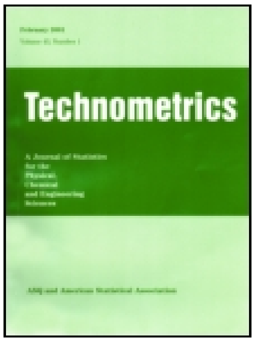

Technometrics

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors

Peter Z. G Qian, Huaiqing Wu & C. F. Jeff Wu

To cite this article: Peter Z. G Qian, Huaiqing Wu & C. F. Jeff Wu (2008) Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors, Technometrics, 50:3, 383-396, DOI: 10.1198/004017008000000262

To link to this article: https://doi.org/10.1198/004017008000000262

Published online: 01 Jan 2012.

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
(*peterq@stat.wisc.edu*)

Huaiqing Wu

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

C. F. Jeff Wu

The H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
(*jeffwu@isye.gatech.edu*)

Modeling experiments with qualitative and quantitative factors is an important issue in computer modeling. We propose a framework for building Gaussian process models that incorporate both types of factors. 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 process; Semidefinite 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 of the factors involved in a computer experiment are quantitative; however, in many situations, some factors are qualitative by nature. Consider, for example, the data center computer experiment described by Schmidt, Cruz, and Iyengar (2005). 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 heat load, and total diffuser flow rate. Examples of qualitative variables are diffuser location, return air vent location, and rack heat load nonuniformity. Computer models with qualitative and quantitative factors occur frequently in business operations applications, in which some socioeconomic factors of the customers, such as gender and commuting method, are inherently qualitative.

In this article we propose new GP models to address this issue. Note that the corresponding problem for physical experimentation is much easier, because a GP model is not involved (Wu and Ding 1998; Wu and Hamada 2000). Quadratic models have long been used for modeling physical experiments involving quantitative and qualitative factors, and such polynomial models can provide reasonable approximations to physical phenomena. But computer experiments often include many factors and can have highly nonlinear 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. Whereas 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 we discuss in Section 4.2. As a key to the development of the new GP models, we propose a general approach for constructing correlation functions with both types of factors. We develop an iterative estimation procedure 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

TECHNOMETRICS, AUGUST 2008, VOL. 50, NO. 3
DOI 10.1198/004017008000000262

some restrictive correlation matrixes 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 a real example for modeling temperature in a data center. Section 8 provides some discussion and concluding remarks.

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) $\mathbf{x} = (x_1, \dots, x_I)^t$, with the data comprising an $n \times I$ matrix of input values $\mathbf{X} = (\mathbf{x}_1^0, \dots, \mathbf{x}_n^0)^t$ and the corresponding $n \times 1$ vector of response values $\mathbf{y} = (y_1, \dots, y_n)^t$. The GP model assumes

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

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_l(\mathbf{x}))^t$ is the vector of l prespecified functions and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_l)^t$ is the vector of unknown coefficients. The residual $\epsilon(\mathbf{x})$ is assumed to be a stationary GP with mean 0, variance σ^2, and some correlation function $\mathrm{cor}(\epsilon(\mathbf{x}_1), \epsilon(\mathbf{x}_2)) = K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2)$, where $\boldsymbol{\phi}$ is the vector of 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 et al. 2003),

$$\begin{aligned} K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2) &= \prod_{i=1}^{I} \exp\{-\phi_i (x_{i1} - x_{i2})^p\} \\ &= \exp\left\{-\sum_{i=1}^{I} \phi_i (x_{i1} - x_{i2})^p\right\}, \end{aligned} \tag{2}$$

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 often is fixed at 2, giving the *Gaussian correlation function*, which we use 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.

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 often are needed to specify the underlying physics or mathematics of the experiment. Although the number of qualitative factors usually is not large, they can determine some important properties of the experiment. For example, in the data center experiment described by Schmidt et al. (2005), diffuser location, return air vent location, and rack heat load nonuniformity are qualitative factors. To take into account the distinct natures and roles of quantitative and qualitative factors in computer modeling, we describe two analysis approaches.

The first approach is the *independent analysis*, in which distinct GP models are used to model 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 four-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). Accurately estimating these 576 parameters would require a large number of observations, which generally is not feasible.

In view of the shortcomings of the foregoing approach, we introduce an integrated analysis that assumes a single GP model across different values of qualitative and quantitative factors so as to borrow strength from all the observations. Suppose that a computer experiment involves factors $\mathbf{w} = (\mathbf{x}^t, \mathbf{z}^t)^t$, where the factors in $\mathbf{x} = (x_1, \dots, x_I)^t$ are quantitative, and the factors in $\mathbf{z} = (z_1, \dots, z_J)^t$ are qualitative. Throughout the article, the factors in $\mathbf{z}$ are assumed to be categorical but not ordinal, unless described otherwise.

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

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

where $\mathbf{f}(\mathbf{w}) = (f_1(\mathbf{w}), \dots, f_l(\mathbf{w}))^t$ is the vector of l prespecified functions (e.g., polynomials) and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_l)^t$ is the vector of unknown coefficients. The residual $\epsilon(\mathbf{w})$ is assumed to be a GP with mean 0, 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 and other distance-based correlation functions (Santner et al. 2003) are not applicable, because of 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 thus 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-0 m_1-variate process

$$\boldsymbol{\epsilon}^*(\mathbf{x}) = \left(\epsilon_1(\mathbf{x}), \dots, \epsilon_{m_1}(\mathbf{x})\right)^t.$$

Then we need only 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 $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_{m_1})^t$ is an $m_1 \times m_1$ nonsingular matrix with unit row vectors (i.e., $\mathbf{a}_u^t \mathbf{a}_u = 1$ for $u = 1, \ldots, m_1$) and $\boldsymbol{\eta}(\mathbf{x}) = (\eta_1(\mathbf{x}), \ldots, \eta_{m_1}(\mathbf{x}))^t$, where $\eta_1(\mathbf{x}), \ldots, \eta_{m_1}(\mathbf{x})$ are independent stochastic processes with the same variance σ^2 and correlation function $K_{\boldsymbol{\phi}}$. Thus $\operatorname{cor}(\boldsymbol{\eta}(\mathbf{x}_1), \boldsymbol{\eta}(\mathbf{x}_2)) = K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{I}_{m_1}$, where $\mathbf{I}_{m_1}$ is 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

$$\begin{aligned}\operatorname{cor}(\epsilon(\mathbf{w}_1), \epsilon(\mathbf{w}_2)) &= \operatorname{cor}\big(\epsilon_{z_{11}}(\mathbf{x}_1), \epsilon_{z_{12}}(\mathbf{x}_2)\big) \\ &= \operatorname{cor}\big(\mathbf{a}_{z_{11}}^t \boldsymbol{\eta}(\mathbf{x}_1), \mathbf{a}_{z_{12}}^t \boldsymbol{\eta}(\mathbf{x}_2)\big) \\ &= \mathbf{a}_{z_{11}}^t \mathbf{a}_{z_{12}} K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2). \qquad (4)\end{aligned}$$

Let $\tau_{r,s} = \mathbf{a}_r^t \mathbf{a}_s$, where $r, s = 1, \ldots, m_1$. 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 (PDUDE). Any PDUDE can be written as $\mathbf{B}\mathbf{B}^t$, where $\mathbf{B}$ is a nonsingular matrix with unit row vectors. Thus the foregoing construction shows that for any PDUDE $\mathbf{T}_1 = (\tau_{r,s})$ and any correlation function $K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2)$, $\operatorname{cor}(\epsilon(\mathbf{w}_1), \epsilon(\mathbf{w}_2)) = \tau_{z_{11}, z_{12}} K_{\boldsymbol{\phi}}(\mathbf{x}_1, \mathbf{x}_2)$ is a valid correlation function. Similar correlation functions were used by Mardia and Goodall (1993) for kriging, by Brown, Le, and Zidek (1994) for assigning a prior to a covariance matrix, and by Banerjee and Gelfand (2002) for modeling a cross-covariance matrix.

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

$$\operatorname{cor}(\epsilon(\mathbf{w}_1), \epsilon(\mathbf{w}_2)) = \prod_{j=1}^{J} \big[\tau_{j, z_{j1}, z_{j2}} K_{\boldsymbol{\phi}_j}(\mathbf{x}_1, \mathbf{x}_2)\big], \qquad (5)$$

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

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

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

where $\phi_i = \sum_{j=1}^{J} \phi_{ij}$ for $i = 1, \ldots, I$. Note that (6) bears some resemblance to (2) for the GP model with quantitative factors. This similarity in part motivates the inference procedures in Section 5. The parameter $\tau_{j, z_{j1}, z_{j2}}$ measures the correlation between the responses at any two input values $\mathbf{w}_1$ and $\mathbf{w}_2$ that differ only in terms of the values of z_j, at levels z_{j1} and z_{j2}. When using (6) for modeling a computer experiment, it is sometimes desirable to assume that all of 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 MATRIXES FOR QUALITATIVE FACTORS

For flexible modeling, unrestrictive correlation matrixes (i.e., PDUDEs) should be used in (6) for the qualitative factors. Sometimes it may be desirable to use *restrictive* correlation matrixes that assume some parametric relationships among the factor levels. In this section we consider several such correlation matrixes. Although substantial simplification in estimating these 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 Exchangeable Correlation Functions

The exchangeable correlation function assumes that the m levels of z are of an *exchangeable* 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 $\mathbf{1} = (1, \ldots, 1)^t$. 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 nonzero $m \times 1$ 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 also was used by Joseph and Delaney (2007) for modeling some physical experiments. Using it in (6) leads to

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

where $0 < p \leq 2$, $\phi_i \geq 0$ for $i = 1, \ldots, I$ and $\theta_j > 0$ for $j = 1, \ldots, J$. On a logarithmic scale, (7) uses the pth power of 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 the levels of z to have different correlations:

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

where θ_r and θ_s are positive and 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 et al. (1999) to model correlations of unordered categorical variables using GPs. But 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 $(\theta_1, \ldots, \theta_m)$ instead of $m(m-1)/2$ parameters for an unrestrictive correlation matrix. In particular, we observe that for $m \geq 4$ and any four levels of z (say 1, 2, 3, and 4), by (8),

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

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 that the m levels of z form K groups: $g_1, \ldots, g_K$, where g_k includes b_k levels. For simplicity, we assume that the correlation between any two levels in g_k is α_k $(0 < \alpha_k < 1)$ and that 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 $(k = 1, \ldots, K)$ 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 must 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 level forming another group, then the constraint is $\gamma < ((1 + \alpha_1)/2)^{1/2}$.

4.4 Correlation Functions for Ordinal Qualitative Factors

Sections 4.1–4.3 focus on correlation functions for categorical, but not ordinal 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 $1, \ldots, m$ in an increasing order.

The first method assumes that

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

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 $\gamma_1, \ldots, \gamma_{m-1}$. For $m = 3$, the constraint is $\gamma_1 < ((1 + \gamma_2)/2)^{1/2}$. 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).

The second method, called the *transformation method*, 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 equation $F(v_k) = F(0) + (F(1) - F(0))(k-1)/(m-1)$, which has a unique solution, with $0 = v_1 < \cdots < v_m = 1$. We then define $\tau_{r,s} = K(v_r, v_s)$, where $K(v_r, v_s)$ is a correlation function for v. Selecting an appropriate F may vary depending on the application and may require subject matter knowledge.

5. ESTIMATION AND PREDICTION

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

Here we focus mainly on the estimation for the general case of $\mathbf{T}$ without assuming any special correlation structures (such as those discussed in Sec. 4) for the qualitative factors $z_1, \ldots, z_J$. We also briefly discuss 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 matrixes, whereas in other applications, such as in a Gaussian graphical model, usually a single correlation matrix is involved. Second, sometimes computer experiments can be run 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 we discuss 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 of 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 the approach 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 semidefinite programming technique in optimization. A brief introduction of semidefinite programming is given in Section 5.1. The estimation procedures are developed in Sections 5.2–5.4, and the prediction procedure is provided in Section 5.5.

5.1 Semidefinite Programming

Consider the optimization problem

$$\begin{aligned} &\min_{\mathbf{X}} \mathbf{C} \bullet \mathbf{X} \\ &\text{subject to} \quad \mathbf{A}_s \bullet \mathbf{X} = \mathbf{b}_s, \qquad s = 1, \ldots, S, \\ &\qquad\qquad\quad \mathbf{X} \succ 0\ (\succeq 0), \end{aligned} \qquad (10)$$

where $\mathbf{C}$ is an $n \times n$ symmetric real matrix and the optimization variable is $\mathbf{X}$ in the space of $n \times n$ real symmetric matrixes. The inequalities $\mathbf{X} \succ 0$ and $\mathbf{X} \succeq 0$ mean that $\mathbf{X}$ is positive definite and positive semidefinite. The problem (10) is referred to as a semidefinite program in optimization (Vandenberghe and Boyd 1996; Wolkowicz, Saigal, and Vandenberghe 2000). The notation $\mathbf{C} \bullet \mathbf{X}$ represents the inner product of the matrixes $\mathbf{C}$ 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 $\mathbf{X}$. Equivalently, $\mathbf{C} \bullet \mathbf{X}$ also can be written as $\mathrm{tr}(\mathbf{CX})$. Throughout the 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 semidefinite programming (SP) problem is a convex problem, which can be solved efficiently by *interior point algorithms* (Nesterov and Nemirovskii 1994; Wolkowicz et al. 2000). These algorithms compute the solution to (10) within a cone formed by positive definite matrixes 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 $x_1, \ldots, x_I$ and J qualitative factors $z_1, \ldots, z_J$, 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

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

where $\mathbf{F} = (\mathbf{f}(\mathbf{w}_1^0), \ldots, \mathbf{f}(\mathbf{w}_n^0))^t$ is an $n \times l$ 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 $\mathrm{cor}(\epsilon(\mathbf{w}_i^0), \epsilon(\mathbf{w}_j^0))$ defined in (6), with $p = 2$.

Given $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, and $\mathbf{T}$, the maximum likelihood estimate of σ^2 is $\widehat{\sigma}^2 = (1/n)(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^t \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$. Substituting $\widehat{\sigma}^2$ 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 $\widehat{\boldsymbol{\beta}}$, let $\mathbf{e} = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}$ and $\mathbf{E} = \mathbf{e}\mathbf{e}^t$ throughout the article. Thus

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

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

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

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

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

where the optimization variables are $\boldsymbol{\phi}$ and $\mathbf{T}$. Throughout "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 $\widehat{\mathbf{T}}$, $\widehat{\boldsymbol{\phi}}$ is obtained as follows:

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

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

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

In optimization, such an iterative algorithm is called the *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. We give the details for implementing the T-step. Let $f(\mathbf{T})$ denote the objective function in (14), that is,

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

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

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

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

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

The formulas for $\partial\, \mathrm{tr}(\mathbf{E}\mathbf{R}^{-1})/\partial \mathbf{T}_j$ and $\partial|\mathbf{R}|/\partial \mathbf{T}_j$ are given in Appendix A. Such a linear approximation has been shown to be reasonable and is widely used to approximate SP problems with nonlinear objective functions (Wolkowicz et al. 2000). The solution $\widehat{\mathbf{T}}$ to (15) can be used 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 matrixes:

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

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

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

In summary, the foregoing algorithm consists of an outer loop [the β-step and (ϕ, T)-step] and an inner loop (the ϕ-step and T-step). The outer loop is repeated M times, and for each iteration in the outer loop, the inner loop is repeated N times until convergence.

When $\boldsymbol{\beta}^t\mathbf{f}(\mathbf{w})$ 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 $\widehat{\boldsymbol{\phi}}$ and $\widehat{\mathbf{T}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are obtained as follows:

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

Correlation fitting: Given $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$, let $u_i = (y_i - \widehat{\boldsymbol{\beta}}^t \times \mathbf{f}(\mathbf{w}_i))/\widehat{\sigma}$, for $i = 1, \ldots, n$. Then the proposed GP model with mean 0, variance 1, and correlation matrix $\mathbf{R}$ is fitted to the data $\mathbf{u} = (u_1, \ldots, u_n)^t$, and $\boldsymbol{\phi}$ and $\mathbf{T}$ are estimated using the method of maximum likelihood.

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

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

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 $\widehat{\mathbf{T}}$, $\widehat{\boldsymbol{\phi}}$ is obtained as follows:

$$\widehat{\boldsymbol{\phi}} = \text{argmin}_{\boldsymbol{\phi}}[\text{tr}(\mathbf{G}\mathbf{R}^{-1}) + \ln|\mathbf{R}|]$$
$$\text{subject to} \quad \phi_i \geq 0, \qquad i = 1, \ldots, I.$$

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

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

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

$$\widehat{\mathbf{T}} = \text{argmin}_{\mathbf{T}}\left[f(\mathbf{T}_0) + \sum_{j=1}^{J}\left(\frac{\partial f(\mathbf{T}_0)}{\partial \mathbf{T}_j} \bullet (\mathbf{T}_j - \mathbf{T}_{0,j})\right)\right]$$
$$\text{subject to} \quad \mathbf{T}_j \succ 0, \qquad j = 1, \ldots, J; \tag{17}$$
$$\text{diag}(\mathbf{T}_j) = \mathbf{1}, \qquad j = 1, \ldots, J,$$

with

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

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

5.3 Simplification of Estimation With Structured Design

Some structured designs $\mathbf{D}_w$ can significantly simplify the computation in the T-step. This convenience is possible only 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 $1, \ldots, m_1$. Assume that $\mathbf{D}_w$ is a cross-array (Wu and Hamada 2000) of $\mathbf{D}_x$ and $\mathbf{D}_z$, where $\mathbf{D}_x$ is a $b \times I$ design matrix for the quantitative factors $\mathbf{x}$ and $\mathbf{D}_z = (1, \ldots, m_1)^t$ is an $m_1 \times 1$ design matrix for the qualitative factor z_1. Consequently, $\mathbf{D}_w$ consists of all level combinations between those in $\mathbf{D}_x$ and those in $\mathbf{D}_z$. Thus $\mathbf{D}_w$ has $n = bm_1$ rows (runs). As shown in the following proposition, this cross-array structure of $\mathbf{D}_w$ simplifies the optimization problem in (14) and also makes it free of $\widehat{\boldsymbol{\phi}}$. Consequently, estimating $\boldsymbol{\phi}$ 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 $\mathbf{H} = (h_{j_1j_2})$ denote a $b \times b$ matrix with its (j_1, j_2)th entry given as

$$h_{j_1j_2} = \exp\left\{-\sum_{i=1}^{I}\phi_i\left(x_{ij_1} - x_{ij_2}\right)^2\right\}.$$

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

Proposition 1. Suppose that $\mathbf{D}_w$ is a cross-array of $\mathbf{D}_x$ and $\mathbf{D}_z$. Then the problem in (14) is equivalent to

$$\widehat{\mathbf{T}}_1 = \text{argmin}_{\mathbf{T}_1}\left(m_1 \ln[\text{tr}(\mathbf{T}_1^{-1})] + \ln|\mathbf{T}_1|\right)$$
$$\text{subject to} \quad \mathbf{T}_1 \succ 0, \tag{18}$$
$$\text{diag}\,(\mathbf{T}_1) = \mathbf{1}.$$

The proof is given in Appendix B. With this proposition, the linear approximation (15) becomes

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

where

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

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

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

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

We now consider the general case with J qualitative factors $z_1, \ldots, z_J$. Assume that $\mathbf{D}_w$ is a cross-array of $\mathbf{D}_x$, the $b \times I$ design matrix for $\mathbf{x}$, and $\mathbf{D}_z$, the $q \times J$ design matrix for $\mathbf{z}$. Thus $\mathbf{D}_w$ has $n = bq$ rows (runs). Let $\mathbf{z}_1^0, \ldots, \mathbf{z}_q^0$ denote the q input values for $\mathbf{z}$, and let $\mathbf{T}^*$ 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 that $\mathbf{D}_w$ is a cross-array of $\mathbf{D}_x$ and $\mathbf{D}_z$, where $\mathbf{D}_x$ is a $b \times I$ design matrix for $\mathbf{x}$ and $\mathbf{D}_z$ is a $q \times J$ design matrix for $\mathbf{z}$. Then the problem in (14) is equivalent to

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

Again, the cross-array structure of $\mathbf{D}_w$ reduces the estimation of $\boldsymbol{\phi}$ 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 $\mathbf{D}_z$ also has some cross-array structure, then the optimization problem in (20) can be further simplified. Suppose that the qualitative factors $\mathbf{z}$ are grouped into $d \geq 2$ disjoint sets, $\{z_j : j \in A_k\}$, for $k = 1, \ldots, d$, where $\bigcup_{j=1}^{d} A_j = \{1, \ldots, J\}$ and the size of A_k is $J_k \geq 1$. Suppose further that $\mathbf{D}_z$ is a cross-array of $\mathbf{D}_1, \ldots, \mathbf{D}_d$, where $\mathbf{D}_k$ is a $q_k \times J_k$ design matrix for the factors in $\{z_j : j \in A_k\}$. (However, $\mathbf{D}_k$ 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 let $\mathbf{T}_k^*$ 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 that $\mathbf{D}_z$ in Proposition 2 is a cross-array of $\mathbf{D}_1, \ldots, \mathbf{D}_d$, where $\mathbf{D}_k$ 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:

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

for $k = 1, \ldots, d$. 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

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

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

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

Proposition 4. Suppose that $\mathbf{D}_w$ is a cross-array of $\mathbf{D}_x$ and $\mathbf{D}_z$, where $\mathbf{D}_x$ is a $b \times I$ design matrix for $\mathbf{x}$ and $\mathbf{D}_z = (1, \ldots, m_1)^t$ is an $m_1 \times 1$ design matrix for the qualitative factor z_1. Then the problem in (16) is equivalent to

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

With this proposition, the optimization problem (17) becomes

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

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

Proposition 5. Suppose that $\mathbf{D}_w$ is a cross-array of $\mathbf{D}_x$ and $\mathbf{D}_z$, where $\mathbf{D}_x$ is a $b \times I$ design matrix for $\mathbf{x}$ and $\mathbf{D}_z$ is a $q \times J$ design matrix for $\mathbf{z}$. Then the problem in (16) is equivalent to

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

5.4 Estimation When Restrictive Correlation Matrixes Are Used for Qualitative Factors

When restrictive correlation matrixes (as discussed in Sec. 4) are used for the qualitative factors $\mathbf{z}$, estimating the unknown parameters is easier and becomes that of $\boldsymbol{\phi}$ and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^t, \ldots, \boldsymbol{\psi}_J^t)^t$, where $\boldsymbol{\psi}_j$ $(j = 1, \ldots, J)$ 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 $\boldsymbol{\psi}_j$ such that $\mathbf{T}_j$ is a valid correlation matrix. It follows from the log-likelihood (11) that for given $\boldsymbol{\phi}$ and $\mathbf{T}$, $\widehat{\boldsymbol{\beta}} = (\mathbf{F}^t\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^t\mathbf{R}^{-1}\mathbf{y}$ and $\widehat{\sigma}^2 = (1/n)(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})^t\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})$. Substituting $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ into (11), this simplifies to (up to a negative constant) $n \ln(\widehat{\sigma}^2) + \ln|\mathbf{R}|$, where $\widehat{\sigma}^2$ and $\mathbf{R}$ depend on $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, and the data. Then $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\psi}}$ are obtained as follows:

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

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 eq. (7) of Sacks, Welch, Mitchell, and Wynn (1989), the empirical best linear unbiased predictor (BLUP) of y at the point $\mathbf{w}_0$ is

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

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

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

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

To visualize the features of the function $y(\mathbf{w})$ through 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, evaluating an integral for a qualitative factor simplifies to averaging over the predicted response values for all of the levels of that factor (see sec. 1 of their article for details).

5.6 Bayesian Methods

As an alternative, Bayesian methods also can be used for the proposed GP model, but this will require more computational effort. It is beyond the scope of this article to provide details for these methods. We give only a brief description here. A different Bayesian method was proposed by Han, Santner, Notz, and Bartel (2007), based on a model different from the one proposed in this article. The model in (3) can be formulated as a *hierarchical Bayesian model* (Gelman, Carlin, Stern, and Rubin 2004). As is often assumed in Bayesian statistics, the priors for the parameters $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}$, and $\mathbf{T}$ take the form

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \mathbf{T}) &= p(\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\phi}, \mathbf{T}) \\ &= p(\boldsymbol{\beta}|\sigma^2) p(\sigma^2) \prod_{i=1}^{I} p(\phi_i) \prod_{j=1}^{J} p(\mathbf{T}_j). \end{aligned}$$

One possible choice for the priors $p(\boldsymbol{\beta}|\sigma^2)$, $p(\sigma^2)$, and $p(\phi_i)$ is

$$p(\boldsymbol{\beta}|\sigma^2) \sim \mathrm{N}(\boldsymbol{\mu}, v\sigma^2 \mathbf{I}_l), \qquad p(\sigma^2) \sim \mathrm{IG}(\alpha, \gamma), \qquad \text{and}$$
$$p(\phi_i) \sim \mathrm{G}(a, b).$$

Here $\mathrm{N}(\boldsymbol{\mu}, v\sigma^2\mathbf{I}_l)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $v\sigma^2\mathbf{I}_l$, where $\mathbf{I}_l$ is the $l \times l$ identity matrix, $\mathrm{IG}(\alpha, \gamma)$ (where $\alpha > 0$ and $\gamma > 0$) is the inverse-gamma distribution, with probability density function (pdf)

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

and $\Gamma(a, b)$ (where $a > 0$ and $b > 0$) is the gamma distribution, with pdf

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

The choice of prior for $\mathbf{T}_j$ is more complicated; detailed discussions have been provided by Boscardin and Zhang (2004) and Gelman et al. (2004, sec. 19.2).

An empirical Bayes approach can be used for predicting the response y at a new point $\mathbf{w}_0$. This approach comprises two steps. In the first step, the correlation parameters $\boldsymbol{\phi}$ 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 $p(\boldsymbol{\phi}, \mathbf{T}|\mathbf{y})$ can be obtained from

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

where the integrand is determined by $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \mathbf{T}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \mathbf{T}) p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \mathbf{T})$. The posterior modes of $\boldsymbol{\phi}$ and $\mathbf{T}$, denoted by $\widehat{\boldsymbol{\phi}}$ and $\widehat{\mathbf{T}}$, are given by an optimal solution to the optimization problem $\max_{\boldsymbol{\phi}, \mathbf{T}} p(\boldsymbol{\phi}, \mathbf{T}|\mathbf{y})$. If the integral in (23) does not have an analytic form, then iterative methods such as the EM algorithm (Dempster, Laird, and Rubin 1977) and the stochastic programming method (Ruszczynski and Shapiro 2003) need to be used to solve this optimization problem. For the second step, assuming that $p(\boldsymbol{\beta}|\sigma^2) \sim \mathrm{N}(\boldsymbol{\mu}, v\sigma^2\mathbf{I}_l)$ and $p(\sigma^2) \sim \mathrm{IG}(\alpha, \gamma)$, it can be shown (Santner et al. 2003) that the conditional distribution of y at $\mathbf{w}_0$, given the observed $\mathbf{y}$ and $\boldsymbol{\phi}$ and $\mathbf{T}$, is the noncentral t distribution $T_1(n + 2a, \mu_1^*, \sigma_1^2)$, where

$$\begin{aligned} \mu_1^* &= \mathbf{f}_0^t \boldsymbol{\mu}_{\boldsymbol{\beta}|n}^* + \mathbf{r}_0^t \mathbf{R}^{-1}\big(\mathbf{y} - \mathbf{F}\boldsymbol{\mu}_{\boldsymbol{\beta}|n}^*\big); \\ \boldsymbol{\mu}_{\boldsymbol{\beta}|n}^* &= (\mathbf{F}^t\mathbf{R}^{-1}\mathbf{F} + v^{-1}\mathbf{I}_l)^{-1}(\mathbf{F}^t\mathbf{R}^{-1}\mathbf{y} + v^{-1}\boldsymbol{\mu}); \\ \sigma_1^2 &= \frac{Q_1^2}{n + 2a}\left\{1 - (\mathbf{f}_0^t, \mathbf{r}_0^t)\begin{bmatrix} -v^{-1}\mathbf{I}_l & \mathbf{F}^t \\ \mathbf{F} & \mathbf{R} \end{bmatrix}^{-1} \begin{pmatrix} c\mathbf{f}_0 \\ \mathbf{r}_0 \end{pmatrix}\right\}; \\ Q_1^2 &= (b/a)^{1/2} + \mathbf{y}^t[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{F}(\mathbf{F}^t\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^t\mathbf{R}^{-1}]\mathbf{y} \\ &\quad + (\boldsymbol{\mu} - \widehat{\boldsymbol{\beta}})^t[v\mathbf{I}_l + (\mathbf{F}^t\mathbf{R}^{-1}\mathbf{F})^{-1}]^{-1}(\boldsymbol{\mu} - \widehat{\boldsymbol{\beta}}); \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{F}^t\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^t\mathbf{R}^{-1}\mathbf{y}; \end{aligned}$$

$\mathbf{f}_0 = \mathbf{f}(\mathbf{w}_0)$; $\mathbf{r}_0 = (\mathrm{cor}(y(\mathbf{w}_0), y(\mathbf{w}_1^0)), \ldots, \mathrm{cor}(y(\mathbf{w}_0), y(\mathbf{w}_n^0)))^t$; $\mathbf{R}$ is the correlation matrix with entries $\mathrm{cor}(y(\mathbf{w}_i^0), y(\mathbf{w}_j^0))$ for $i, j = 1, \ldots, n$; and $\mathbf{F} = (\mathbf{f}(\mathbf{w}_1^0), \ldots, \mathbf{f}(\mathbf{w}_n^0))^t$.

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

$$\begin{aligned} p(y(\mathbf{w}_0)|\mathbf{y}) &= \int p[y(\mathbf{w}_0), \boldsymbol{\phi}, \mathbf{T}|\mathbf{y}]\, d\boldsymbol{\phi}\, d\mathbf{T} \\ &= \int p[y(\mathbf{w}_0)|\mathbf{y}, \boldsymbol{\phi}, \mathbf{T}] p(\boldsymbol{\phi}, \mathbf{T}|\mathbf{y})\, d\boldsymbol{\phi}\, d\mathbf{T}. \end{aligned} \tag{24}$$

Here $p(\boldsymbol{\phi}, \mathbf{T}|\mathbf{y})$ is given in (23). The integration in (24) can be computationally prohibitive. For example, for six quantitative factors and five qualitative factors, each with four levels, $\boldsymbol{\phi}$ would be a six-dimensional vector and $\mathbf{T}$ would be five 4×4 matrixes. 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 1 depicts the two curves of the function values with $z_1 = 1$ and 2. The overall similarity of the curves suggests that the independent analysis is insufficient for this example; that the integrated analysis can better exploit the common information in the two curves, and is expected to perform better.

Table 1 lists the training data used for model building, including two six-run Latin hypercube samples of x_1, one for $z_1 = 1$ and the other for $z_1 = 2$. For comparison, we analyzed the data using both methods. For the independent analysis, we fitted two separate GP models with means μ_1 and μ_2, one for $z_1 = 1$ and the other for $z_1 = 2$, using the correlation function (2) with $I = 1$ and $p = 2$; the estimated parameters are given in Table 2. For the integrated analysis, we fitted 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 an exchangeable nature, and thus we used (7).

Next, we assessed the prediction accuracy of the two methods. The testing data comprised 40 data points generated as follows. For $z_1 = 1$ and 2, x_1 took 20 equally spaced values .025, .075, ..., .975 in [0, 1]. The root mean squared errors (RMSEs) for the two prediction methods were calculated. The RMSE for the integrated analysis was 1.03, 15% smaller than the RMSE of 1.21 for the independent analysis. This indicates that the average prediction accuracy of the integrated analysis was better than that of the independent analysis. For the integrated analysis, we also fitted a GP with the process mean μ taking two different values for $z_1 = 1$ and 2; the average prediction accuracy was nearly the same as that of the GP model with a constant mean for $z_1 = 1$ and 2.

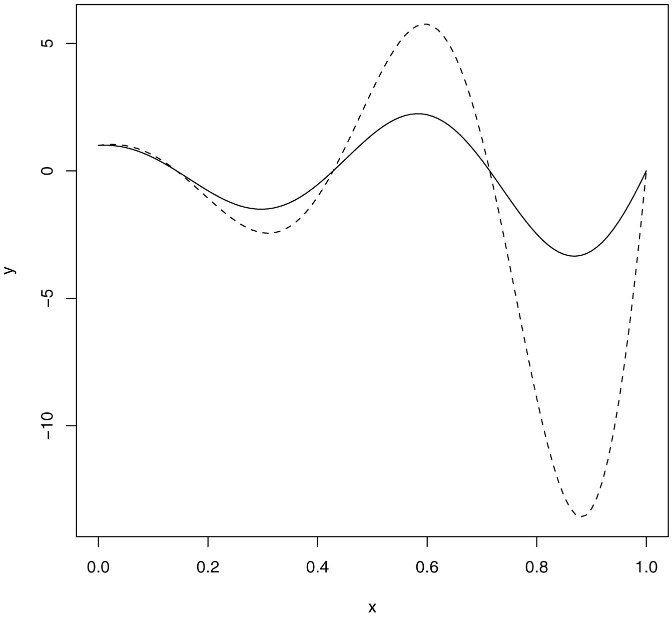

Figure 1. Two curves of the function values with $z_1 = 1$ (—) and 2 (- - -).

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

$z_1 = 1$	x_1	.1232	.2969	.4999	.6179	.7614	.9950
	y	.2548	−1.5039	1.4222	2.0718	−1.4378	−.2213
$z_1 = 2$	x_1	.1136	.3270	.3433	.6119	.7778	.9431
	y	.4446	−2.3970	−2.2578	5.6589	−6.6310	−9.9167

7. A DATA CENTER COMPUTER EXPERIMENT

In this section we illustrate the proposed method using a data center computer experiment from the information technology 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 applications, such as web-hosting Internet, intranet, and telecommunication. 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 a large amount of computer equipment constantly generating heat. Monitoring and studying the temperature of a data center is a difficult task, because how different configurations affect the thermal distribution of the data center is largely unknown. The physical thermal process is complex, depending on many factors, and detailed temperatures at different locations cannot be actually measured. Computer experiments, built on computational fluid-dynamics (CFD) models and implemented in professional software like Flotherm (Flometrics 2005) and FLUENT (Fluent 1998), often are used as proxies to study the air movement and thermal distribution of a data center. More details for the engineering background of data centers have been given by Schmidt (2003) and Schmidt et al. (2005).

The experiment considered in this section modeled an air-cooled cabinet and was implemented in Flotherm for predicting the airflow and heat transfer in the electronic equipment. Each run in this experiment took several days to complete. This example has eight configuration variables. Five of these variables are quantitative factors, denoted by x_1, x_2, x_3, x_4, and x_5, which have 3, 5, 2, 3, and 3 levels. The remaining three variables are 2-, 4-, and 3-level qualitative factors, denoted by z_1, z_2, and z_3. (For confidentiality reasons, we do not provide descriptions or the corresponding names of these eight factors.) The response

Table 2. Estimated parameters of the GP models for the independent analysis

	$\widehat{\phi}_1$	$\widehat{\sigma}^2$	$\widehat{\mu}$
$z_1 = 1$	115.94	1.73	−.002
$z_1 = 2$	25.65	30.16	−2.09

Table 3. Estimated parameters of the GP model for the integrated analysis

$\widehat{\phi}_1$	$\widehat{\theta}_1$	$\widehat{\sigma}^2$	$\widehat{\mu}$
27.48	20.00	16.76	−1.07

Table 5. 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 for the data center example

$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	$\widehat{\phi}_4$	$\widehat{\phi}_5$	$\widehat{\tau}_1$
5.35	1.07	7.71	3.36	1.45	.005

of interest, denoted by y, is the temperature at one selected location of the system.

The five quantitative factors were of distinct scales, and their values were standardized first. The standardization of each variable was 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 had 73 observations. (What design was used for the experiment is not clear, but it was not a space-filling design.) Six of the original observations were removed because of unsuccessful tuning and convergence checking of the CFD algorithms after confirmation from the data center scientists. The subsequent analysis used the remaining 67 observations. There were 24 level combinations for the three qualitative factors. Thus, on average, each of these combinations had fewer than three observations, making the independent analysis infeasible. We analyzed the data using the integrated analysis.

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

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

Here the baseline 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 $\boldsymbol{\beta}$ is

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

The major difficulty with the model fitting is estimating the correlation matrices for z_2 and z_3. The estimation was carried out using the iterative procedure in Section 5, implemented in Matlab (The MathWorks 2006) and making use of a semidefinite programming package CVX (Grant, Boyd, and Ye 2006). Note that this data set does not have a cross-array structure, and that 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.

Table 5 presents the estimated correlation parameters for the quantitative factors x_1, x_2, x_3, x_4, and x_5 along with the estimated correlation for 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 those for the others. The estimated correlation for z_1 (between its two levels) is small (.005), indicating that the responses at the two levels of z_1 are not significantly correlated. This is consistent with the known physics that the two levels of z_1 correspond to distinct data center thermal distributions.

Tables 6 and 7 give the estimated correlation matrixes for z_2 and z_3. Both matrixes are symmetric with unit diagonal elements. Also, their eigenvalues are all positive [(3.11, .58, .17, .14) and (2.38, .45, .17)]. Thus the estimated correlation matrixes are PDUDEs and are indeed valid correlation matrixes.

Following Welch et al. (1992), we plotted the estimated main effects and two-factor interactions. Figure 2 depicts the main-effect functions of the quantitative factors x_1, x_2, x_3, x_4, and x_5; Table 8 lists the main effects of the qualitative factors z_1, z_2, and z_3. Note that the response differs substantially as x_1 is varied from its lower to upper bound, and that the response is quite different at the first two levels of z_3. Figure 3 displays the two-factor interaction plots for some selected pairs of the quantitative factors. Note the large interactions and complex interaction patterns of (x_1, x_2) and (x_1, x_4). Figure 4 shows some two-factor interaction functions between the quantitative and qualitative factors. As the figure shows, such interactions are rather intricate; for example, the response is quite different at the first and fourth levels of z_3 when x_3 is near its lower and upper bounds. As x_2 is varied from its lower bound to its upper bound, the response profiles are similar at the first three levels of z_2 but have a very different pattern at the fourth level of z_2. The observed complex second-order relationships cannot be captured by standard quadratic models.

To assess the prediction accuracy of the fitted GP model, we performed a leave-one-out cross-validation, the same approach used by Welch et al. (1992). Let $\widehat{y}_{-i}(\mathbf{w}_i^0)$ denote the empirical BLUP in (22) of $y(\mathbf{w}_i^0)$ based on all of the data except the observation $y(\mathbf{w}_i^0)$. The cross-validation version of the RMSE was $(\sum_{i=1}^{67}[\widehat{y}_{-i}(\mathbf{w}_i^0) - y(\mathbf{w}_i^0)]^2/67)^{1/2} = 1.88$ (relative to a data range of 6.37–22.08). As done by Welch et al. (1992), to minimize computation, the estimates of the correlation parameters and correlation matrixes were not recomputed for each prediction, but instead were still based on all 67 data points. (Recomputing them for each prediction would be very time-consuming; running the algorithm with 400 iterations would take more than 3 hours in a double-core PC running a Linux system.) The plot

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

$\widehat{\eta}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\alpha}_{12}$	$\widehat{\alpha}_{22}$	$\widehat{\alpha}_{23}$	$\widehat{\alpha}_{24}$	$\widehat{\alpha}_{32}$	$\widehat{\alpha}_{33}$	$\widehat{\sigma}^2$
11.95	6.17	−2.77	3.05	−4.53	.20	.08	−.95	−.72	−1.73	2.66	1.27	2.85

Table 6. Estimated correlation matrix for z_2 for the data center example

1.00	.84	.78	.50
.84	1.00	.82	.54
.78	.82	1.00	.71
.50	.54	.71	1.00

Table 7. Estimated correlation matrix for z_3 for the data center example

1.00	.62	.83
.62	1.00	.61
.83	.61	1.00

of $\widehat{y}_{-i}(\mathbf{w}_i^0)$ in Figure 5 against $y(\mathbf{w}_i^0)$ demonstrates the predictor's decent accuracy. As the figure shows, the prediction accuracy deteriorates moderately when the responses are near the two ends. This is understandable, because the design set for the input values was not a space-filling design, and fewer observations were available for large or small response values.

8. DISCUSSION AND CONCLUDING REMARKS

Beginning with the work of Sacks et al. (1989), GP models have enjoyed great popularity in computer modeling. An important, but still unsettled problem is how to model computer experiments with both qualitative and quantitative factors. Here we have presented a systematic treatment of building GP models with both types of factors. Our proposed methodology makes two major contributions: It is a general method for constructing correlation functions with qualitative and quantitative factors that makes use of some underlying multivariate Gaussian processes, and it is an iterative procedure for estimation. The validity of the constructed correlation functions in the estimation is ensured by some recently developed optimization techniques. The proposed method has been successfully applied to an example involving a known function and a real example for modeling the thermal distribution of a data center.

We also have discussed and proposed some restrictive correlation functions for qualitative factors that may be justifiable in particular applications. In such cases, we have shown that the estimation procedure can be simplified significantly. 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 have focused on the maximum likelihood method for estimation. Although this method is widely used in computer experiments (Sacks et al. 1989; Welch et al. 1992), it has some drawbacks; for example, it sometimes may be difficult to obtain a global maximum of the likelihood, the likelihood function near the optimum may be flat, and the likelihood surface may be difficult to assess or visualize because the parameters include correlation matrixes. Some methods have been proposed 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 a penalized likelihood method to deal with the flatness of the likelihood function near the optimum. Handcock and Stein (1993) and Handcock, Meier, and Nychka (1994) proposed methods for studying the surface of the likelihood function. We plan to explore the general issue

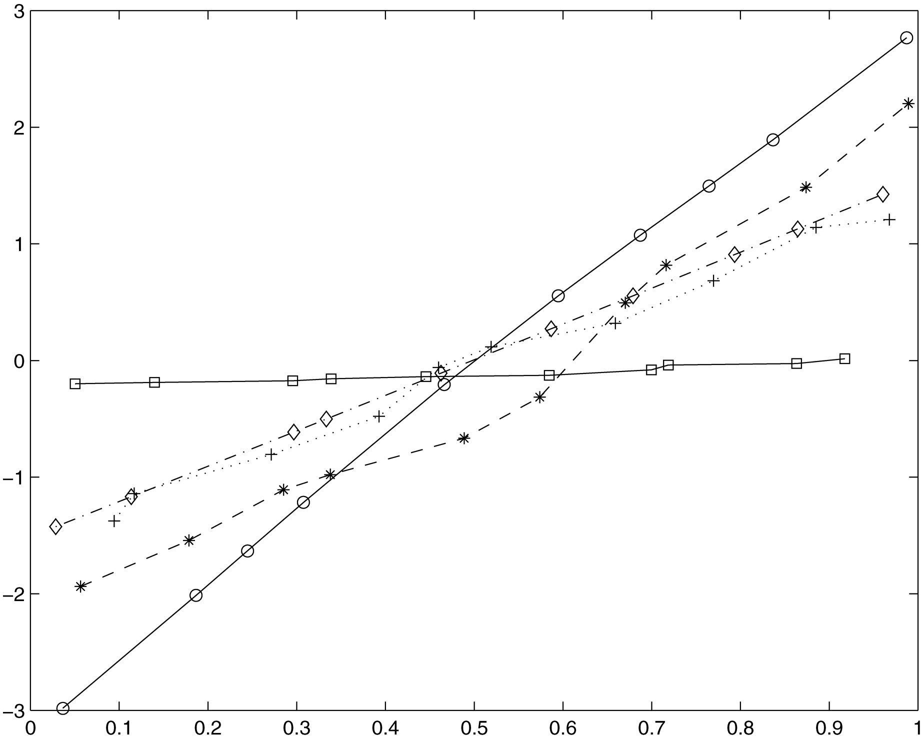

Figure 2. Main-effect functions of x_1 (—○—), x_2 (· · + · ·), x_3 (· −◇− ·), x_4 (− ∗ −), and x_5 (—□—) for the data center example.

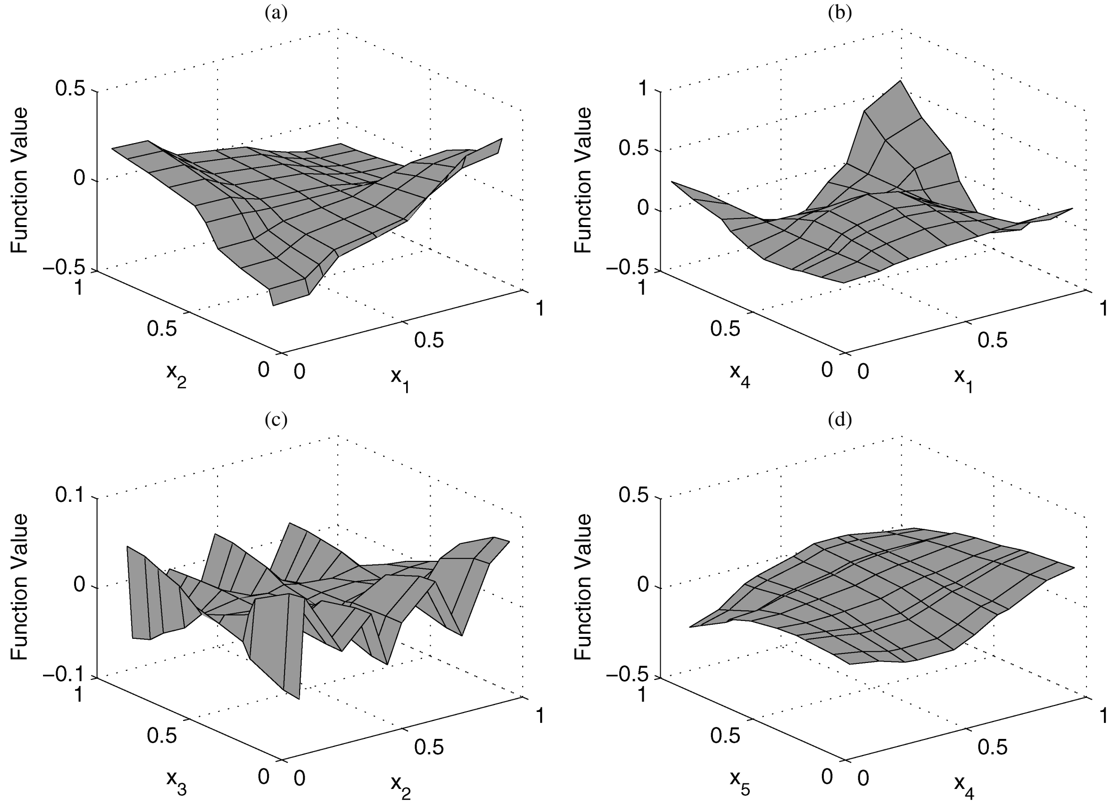

Figure 3. Two-factor interaction functions for some selected pairs of x_1, x_2, x_3, x_4, and x_5 for the data center example. (a) Interaction of x_1 and x_2. (b) Interaction of x_1 and x_4. (c) Interaction of x_2 and x_3. (d) Interaction of x_4 and x_5.

of visualizing and assessing a function of correlation matrixes in a separate research effort. As an alternative to the maximum likelihood method, we may use Bayesian methods for the modeling and estimation. We have briefly discussed such methods and the related computational challenges in this article; further work on the Bayesian methods will be developed and reported elsewhere.

ACKNOWLEDGMENTS

P. Qian is grateful to Dr. Zhaosong Lu and Dr. Renato D. C. Monteiro for useful discussions on some optimization issues. The authors thank the editors, an associate editor, and two referees for their valuable comments and suggestions that significantly improved this article. Qian's research was supported by National Science Foundation (NSF) grant DMS-0705206; C. F. J. Wu's research was supported by NSF grants DMI-0620259 and DMS-0705261. H. Wu's research was supported by Iowa State University's Faculty Professional Development Assignment Program and NSF DMI-0620259 for his visit at Georgia Tech in the fall of 2005.

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

z_1	Level 1: −.14	level 2: .0061		
z_2	Level 1: .80	level 2: −.15	level 3: .08	level 4: −.87
z_3	Level 1: −1.33	level 2: 1.39	level 3: −.09	

APPENDIX A: DEFINITIONS AND FORMULAS FOR $\partial \mathrm{tr}(\mathbf{ER}^{-1})/\partial \mathbf{T}_j$ AND $\partial|\mathbf{R}|/\partial \mathbf{T}_j$

The definitions and results here follow from the work of Graham (1981, chap. 4).

1. Define $\partial\,\mathrm{tr}(\mathbf{ER}^{-1})/\partial\mathbf{T}_j$ as

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

For $1 \le r \le m_j$ and $1 \le s \le m_j$, it is clear that $\partial\,\mathrm{tr}(\mathbf{ER}^{-1})/\partial\tau_{j,r,s} = \mathrm{tr}(\partial(\mathbf{ER}^{-1})/\partial\tau_{j,r,s})$. Furthermore, $\mathrm{tr}(\partial(\mathbf{ER}^{-1})/\partial\tau_{j,r,s}) = \mathrm{tr}(\mathbf{E}(\partial\mathbf{R}^{-1}/\partial\tau_{j,r,s})) = \mathrm{tr}(-\mathbf{ER}^{-1}(\partial\mathbf{R}/\partial\tau_{j,r,s})\mathbf{R}^{-1})$.

2. Define $\partial|\mathbf{R}|/\partial\mathbf{T}_j$ 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 entry of $\mathbf{R}$, and let $\mathbf{R}_{uv}$ be the cofactor of ρ_{uv} in $|\mathbf{R}|$. Then for $1 \le r \le m_j$ and $1 \le s \le m_j$,

$$\frac{\partial|\mathbf{R}|}{\partial\tau_{j,r,s}} = \mathrm{tr}(\mathbf{AB}^t_{jrs}),$$

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

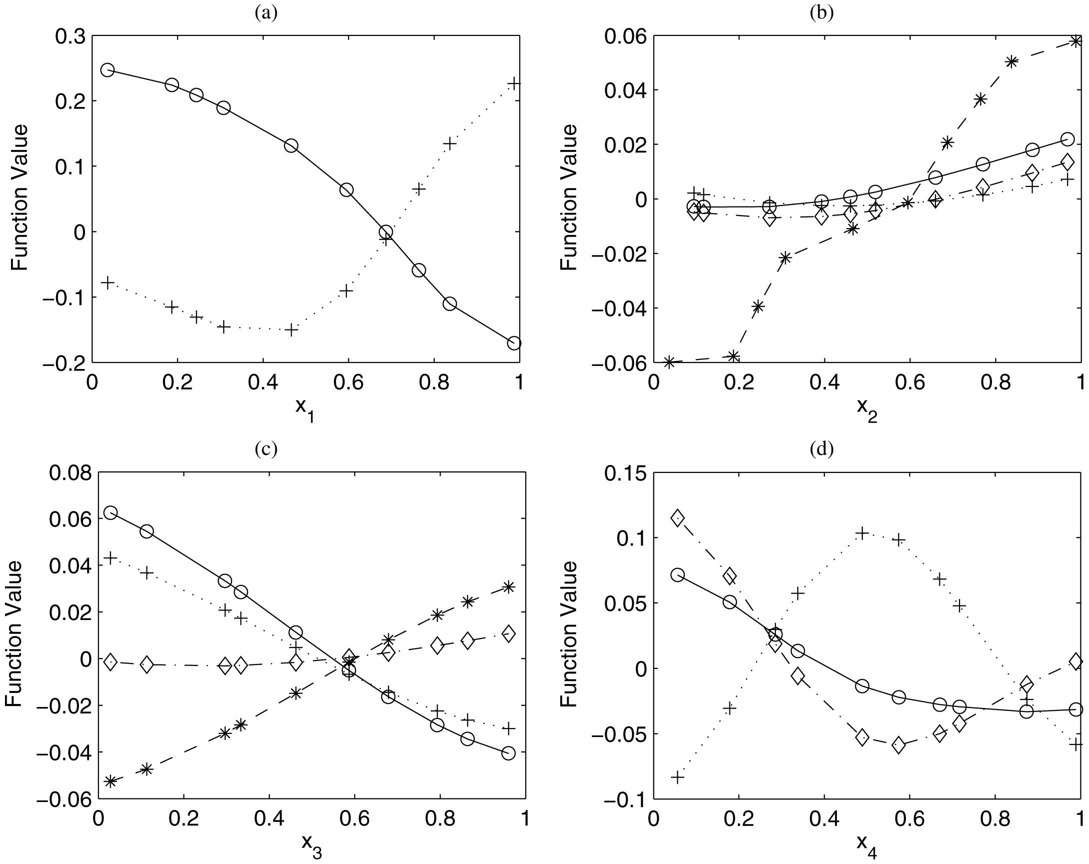

Figure 4. Two-factor interaction functions for some selected pairs of quantitative and qualitative factors for the data center example. (a) Interaction of x_1 and z_1. (b) Interaction of x_2 and z_2. (c) Interaction of x_3 and z_2. (d) Interaction of x_4 and z_3. The solid lines (with $\circ$'s) correspond to responses at the first levels of z_1, z_2, and z_3; $+$'s, to responses at the second levels of z_1, z_2, and z_3; $\diamond$'s, to responses at the third levels of z_2 and z_3; $*$'s, to responses at the fourth level of z_2.

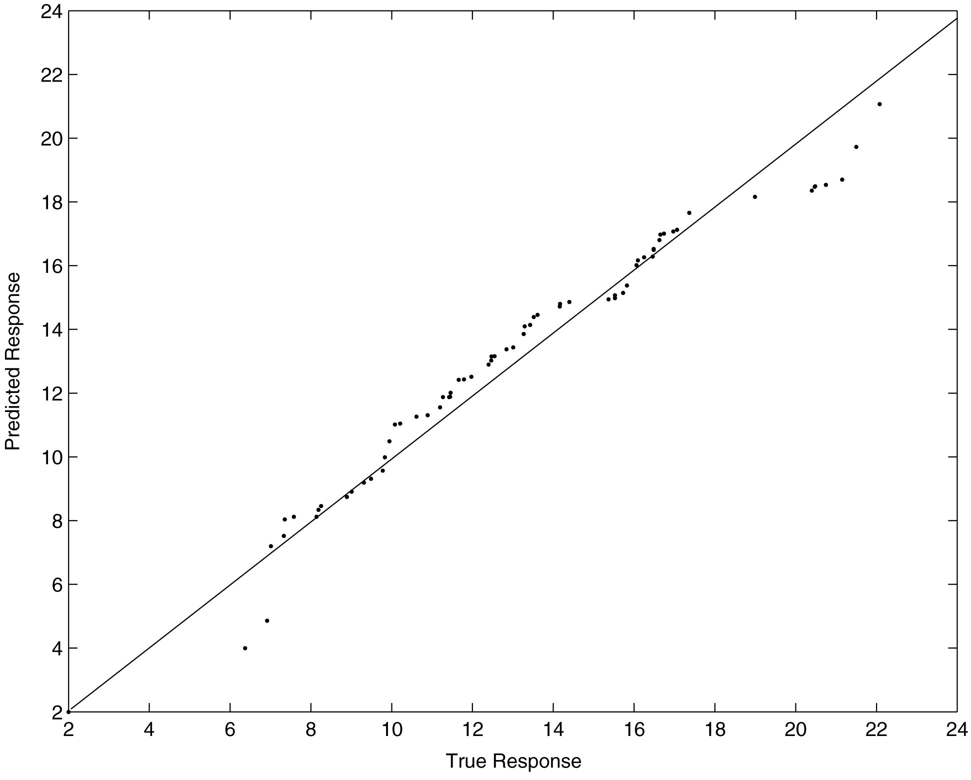

Figure 5. Predicted responses using cross-validation versus the true responses for the data center example.

APPENDIX B: PROOF OF PROPOSITION 1

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

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

and

$$\operatorname{tr}(\mathbf{E}\mathbf{R}^{-1}) = \operatorname{tr}(\mathbf{E}\mathbf{H}^{-1} \otimes \mathbf{T}_1^{-1}) = \operatorname{tr}(\mathbf{E}\mathbf{H}^{-1}) \operatorname{tr}(\mathbf{T}_1^{-1}).$$

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

[Received June 2006. Revised August 2007.]

REFERENCES

Abrahamsen, P. (1997), "A Review of Gaussian Random Fields and Correlation Functions," Report 917, Norwegian Computer Center, available at *http://publications.nr.no/917Rapport.pdf*.

Banerjee, S., and Gelfand, A. E. (2002), "Prediction, Interpolation and Regression for Spatially Misaligned Data," *Sankhya*, 64, 227–245.

Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, London: Arnold.

Bertsekas, D. P. (1999), *Nonlinear Programming*, Nashua, NH: Athena Scientific.

Boscardin, W. J., and Zhang, X. (2004), "Modeling the Covariance and Correlation Matrix of Repeated Measures," in *Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspectives*, eds. A. Gelman and X.-L. Meng, New York: Wiley, pp. 215–226.

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden–Day.

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.

Dattorro, J. (2005), *Convex Optimization and Euclidean Distance Geometry*, Palo Alto, CA: Meboo Publishing.

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*, Ser. B, 39, 1–38.

Edwards, D. M. (2000), *Introduction to Graphical Modeling*, New York: Springer.

Fang, K. F., Li, R. Z., and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, New York: Chapman & Hall/CRC Press.

Flometrics (2005), "Flotherm: Fluid Dynamics–Based Thermal Analysis Software," available at *http://www.flomerics.com/*.

Fluent, Inc. (1998), FLUENT, Release 5.5.14 (3d, segregated, laminar), Lebanon: Author.

Fridlyander, I. N. (2002), "Modern Aluminum and Magnesium Alloys and Composite Materials Based on Them," *Metal Science and Heat Treatment*, 44, 292–296.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC Press.

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computation*, Baltimore: The Johns Hopkins University Press.

Graham, A. (1981), *Kronecker Products and Matrix Calculus With Applications*, Chichester, U.K.: Ellis Horwood Limited.

Grant, M., Boyd, S., and Ye, Y. (2006), *CVX: Matlab Software for Disciplined Convex Programming*, version 1.0 beta 3, Dept. of Managment Science and Engineering, Stanford University, available at *http://www.stanford.edu/~boyd/cvx/*.

Han, G., Santner, T. J., Notz, W. I., and Bartel, D. L. (2007), "Prediction for Computer Experiments Having Quantitative and Qualitative Input Variables," Dept. of Statistics, Ohio State University, *http://www.stat.osu.edu/~hangang/hqqv_09_05.pdf*.

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410.

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.

Joseph, V. R., and Delaney, J. D. (2007), "Functionally Induced Priors for the Analysis of Experiments," *Technometrics*, 49, 1–11.

Katz, M. H. (2006), *Multivariable Analysis: A Practical Guide for Clinicians*, Cambridge, U.K.: Cambridge University Press.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, U.K.: Clarendon Press.

Li, R., and Sudjianto, A. (2005), "Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models," *Technometrics*, 47, 111–120.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer.

Mardia, K. V., and Goodall, C. R. (1993), "Spatial–Temporal Analysis of Multivariate Environmental Monitoring Data," in *Multivariate Environmental Statistics*, eds. G. P. Patil and C. R. Rao, Amsterdam: Elsevier, pp. 347–386.

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.

Nesterov, Y., and Nemirovskii, A. (1994), "Intenor-Point Polynomial Algorithms in Convex Programming," in *SIAM Studies in Applied Mathematics, 13*, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Ruszczynski, A., and Shapiro, A. (eds.) (2003), *Stochastic Programming*, Amsterdam: Elsevier.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–435.

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

Schmidt, R. R. (2003), "Hot Spots in Data Centers," Online Forum of Electrical Cooling, available at *http://www.electronics-cooling.com*.

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.

Stein, M. L. (1999), *Interpolation of Spatial Data*, New York: Springer.

The MathWorks, Inc. (2006), *Matlab: The Language of Technical Computing*, version 6.5.1, Natick, MA: Author.

Vandenberghe, L., and Boyd, S. (1996), "Semidefinite Programming," *SIAM Review*, 38, 49–95.

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.

Wolkowicz, H., Saigal, R., and Vandenberghe, L. (eds.) (2000), *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, Boston: Kluwer Academic.

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.

Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.