

Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors

PETER Z. G. QIAN
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706
(*zhiguang@stat.wisc.edu*)

HUAIQING WU
Department of Statistics
Iowa State University
Ames, IA 50011
(*isuhwu@iastate.edu*)

C. F. JEFF WU
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
(*jeffwu@isye.gatech.edu*)

August 22, 2007

Abstract

Modeling experiments with qualitative and quantitative factors is an important issue in computer modeling. A framework for building Gaussian process models that incorporate both types of factors is proposed. The key to the development of these new models is an approach for constructing correlation functions with qualitative and quantitative factors. An iterative estimation procedure is developed for the proposed models. Modern optimization techniques are used in the estimation to ensure the validity of the constructed correlation functions. The proposed method is illustrated with an example involving a known function and a real example for modeling the thermal distribution of a data center.

KEY WORDS: Cokriging; Design of experiments; Kriging; Multivariate Gaussian processes; Semi-definite programming.

1 INTRODUCTION

In recent years, there has been a growing interest in the use of computer models in sciences, engineering, and business. The corresponding physical experimentation might otherwise be time-consuming, costly, or even impossible to conduct. Because of their many attractive features, Gaussian process (GP) models have been established as a core tool for modeling computer

experiments. (For detailed discussions of such models, see Santner, Williams, and Notz 2003; Fang, Li, and Sudjianto 2005.) An important but unresolved issue is how to model computer experiments with qualitative and quantitative factors. Standard methods assume that all the factors involved in a computer experiment are quantitative. However, in many situations, some factors are qualitative by nature. Consider, for instance, the data-center computer experiment to be discussed in Section 7. The configuration variables that determine the thermal properties of a data center can be either quantitative or qualitative. Examples of quantitative variables are rack temperature rise, rack power, and diffuser flow rate. Examples of qualitative variables are diffuser height (with levels “mid height of the room” and “ceiling height”), mixed power (with levels “uniform,” “alt-zero,” and “alt-half”), and hot-air return-vent location (with levels “perpendicular-bottom,” “perpendicular-top,” “parallel-bottom,” and “parallel-top”) (Schmidt, Cruz, and Iyengar 2005). Computer models with qualitative and quantitative factors occur frequently in business operations applications, where some social economical status of the customers, such as gender and commuting method, is inherently qualitative.

The purpose of this article is to propose new GP models to address this issue. Note that the corresponding problem for physical experimentation is much easier because no GP model is involved (Wu and Ding 1998; Wu and Hamada 2000). Quadratic models have long been used for modeling physical experiments involving quantitative and qualitative factors. While such polynomial models can provide reasonable approximations to physical phenomena, they are inapplicable to computer modeling. Unlike physical experiments, computer experiments often include many factors and are known to have highly non-linear input-output relationships. It is essential to develop data-driven models for computer experiments with qualitative and quantitative factors. Inspired by the success of GP models with quantitative factors, we extend them to accommodate both qualitative and quantitative factors. While McMillan, Sacks, Welch, and Gao (1999) proposed GP models with both types of factors for analyzing protein activity data, their correlation functions for the qualitative factors assume special structures and have limited applicability, as will be discussed in Section 4.2. As a key to the development of the new GP models, a general approach for constructing correlation functions with both types of factors is proposed. An iterative estimation procedure is developed for the proposed model, making use of some modern optimization techniques to ensure the validity of the constructed correlation functions.

The remainder of this article is organized as follows. Section 2 presents the models used throughout the article and the motivation for this study. Section 3 gives a general approach for constructing correlation functions for GP models with qualitative and quantitative factors. Section 4 discusses and proposes some restrictive correlation matrices for qualitative factors that may be justifiable in particular applications. Section 5 presents estimation and prediction procedures. Sections 6 and 7 illustrate the proposed method with an example involving a known function, and with a real example for modeling temperature in a data center. Section 8 provides some discussions and concluding remarks. Proofs and computational details are deferred to the Appendix.

2 MODELS AND MOTIVATION

2.1 Gaussian Process Models With Quantitative Factors

For later development, we first briefly review GP models with quantitative factors. Suppose that an experiment involves I factors (input variables) $\mathbf{x} = (x_1, \dots, x_I)^t$; the data consist of an $n \times I$ matrix of input values $\mathbf{X} = (\mathbf{x}_1^0, \dots, \mathbf{x}_n^0)^t$ and the corresponding response values $\mathbf{y} = (y_1, \dots, y_n)^t$. The GP model assumes the following:

$$y(\mathbf{x}) = \boldsymbol{\beta}^t \mathbf{f}(\mathbf{x}) + \epsilon(\mathbf{x}), \tag{1}$$

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^t$ is a vector of m pre-specified functions, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^t$ is a vector of unknown coefficients. The residual $\epsilon(\mathbf{x})$ is assumed to be a stationary GP with mean zero, variance σ^2 , and correlation function $\text{cor}(\epsilon(\mathbf{x}_1), \epsilon(\mathbf{x}_2)) = K_\phi(\mathbf{x}_1, \mathbf{x}_2)$, where ϕ is the unknown *correlation parameters*.

It is well known that the product of one-dimensional correlation functions is a valid correlation function. This allows each factor to have its own correlation parameters, which can shed light on how response values are correlated among different factors. One popular choice is the *product exponential correlation function* (Santner, Williams, and Notz 2003):

$$K_\phi(\mathbf{x}_1, \mathbf{x}_2) = \prod_{i=1}^I \exp\{-\phi_i(x_{i1} - x_{i2})^p\}, \tag{2}$$

where $\phi_i \geq 0$ for $i = 1, \dots, I$. Here, $\exp\{-\phi_i(x_{i1} - x_{i2})^p\}$ ($0 < p \leq 2$) is a valid correlation function for the variable x_i (Abrahamsen 1997). Note that p is often fixed at 2, giving the *Gaussian correlation function*, which will be used in our examples. This reduces the complication in estimating the correlation parameters, and also makes the sample path of the GP infinitely differentiable, which is a reasonable assumption for many applications. The scale correlation parameters ϕ_1, \dots, ϕ_I measure the ruggedness of the response surface (sample path) of the GP. Larger values of ϕ_i 's imply a more rugged response surface.

2.2 Gaussian Process Models With Qualitative and Quantitative Factors

To develop GP models with qualitative and quantitative factors, we first note that a computer experiment tends to involve more quantitative factors, because they are more informative and more of them are often needed to specify the underlying physics or mathematics of the experiment. Although the number of qualitative factors is usually not large, they can determine some important properties of the experiment. For example, in the data-center experiment to be discussed in Section 7, cooling material, heat-transformation method, and diffuser orientation are qualitative factors. To take into account the distinct natures and roles of quantitative and qualitative factors in computer modeling, we describe below two analysis approaches.

The first is the *independent analysis* in which distinct GP models are used for modeling the data collected at different level combinations of the qualitative factors. This method ignores possible correlations among the responses at the same input values for the quantitative factors. Furthermore, its implementation requires fitting many GP models, with a large number of unknown parameters, even for a small number of qualitative factors. Consider, for example, an experiment with seven quantitative factors and three 4-level qualitative factors. The independent analysis would require fitting 64 ($= 4^3$) models, which involve 64 mean parameters (with a constant mean for each GP), 64 variances, and 448 ($= 64 \times 7$) correlation parameters (when the Gaussian correlation function is used). To accurately estimate these 576 parameters would require a large number of observations, which generally cannot be afforded.

In view of the shortcomings of the above approach, we introduce an *integrated analysis* in this article. It assumes a *single* GP model across different values of qualitative and quantitative factors as to borrow strengths from all the observations. Suppose that a computer experiment involves factors $\mathbf{w} = (\mathbf{x}^t, \mathbf{z}^t)^t$, where $\mathbf{x} = (x_1, \dots, x_I)^t$ are quantitative factors, and $\mathbf{z} = (z_1, \dots, z_J)^t$ are qualitative factors. Throughout this article, \mathbf{z} are assumed to be *unordered* qualitative factors unless described otherwise.

Similar to (1), the response $y(\mathbf{w})$ at the input value \mathbf{w} is assumed to be

$$y(\mathbf{w}) = \boldsymbol{\beta}^t \mathbf{f}(\mathbf{w}) + \epsilon(\mathbf{w}), \quad (3)$$

where $\mathbf{f}(\mathbf{w}) = (f_1(\mathbf{w}), \dots, f_m(\mathbf{w}))^t$ is a vector of m pre-specified functions (e.g., polynomials), and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^t$ is a vector of unknown coefficients. The residual $\epsilon(\mathbf{w})$ is assumed to be a GP with mean zero, variance σ^2 , and some correlation function. Construction of a “valid” correlation function for $\epsilon(\mathbf{w})$ is not straightforward because such a function needs to be defined in the space involving both qualitative and quantitative factors. The Gaussian correlation function used in Section 2.1 or other distance-based correlation functions (Santner, Williams, and Notz 2003) are not applicable due to the absence of the notion of “distance” for qualitative factors. A general method for constructing valid correlation functions is developed in Section 3.

3 CONSTRUCTION OF CORRELATION FUNCTIONS FOR GAUSSIAN PROCESSES WITH QUALITATIVE AND QUANTITATIVE FACTORS

In this section, we propose a general method for constructing valid correlation functions for $\epsilon(\mathbf{w})$ in model (3). The method does not use the normality assumption of GPs and hence applies to general stochastic processes with qualitative and quantitative factors.

First, consider the simple case with one qualitative factor, z_1 , with m_1 levels, denoted by $1, \dots, m_1$. To define the correlation function of $\epsilon(\mathbf{w})$, where $\mathbf{w} = (\mathbf{x}^t, z_1)^t$, let $\epsilon_u(\mathbf{x}) = \epsilon((\mathbf{x}^t, u)^t)$,

for $u = 1, \dots, m_1$, and envision a mean-zero m_1 -variate process

$$\boldsymbol{\epsilon}^*(\mathbf{x}) = (\epsilon_1(\mathbf{x}) \cdots \epsilon_{m_1}(\mathbf{x}))^t.$$

Then we only need to define correlation and cross-correlation functions for $\boldsymbol{\epsilon}^*(\mathbf{x})$. A convenient approach is to assume that $\boldsymbol{\epsilon}^*(\mathbf{x}) = \mathbf{A}\boldsymbol{\eta}(\mathbf{x})$, where $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_{m_1})^t$ is an $m_1 \times m_1$ nonsingular matrix with unit row vectors (i.e., $\mathbf{a}_u^t \mathbf{a}_u = 1$ for $u = 1, \dots, m_1$), and $\boldsymbol{\eta}(\mathbf{x}) = (\eta_1(\mathbf{x}), \dots, \eta_{m_1}(\mathbf{x}))^t$, where $\eta_1(\mathbf{x}), \dots, \eta_{m_1}(\mathbf{x})$ are independent stochastic processes with the same variance σ^2 and correlation function K_ϕ . Thus, $\text{cor}(\boldsymbol{\eta}(\mathbf{x}_1), \boldsymbol{\eta}(\mathbf{x}_2)) = K_\phi(\mathbf{x}_1, \mathbf{x}_2) \mathbf{I}_{m_1}$, with \mathbf{I}_{m_1} being the $m_1 \times m_1$ identity matrix. Then for input values $\mathbf{w}_i = (\mathbf{x}_i^t, z_{1i})^t$ ($i = 1, 2$), the correlation function is

$$\text{cor}(\boldsymbol{\epsilon}(\mathbf{w}_1), \boldsymbol{\epsilon}(\mathbf{w}_2)) = \text{cor}(\epsilon_{z_1}(\mathbf{x}_1), \epsilon_{z_2}(\mathbf{x}_2)) = \text{cor}(\mathbf{a}_{z_{11}}^t \boldsymbol{\eta}(\mathbf{x}_1), \mathbf{a}_{z_{12}}^t \boldsymbol{\eta}(\mathbf{x}_2)) = \mathbf{a}_{z_{11}}^t \mathbf{a}_{z_{12}} K_\phi(\mathbf{x}_1, \mathbf{x}_2). \quad (4)$$

Let $\tau_{r,s} = \mathbf{a}_r^t \mathbf{a}_s$, where $r, s = 1, \dots, m_1$. Then $\mathbf{T}_1 = (\tau_{r,s}) = \mathbf{A}\mathbf{A}^t$ is an $m_1 \times m_1$ *positive definite matrix with unit diagonal elements* (abbreviated to PDUDE). In fact, any PDUDE can be written as $\mathbf{B}\mathbf{B}^t$, where \mathbf{B} is a nonsingular matrix with unit row vectors. Thus, the above construction shows that, for any PDUDE $\mathbf{T}_1 = (\tau_{r,s})$ and any correlation function $K_\phi(\mathbf{x}_1, \mathbf{x}_2)$, $\text{cor}(\boldsymbol{\epsilon}(\mathbf{w}_1), \boldsymbol{\epsilon}(\mathbf{w}_2)) = \tau_{z_{11}, z_{12}} K_\phi(\mathbf{x}_1, \mathbf{x}_2)$ is a valid correlation function. Similar correlation functions are used in Mardia and Goodall (1993) for kriging, in Brown, Le, and Zidek (1994) for assigning a prior to a covariance matrix, and in Banerjee and Gelfand (2002) for modeling a cross-covariance matrix.

Now consider the general case with J qualitative factors $\mathbf{z} = (z_1, \dots, z_J)^t$, where z_j has m_j levels, denoted by $1, \dots, m_j$, for $j = 1, \dots, J$. As an extension to (4), a correlation function for $\boldsymbol{\epsilon}(\mathbf{w})$ can be constructed as

$$\text{cor}(\boldsymbol{\epsilon}(\mathbf{w}_1), \boldsymbol{\epsilon}(\mathbf{w}_2)) = \prod_{j=1}^J \left[\tau_{j, z_{j1}, z_{j2}} K_{\phi_j}(\mathbf{x}_1, \mathbf{x}_2) \right], \quad (5)$$

where $\mathbf{T}_j = (\tau_{j,r,s})$ is an $m_j \times m_j$ PDUDE. This is a valid correlation function as it is the product of J valid correlation functions $\tau_{j, z_{j1}, z_{j2}} K_{\phi_j}(\mathbf{x}_1, \mathbf{x}_2)$ ($j = 1, \dots, J$) in (4) for the qualitative factors z_1, \dots, z_J (Santner, Williams, and Notz 2003).

In particular, if $K_{\phi_j}(\mathbf{x}_1, \mathbf{x}_2)$ takes the form of $\exp\{-\sum_{i=1}^I \phi_{ij}(x_{i1} - x_{i2})^p\}$ in (2), it is a valid correlation function as discussed in Section 2.1. Then, (5) becomes

$$\text{cor}(\boldsymbol{\epsilon}(\mathbf{w}_1), \boldsymbol{\epsilon}(\mathbf{w}_2)) = \left[\prod_{j=1}^J \tau_{j, z_{j1}, z_{j2}} \right] \exp \left\{ - \sum_{i=1}^I \phi_i(x_{i1} - x_{i2})^p \right\}, \quad (6)$$

where $\phi_i = \sum_{j=1}^J \phi_{ij}$ for $i = 1, \dots, I$. Note that (6) bears some resemblance to (2) for the GP model with quantitative factors. This similarity will, in part, motivate the inference procedures in Section 5. The parameter $\tau_{j, z_{j1}, z_{j2}}$ measures the correlation (similarity) between the responses

at any two input values \mathbf{w}_1 and \mathbf{w}_2 that differ only on the values of z_j —at levels z_{j1} and z_{j2} , respectively. Similar to the GP model with quantitative factors, the scale correlation parameters ϕ_1, \dots, ϕ_I measure the ruggedness of the response surface (sample path) of the GP with the same values of \mathbf{z} for the qualitative factors. Larger values of ϕ_i 's imply a more rugged response surface. When (6) is used for modeling a computer experiment, it is sometimes desirable to assume that all the elements in \mathbf{T}_j are *positive* to ensure that the responses of the experiment at any two input values are positively correlated.

4 RESTRICTIVE CORRELATION MATRICES FOR QUALITATIVE FACTORS

For flexible modeling, unrestrictive correlation matrices (i.e., PDUDEs) should be used in (6) for the qualitative factors. Sometimes it may be desirable to use *restrictive* correlation matrices that assume some parametric relationships among the factor levels. In this section, we shall consider several such correlation matrices. Although substantial simplification in estimating them may be realized, they should be used with justification of the assumed relationships. Throughout this section, we consider modeling an $m \times m$ correlation matrix $\mathbf{T} = (\tau_{r,s})$ for a qualitative factor z with m levels.

4.1 Isotropic Correlation Functions

The isotropic correlation function assumes that the m levels of z are of *isotropic* nature (Stein 1999); that is, $\tau_{r,s} = c$ for all $r \neq s$, which holds automatically if z has two levels. Then $\mathbf{T} = (1 - c)\mathbf{I}_m + c\mathbf{1}\mathbf{1}^t$, where $\mathbf{1} = (1, \dots, 1)^t$. For $0 < c < 1$, $\mathbf{a}^t\mathbf{T}\mathbf{a} = (1 - c)\mathbf{a}^t\mathbf{a} + c(\mathbf{a}^t\mathbf{1})^2 > 0$, for any non-zero $m \times 1$ vector \mathbf{a} . Thus, for $0 < c < 1$, \mathbf{T} is a PDUDE and is indeed a legitimate correlation matrix. In this case, $\tau_{r,s} = \exp\{-\theta I[r \neq s]\}$, where $\theta = \ln(1/c) > 0$, and $I[r \neq s]$ is the indicator function that takes 1 if $r \neq s$ and 0 otherwise. This correlation matrix is called the *compound symmetric correlation matrix* in multivariate analysis (Katz 2006). It was also used by Joseph and Delaney (2007) for modeling some physical experiments. Using it in (6) leads to

$$\text{cor}(\epsilon(\mathbf{w}_1), \epsilon(\mathbf{w}_2)) = \exp \left\{ - \sum_{i=1}^I \phi_i (x_{i1} - x_{i2})^p - \sum_{j=1}^J \theta_j I[z_{j1} \neq z_{j2}] \right\}, \quad (7)$$

where $0 < p \leq 2$ and $\theta_j > 0, j = 1, \dots, J$. On a logarithmic scale, (7) uses the L_p distance for the quantitative factors and the 0-1 distance for the qualitative factors.

4.2 Multiplicative Correlation Functions

The following $\tau_{r,s}$ allows different pairs of z levels to have different correlations:

$$\tau_{r,s} = \exp\{-\theta_{r,s}\} = \exp\{-(\theta_r + \theta_s)I[r \neq s]\}, \quad (8)$$

where $\theta_r, \theta_s > 0$ determine the respective contributions of levels r and s to $\theta_{r,s}$ ($r \neq s$). We call (8) a multiplicative correlation function because $\tau_{r,s}$ is the product of $\exp\{-\theta_r\}$ and $\exp\{-\theta_s\}$ for $r \neq s$. It was used by McMillian, Sacks, Welch, and Gao (1999) to model correlations of unordered categorical variables using GPs. However, its applicability is limited by the multiplicative structure in (8), which is difficult to interpret and justify for many computer experiments. Moreover, as McMillian et al. (1999) pointed out, (8) is restricted for $m \geq 4$, with m parameters $(\theta_1, \dots, \theta_m)$, instead of $m(m-1)/2$ parameters for an unrestrictive correlation matrix. In particular, we observe that, by (8), for $m \geq 4$ and any four levels of z , say 1, 2, 3, and 4,

$$\tau_{1,2} \cdot \tau_{3,4} = \tau_{1,3} \cdot \tau_{2,4} = \tau_{1,4} \cdot \tau_{2,3} = \exp\{-(\theta_1 + \theta_2 + \theta_3 + \theta_4)\},$$

implying that it is impossible to independently specify or estimate the parameters $\tau_{1,2}$, $\tau_{3,4}$, $\tau_{1,3}$, $\tau_{2,4}$, $\tau_{1,4}$, and $\tau_{2,3}$.

4.3 Group Correlation Functions

Natural grouping among levels of a qualitative factor may occur in some computer experiments. For example, four types of structural materials in aircraft design (Fridlyander 2002) may be grouped as metals (aluminum and magnesium alloys) and composites (carbon fiber and fiber glass). We propose the *group correlation function* for such a factor. Suppose the m levels of z form K groups: g_1, \dots, g_K , where g_k includes b_k levels. For simplicity, we assume the correlation between any two levels in g_k is α_k ($0 < \alpha_k < 1$), and the correlation between any two levels in two groups is γ ($0 < \gamma < 1$). That is,

$$\mathbf{T} = \begin{pmatrix} A_1 & * & * \\ * & \ddots & * \\ * & * & A_K \end{pmatrix},$$

where A_k ($k = 1, \dots, K$) is a $b_k \times b_k$ matrix with unit diagonal elements and off-diagonal elements α_k , and all other elements of \mathbf{T} are γ . For \mathbf{T} to be a valid correlation matrix, further constraints need to be imposed on γ and α_k . For example, if z has three levels, with its first two levels forming one group and the third another group, the constraint is $\gamma < \sqrt{(1 + \alpha_1)/2}$.

4.4 Correlation Functions for Ordinal Qualitative Factors

Sections 4.1 to 4.3 focus on correlation functions for *unordered* qualitative factors. We now propose two methods for constructing correlation functions for *ordinal* qualitative factors (e.g., customer satisfaction with levels “unsatisfied,” “barely satisfied,” “nearly satisfied,” “satisfied,” and “very satisfied” in agent-based models used in marketing research). For convenience, denote the m levels of z by $1, \dots, m$ in an *increasing* order.

The first is called the *transformation method*, which transforms z to a quantitative factor v and then defines a correlation function for v . This method is flexible and conceptually simple. After selecting a strictly increasing continuous function $F(t)$ on $[0, 1]$, we transform level k of z to level v_k of v by solving the following equation:

$$F(v_k) = F(0) + (k - 1) \frac{F(1) - F(0)}{m - 1},$$

which has a unique solution, with $0 = v_1 < \dots < v_m = 1$. We then define $\tau_{r,s} = K(v_r, v_s)$, where $K(v_r, v_s)$ is a correlation function for v . Figure 1 gives three transformation functions F for z with six levels, based on the cumulative distribution function (CDF) of (a) uniform $[0, 1]$, (b) normal $(0.5, 0.25^2)$, and (c) lognormal $(-1.5, 0.75^2)$. Selecting an appropriate F may vary with applications and require subject-matter knowledge. (Nevertheless, performance in modeling/estimation may be insensitive to the choice of F .) We discuss below how to select F for two special types of z .

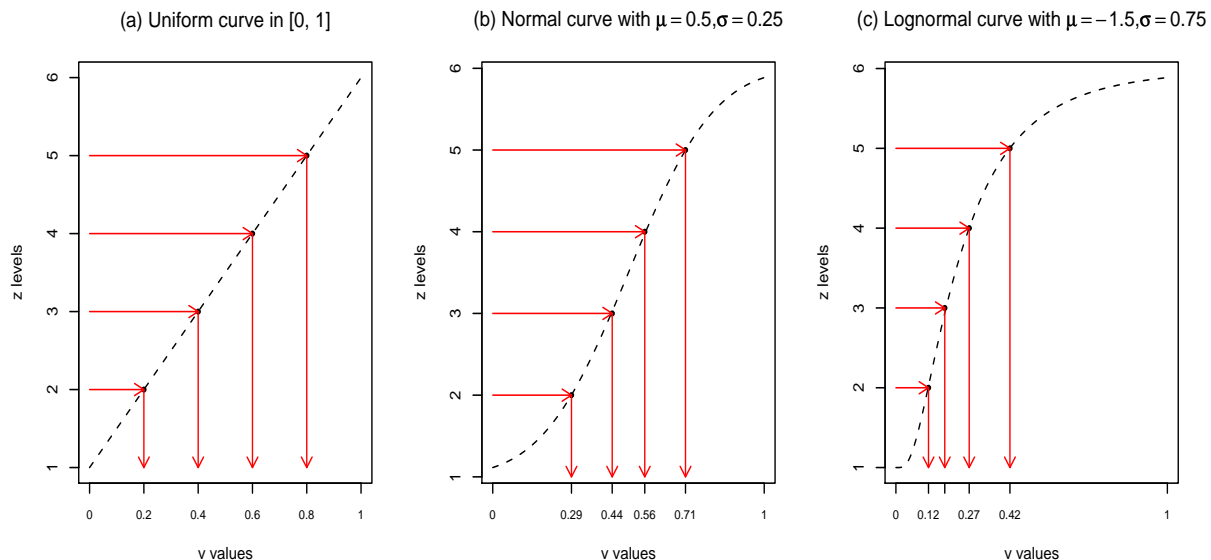


Figure 1: Three transformation functions for an ordinal qualitative factor with six levels.

We define z to be ordinal with an *increasing change rate* (ICR) if $\tau_{r,r+1}$ increases on $r \in \{1, \dots, m - 1\}$, and with a *decreasing change rate* (DCR) if $\tau_{r,r+1}$ decreases on $\{1, \dots, m - 1\}$.

For example, an ordinal qualitative factor with ICR is the education level (Less than high school, High school, Associate’s, Bachelor’s, Master’s, Professional, and Ph.D.’s). For an ordinal z with ICR, we suggest selecting a convex F (i.e., $F'' > 0$), and for z with DCR, a concave F (i.e., $F'' < 0$). For example, $F(t) = \exp\left(\frac{t}{\eta}\right) - \frac{t}{\gamma} - 1$ ($\gamma \geq \eta > 0$) is convex on $[0, 1]$, and the CDF of the Weibull distribution, $F(t) = 1 - \exp\left(-\left(\frac{t}{\eta}\right)^\beta\right)$, is concave on $[0, 1]$ (for $0 < \beta \leq 1$ and $\eta > 0$). (Both functions are strictly increasing on $[0, 1]$.)

An alternative to the transformation method assumes that

$$\tau_{r,s} = \gamma_{|r-s|}, \quad \text{where } \gamma_0 = 1, 0 < \gamma_k < 1, \text{ for } k = 1, \dots, m-1. \quad (9)$$

The matrix $\mathbf{T} = (\tau_{r,s})$ has a *Toeplitz form*; that is, its entries are constant along the diagonals parallel to the main diagonal (Golub and Van Loan 1996). For \mathbf{T} to be a valid correlation matrix, further constraints need to be imposed on $\gamma_1, \dots, \gamma_{m-1}$. For $m = 3$, the constraint is $\gamma_1 < \sqrt{(1 + \gamma_2)/2}$. A special case of (9), $\tau_{r,s} = \rho^{|r-s|}$ ($0 < \rho < 1$), always gives a valid correlation matrix $\mathbf{T} = (\tau_{r,s})$, because \mathbf{T} can be viewed as the correlation matrix of a first-order autoregressive process (Box and Jenkins 1976).

5 ESTIMATION AND PREDICTION

Suppose the data consist of n different input values, $D_w = (\mathbf{w}_1^0, \dots, \mathbf{w}_n^0)^t$, and the corresponding responses, $\mathbf{y} = (y_1, \dots, y_n)^t$. Consider model (3) with the correlation function in (6), with $p = 2$. The parameters to be estimated are $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^t$, σ^2 , $\boldsymbol{\phi} = (\phi_1, \dots, \phi_I)^t$, and $\mathbf{T} = \{\mathbf{T}_1, \dots, \mathbf{T}_J\}$. We use the maximum likelihood method for the estimation, and denote the resulting estimators by $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{T}}$. We shall also briefly discuss Bayesian methods in Section 5.6.

Here we mainly focus on the estimation for the general case of \mathbf{T} without assuming any special correlation structures, such as those discussed in Section 4, for the qualitative factors z_1, \dots, z_J . We shall also briefly discuss the estimation for models with restrictive correlation matrices in Section 5.4. For the general case of \mathbf{T} , the validity of (6) as a correlation function requires that all the \mathbf{T}_j 's be valid correlation matrices (i.e., PDUDEs). The problem of estimating a positive definite matrix occurs in many applications in statistics, including factor analysis (Bartholomew and Knott 1999) and Gaussian graphical models (Lauritzen 1996; Edwards 2000). The present application has two distinct features. First, our problem can be more challenging because it entails estimating *multiple* correlation matrices whereas in other applications, such as in a Gaussian graphical model, usually a *single* correlation matrix is involved. Second, sometimes one can run computer experiments using selected designs for input factor values. This flexibility is not shared by the observational studies to which factor analysis and Gaussian graphical models are usually applied. As will be discussed in Section 5.3, the use of “appropriate” experimental designs for input factors can significantly simplify the estimation procedure.

Standard methods used in statistics for maximizing a likelihood function involving a positive definite matrix work in the following manner. First, note that a matrix is positive definite if and only if all its leading principle minors are positive. These constraints then transfer to a series of nonlinear inequalities involving the elements of the matrix. Finally, an optimization problem is solved with the resulting nonlinear inequalities as the constraints and the elements of the matrix as the optimization variables. This “element-oriented” approach involves many complicated nonlinear inequalities and a huge number of optimization variables even when the dimension of the matrix is not very large, making it computationally infeasible. To better address the optimization problem with positive-definiteness constraints on the \mathbf{T}_j 's, we make use of the recently developed semi-definite programming technique in optimization. A brief introduction of semi-definite programming is given in Section 5.1. The estimation procedures are developed in Sections 5.2 to 5.4, and the prediction procedure is provided in Section 5.5.

5.1 Semi-definite Programming

Consider the optimization problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathbf{C} \bullet \mathbf{X} \\ \text{subject to} \quad & \mathbf{A}_i \bullet \mathbf{X} = \mathbf{b}_i, \quad i = 1, \dots, m, \\ & \mathbf{X} \succ 0 \ (\succeq 0), \end{aligned} \tag{10}$$

where C is an $n \times n$ real matrix, and the optimization variable is \mathbf{X} in the space of $n \times n$ real symmetric matrices. The inequalities $\mathbf{X} \succ 0$ and $\mathbf{X} \succeq 0$ mean that \mathbf{X} is positive definite and positive semi-definite, respectively. The problem (10) is referred to as a *semi-definite programming* (SP) in optimization (Vandenberghe and Boyd 1996; Wolkowicz, Saigal, and Vandenberghe 2000). The notation $\mathbf{C} \bullet \mathbf{X}$ represents the inner product of the matrices \mathbf{C} and \mathbf{X} :

$$\mathbf{C} \bullet \mathbf{X} = \sum_{i=1}^n \sum_{j=1}^n c_{ij} x_{ij},$$

where c_{ij} and x_{ij} are the (i, j) th entries of \mathbf{C} and \mathbf{X} , respectively. Equivalently, $\mathbf{C} \bullet \mathbf{X}$ can also be written as $\text{tr}(\mathbf{C}\mathbf{X})$. Throughout this article, “tr” stands for the trace of a square matrix. This type of optimization problem arises in many fields, including statistics, communication theory, and machine learning. The SP problem is a convex problem, which can be solved efficiently by *interior point algorithms* (Wolkowicz, Saigal, and Vandenberghe 2000). These algorithms compute the solution to (10) within a *cone* formed by positive definite matrices and can lead to significant computational savings, especially for large scale problems.

5.2 Estimation Procedures

The general case under consideration involves I quantitative factors x_1, \dots, x_I and J qualitative factors z_1, \dots, z_J , with no special correlation structures imposed on the z_j 's. Without loss of generality, the number of levels of z_j , denoted by m_j , is assumed to be three or higher. If a qualitative factor has two levels, it can be grouped with the quantitative factors in the estimation because there is no need to impose positive-definiteness conditions on it.

Up to an additive constant, the log-likelihood of \mathbf{y} is

$$-\frac{1}{2} \left[n \ln \sigma^2 + \ln |\mathbf{R}| + \frac{(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})}{\sigma^2} \right], \quad (11)$$

where $\mathbf{F} = (\mathbf{f}(\mathbf{w}_1^0), \dots, \mathbf{f}(\mathbf{w}_n^0))^t$ is an $n \times m$ matrix; \mathbf{R} is the correlation matrix, which depends on the correlation parameters $\boldsymbol{\phi}$ and \mathbf{T} , and its (i, j) th entry is $\text{cor}(\epsilon(\mathbf{w}_i^0), \epsilon(\mathbf{w}_j^0))$ defined in (6), with $p = 2$.

Given $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, and \mathbf{T} , the maximum likelihood estimate (MLE) of σ^2 is $\hat{\sigma}^2 = (1/n)(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$. Substituting $\hat{\sigma}^2$ into the log-likelihood (11), the problem is to numerically minimize

$$n \ln[(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})] + \ln |\mathbf{R}|, \quad (12)$$

which is a function of $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, \mathbf{T} , and the data. For easier presentation, for given $\hat{\boldsymbol{\beta}}$, let $\mathbf{e} = \mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}$ and $\mathbf{E} = \mathbf{e}\mathbf{e}^t$ throughout this article. Thus,

$$(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) = \text{tr}(\mathbf{e}^t \mathbf{R}^{-1} \mathbf{e}) = \text{tr}(\mathbf{e}\mathbf{e}^t \mathbf{R}^{-1}) = \text{tr}(\mathbf{E}\mathbf{R}^{-1}).$$

Then the problem in (12) can be solved by iterating between the following β -step and (ϕ, T) -step:

β -step: Given $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{T}}$, obtain $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}} = (\mathbf{F}^t [\mathbf{R}(\hat{\boldsymbol{\phi}}, \hat{\mathbf{T}})]^{-1} \mathbf{F})^{-1} \mathbf{F}^t [\mathbf{R}(\hat{\boldsymbol{\phi}}, \hat{\mathbf{T}})]^{-1} \mathbf{y}$.

(ϕ, T) -step: Given $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{T}}$ can be obtained as follows:

$$\begin{aligned} (\hat{\boldsymbol{\phi}}, \hat{\mathbf{T}}) &= \underset{(\boldsymbol{\phi}, \mathbf{T})}{\text{argmin}} && [n \ln(\text{tr}(\mathbf{E}\mathbf{R}^{-1})) + \ln |\mathbf{R}|] \\ &\text{subject to} && \phi_i \geq 0, \quad i = 1, \dots, I, \\ &&& \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ &&& \text{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J, \end{aligned}$$

where the optimization variables are $\boldsymbol{\phi}$ and \mathbf{T} . Throughout this article, ‘‘diag’’ stands for the diagonal elements of a square matrix and $\mathbf{1}$ stands for a column vector of 1's. Estimating $\boldsymbol{\phi}$ and \mathbf{T} can be carried out by iterating between the following ϕ -step and T -step:

ϕ -step: Given $\widehat{\mathbf{T}}$, $\widehat{\phi}$ is obtained as follows:

$$\begin{aligned} \widehat{\phi} &= \operatorname{argmin}_{\phi} [n \ln(\operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})) + \ln |\mathbf{R}|] \\ \text{subject to} \quad & \phi_i \geq 0, \quad i = 1, \dots, I. \end{aligned} \quad (13)$$

T -step: Given $\widehat{\phi}$, $\widehat{\mathbf{T}}$ is obtained as follows:

$$\begin{aligned} \widehat{\mathbf{T}} &= \operatorname{argmin}_{\mathbf{T}} [n \ln(\operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})) + \ln |\mathbf{R}|] \\ \text{subject to} \quad & \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ & \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J. \end{aligned} \quad (14)$$

In optimization, such an iterative algorithm is called *block coordinate descent* or *nonlinear Gaussian-Seidel* method (Bertsekas 1999). It is well known that this type of algorithm will converge under mild conditions. The optimization in (13) is a standard nonlinear problem, which can be solved by quasi-Newton algorithms. The major difficulty lies in the T -step because of the complex objective function and constraints involved. The details for implementing the T -step are given below. Let $f(\mathbf{T})$ denote the objective function in (14), that is,

$$f(\mathbf{T}) = n \ln[\operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})] + \ln |\mathbf{R}|.$$

For computational convenience, we need to linearize the optimization problem in (14) as follows:

$$\begin{aligned} \widehat{\mathbf{T}} &= \operatorname{argmin}_{\mathbf{T}} \left[f(\mathbf{T}_0) + \sum_{j=1}^J \left(\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j} \bullet \mathbf{T}_j \right) \right] \\ \text{subject to} \quad & \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ & \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J, \end{aligned} \quad (15)$$

where $\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j}$ is the partial derivative of $f(\mathbf{T})$ with respect to \mathbf{T}_j , evaluated at some given value of \mathbf{T} , $\mathbf{T}_0 = \{\mathbf{T}_{0,1}, \dots, \mathbf{T}_{0,J}\}$. That is,

$$\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j} = \left[\frac{n}{\operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})} \frac{\partial \operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \mathbf{T}_j} + \frac{1}{|\mathbf{R}|} \frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j} \right] \Big|_{\mathbf{T}=\mathbf{T}_0}.$$

The formulas for $\frac{\partial \operatorname{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \mathbf{T}_j}$ and $\frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j}$ are given in the Appendix. Such a linear approximation has been shown to be reasonable and is widely used to approximate SP problems with nonlinear objective function constraints (Wolkowicz, Saigal, and Vandenberghe 2000). We can use the solution $\widehat{\mathbf{T}}$ to (15) to replace \mathbf{T}_0 and solve (15) again. This process can be repeated multiple times until the solutions to (15) converge. Now define the following block diagonal matrices

$$\mathbf{W} = \operatorname{bkdiag}(\mathbf{T}_1, \dots, \mathbf{T}_J), \quad \text{and} \quad \mathbf{C} = \operatorname{bkdiag} \left(\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_1}, \dots, \frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_J} \right).$$

Note that $\mathbf{W} \succ 0$ if and only if $\mathbf{T}_1 \succ 0, \dots, \mathbf{T}_J \succ 0$. Then, the optimization problem in (15) can be recast as the following SP problem:

$$\begin{aligned} \widehat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W}} \quad & \mathbf{C} \bullet \mathbf{W} \\ \text{subject to} \quad & \mathbf{W} \succ 0, \\ & \operatorname{diag}(\mathbf{W}) = \mathbf{1}. \end{aligned}$$

In summary, the above algorithm consists of an outer loop (the β -step and (ϕ, T) -step) and an inner loop (the ϕ -step and T -step). We should repeat the outer loop M times, and for each iteration in the outer loop repeat the inner loop N times until convergence.

When $\beta^t \mathbf{f}(\mathbf{w})$ in (3) is not very simple (e.g., as an additive or interaction model of \mathbf{x} and \mathbf{z}), it may be desirable to consider an alternative algorithm for the estimation. The basic idea is to iterate between a *regression fitting* and a *correlation fitting* as follows:

Regression fitting: Given $\widehat{\phi}$ and $\widehat{\mathbf{T}}$, obtain $\widehat{\beta}$ and $\widehat{\sigma}^2$ by

$$\widehat{\beta} = (\mathbf{F}^t [\mathbf{R}(\widehat{\phi}, \widehat{\mathbf{T}})]^{-1} \mathbf{F})^{-1} \mathbf{F}^t [\mathbf{R}(\widehat{\phi}, \widehat{\mathbf{T}})]^{-1} \mathbf{y} \quad \text{and} \quad \widehat{\sigma}^2 = (1/n) (\mathbf{y} - \mathbf{F} \widehat{\beta})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F} \widehat{\beta}).$$

Correlation fitting: Given $\widehat{\beta}$ and $\widehat{\sigma}^2$, let $u_i = (y_i - \widehat{\beta}^t \mathbf{f}(\mathbf{w}_i)) / \widehat{\sigma}$, for $i = 1, \dots, n$. Then fit the proposed GP model with mean zero, variance one, and correlation matrix \mathbf{R} to the data $\mathbf{U} = (u_1, \dots, u_n)^t$, and estimate ϕ and \mathbf{T} using the maximum likelihood method.

Up to an additive constant, the log-likelihood of \mathbf{U} is

$$(-1/2) [\ln |\mathbf{R}| + \mathbf{U}^t \mathbf{R}^{-1} \mathbf{U}] = (-1/2) [\ln |\mathbf{R}| + \operatorname{tr}(\mathbf{G} \mathbf{R}^{-1})], \quad \text{where} \quad \mathbf{G} = \mathbf{U} \mathbf{U}^t.$$

Thus, similar to the (ϕ, T) -step in the previous algorithm, the correlation fitting is done by iterating between the following ϕ -step and T -step:

ϕ -step: Given $\widehat{\mathbf{T}}$, $\widehat{\phi}$ is obtained as follows:

$$\begin{aligned} \widehat{\phi} = \operatorname{argmin}_{\phi} \quad & [\operatorname{tr}(\mathbf{G} \mathbf{R}^{-1}) + \ln |\mathbf{R}|] \\ \text{subject to} \quad & \phi_i \geq 0, \quad i = 1, \dots, I. \end{aligned}$$

T -step: Given $\widehat{\phi}$, $\widehat{\mathbf{T}}$ is obtained as follows:

$$\begin{aligned} \widehat{\mathbf{T}} = \operatorname{argmin}_{\mathbf{T}} \quad & [\operatorname{tr}(\mathbf{G} \mathbf{R}^{-1}) + \ln |\mathbf{R}|] \\ \text{subject to} \quad & \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ & \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J. \end{aligned} \tag{16}$$

Again, the optimization problem in (16) needs to be linearized as follows:

$$\begin{aligned} \hat{\mathbf{T}} = \operatorname{argmin}_{\mathbf{T}} & \left[f(\mathbf{T}_0) + \sum_{j=1}^J \left(\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j} \bullet \mathbf{T}_j \right) \right] \\ \text{subject to} & \quad \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ & \quad \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J, \end{aligned} \quad (17)$$

with

$$\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j} = \left[\frac{\partial \operatorname{tr}(\mathbf{G}\mathbf{R}^{-1})}{\partial \mathbf{T}_j} + \frac{1}{|\mathbf{R}|} \frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j} \right] \Big|_{\mathbf{T}=\mathbf{T}_0}.$$

Other aspects of this T -step are the same as those in the previous algorithm.

5.3 Simplification of Estimation With Structured D_w

Some structures of D_w can significantly simplify the computation in the T -step. This convenience is only possible for the present experimental situation but not for an observational study. Consider first the simple case involving *one* qualitative factor z_1 with more than two levels, denoted by $1, \dots, m_1$. Assume D_w is a *cross array* (Wu and Hamada 2000) of D_x and D_z , where D_x is a $b \times I$ design matrix for the quantitative factors \mathbf{x} , and $D_z = (1, \dots, m_1)^t$ is an $m_1 \times 1$ design matrix for the qualitative factor z_1 . Consequently, D_w consists of all level combinations between those in D_x and those in D_z . Hence D_w has $n = bm_1$ rows (runs). As shown in the following proposition, this cross-array structure of D_w simplifies the optimization problem in (14) and also makes it free of $\hat{\phi}$. Consequently, estimating ϕ and \mathbf{T}_1 can be done *separately* by carrying out a simplified T -step and then the ϕ -step. This is much simpler than the general estimation procedure, which iterates between the ϕ -step and the T -step.

Let $\mathbf{H} = (h_{j_1 j_2})$ denote a $b \times b$ matrix with its (j_1, j_2) th entry given as

$$h_{j_1 j_2} = \exp \left\{ - \sum_{i=1}^I \phi_i (x_{ij_1} - x_{ij_2})^2 \right\}.$$

With the above assumption on experimental design, the optimization problem in (14) can be simplified as follows:

Proposition 1. *Suppose D_w is a cross array of D_x and D_z . Then, the problem in (14) is equivalent to*

$$\begin{aligned} \hat{\mathbf{T}}_1 = \operatorname{argmin}_{\mathbf{T}_1} & \quad (m_1 \ln[\operatorname{tr}(\mathbf{T}_1^{-1})] + \ln |\mathbf{T}_1|) \\ \text{subject to} & \quad \mathbf{T}_1 \succ 0, \\ & \quad \operatorname{diag}(\mathbf{T}_1) = \mathbf{1}. \end{aligned} \quad (18)$$

The proof of this proposition is given in the Appendix. With the proposition, the linear approx-

imation (15) becomes

$$\begin{aligned} \widehat{\mathbf{T}}_1 &= \operatorname{argmin}_{\mathbf{T}_1} [f(\mathbf{T}_{1,0}) + \nabla_{\mathbf{T}_1} f(\mathbf{T}_{1,0}) \bullet \mathbf{T}_1] \\ &\text{subject to} \quad \mathbf{T}_1 \succ 0, \\ &\quad \operatorname{diag}(\mathbf{T}_1) = \mathbf{1}, \end{aligned} \tag{19}$$

where

$$f(\mathbf{T}_{1,0}) = m_1 \ln[\operatorname{tr}(\mathbf{T}_{1,0}^{-1})] + \ln |\mathbf{T}_{1,0}|,$$

and (from Dattorro 2005, App. D.2.3 and D.2.4)

$$\nabla_{\mathbf{T}_1} f(\mathbf{T}_{1,0}) = -\frac{m_1}{\operatorname{tr}(\mathbf{T}_{1,0}^{-1})} \mathbf{T}_{1,0}^{-2} + \mathbf{T}_{1,0}^{-1}.$$

These expressions significantly simplify the optimization problem in (19). This is a SP described in Section 5.1 and can be solved by efficient interior point algorithms.

We now consider the general case with J qualitative factors z_1, \dots, z_J . Assume D_w is a cross array of D_x , the $b \times I$ design matrix for \mathbf{x} , and D_z , the $q \times J$ design matrix for \mathbf{z} . Hence D_w has $n = bq$ rows (runs). Let $\mathbf{z}_1^0, \dots, \mathbf{z}_q^0$ denote the q input values for \mathbf{z} , and \mathbf{T}^* be a $q \times q$ matrix with its (r, s) th entry given as

$$t_{r,s} = \prod_{j=1}^J \tau_{j, z_{jr}^0, z_{js}^0}.$$

Using the argument for establishing Proposition 1, we have the following result:

Proposition 2. *Suppose D_w is a cross array of D_x and D_z , where D_x is a $b \times I$ design matrix for \mathbf{x} and D_z is a $q \times J$ design matrix for \mathbf{z} . Then, the problem in (14) is equivalent to*

$$\begin{aligned} \widehat{\mathbf{T}} &= \operatorname{argmin}_{\mathbf{T}} (q \ln[\operatorname{tr}((\mathbf{T}^*)^{-1})] + \ln |\mathbf{T}^*|) \\ &\text{subject to} \quad \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ &\quad \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J. \end{aligned} \tag{20}$$

Again, the cross-array structure of D_w reduces the estimation of ϕ and \mathbf{T} to separately carrying out the simplified T -step and then the ϕ -step. This is much simpler than the general estimation procedure that iterates between the ϕ -step and the T -step.

If D_z also has some cross-array structure, the optimization problem in (20) can be further simplified. Suppose the qualitative factors \mathbf{z} are grouped into $d \geq 2$ disjoint sets, $\{z_j : j \in A_k\}$, for $k = 1, \dots, d$, where $\bigcup_{j=1}^d A_j = \{1, \dots, J\}$ and the size of A_k is $J_k \geq 1$. Suppose further that D_z is a cross array of D_1, \dots, D_d , where D_k is a $q_k \times J_k$ design matrix for the factors in $\{z_j : j \in A_k\}$. (However, D_k is not required to be a cross array among its constituent factors $z_j, j \in A_k$.) Thus, $\prod_{k=1}^d q_k = q$ and $\sum_{k=1}^d J_k = J$. Let $\mathbf{T}_{A_k} = \{\mathbf{T}_j : j \in A_k\}$ and \mathbf{T}_k^* be a $q_k \times q_k$

matrix with its (r, s) th entry given as

$$t_{r,s}^{(k)} = \prod_{j \in A_k} \tau_{j, z_{jr}^0, z_{js}^0}.$$

Again, using the argument for establishing Proposition 1, we have the following result:

Proposition 3. *Suppose D_z in Proposition 2 is a cross array of D_1, \dots, D_d , where D_k is a $q_k \times J_k$ design matrix for the factors in $\{z_j : j \in A_k\}$. Then, solving the problem (20) is equivalent to solving the following d simpler problems separately:*

$$(P_{A_k}) : \quad \begin{aligned} \widehat{\mathbf{T}}_{A_k} &= \operatorname{argmin}_{\mathbf{T}_{A_k}} (q_k \ln[\operatorname{tr}((\mathbf{T}_k^*)^{-1})] + \ln |\mathbf{T}_k^*|) \\ &\text{subject to} \quad \mathbf{T}_j \succ 0, \quad j \in A_k, \\ &\quad \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j \in A_k, \end{aligned} \quad (21)$$

for $k = 1, \dots, d$. In particular, if $d = J$, $A_k = \{k\}$, and $q_k = m_k$, then $\mathbf{T}_{A_k} = \mathbf{T}_k$ and $\mathbf{T}_k^* = \mathbf{T}_k$. Then (21) simplifies to

$$(P_k) : \quad \begin{aligned} \widehat{\mathbf{T}}_k &= \operatorname{argmin}_{\mathbf{T}_k} (m_k \ln[\operatorname{tr}((\mathbf{T}_k)^{-1})] + \ln |\mathbf{T}_k|) \\ &\text{subject to} \quad \mathbf{T}_k \succ 0, \\ &\quad \operatorname{diag}(\mathbf{T}_k) = \mathbf{1}. \end{aligned}$$

The method proposed to tackle the problem in (14) can be used to solve the problems in the above two propositions.

For the alternative algorithm in Section 5.2, similar to Propositions 1 and 2, we have the following results. (No counterpart of Proposition 3 holds here.)

Proposition 4. *Suppose D_w is a cross array of D_x and D_z , where D_x is a $b \times I$ design matrix for \mathbf{x} and $D_z = (1, \dots, m_1)^t$ is an $m_1 \times 1$ design matrix for the qualitative factor z_1 . Then, the problem in (16) is equivalent to*

$$\begin{aligned} \widehat{\mathbf{T}}_1 &= \operatorname{argmin}_{\mathbf{T}_1} (\operatorname{tr}(\mathbf{GH}^{-1})\operatorname{tr}(\mathbf{T}_1^{-1}) + b \ln |\mathbf{T}_1|) \\ &\text{subject to} \quad \mathbf{T}_1 \succ 0, \\ &\quad \operatorname{diag}(\mathbf{T}_1) = \mathbf{1}. \end{aligned}$$

With this proposition, the optimization problem (17) becomes

$$\begin{aligned} \widehat{\mathbf{T}}_1 &= \operatorname{argmin}_{\mathbf{T}_1} [f(\mathbf{T}_{1,0}) + \nabla_{\mathbf{T}_1} f(\mathbf{T}_{1,0}) \bullet \mathbf{T}_1] \\ &\text{subject to} \quad \mathbf{T}_1 \succ 0, \\ &\quad \operatorname{diag}(\mathbf{T}_1) = \mathbf{1}, \end{aligned}$$

where $f(\mathbf{T}_{1,0}) = \operatorname{tr}(\mathbf{GH}^{-1})\operatorname{tr}(\mathbf{T}_{1,0}^{-1}) + b \ln |\mathbf{T}_{1,0}|$, and $\nabla_{\mathbf{T}_1} f(\mathbf{T}_{1,0}) = -\operatorname{tr}(\mathbf{GH}^{-1})\mathbf{T}_{1,0}^{-2} + b\mathbf{T}_{1,0}^{-1}$.

Proposition 5. *Suppose D_w is a cross array of D_x and D_z , where D_x is a $b \times I$ design matrix for \mathbf{x} and D_z is a $q \times J$ design matrix for \mathbf{z} . Then, the problem in (16) is equivalent to*

$$\begin{aligned} \widehat{\mathbf{T}} &= \operatorname{argmin}_{\mathbf{T}} \quad (tr(\mathbf{G}\mathbf{H}^{-1})tr((\mathbf{T}^*)^{-1}) + b \ln |\mathbf{T}^*|) \\ &\text{subject to} \quad \mathbf{T}_j \succ 0, \quad j = 1, \dots, J, \\ &\quad \quad \quad \operatorname{diag}(\mathbf{T}_j) = \mathbf{1}, \quad j = 1, \dots, J. \end{aligned}$$

5.4 Estimation When Restrictive Correlation Matrices are Used for Qualitative Factors

When restrictive correlation matrices (as discussed in Section 4) are used for the qualitative factors \mathbf{z} , the estimation of the unknown parameters is easier and becomes that of ϕ and $\psi = (\psi_1^t, \dots, \psi_J^t)^t$, where ψ_j ($j = 1, \dots, J$) is a column vector of the parameters involved in the restrictive correlation matrix \mathbf{T}_j . Let C_j denote the set of values of ψ_j such that \mathbf{T}_j is a valid correlation matrix. It follows from the log-likelihood (11) that, for given ϕ and \mathbf{T} , $\widehat{\beta} = (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^t \mathbf{R}^{-1} \mathbf{y}$, and $\widehat{\sigma}^2 = (1/n)(\mathbf{y} - \mathbf{F}\widehat{\beta})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\widehat{\beta})$. Substituting $\widehat{\beta}$ and $\widehat{\sigma}^2$ into (11), it simplifies to (up to a negative constant) $n \ln(\widehat{\sigma}^2) + \ln |\mathbf{R}|$, where $\widehat{\sigma}^2$ and \mathbf{R} depend on ϕ , ψ , and the data. Then $\widehat{\phi}$ and $\widehat{\psi}$ can be obtained as follows:

$$\begin{aligned} (\widehat{\phi}, \widehat{\psi}) &= \operatorname{argmin}_{(\phi, \psi)} \quad [n \ln(\widehat{\sigma}^2) + \ln |\mathbf{R}|] \\ &\text{subject to} \quad \phi_i \geq 0, \quad i = 1, \dots, I, \\ &\quad \quad \quad \psi_j \in C_j, \quad j = 1, \dots, J. \end{aligned}$$

5.5 Prediction

The fitted GP model can be used in predicting the response value y at any untried point in the design space. Similar to equation (7) of Sacks, Welch, Mitchell, and Wynn (1989), the empirical best linear unbiased predictor (BLUP) of y at the point \mathbf{w}_0 is

$$\widehat{y}(\mathbf{w}_0) = \widehat{\beta}^t \mathbf{f}(\mathbf{w}_0) + \widehat{\mathbf{r}}_0^t \widehat{\mathbf{R}}^{-1} (\mathbf{y} - \mathbf{F}\widehat{\beta}), \quad (22)$$

where $\widehat{\beta}$ is the estimate of β , $\widehat{\mathbf{R}}$ is the estimated correlation matrix of \mathbf{y} and

$$\widehat{\mathbf{r}}_0 = (\widehat{\operatorname{cor}}(y(\mathbf{w}_0), y(\mathbf{w}_1^0)), \dots, \widehat{\operatorname{cor}}(y(\mathbf{w}_0), y(\mathbf{w}_n^0)))^t.$$

The empirical BLUP in (22) enjoys such nice properties as the smooth interpolation of all the observed data points, similar to the empirical BLUP for the GP model with quantitative factors in Section 2.1.

To visualize the features of the function $y(\mathbf{w})$ via the predictor $\widehat{y}(\mathbf{w})$, we can use the approach of Welch et al. (1992) and plot the estimated main effects and interactions. In calculating these

effects using their definitions, evaluation of an integral for a qualitative factor simplifies to averaging over the predicted response values for all the levels of that factor (see Sec. 1 of their article for details).

5.6 Bayesian Methods

As an alternative, Bayesian methods can also be used for the proposed GP model but will require more computational effort. It is beyond the scope of this article to provide details for these methods. Only a brief description is given below. The model in (3) can be formulated as a *hierarchical Bayesian model* (Gelman, Carlin, Stern, and Rubin 2004). As often assumed in Bayesian statistics, the priors for the parameters β , σ^2 , ϕ , and \mathbf{T} take the form of $p(\beta, \sigma^2, \phi, \mathbf{T}) = p(\beta, \sigma^2)p(\phi, \mathbf{T}) = p(\beta|\sigma^2)p(\sigma^2)p(\phi)p(\mathbf{T})$. One possible choice for the priors is:

$$\begin{aligned} p(\sigma^2) &\sim IG(\alpha, \gamma), \\ p(\beta|\sigma^2) &\sim N(\mathbf{u}, v\sigma^2\mathbf{I}_m), \\ \phi_i &\sim G(a, b), \text{ for } i = 1, \dots, I, \\ \mathbf{T}_j &\sim \text{Inv-Wishart}_{\eta_j}(\Lambda_j), \text{ for } j = 1, \dots, J. \end{aligned}$$

Here $IG(\alpha, \gamma)$ denotes the inverse gamma distribution and its probability density function (pdf) is

$$p(z, \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} \exp\left\{-\frac{\gamma}{z}\right\}, \text{ for } z > 0 (\alpha > 0, \gamma > 0);$$

$G(a, b)$ is the gamma distribution and its pdf is

$$p(z, a, b) = \frac{b^a}{\Gamma(a)} z^{a-1} e^{-bz}, \text{ } z > 0 (a > 0, b > 0);$$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, and \mathbf{I}_m is the $m \times m$ identity matrix; $\text{Inv-Wishart}_{\eta_0}(\Lambda_0)$ is the inverse-Wishart distribution and its pdf is

$$p(\mathbf{Z}, \eta_0, \Lambda_0) = \frac{|\Lambda_0|^{\eta_0/2}}{2^{\eta_0 k/2} \pi^{k(k-1)/4} \prod_{i=1}^k \Gamma((\eta_0 + 1 - i)/2)} \cdot |\mathbf{Z}|^{-(\eta_0+k+1)/2} \exp(-\text{tr}(\Lambda_0 \mathbf{Z}^{-1})/2),$$

where \mathbf{Z} is a $k \times k$ matrix.

An empirical Bayes approach can be used for predicting the response y at a new point \mathbf{w}_0 . It consists of two steps. In the first step, the correlation parameters ϕ and \mathbf{T} are fixed at their posterior modes. In the second step, prediction is made conditionally on the estimated correlation parameters. For the first step, note that the posterior $p(\phi, \mathbf{T}|\mathbf{y})$ can be obtained from

$$p(\phi, \mathbf{T}|\mathbf{y}) = \int p(\beta, \sigma^2, \phi, \mathbf{T}|\mathbf{y}) d\beta d\sigma^2, \quad (23)$$

where the integrand is determined by $p(\beta, \sigma^2, \phi, \mathbf{T}|\mathbf{y}) \propto p(\mathbf{y}|\beta, \sigma^2, \mathbf{T}, \phi)p(\beta, \sigma^2, \mathbf{T}, \phi)$. The

posterior modes of ϕ and \mathbf{T} , denoted by $\hat{\phi}$ and $\hat{\mathbf{T}}$, are given by an optimal solution to the optimization problem $\max_{\phi, \mathbf{T}} p(\phi, \mathbf{T} | \mathbf{y})$. If the integral in (23) does not have an analytic form, iterative methods such as the EM algorithm (Dempster, Laird, and Rubin 1977) and stochastic programming method (Ruszczynski and Shapiro 2003) need to be used to solve this optimization problem. For the second step, it can be shown (Santner, Williams, and Notz 2003) that, with assuming $p(\sigma^2) \sim IG(\alpha, \gamma)$ and $p(\beta | \sigma^2) \sim N(\mathbf{u}, v\sigma^2 \mathbf{I}_m)$, the conditional distribution of y at \mathbf{w}_0 , given the observed \mathbf{y} and $\hat{\phi}$ and $\hat{\mathbf{T}}$, is the non-central t distribution $T_1(n + \nu_0, \mu_1, \sigma_1^2)$, where

$$\begin{aligned} \mu_1 &= \mathbf{f}_0^t \boldsymbol{\mu}_{\beta|n} + \mathbf{r}_0^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F} \boldsymbol{\mu}_{\beta|n}), \\ \boldsymbol{\mu}_{\beta|n} &= (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{F} + v^{-1} \mathbf{I}_m)^{-1} (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{y} + v^{-1} \mathbf{u}), \\ \hat{\boldsymbol{\beta}} &= (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^t \mathbf{R}^{-1} \mathbf{y}, \\ \sigma_1^2 &= \frac{Q_1^2}{\nu_1} \left\{ 1 - (\mathbf{f}_0^t, \mathbf{r}_0^t) \begin{bmatrix} -v^{-1} \mathbf{I}_m & \mathbf{F}^t \\ \mathbf{F} & \mathbf{R} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{f}_0 \\ \mathbf{r}_0 \end{pmatrix} \right\}, \\ Q_1^2 &= c_0 + \mathbf{y}^t [\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{F} (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^t \mathbf{R}^{-1}] \mathbf{y} \\ &\quad + (\mathbf{u} - \hat{\boldsymbol{\beta}})^t [v \mathbf{I}_m + (\mathbf{F}^t \mathbf{R}^{-1} \mathbf{F})^{-1}]^{-1} (\mathbf{u} - \hat{\boldsymbol{\beta}}), \end{aligned}$$

$\nu_0 = 2a$, $\nu_1 = n + 2a$, $c_0 = (b/a)^{1/2}$, $\mathbf{f}_0 = \mathbf{f}(\mathbf{w}_0)$, $\mathbf{r}_0 = (\text{cor}(y(\mathbf{w}_0), y(\mathbf{w}_1^0)), \dots, \text{cor}(y(\mathbf{w}_0), y(\mathbf{w}_n^0)))^t$, \mathbf{R} is the correlation matrix with entry $\text{cor}(y(\mathbf{w}_i^0), y(\mathbf{w}_j^0))$ for $i, j = 1, \dots, n$, and $\mathbf{F} = (\mathbf{f}(\mathbf{w}_1^0), \dots, \mathbf{f}(\mathbf{w}_n^0))^t$.

To accommodate the uncertainty in ϕ and \mathbf{T} , the following fully Bayesian approach can be used. In this approach, prediction of y at \mathbf{w}_0 is based on the posterior distribution

$$p(y(\mathbf{w}_0) | \mathbf{y}) = \int p[y(\mathbf{w}_0), \phi, \mathbf{T} | \mathbf{y}] d\phi d\mathbf{T} = p[y(\mathbf{w}_0) | \mathbf{y}, \phi, \mathbf{T}] p(\phi, \mathbf{T} | \mathbf{y}) d\phi d\mathbf{T}. \quad (24)$$

Here $p(\phi, \mathbf{T} | \mathbf{y})$ is given in (23). The integration in (24) can be computationally prohibitive. For six quantitative factors and four qualitative factors, each with six levels, ϕ would be 12-dimensional (with the exponential correlation function in (6)) and \mathbf{T} would be four 6×6 matrices. Advanced Markov chain Monte Carlo methods (Liu 2001) need to be used to mitigate this difficulty.

6 AN EXAMPLE INVOLVING A KNOWN FUNCTION

Here, we consider an experiment involving one qualitative factor, z_1 , and one quantitative factor, x_1 , with the following function:

$$y = \begin{cases} \exp(1.4x_1) \cos(7\pi x_1/2) & \text{if } z_1 = 1, \\ \exp(3x_1) \cos(7\pi x_1/2) & \text{if } z_1 = 2. \end{cases}$$

Figure 2 depicts the two curves of the function values with $z_1 = 1$ and 2, respectively. The overall similarity of the curves suggests that the independent analysis would be insufficient for this example and the integrated analysis can better exploit the common information in the two curves and is expected to perform better.

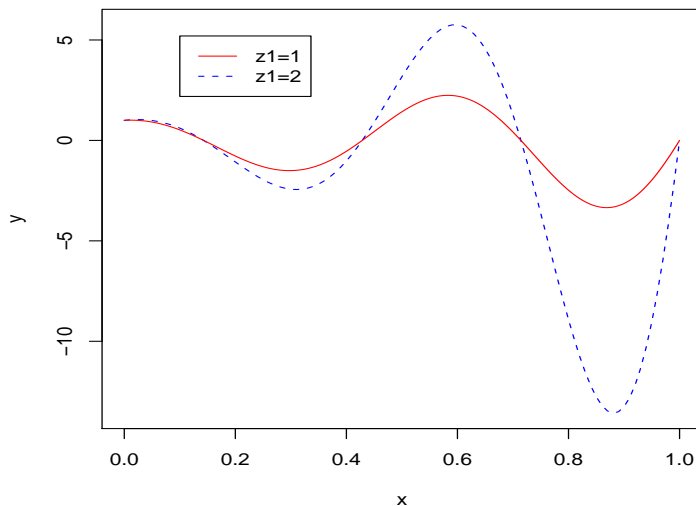


Figure 2: Two curves of the function values with $z_1 = 1$ and 2.

Table 1 lists the training data used for model building, including two 6-run Latin hypercube samples of x_1 , one for $z_1 = 1$ and another for $z_1 = 2$. For comparison, the data are analyzed using both methods. The independent analysis fits two separate GP models with means μ_1 and μ_2 , one for $z_1 = 1$ and another for $z_1 = 2$, using the correlation function (2) with $I = 1$ and $p = 2$; the estimated parameters are given in Table 2. The integrated analysis fits a GP model with mean μ that incorporates both x_1 and z_1 , using the correlation function (7) with $I = J = 1$ and $p = 2$; the estimated parameters are given in Table 3. (Because z_1 has two levels, it is automatically of isotropic nature, and thus (7) is used.) Note that $\hat{\phi} = 115.94$ in Table 2 for the GP model with $z_1 = 1$ is so large that its prediction will be rugged, potentially producing inaccurate results for the independent analysis; while $\hat{\phi} = 27.48$ and $\hat{\theta}_1 = 20.00$ in Table 3 are much smaller and should lead to better prediction results for the integrated analysis.

$z_1 = 1$	x_1	0.1232	0.2969	0.4999	0.6179	0.7614	0.9950
	y	0.2548	-1.5039	1.4222	2.0718	-1.4378	-0.2213
$z_1 = 2$	x_1	0.1136	0.3270	0.3433	0.6119	0.7778	0.9431
	y	0.4446	-2.3970	-2.2578	5.6589	-6.631	-9.9167

Table 1: The training data for the example involving a known function

Next we assess the prediction accuracy of the two methods. The testing data consists of 40 data points generated as follows. For $z_1 = 1$ and 2, x_1 takes 20 equally-spaced values 0.025, 0.075, \dots , 0.975 in $[0, 1]$. The root mean squared errors (RMSEs) for the two prediction

	$\hat{\phi}_1$	$\hat{\sigma}^2$	$\hat{\mu}$
$z_1 = 1$	115.94	1.73	-0.002
$z_1 = 2$	25.65	30.16	-2.09

Table 2: Estimated parameters of the GP models in the independent analysis

$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}^2$	$\hat{\mu}$
27.48	20.00	16.76	-1.07

Table 3: Estimated parameters of the GP model in the integrated analysis

methods are calculated. The RMSE for the integrated analysis is 1.03, which is 15% smaller than the RMSE (1.21) for the independent analysis. Clearly, the average prediction accuracy of the integrated analysis is much better than that of the independent analysis. For the integrated analysis, we also fit a GP with the process mean μ taking two different values for $z_1 = 1$ and 2, and the average prediction accuracy turned out to be nearly the same as that of the GP model with a constant mean for $z_1 = 1$ and 2.

7 A DATA-CENTER COMPUTER EXPERIMENT

In this section, the proposed method is illustrated using a data-center computer experiment from the IT industry. With the increasing need for storing, manipulating, and managing data sets, data centers are widely used to provide application services or management for various data processing, such as web hosting internet, intranet, telecommunication, and information technology. Figure 3 shows a schematic layout of an internet data center using Sun Microsystems (Lawrence Berkeley National Laboratory 2002). Driven by advances in hardware and data-storage techniques, data centers now can be very large, sprawling over thousands of square feet.

In designing and running a reliable data center, it is essential to maintain the system operating environment at a temperature within a functional range. Data-center facilities are extremely energy intensive, with many computer equipments constantly generating heat. Monitoring and studying the temperature of a data center is a difficult task, because it is largely unknown how different configurations affect the thermal distribution of the data center. The physical thermal process is complex, depending on many factors, and detailed temperatures at different locations cannot be actually measured. Computer experiment, built on computational fluid-dynamics (CFD) models, implemented in professional software like Flotherm (Flometrics 2005) and FLU-ENT (Fluent 1998), is often used as a proxy to study the air movement and thermal distribution of a data center. More details for the engineering background of data centers can be found in Schmidt (2003) and Schmidt, Cruz, and Iyengar (2005).

The experiment considered in this section models an air-cooled cabinet, implemented in Flotherm (Flometrics 2005), for predicting the airflow and heat transfer in the electronic equip-

ments. Each run in this experiment takes several days to complete. This example has eight configuration variables. Five of them are quantitative factors, denoted by x_1 , x_2 , x_3 , x_4 , and x_5 . The numbers of levels for these quantitative factors are 3, 5, 2, 3, and 3, respectively. The remaining three variables are 2-, 4- and 3-level qualitative factors, denoted by z_1 , z_2 , and z_3 . The response of interest, denoted by y , is the temperature at one selected location of the system.

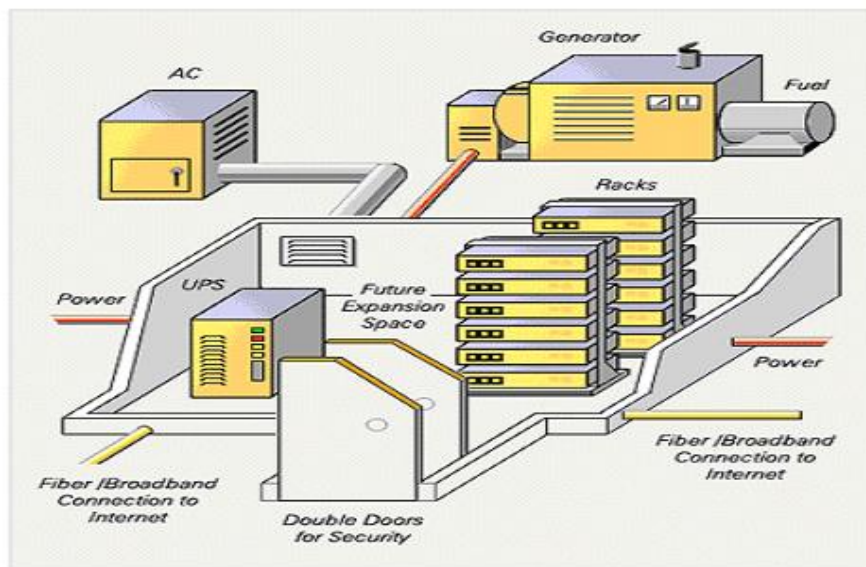


Figure 3: Schematic layout of an internet data center (Sun Microsystems) (Lawrence Berkeley National Laboratory 2002).

The five quantitative factors are of distinct scales, and their values are standardized first. The standardization of each variable is carried out by subtracting its lower design bound from its values, and then dividing the results by its design range. All results and plots given hereafter are associated with the standardized variables, which take values in $[0, 1]$. The original experiment has 73 observations. Six of them are removed due to unsuccessful tuning and convergence checking of the CFD algorithms after confirming from the data center scientists. The subsequent analysis uses the remaining 67 observations. There are 24 level combinations for the three qualitative factors. Hence, on average, each of these combinations has fewer than three observations, making the independent analysis infeasible. The data will be analyzed using the integrated analysis.

For the integrated analysis, it is found reasonable to use the following function for $\beta^t \mathbf{f}(\mathbf{w})$:

$$\eta + \sum_{i=1}^5 \beta_i x_i + \alpha_{12} I\{z_1 = 2\} + \sum_{j=2}^4 \alpha_{2j} I\{z_2 = j\} + \sum_{j=2}^3 \alpha_{3j} I\{z_3 = j\}.$$

Here the base-line constraints (using the first levels) are imposed on z_1 , z_2 , and z_3 to get an identifiable model (Wu and Hamada 2000, Sec. 2.3), and β is

$$\beta = (\eta, \beta_1, \dots, \beta_5, \alpha_{12}, \alpha_{22}, \alpha_{23}, \alpha_{24}, \alpha_{32}, \alpha_{33})^t.$$

The major difficulty with the model fitting is to estimate the correlation matrices for z_2 and z_3 . The estimation is carried out using the iterative procedure in Section 5, implemented in *Matlab* (The MathWorks 2006) and making use of a semi-definite programming package *CVX* (Grant, Boyd, and Ye 2006). Note that this data set does not have a cross-array structure and the GP model has a nontrivial mean part. Thus the alternative estimation procedure in Section 5.2 was used. The procedure was found to converge after 400 iterations with $M = 20$ and $N = 20$ for the two loops involved. Table 4 lists the estimated mean parameters and variance.

$\hat{\eta}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\alpha}_{12}$	$\hat{\alpha}_{22}$	$\hat{\alpha}_{23}$	$\hat{\alpha}_{24}$	$\hat{\alpha}_{32}$	$\hat{\alpha}_{33}$	$\hat{\sigma}^2$
11.95	6.17	-2.77	3.05	-4.53	0.20	0.08	-0.95	-0.72	-1.73	2.66	1.27	2.85

Table 4: Estimated mean parameters and variance for the data center example

Table 5 presents the estimated correlation parameters for the quantitative factors x_1, x_2, x_3, x_4, x_5 , and the two-level qualitative factor z_1 . As shown in the table, the estimated correlation parameters vary significantly from one quantitative factor to another, and the values for x_3 and x_1 are much larger than the rest, indicating that the responses may be rugged in the dimensions of x_3 or x_1 . The estimated correlation for z_1 (between its two levels) is small (0.005), indicating that the responses at the two levels of z_1 are not significantly correlated. This is consistent with the known physics that two levels of z_1 correspond to distinct data-center thermal distributions.

$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$	$\hat{\phi}_5$	$\hat{\tau}_1$
5.35	1.07	7.71	3.36	1.45	0.005

Table 5: The estimated correlation parameters for the quantitative factors x_1, x_2, x_3, x_4 , and x_5 , and the estimated correlation for z_1 in the data center example

Tables 6 and 7 give the estimated correlation matrices for z_2 and z_3 . Both matrices are symmetric with unit diagonal elements. Also, their eigenvalues are all positive [(3.11, 0.58, 0.17, 0.14) and (2.38, 0.45, 0.17), respectively]. Thus, the estimated correlation matrices are PDUDEs, and are indeed valid correlation matrices.

1.00	0.84	0.78	0.50
0.84	1.00	0.82	0.54
0.78	0.82	1.00	0.71
0.50	0.54	0.71	1.00

Table 6: The estimated correlation matrix for z_2 in the data center example

1.00	0.62	0.83
0.62	1.00	0.61
0.83	0.61	1.00

Table 7: The estimated correlation matrix for z_3 in the data center example

Following Welch et al. (1992), we plot the estimated main effects and two-factor interactions. Figure 4 depicts the main-effect functions of the quantitative factors $x_1, x_2, x_3, x_4,$ and $x_5,$ and Table 8 lists the main effects of the qualitative factors $z_1, z_2,$ and $z_3.$ Note that the main effects of x_1 differ a lot near the two ends, and the main effects of z_3 differ a lot at levels 1 and 2 (-1.33 versus 1.39). Figure 5 displays the two-factor interaction plots for some selected pairs of the quantitative factors. Note the large and complex interaction patterns of (x_1, x_2) and $(x_1, x_4).$ Figure 6 shows some two-factor interaction functions between the quantitative and qualitative factors. As illustrated by this figure, such interactions are rather intricate. For example, the interactions between x_3 and z_2 become larger as the values of x_3 move away from the middle. At levels 1, 2, and 3 of $z_2,$ the main effects of x_2 profile similarly but have a very different pattern at level 4 of $z_2.$ The observed complex second-order relationships cannot be captured by standard quadratic models.

z_1	level 1: -0.14	level 2: 0.0061		
z_2	level 1: 0.80	level 2: -0.15	level 3: 0.08	level 4: -0.87
z_3	level 1: -1.33	level 2: 1.39	level 3: -0.09	

Table 8: Estimated main effects of $z_1, z_2,$ and z_3 for the data center example

To assess the prediction accuracy of the fitted GP model, we perform a *leave-one-out* cross-validation, the same approach used by Welch et al. (1992). Denote by $\hat{y}_{-i}(\mathbf{w}_i^0)$ the empirical BLUP in (22) of $y(\mathbf{w}_i^0)$ based on all the data except the observation $y(\mathbf{w}_i^0).$ The cross-validation version of the RMSE is $(\sum_{i=1}^{67} [\hat{y}_{-i}(\mathbf{w}_i^0) - y(\mathbf{w}_i^0)]^2 / 67)^{1/2} = 1.88$ (relative to a data range of 6.37 to 22.08). Again, following Welch et al. (1992), to minimize computation, the estimates of the correlation parameters and correlation matrices are not recomputed for each prediction and instead are still based on all 67 data points. (Recomputing them for each prediction is very time consuming, since running the algorithm with 400 iterations would take more than 3 hours in a double-core PC running a Linux system.) The plot in Figure 7 of $\hat{y}_{-i}(\mathbf{w}_i^0)$ against $y(\mathbf{w}_i^0)$ demonstrates the decent accuracy of the predictor. As shown in the figure, the prediction accuracy deteriorates moderately when the responses are near the two ends. This is understandable since the design set for the input values is not a space-filling design and fewer observations are available for large or small response values.

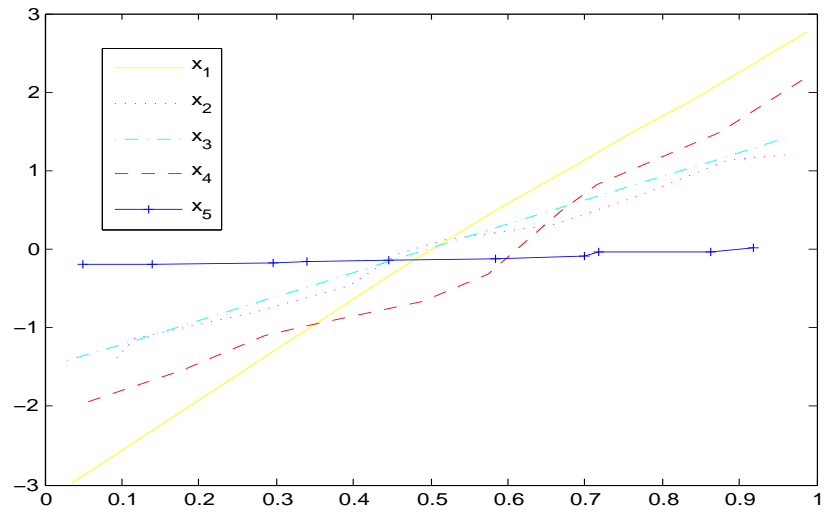


Figure 4: The main-effect functions of x_1 , x_2 , x_3 , x_4 , and x_5 for the data center example.

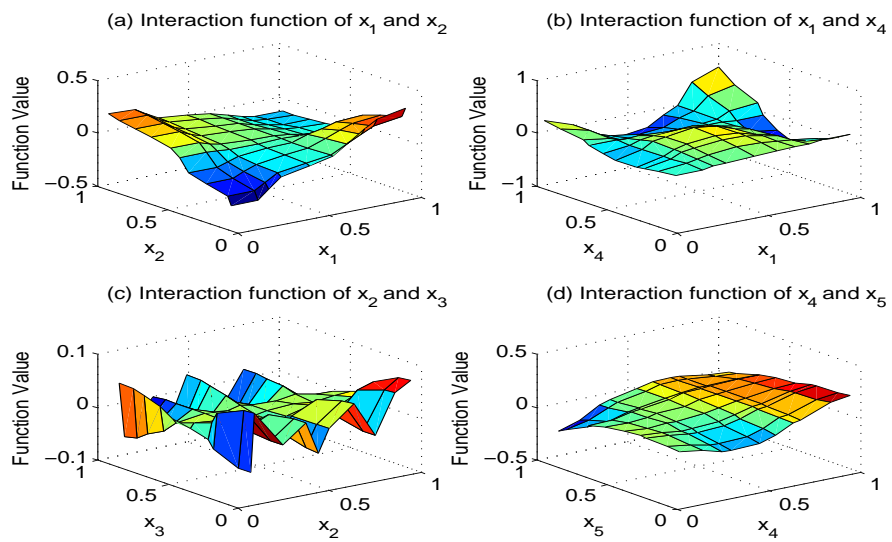


Figure 5: The two-factor interaction functions for some selected pairs of x_1 , x_2 , x_3 , x_4 , and x_5 for the data center example.

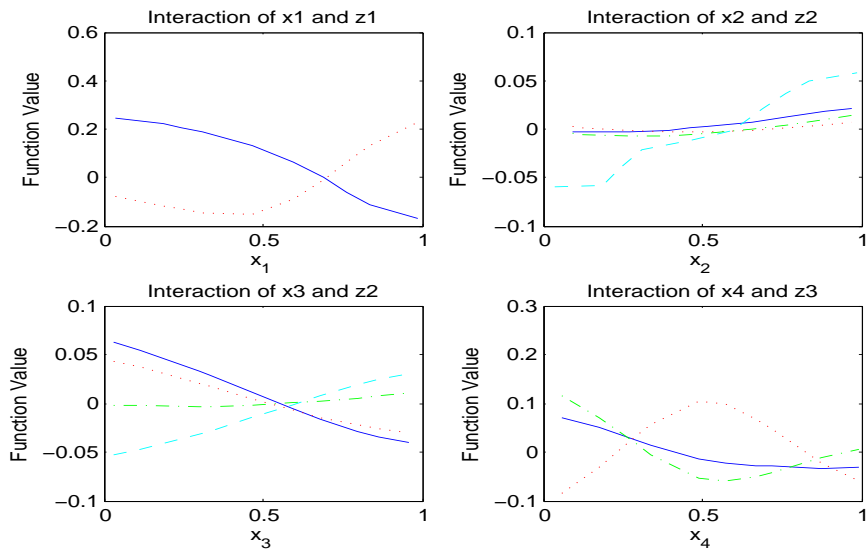


Figure 6: The two-factor interaction functions for some selected pairs of quantitative and qualitative factors, where the blue solid lines are the responses at the first levels of z_1 , z_2 , and z_3 , the red dotted lines are the responses at the second levels of z_1 , z_2 , and z_3 , the green dash-dot lines are the responses at the third levels of z_2 and z_3 , and the cyan dashed lines are the responses at the fourth level of z_2 for the data center example.

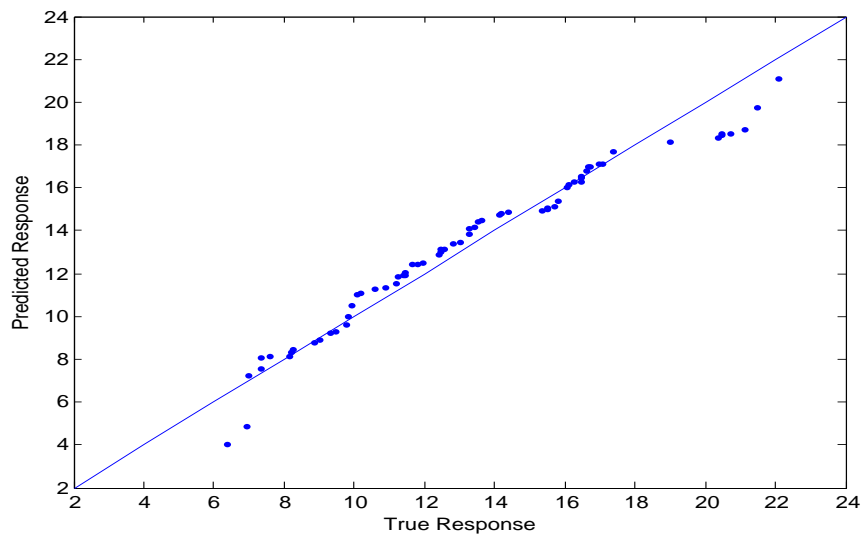


Figure 7: The predicted responses using cross validation versus the true responses for the data center example.

8 Discussions and Concluding Remarks

Ever since the publication of Sacks, Welch, Mitchell, and Wynn (1989), GP models have enjoyed great popularity in computer modeling. To date, one important but still unsettled problem is how to model computer experiments with qualitative and quantitative factors. Here, we give a systematic treatment of building GP models with both types of factors. The proposed methodology has two major contributions. First, it is a general method for constructing correlation functions with qualitative and quantitative factors. It makes use of some underlying multivariate Gaussian processes. The second is an iterative procedure used for estimation. The validity of the constructed correlation functions in the estimation is ensured by some recently developed optimization techniques. The proposed method is successfully applied to an example involving a known function and a real example for modeling the thermal distribution of a data center.

We also discuss and propose some restrictive correlation functions for qualitative factors that may be justifiable in particular applications. In such cases, we show that the estimation procedure can be significantly simplified. Although the primary focus is on modeling and estimation, some suggestions for selecting designs for computer experiments with qualitative and quantitative factors are also given. Research on the design issue is currently ongoing and will be reported elsewhere.

In this article, we focus on the method of maximum likelihood for estimation. While the method is widely used in computer experiments (Sacks, Welch, Mitchell, and Wynn 1989; Welch et al. 1992), it also has some drawbacks. For example, it may sometimes be difficult to obtain a global maximum of the likelihood; the likelihood function near the optimum may be flat; the likelihood surface may be difficult to assess or visualize because the parameters include correlation matrices. Some methods have been proposed in the literature to mitigate these problems. Welch et al. (1992) suggested making several, usually five, tries from different starting points to improve the chance of getting a global maximum. Li and Sudjianto (2005) proposed penalized likelihood method to deal with the flatness of the likelihood function near the optimum. For studying the surface of the likelihood function, we refer to Handcock and Stein (1993) and Handcock, Meier, and Nychka (1994). We also plan to explore the general issue of visualizing and assessing a function of correlation matrices in a separate research effort. As an alternative to the maximum likelihood method, we may use Bayesian methods for the modeling and estimation. We briefly discuss such methods and the related computational challenges in this article. Further work on the Bayesian methods will be developed and reported elsewhere.

Acknowledgments

Z. Qian is grateful to Dr. Zhaosong Lu and Dr. Renato D.C. Monteiro for useful discussions on some optimization issues. The authors thank the editor, an associate editor, and two referees for their valuable comments and suggestions, which led to significant improvements in the paper.

Qian is supported by NSF DMS-0705206 and C. F. J. Wu is supported by NSF grants DMI-0620259 and DMS-0705261. H. Wu was supported by NSF DMI-0620259 while visiting Professor C. F. J. Wu at Georgia Tech in the fall of 2005.

Appendix: Proofs and Computational Details

Definitions and Formulas for $\frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \mathbf{T}_j}$ and $\frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j}$

The definitions and results below follow from Graham (1981, Chap. 4).

(1): Define $\frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \mathbf{T}_j}$ as

$$\frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \mathbf{T}_j} = \begin{pmatrix} \frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,1,1}} & \cdots & \frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,1,m_j}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,m_j,1}} & \cdots & \frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,m_j,m_j}} \end{pmatrix}.$$

For $1 \leq r \leq m_j$, $1 \leq s \leq m_j$, it is clear that $\frac{\partial \text{tr}(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,r,s}} = \text{tr}\left(\frac{\partial(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,r,s}}\right)$. Furthermore, $\text{tr}\left(\frac{\partial(\mathbf{E}\mathbf{R}^{-1})}{\partial \tau_{j,r,s}}\right) = \text{tr}\left(\mathbf{E} \frac{\partial \mathbf{R}^{-1}}{\partial \tau_{j,r,s}}\right) = \text{tr}\left(-\mathbf{E}\mathbf{R}^{-1} \frac{\partial \mathbf{R}}{\partial \tau_{j,r,s}} \mathbf{R}^{-1}\right)$.

(2): Define $\frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j}$ as

$$\frac{\partial |\mathbf{R}|}{\partial \mathbf{T}_j} = \begin{pmatrix} \frac{\partial |\mathbf{R}|}{\partial \tau_{j,1,1}} & \cdots & \frac{\partial |\mathbf{R}|}{\partial \tau_{j,1,m_j}} \\ \vdots & \ddots & \vdots \\ \frac{\partial |\mathbf{R}|}{\partial \tau_{j,m_j,1}} & \cdots & \frac{\partial |\mathbf{R}|}{\partial \tau_{j,m_j,m_j}} \end{pmatrix}.$$

Let ρ_{uv} be the (u, v) th element of \mathbf{R} and \mathbf{R}_{uv} be the cofactor of element ρ_{uv} in $|\mathbf{R}|$. Then, for $1 \leq r \leq m_j$, $1 \leq s \leq m_j$,

$$\frac{\partial |\mathbf{R}|}{\partial \tau_{j,r,s}} = \text{tr}(\mathbf{A}\mathbf{B}_{jrs}^t),$$

where $\mathbf{A} = [\mathbf{R}_{uv}]$ and $\mathbf{B}_{jrs} = [b_{uv}^{(jrs)}]$ are $n \times n$ matrices. Here, $b_{uv}^{(jrs)} = \frac{\partial \rho_{uv}}{\partial \tau_{j,r,s}}$, for $1 \leq u \leq n$ and $1 \leq v \leq n$.

Proof of Proposition 1

Using the Kronecker product notation (Graham 1981), we have $\mathbf{R} = \mathbf{H} \otimes \mathbf{T}_1$. Basic facts on Kronecker product (Graham 1981, Chap. 2) imply that

$$\begin{aligned} |\mathbf{H} \otimes \mathbf{T}_1| &= |\mathbf{H}|^{m_1} |\mathbf{T}_1|^b, \\ \mathbf{E}\mathbf{R}^{-1} &= \mathbf{E}(\mathbf{H}^{-1} \otimes \mathbf{T}_1^{-1}) = (\mathbf{E} \otimes 1)(\mathbf{H}^{-1} \otimes \mathbf{T}_1^{-1}) = (\mathbf{E}\mathbf{H}^{-1}) \otimes \mathbf{T}_1^{-1}, \\ \text{tr}(\mathbf{E}\mathbf{R}^{-1}) &= \text{tr}(\mathbf{E}\mathbf{H}^{-1} \otimes \mathbf{T}_1^{-1}) = \text{tr}(\mathbf{E}\mathbf{H}^{-1})\text{tr}(\mathbf{T}_1^{-1}). \end{aligned}$$

Since \mathbf{E} and \mathbf{H} are independent of \mathbf{T}_1 , the problem (14) simplifies to (18).

References

- [1] Abrahamsen, P. (1997), “A Review of Gaussian Random Fields and Correlation Functions,” Report No. 917, <http://publications.nr.no/917Rapport.pdf>, Norwegian Computer Center, Oslo, Norway.
- [2] Banerjee, S., and Gelfand, A. E. (2002), “Prediction, Interpolation and Regression for Spatially Misaligned Data,” *Sankhya*, 64, 227-245.
- [3] Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, London: Arnold.
- [4] Bertsekas, D. P. (1999), *Nonlinear Programming*, Nashua, NH: Athena Scientific.
- [5] Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.
- [6] Brown, P. J., Le, N. D., and Zidek, J. V. (1994), “Multivariate Spatial Interpolation and Exposure to Air Pollutants,” *The Canadian Journal of Statistics*, 22, 489-509.
- [7] Dattorro, J. (2005), *Convex Optimization and Euclidean Distance Geometry*, Palo Alto, CA: Meboo Publishing.
- [8] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- [9] Edwards, D. M. (2000), *Introduction to Graphical Modeling*, New York: Springer.
- [10] Fang, K. F., Li, R. Z., and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, New York: Chapman & Hall/CRC Press.
- [11] Flometrics (2005), “Flotherm: Fluid Dynamics Based Thermal Analysis Software,” <http://www.flometrics.com/>.
- [12] Fluent, Inc. (1998), FLUENT, Release 5.5.14 (3d, segregated, laminar).
- [13] Fridlyander, I. N. (2002), “Modern Aluminum and Magnesium Alloys and Composite Materials Based on Them,” *Metal Science and Heat Treatment*, 44, 292-296.
- [14] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton: Chapman & Hall/CRC Press.
- [15] Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computation*, Baltimore: The Johns Hopkins University Press.

- [16] Graham, A. (1981), *Kronecker Products and Matrix Calculus With Applications*, Chichester, England: Ellis Horwood Limited.
- [17] Grant, M., Boyd, S., and Ye, Y. (2006), *CVX: Matlab Software for Disciplined Convex Programming*, Version 1.0 beta 3, <http://www.stanford.edu/boyd/cvx/>.
- [18] Handcock, M. S., Meier, K., and Nychka, D. (1994), Comment on “Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications” by G. M. Laslett, *Journal of the American Statistical Association*, 89, 401-403.
- [19] Handcock, M. S., and Stein, M. L. (1993), “A Bayesian Analysis of Kriging,” *Technometrics*, 35, 403-410.
- [20] Joseph, V. R., and Delaney, J. D. (2007), “Functionally Induced Priors for the Analysis of Experiments,” *Technometrics*, 49, 1-11.
- [21] Katz, M. H. (2006), *Multivariable Analysis: A Practical Guide for Clinicians*, Cambridge, U.K.: Cambridge University Press.
- [22] Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press.
- [23] Lawrence Berkeley National Laboratory (2002), “Data Center Energy Use: Truth Versus Myth,” <http://www.lbl.gov/Science-Articles/Archive/data-center-energy-myth.html>.
- [24] Li, R., and Sudjianto, A. (2005), “Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models,” *Technometrics*, 47, 111-120.
- [25] Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer.
- [26] Mardia, K. V., and Goodall, C. R. (1993), “Spatial-Temporal Analysis of Multivariate Environmental Monitoring Data,” In: *Multivariate Environmental Statistics*, G. P. Patil and C. R. Rao, eds., 347-386, Amsterdam: Elsevier.
- [27] McMillian, N. J., Sacks, J., Welch, W. J., and Gao, F. (1999), “Analysis of Protein Activity Data by Gaussian Stochastic Process Models,” *Journal of Biopharmaceutical Statistics*, 9, 145-160.
- [28] Ruszczyński, A., and Shapiro, A. (eds) (2003), *Stochastic Programming. Handbooks in Operations Research and Management Science*, 10, Elsevier.
- [29] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), “Design and Analysis of Computer Experiments,” *Statistical Science*, 4, 409-435.
- [30] Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.
- [31] Schmidt, R. (2003), “Hot Spots in Data Centers,” Online Forum of Electrical Cooling, www.electronics-cooling.com.

- [32] Schmidt, R. R., Cruz, E. E., and Iyengar, M. K. (2005), “Challenges of Data Center Thermal Management,” *IBM Journal of Research and Development*, 49,709-723.
- [33] Stein, M. L. (1999), *Interpolation of Spatial Data*, New York: Springer.
- [34] The MathWorks, Inc. (2006), *Matlab: The Language of Technical Computing*, Version 6.5.1.
- [35] Vandenberghe, L., and Boyd, S. (1996), “Semidefinite Programming,” *SIAM Review*, 38, 49-95.
- [36] Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992), “Screening, Predicting and Computer Experiments,” *Technometrics*, 34, 15-25.
- [37] Wolkowicz, H., Saigal, R., and Vandenberghe, L.(eds.) (2000), *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, Boston: Kluwer Academic.
- [38] Wu, C. F. J., and Ding, Y. (1998), “Construction of Response Surface Designs for Qualitative and Quantitative Factors,” *Journal of Statistical Planning and Inference*, 71, 331-348.
- [39] Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.