbf{0t}})c_{2i}c_{2j}c_{2l}}{\Phi^3(\widehat{x}_{\mathbf{0t}})} \end{pmatrix}\left(1+\frac{2}{\Phi\left(\widehat{x}_{\mathbf{0t}}\right)}\right) \\
& +\frac{2\Phi\left(\widehat{y}_{n\mathbf{dt}}\right)\phi^3\left(\widehat{x}_{\mathbf{0t}}\right)c_{2i}c_{2j}c_{2l}}{\Phi^4\left(\widehat{x}_{\mathbf{0t}}\right)},
\end{aligned}
\tag{B.19}
$$

where $\frac{\mathbf{a}_{n\mathbf{t}}^T(\mathbf{d})\mathbf{M}_\xi}{\sigma_{n\mathbf{t}}(\mathbf{d})} =: \widetilde{\mathbf{c}}_1^T, \frac{\mathbf{e}_{\mathbf{t}}^T(\mathbf{t})}{\sigma_{\mathbf{t}}(\mathbf{d})} =: \widetilde{\mathbf{c}}_2^T$, $\widetilde{\mathbf{c}}_1^T = (c_{11}, ..., c_{1N})$ and $\widetilde{\mathbf{c}}_2^T = (c_{21}, ..., c_{2N})$, and $\widehat{y}_{n\mathbf{dt}} = \widetilde{y}_{n\mathbf{dt}}\sigma_{n\mathbf{t}}(0)\sigma_{n\mathbf{t}}^{-1}(\mathbf{d}) + \widetilde{\mathbf{c}}_1^T\widehat{\mathbf{\Psi}}_{\mathbf{N}}$ and $\widehat{x}_{\mathbf{0t}} = x_{\mathbf{0t}}\sigma_{\mathbf{t}}(0)\sigma_{\mathbf{t}}^{-1}(\mathbf{d}) + \widetilde{\mathbf{c}}_2^T\widehat{\mathbf{\Psi}}_{\mathbf{N}}$.