EXPERIMENTAL DESIGNS FOR ESTIMATION OF HYPERPARAMETERS IN HIERARCHICAL LINEAR MODELS ⋆

Qing Liu

Department of Statistics, The Ohio State University

Angela M. Dean *

Department of Statistics, The Ohio State University

Greg M. Allenby

Department of Marketing, The Ohio State University

Abstract

Optimal design for the joint estimation of the mean and covariance matrix of the random effects in hierarchical linear models is discussed. A criterion is derived under a Bayesian formulation which requires the integration over the prior distribution of the covariance matrix of the random effects. A theoretical optimal design structure is obtained for the situation of independent and homoscedastic random effects. For both the situation of independent and heteroscedastic random effects and that of correlated random effects, optimal designs are obtained through computer search. It is shown that orthogonal designs, if they exist, are optimal when the random effects are believed to be independent. When the random effects are believed to be correlated, it is shown by example that nonorthogonal designs tend to be more efficient than orthogonal designs. In addition, design robustness is studied under various prior mean specifications of the random effects covariance matrix.

Key words:
Bayesian Design, Optimal Design, Hierarchical Linear Model, Hyperparameter, Random Effects Model

1 Introduction

Hierarchical models are also known as "multi-level models", "mixed-effects models", "random-effects models", "population models", "random coefficient regression models" and "covariance components models" (see Raudenbush and Bryk, 2002). These models have been applied in a wide variety of fields including the social and behavioral sciences, agriculture, education, medicine, healthcare studies, and marketing. For example,in educational research, data may contain repeated measurements over time for each individual within an institution, and hierarchical models have been used to analyze individuals' learning curves over time, and to discover how the learning rate is affected by individual characteristics, and how the effects are influenced by institutional characteristics (see, for example, Raudenbush, 1993; Draper, 1995; Goldstein, 2003). In marketing research, hierarchical models are often the models of choice for marketing studies in which the learning of effect sizes and the determination of conditions for maximization or minimization of effect sizes are of importance (see, for example, Allenby and Lenk, 1994; Bradlow and Rao, 2000; Montgomery et al., 2004). In pharmacokinetics, toxicokinetics and pharmacodynamics, hierarchical models are often used to describe the characteristics of a whole population while taking into consideration the heterogeneity among subjects (see Yuh et al., 1994, for a bibliography).

⋆ Supported in part by NSF Grant SES-0437251
* Corresponding Author.
Email addresses: liu.124@osu.edu (Qing Liu), dean.9@osu.edu (Angela M. Dean), allenby.1@osu.edu (Greg M. Allenby).

Hierarchical models, linear or nonlinear, often consist of two levels, where parameters in the first-level of the hierarchy reflect individual-level effects, which are assumed to be random effects and distributed according to a probability distribution characterized by the "hyperparameters" in the second-level of the hierarchy. Hyperparameters capture the variation of the individual-level random effects, and the mean of the random effects when there are no covariates in the second-level of the model, or how the covariates drive the sizes of the individual-level effects when there are covariates.

Experimental designs for efficient estimation of the individual-level random effects have been investigated in the literature under hierarchical models (See, for example, Smith and Verdinelli, 1980; Arora and Huber, 2001; Sándor and Wedel, 2001, 2005; Kessels et al., 2006). While it is important to have accurate information on the individual-level effects in situations such as direct marketing, which focuses on individual customization of products, in other situations, which focuses on population characteristics or prediction to new contexts, accurate information on hyperparameters is important. These situations include those in pharmacokinetics where the mean and/or covariance matrix of the individual-level random effects ("population parameters") are of interest, or in situations where predictions of consumer preferences in a new target population are required using information on covariates.

A few researchers have proposed pragmatic approaches to finding efficient designs for the estimation of hyperparameters under a hierarchical nonlinear model. For example, the swapping, relabeling and cycling heuristic by Sándor and Wedel (2002); the linearization approach by Mentré et al. (1997); the stochastic gradient search by Tod et al. (1998), and the "MCMC nested within Monte Carlo" approach by Han and Chaloner (2004).

Under a hierarchical linear model, Fedorov and Hackl (1997, pg. 78), Entholzner et al. (2005), and Liu et al. (2007) studied optimal designs for the

estimation of hyperparameters that capture the mean of the individual-level random effects. For the joint estimation of both the mean and the variance of the independent and identically distributed individual-level random effects, Lenk et al. (1996) analytically investigated, in the survey setting, the trade-off between the number of subjects and the number of questions per subject under a cost constraint and an orthogonal design structure.

In this paper, we focus on experimental designs for hyperparameter estimation under hierarchical linear models. Building on the results obtained by Liu et al. (2007) where the primary interest is on the estimation of the mean of the individual-level random effects, we make the extension here and investigate optimal designs where the interest is on the joint estimation of both the mean and the covariance matrix of the individual-level random effects. We derive a design criterion for both the situation of independent random effects and that of correlated random effects. We prove that, orthogonal designs, if they exist, are optimal when the random effects are believed to be independent and homoscedastic. When the random effects are independent but heteroscedastic, or the random effects are correlated, we obtain efficient designs through computer search and show by example that nonorthogonal designs tend to be superior to orthogonal designs.

This paper is organized as follows. In Section 2, we introduce a hierarchical Bayesian linear model and derive a design criterion under the Bayesian formulation. An optimal design is dependent upon the prior probability distribution of the unknown random effects covariance matrix. In Section 3, we examine the situation when the random effects are believed to be independent. We obtain the theoretical optimal design structure for the situation of independent and homoscedastic random effects, and use computer search to obtain optimal designs for the situation of independent and heteroscedastic random effects. In both cases, orthogonal designs, if they exist, are found to be optimal. In Section 4, we focus on the situation when the random effects are believed to

be correlated. We show by example that nonorthogonal designs tend to be more efficient than orthogonal designs. In addition, we investigate design robustness to different prior mean specifications of the random effect covariance matrix and make a recommendation for the specification of the prior mean in the search for optimal designs. A summary and conclusions are provided in Section 5.

2 Optimal designs for hyperparameter estimation

Consider a consumer survey in which respondent i is given a set of m_i questions ($i = 1, \ldots, n$). The questions contain information on various levels of marketing variables (treatment factors), such as price, product attributes or possibly aspects of advertisements. Suppose treatment factor k contains h_k levels ($k = 1, \ldots, t$) and only main effects are considered. The model matrix $\mathbf{X}_i$ includes a column of ones that corresponds to the general mean and $h_k - 1$ columns that correspond to the coefficients of contrasts for factor k, $k = 1, \ldots, K$. Thus the design matrix $\mathbf{X}_i$ is of size $m_i \times p$ where $p = 1 + \sum_{k=1}^{K}(h_k - 1)$. The responses of subject i to the set of questions are represented by the vector $\mathbf{y}_i$ of length m_i. The effects of the variables on respondent i are captured by the p elements in vector $\boldsymbol{\beta}_i$, which are assumed to be random effects that are distributed according to a multivariate normal distribution with $p \times p$ variance-covariance matrix $\boldsymbol{\Lambda}$, and mean $\mathbf{Z}_i\boldsymbol{\theta}$ where $\mathbf{Z}_i$ is a matrix $(p \times q)$ of covariates, such as household income or age, and $\boldsymbol{\theta}$ is a parameter vector of length q. Thus, the hierarchical linear model is of the following form:

$$\mathbf{y}_i|\boldsymbol{\beta}_i, \sigma^2 = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

$$\boldsymbol{\beta}_i|\boldsymbol{\theta}, \boldsymbol{\Lambda} = \mathbf{Z}_i\boldsymbol{\theta} + \boldsymbol{\delta}_i \tag{2.2}$$

The error vector $\boldsymbol{\epsilon}_i$ of length m_i in the first level of the hierarchy captures consumer i's response variability to the set of questions, and is assumed to have a multivariate normal distribution with mean vector $\mathbf{0}$ of size m_i and variance-covariance matrix $\sigma^2\mathbf{I}_{m_i}$ if the response errors are believed to be homoscedastic. The error vector $\boldsymbol{\delta}_i$ of length p in the second level of the hierarchy captures the variation of individual-level effects $\boldsymbol{\beta}_i$, and is assumed to be multivariate normal with mean vector $\mathbf{0}$ of size p and variance-covariance matrix $\boldsymbol{\Lambda}$ of size $p \times p$. When the prior knowledge at the second level is weak, the following priors are usually assumed for $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ (see, for example, Rossi et al., 2005).

$$\boldsymbol{\theta} \sim \text{Normal}(\mathbf{0}_q, 100\mathbf{I}_q), \tag{2.3}$$

$$\boldsymbol{\Lambda} \sim \text{Inverted Wishart}(\nu_0 = p + 3, \nu_0\mathbf{I}_p), \tag{2.4}$$

These are replaced by more informative priors when information is available. In this paper, we consider the estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ given known σ^2. For example, a retailer is interested in learning about the mean consumer preference and the dispersion of individual consumer preferences. The two layers, (2.1) and (2.2), of the hierarchical model can be combined to obtain

$$\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\Lambda} \sim N_{m_i}(\mathbf{X}_i\mathbf{Z}_i\boldsymbol{\theta},\ \boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{m_i} + \mathbf{X}_i\boldsymbol{\Lambda}\mathbf{X}_i'), \tag{2.5}$$

(see Lenk et al., 1996, pg 187), with proper priors (2.3) and (2.4) assumed for $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$.

Let $\mathcal{D}(m_1, \ldots, m_n)$ be a class of designs $\tilde{d} = (d_1, \ldots, d_n) \in \mathcal{D}(m_1, \ldots, m_n)$, where d_i is the m_i-point sub-design allocated to subject i. For a given $\tilde{d} = (d_1, \ldots, d_n)$, let $\widetilde{\mathbf{X}}' = (\mathbf{X}_1', \ldots, \mathbf{X}_n')$ where $\mathbf{X}_i$ of size $m_i \times p$ is the corresponding model matrix of d_i. Following Chaloner and Verdinelli (1995, page 277), we seek an optimal design $\tilde{d}^*$ in $\mathcal{D}(m_1, \ldots, m_n)$ that maximizes the expected gain in Shannon Information, that is, we seek a design that gives maximum

$$\int \left[\log p(\boldsymbol{\theta}, \boldsymbol{\Lambda}|\mathbf{y}, \widetilde{\mathbf{X}})\right] p(\boldsymbol{\theta}, \boldsymbol{\Lambda}|\mathbf{y}, \widetilde{\mathbf{X}}) p(\mathbf{y}|\widetilde{\mathbf{X}}) d\boldsymbol{\theta} d\boldsymbol{\Lambda} d\mathbf{y}, \tag{2.6}$$

where $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ and where $p(\boldsymbol{\theta}, \boldsymbol{\Lambda}|\mathbf{y}, \widetilde{\mathbf{X}})$ is

$$p(\boldsymbol{\theta}, \boldsymbol{\Lambda}|\mathbf{y}, \widetilde{\mathbf{X}}) = \frac{p(\mathbf{y}|\widetilde{\mathbf{X}}, \boldsymbol{\theta}, \boldsymbol{\Lambda})p(\boldsymbol{\theta})p(\boldsymbol{\Lambda})}{\int p(\mathbf{y}|\widetilde{\mathbf{X}}, \boldsymbol{\theta}, \boldsymbol{\Lambda})p(\boldsymbol{\theta})p(\boldsymbol{\Lambda})d\boldsymbol{\theta}d\boldsymbol{\Lambda}}\,. \tag{2.7}$$

Since (2.7) is not of closed-form, a normal approximation is used as follows. First, let $\boldsymbol{\zeta}$ be the vector that includes all the p parameters in $\boldsymbol{\theta}$ and the $p(p+1)/2$ parameters in $\boldsymbol{\Lambda}$, then according to Berger (1985, page 224, (iv)), $\boldsymbol{\zeta}|\mathbf{y}, \widetilde{\mathbf{X}}$ has the following approximate distribution

$$\boldsymbol{\zeta}|\mathbf{y}, \widetilde{\mathbf{X}} \;\sim\; N(\hat{\boldsymbol{\zeta}}, \mathbf{I}(\hat{\boldsymbol{\zeta}})^{-1}), \tag{2.8}$$

where $\mathbf{I}(\hat{\boldsymbol{\zeta}})$ is the expected Fisher information matrix evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\zeta}}$. Now partition $\mathbf{I}(\boldsymbol{\zeta})$ as

$$\mathbf{I}(\boldsymbol{\zeta}) = \begin{pmatrix} FI(\boldsymbol{\theta}, \boldsymbol{\theta}) & FI(\boldsymbol{\theta}, \boldsymbol{\Lambda}) \\ FI(\boldsymbol{\theta}, \boldsymbol{\Lambda}) & FI(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}) \end{pmatrix}. \tag{2.9}$$

Let $[\boldsymbol{\Lambda}]_{uv} = \lambda_{u,v}$ denote the $(u,v)^{th}$ element of $\boldsymbol{\Lambda}$. Then, as shown by Lenk et al. (1996),

$$FI(\boldsymbol{\theta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{Z}_i'\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\mathbf{Z}_i \quad \text{where} \quad \boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{m_i} + \mathbf{X}_i\boldsymbol{\Lambda}\mathbf{X}_i'\,, \tag{2.10}$$

$$FI(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \mathbf{0}, \quad \text{and}$$

$$FI(\lambda_{u,v}, \lambda_{r,s}) = \frac{1}{2}\sum_{i=1}^{n} Tr\left(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\lambda_{u,v}}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\lambda_{r,s}}\right). \tag{2.11}$$

Using the facts that $FI(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \mathbf{0}$, we obtain

$$\left|\mathbf{I}(\hat{\boldsymbol{\zeta}})\right| = \left|FI(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\right|\left|FI(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}})\right|,$$

Therefore using the normal approximation for the posterior distribution of $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\Lambda})$ as shown in (2.8), the integral (2.6) can be approximated by

$$\int \Bigg\{ -\frac{p}{2}\log(2\pi) - \frac{p}{2} + \frac{1}{2}\log\left|\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i'(\sigma^2\mathbf{I} + \mathbf{X}_i\hat{\boldsymbol{\Lambda}}\mathbf{X}_i')^{-1}\mathbf{X}_i\mathbf{Z}_i\right|\left|FI(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}})\right| \Bigg\} p(\mathbf{y}|\widetilde{\mathbf{X}})d\mathbf{y}. \tag{2.12}$$

The integrand depends on $\mathbf{y}$ only through the consistent maximum likelihood estimates of $\mathbf{\Lambda}$. Following Chaloner and Verdinelli (1995, page 286), a further approximation can be taken where the prior distribution of $\mathbf{\Lambda}$ is used to approximate the distributions of $\hat{\mathbf{\Lambda}}$, that is, (2.6) is further approximated by

$$-\frac{p}{2}\log(2\pi)-\frac{p}{2} + \int\left\{\frac{1}{2}\log\left|\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i'(\sigma^2\mathbf{I}+\mathbf{X}_i\mathbf{\Lambda}\mathbf{X}_i')^{-1}\mathbf{X}_i\mathbf{Z}_i\right|\left|FI(\mathbf{\Lambda},\mathbf{\Lambda})\right|\right\}p(\mathbf{\Lambda})d\mathbf{\Lambda}. \tag{2.13}$$

Thus, we seek an optimal design over the class of designs in $\mathcal{D}(m_1,\ldots,m_n)$ that maximizes (2.13), that is, we seek a design with corresponding model matrix $\widetilde{\mathbf{X}}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ that maximizes the integral

$$\int\left\{\log\left|\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i'(\sigma^2\mathbf{I}+\mathbf{X}_i\mathbf{\Lambda}\mathbf{X}_i')^{-1}\mathbf{X}_i\mathbf{Z}_i\right|\left|FI(\mathbf{\Lambda},\mathbf{\Lambda})\right|\right\}p(\mathbf{\Lambda})d\mathbf{\Lambda}. \tag{2.14}$$

For the rest of the paper, we focus on the special case of designs in $\mathcal{D}(m)\in\mathcal{D}(m_1,\ldots,m_n)$ where

(i) every subject receives the same design so that $\mathbf{X}_i=\mathbf{X}$, and $m_i=m$,

(ii) $\mathbf{Z}_i=\mathbf{I}_p$ so that $\boldsymbol{\theta}$ captures the population characteristics.

Under assumptions (i) and (ii), the maximization of (2.14) simplifies to the maximization of

$$\int\left\{\log\left|\mathbf{X}'(\sigma^2\mathbf{I}+\mathbf{X}\mathbf{\Lambda}\mathbf{X}')^{-1}\mathbf{X}\right|\left|FI(\mathbf{\Lambda},\mathbf{\Lambda})\right|\right\}p(\mathbf{\Lambda})d\mathbf{\Lambda}. \tag{2.15}$$

We call this criterion the ψ^J criterion, where the superscript J indicates "joint estimation of $\boldsymbol{\theta}$ and $\mathbf{\Lambda}$". Note that the ψ^J criterion is independent of $\boldsymbol{\theta}$, but requires integration over the prior distribution of $\mathbf{\Lambda}$.

3 Independent random effects

When the random effects are believed to be independent, the covariance matrix $\mathbf{\Lambda}$ is diagonal. When the diagonal elements in $\mathbf{\Lambda}$ are equal, we show in Section 3.1 that when orthogonal designs exist, they are ψ^J-optimal. When the diagonal elements in $\mathbf{\Lambda}$ are not equal, orthogonal designs are still found to be ψ^J-optimal through computer search in Section 3.2.

3.1 Independent and homoscedastic random effects

We first examine the situation in which the random effects are independently distributed with equal variances, i.e., $\mathbf{\Lambda} = \lambda \mathbf{I}_p$ with $\lambda > 0$, and from (2.5) $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_m + \lambda \mathbf{X}\mathbf{X}'$ for all i ($i = 1, \ldots, n$). So from (2.11),

$$\begin{aligned} FI(\lambda, \lambda) &= \frac{n}{2} Tr\left[(\sigma^2 \mathbf{I}_m + \lambda \mathbf{X}\mathbf{X}')^{-1} \mathbf{X}\mathbf{X}' (\sigma^2 \mathbf{I}_m + \lambda \mathbf{X}\mathbf{X}')^{-1} \mathbf{X}\mathbf{X}'\right] \\ &= \frac{n}{2} Tr[\mathbf{X}'\mathbf{X}(\sigma^2 \mathbf{I}_p + \lambda \mathbf{X}'\mathbf{X})^{-1}]^2, \end{aligned}$$

where the second equality follows from the proof of Lemma 1 in Liu et al. (2007). By (2.10) and the same Lemma,

$$|FI(\boldsymbol{\theta}, \boldsymbol{\theta})| = |\mathbf{X}'\mathbf{X}(\sigma^2 \mathbf{I}_p + \mathbf{\Lambda}\mathbf{X}'\mathbf{X})^{-1}| = \frac{|\mathbf{X}'\mathbf{X}|}{|\sigma^2 \mathbf{I}_p + \mathbf{\Lambda}\mathbf{X}'\mathbf{X}|}\,.$$

The maximization of (2.15) now simplifies to the maximization of

$$\int \left\{ \log \frac{|\mathbf{X}'\mathbf{X}|}{|\sigma^2 \mathbf{I}_p + \lambda \mathbf{X}'\mathbf{X}|} Tr[\mathbf{X}'\mathbf{X}(\sigma^2 \mathbf{I}_p + \lambda \mathbf{X}'\mathbf{X})^{-1}]^2 \right\} p(\lambda) d\lambda. \tag{3.1}$$

Let η be a continuous design measure in the class of probability distributions $\mathcal{H}$ on the Borel sets of $\mathcal{X}$, a compact subset of Euclidean p-space ($\mathcal{R}^p$) that contains all possible design points, and let $\mathbf{M}(\eta) = \frac{1}{m}\mathbf{X}'\mathbf{X}$. Following Silvey (1980), to obtain an upper bound for (3.1), we look for a continuous design

that maximizes the continuous analog of (3.1), namely

$$\int \left\{ \log \frac{|\mathbf{M}(\eta)|}{|\frac{\sigma^2}{m}\mathbf{I}_p + \lambda \mathbf{M}(\eta)|} Tr[\mathbf{M}(\eta)(\tfrac{\sigma^2}{m}\mathbf{I}_p + \lambda \mathbf{M}(\eta))^{-1}]^2 \right\} p(\lambda) d\lambda,$$

which is equivalent to the maximization of

$$\int \left\{ \log \frac{|\mathbf{M}(\eta)|}{|\mathbf{I}_p + c\mathbf{M}(\eta)|} Tr[\mathbf{M}(\eta)(\mathbf{I}_p + c\mathbf{M}(\eta))^{-1}]^2 \right\} p(\lambda) d\lambda, \tag{3.2}$$

where $c = m\lambda/\sigma^2$.

Under the main effects hierarchical linear model, Theorem 2 in Liu et al. (2007) shows that, for any given λ, the maximization of the first term in the log function in (3.2) is achieved by design η^* that satisfies $\mathbf{M}(\eta^*) = \mathbf{I}_p$. We next prove that design η^* with $\mathbf{M}(\eta^*) = \mathbf{I}_p$ also maximizes the second term in the log function in (3.2), for any given λ. To do this, we need the following lemma and the subsequent Theorem 2, whose proofs are given in the Appendix.

Lemma 1 *The function*

$$\xi = \begin{cases} Tr[\mathbf{M}(\eta)(\mathbf{I}_p + c\mathbf{M}(\eta))^{-1}]^2, & \text{if } \mathbf{M}(\eta) \text{ is nonsingular} \\ -\infty & \text{if } \mathbf{M}(\eta) \text{ is singular} \end{cases} \tag{3.3}$$

is concave and increasing in $\mathcal{M}$ where $\mathcal{M} = \{\mathbf{M}(\eta) : \eta \in \mathcal{H}\}$.

Theorem 2 *Let η be a design measure in the class of probability distributions H on the Borel sets of a compact design space $\mathcal{X} \subseteq \mathcal{R}^p$. A necessary and sufficient condition for a design η to maximize ξ is as follows:*

$$\mathbf{x}'(c\mathbf{M} + \mathbf{I})^{-1}\mathbf{M}(c\mathbf{M} + \mathbf{I})^{-2}\mathbf{x} \leq Tr\left[\mathbf{M}(c\mathbf{M} + \mathbf{I})^{-1}\mathbf{M}(c\mathbf{M} + \mathbf{I})^{-2}\right], \tag{3.4}$$

for all $\mathbf{x} \in \mathcal{X}$, where $\mathbf{M}$ in (3.4) stands for $\mathbf{M}(\eta)$.

Following Liu et al. (2007), for the contrast coefficients in model matrix $\mathbf{X}$ under the main effects model, we use the coefficients of the "standardized orthogonal main effect contrasts" and define a compact continuous design

space $\mathcal{X} \subseteq \mathcal{R}^p$, in which the first coordinate of all points $\mathbf{x} \in \mathcal{X}$ is constrained to be 1, that is,

$$\mathcal{X} = \Big\{ \mathbf{x} = [1, \ldots, x_{k_1}, \ldots, x_{k_{(h_k-1)}}, \ldots, x_{K_1}, \ldots, x_{K_{(h_K-1)}}]' \quad \text{such that} \quad \sum_{s=1}^{h_k-1} x_{k_s}^2 \leq h_k - 1, k = 1, \ldots, K. \Big\}. \quad (3.5)$$

Lemma 4 in Liu et al. (2007) shows that, for every design point $\mathbf{x}$ in $\mathcal{X}$,

$$\mathbf{x}'\mathbf{x} = 1 + \sum_{k=1}^{K} \sum_{s=1}^{h_k-1} x_{k_s}^2 \leq 1 + \sum_{k=1}^{K} (h_k - 1) = p. \quad (3.6)$$

With the design space $\mathcal{X}$ defined as (3.5), we now seek an optimal continuous design η^* over $\mathcal{X}$ that maximizes ξ in (3.3). We note that any design η^* that maximizes ξ under the standardized orthogonal main effect contrasts coding of model matrix $\mathbf{X}$ also maximizes ξ under any other model matrix $\breve{\mathbf{X}}$ such that

$$\mathbf{X} = \breve{\mathbf{X}}\mathbf{T}, \ \boldsymbol{\theta} = \mathbf{T}^{-1}\breve{\boldsymbol{\theta}}, \ \ \boldsymbol{\Lambda} = \mathbf{T}^{-1}\breve{\boldsymbol{\Lambda}}\mathbf{T}^{-1\,\prime},$$

and $\mathbf{T}$ is a $p \times p$ non-singular transformation matrix (c.f. Scheffé, 1959, page 31-32).

Theorem 3 shows that, under the main effects model, a continuous design η^* with matrix $\mathbf{M}(\eta^*) = \mathbf{I}$ maximizes ξ when the random effects $\boldsymbol{\beta}_i$ in (2.1) are independent and homoscedastic. The proof follows directly from Theorem 2 and (3.6).

Theorem 3 *Let η be a design measure in the class of probability distributions H on the Borel sets of $\mathcal{X}$ where $\mathcal{X}$ is a compact subspace of $\mathcal{R}^p$ defined in (3.5). When the random effects are independent and homoscedastic, that is, $\boldsymbol{\Lambda}$ is of the form $\boldsymbol{\Lambda} = \lambda\mathbf{I}$ with $\lambda > 0$, a design η^* with $\mathbf{M}(\eta^*) = \mathbf{I}$ maximizes ξ in* (3.3) *for any given λ.*

Therefore, by Theorem 3 above and Theorem 2 in Liu et al. (2007), a design η^* with $\mathbf{M}(\eta^*) = \mathbf{I}$ maximizes both the first and the second terms of the log function in (3.2) for any given λ. This leads to the following theorem.

Theorem 4 *Let η be a design measure in the class of probability distributions H on the Borel sets of $\mathcal{X}$ where $\mathcal{X}$ is a compact subspace of $\mathcal{R}^p$ defined in (3.5). When the random effects are independent and homoscedastic, that is, $\mathbf{\Lambda}$ is of the form $\mathbf{\Lambda} = \lambda\mathbf{I}$ with $\lambda > 0$, a design η^* with $\mathbf{M}(\eta^*) = \mathbf{I}$ is ψ^J-optimal such that η^* maximizes* (3.2).

The following corollary follows directly from Theorem 4 by noting that, for a level-balanced orthogonal design, $\mathbf{X}'\mathbf{X} = m\mathbf{I}$, and so $\mathbf{M}(\eta) = \mathbf{I}$.

Corollary 5 *Under the conditions of Theorem 4, if a level-balanced orthogonal design exists, it is ψ^J-optimal.*

3.2 Independent and heteroscedastic random effects

We now consider the situation when the random effects are believed to be independent but heteroscedastic, that is, $\mathbf{\Lambda} = Diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ where $\lambda_i > 0$ for $i = 1, \ldots, p$. Let $\mathbf{x}_i^{(c)}$ denote the ith column of the model matrix $\mathbf{X}$. From (2.11), it can be shown that the (i, j)th element of the $p \times p$ matrix $FI(\mathbf{\Lambda}, \mathbf{\Lambda})$ is equal to

$$FI(\lambda_i, \lambda_j) = \frac{1}{2} n (\mathbf{x}_i^{(c)\prime} \mathbf{\Sigma}^{-1} \mathbf{x}_j^{(c)})^2$$

(see Lenk et al. 1996), where $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_m + \mathbf{X}\mathbf{\Lambda}\mathbf{X}'$. Note that $\mathbf{x}_i^{(c)\prime} \mathbf{\Sigma}^{-1} \mathbf{x}_j^{(c)}$ is the (i, j)th element of

$$\mathbf{X}'(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Lambda}\mathbf{X}')^{-1}\mathbf{X} = \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}.$$

Therefore, a ψ^J-optimal design that maximizes (2.15) is the design with the model matrix $\mathbf{X} = [\mathbf{x}_1^{(c)}, \ldots, \mathbf{x}_p^{(c)}]$ that maximizes

$$\int \log \left| \begin{matrix} \mathbf{x}_1^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1^{(c)} & \ldots & \mathbf{x}_1^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_p^{(c)} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_p^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1^{(c)} & \ldots & \mathbf{x}_p^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_p^{(c)} \end{matrix} \right| \left| \begin{matrix} (\mathbf{x}_1^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1^{(c)})^2 & \ldots & (\mathbf{x}_1^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_p^{(c)})^2 \\ \vdots & \ddots & \vdots \\ (\mathbf{x}_p^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1^{(c)})^2 & \ldots & (\mathbf{x}_p^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_p^{(c)})^2 \end{matrix} \right| p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}, \quad (3.7)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$. We used a computer search to obtain ψ^J-optimal designs that maximize (3.7) for up to 10 treatment factors, each having 2, 3 or 4 levels, for various numbers of observations. We found that, without exception, when a level-balanced orthogonal design exists, it was ψ^J-optimal. However, as with any optimal designs obtained in computer search, the optimal designs may be locally, rather than globally, optimal. The codes used for the search, together with those used for the search of optimal designs in Section 4, are available at `http://www.stat.osu.edu/~amd/dissertations.html`.

4 Correlated random effects

4.1 Design efficiency

When the random effects are correlated, the off-diagonal terms in $\boldsymbol{\Lambda}$ are non-zero. From (2.11), the elements in the Fisher information matrix $FI(\boldsymbol{\Lambda}, \boldsymbol{\Lambda})$ are

$$FI(\lambda_{u,u}, \lambda_{r,r}) = \frac{1}{2} n \left(\mathbf{x}_u^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_r^{(c)} \right)^2 \quad (4.1)$$

$$FI(\lambda_{u,u}, \lambda_{r,s}) = n \left(\mathbf{x}_u^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_r^{(c)} \right) \left(\mathbf{x}_u^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_s^{(c)} \right) \quad (4.2)$$

$$FI(\lambda_{u,v}, \lambda_{r,s}) = n \left[\left(\mathbf{x}_u^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_r^{(c)} \right) \left(\mathbf{x}_v^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_s^{(c)} \right) + \left(\mathbf{x}_u^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_s^{(c)} \right) \left(\mathbf{x}_v^{(c)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}_r^{(c)} \right) \right] \quad (4.3)$$

(see Lenk et al. 1996), where $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_m + \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}'$. Therefore, a ψ^J-optimal design is a design with model matrix $\mathbf{X}^*$ that maximizes (2.15), where elements of $FI(\boldsymbol{\Lambda}, \boldsymbol{\Lambda})$ are given in (4.1), (4.2) and (4.3).

As we do not know the upper bound for the ψ^J criterion, we use an orthogonal design $d_0 \in \mathcal{D}(m)$ with model matrix $\mathbf{X}_0$ as the base design and define the relative ψ^J-efficiency of an exact design $d \in \mathcal{D}(m)$ with model matrix $\mathbf{X}$ as

$$\text{rel. } \psi^J\text{-eff} = \exp\left\{\frac{1}{p}\int \log\left(\frac{|\mathbf{X}'(\sigma^2\mathbf{I}+\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}')^{-1}\mathbf{X}|\,|FI(\boldsymbol{\Lambda},\boldsymbol{\Lambda};\mathbf{X})|}{|\mathbf{X}_0'(\sigma^2\mathbf{I}+\mathbf{X}_0\boldsymbol{\Lambda}\mathbf{X}_0')^{-1}\mathbf{X}_0|\,|FI(\boldsymbol{\Lambda},\boldsymbol{\Lambda};\mathbf{X}_0)|}\right)p(\boldsymbol{\Lambda})d\boldsymbol{\Lambda}\right\}, \tag{4.4}$$

where $FI(\boldsymbol{\Lambda},\boldsymbol{\Lambda};\mathbf{X})$ denotes information matrix $FI(\boldsymbol{\Lambda},\boldsymbol{\Lambda})$ of the design with model matrix $\mathbf{X}$.

EXAMPLE 4.1: Consider an experiment with two treatment factors, each having two levels, under a hierarchical linear model. No covariates are present ($\mathbf{Z}_i = \mathbf{I}$), response errors are assumed to be known ($\sigma^2 = 1$), and each subject i (i=1, . . . , n) receives the same treatment allocation ($\mathbf{X}_i = \mathbf{X}$). Let the number of observations per subject be $m = 12$. The individual-level random effects $\boldsymbol{\beta}_i$ in (2.1) consists of the general mean, and the main effects of factors 1 and 2, for subject i. The vector $\boldsymbol{\beta}_i$ is assumed to be randomly distributed according to a multivariate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Lambda}$ as in (2.2). Of interest is the joint estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ in (2.2).

Table 4.1 reports ψ^J-optimal designs obtained from a computer search under various mean specification $E(\boldsymbol{\Lambda})$ of the prior Inverted Wishart distribution of the random effects covariance matrix $\boldsymbol{\Lambda}$. In the first three rows of the table, $E(\boldsymbol{\Lambda})$ is specified to be of form $\mathbf{I}_p + b\mathbf{J}_p$, where $\mathbf{J}_p$ is a matrix of ones. The constant b is set to be 0.5, 2, or -0.25 in Table 4.1, that is, all pairs of random effects are expected to be positively ($b = 0.5, 2$) or negatively ($b = -0.25$) correlated with equal variances and covariances. In the last row of the table, a more complicated $E(\boldsymbol{\Lambda})$ is specified. The ψ^J-optimal designs obtained through computer search are expressed as $(m_{11}, m_{12}, m_{21}, m_{22})$, where m_{ij} is the number of times level i of factor 1 and level j of factor 2 occur together in the design. The corresponding matrix $\mathbf{X}'\mathbf{X}$ under the standardized orthogonal main effect contrast coding of the model matrix $\mathbf{X}$ of each design is also

reported. The relative ψ^J-efficiency values show that when the random effects are correlated, nonorthogonal designs tend to be more efficient than orthogonal designs.

Table 4.1: ψ^J-optimal 12-run designs of Example 4.1

$E(\mathbf{\Lambda})$	Design $(m_{11}, m_{12}, m_{21}, m_{22})$	Matrix $\mathbf{X}'\mathbf{X}$	Relative ψ^J-Efficiency
$\mathbf{I}_3 + 0.5\mathbf{J}_3$	(4,3,3,2)	$\begin{pmatrix} 12 & -2 & -2 \\ -2 & 12 & 0 \\ -2 & 0 & 12 \end{pmatrix}$	1.019(=1/0.981)
$\mathbf{I}_3 + 2\mathbf{J}_3$	(4,4,3,1)	$\begin{pmatrix} 12 & -4 & -2 \\ -4 & 12 & -2 \\ -2 & -2 & 12 \end{pmatrix}$	1.084(=1/0.922)
$\mathbf{I}_3 - 0.25\mathbf{J}_3$	(2,2,3,5)	$\begin{pmatrix} 12 & 4 & 2 \\ 4 & 12 & 2 \\ 2 & 2 & 12 \end{pmatrix}$	1.076(=1/0.929)
$\begin{pmatrix} 0.47 & 0.19 & 0.46 \\ 0.19 & 5.50 & 0.24 \\ 0.46 & 0.24 & 0.48 \end{pmatrix}$	(3,2,5,2)	$\begin{pmatrix} 12 & 0 & -6 \\ 0 & 12 & -2 \\ -6 & -2 & 12 \end{pmatrix}$	1.311(=1/0.763)

4.2 Design robustness when covariances are all positive

In practical applications, it is seldom the case that the experimenter has complete knowledge of the mean of the covariance matrix $\mathbf{\Lambda}$. In this section, we examine the situation when all pairs of random effects are expected to be positive but the approximate sizes of the variances and covariances are unknown. We show through simulation that a ψ^J-efficient design is likely to be achieved if it is obtained under a positively correlated $E(\mathbf{\Lambda})$ of the form $\mathbf{I}_p + b\mathbf{J}_p$ with moderate sized correlations ($b = 0.5$ or 2).

Let D_{05} and D_2 denote, respectively, the ψ^J-optimal designs in Table 4.1 with respective treatment allocation (4,3,3,2) and (4,4,3,1), obtained under positively correlated $E(\mathbf{\Lambda})$, that is, $E(\mathbf{\Lambda}) = \mathbf{I}_3 + 0.5\mathbf{J}_3$ and $E(\mathbf{\Lambda}) = \mathbf{I}_3 + 2\mathbf{J}_3$. Similarly, let T_{25} denotes the ψ^J-optimal design in Table 4.1 obtained under negatively correlated $E(\mathbf{\Lambda}) = \mathbf{I}_3 - 0.25\mathbf{J}_3$, with treatment allocation (2,2,3,5). Using the orthogonal design with treatment allocation (3,3,3,3) as the base design, we examine the range of relative ψ^J-efficiencies in (4.4) of each of these designs under different specifications of $E(\mathbf{\Lambda})$.

We generate the variances (diagonal elements) in $E(\mathbf{\Lambda})$ independently from a uniform (0,10) distribution. For the covariances (off-diagonal elements) in $E(\mathbf{\Lambda})$, we generate correlation values from a uniform (0,1) distribution, and multiply these with the square root of the corresponding variances to obtain the covariances. The generation of $E(\mathbf{\Lambda})$ is done 10,000 times, and for each $E(\mathbf{\Lambda})$ the relative ψ^J-efficiency in (4.4) is calculated for each of the designs D_{05}, D_2 and T_{25}, and boxplots of the respective distributions are shown in Figure 4.1. The boxplots show that, over the 10,000 simulated values of $E(\mathbf{\Lambda})$, the nonorthogonal and unbalanced designs D_{05} and D_2, obtained under positively correlated $E(\mathbf{\Lambda})$ of form $\mathbf{I}_p + b\mathbf{J}_p$ with moderate and equal correlations ($b = 0.5$ or 2), are more likely to be ψ^J-efficient than the orthogonal design, whereas T_{25} is less likely to be as ψ^J-efficient as the orthogonal design. Specifically, D_{05} is superior to the orthogonal design 77.5% of the time and is never below 89.8% efficiency. D_2 is superior to the orthogonal design 64.7% of the time. On the other hand, design T_{25} is inferior to the orthogonal 100% of the time.

Fig. 4.1. Relative ψ^J-efficiency under 10,000 different specifications of $E(\mathbf{\Lambda})$ in Example 4.1 where all covariance terms in $E(\mathbf{\Lambda})$ are positive

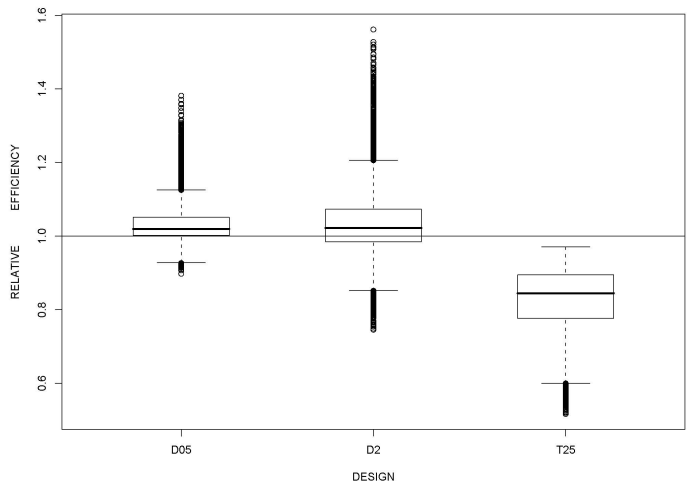
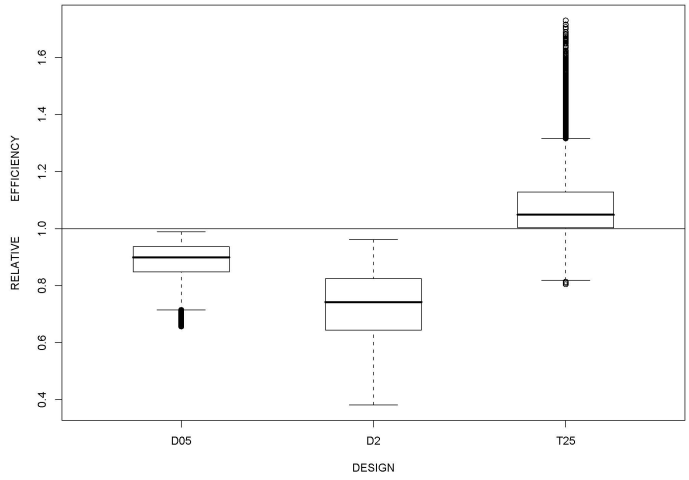


Fig. 4.2. Relative ψ^J-efficiency under 10,000 different specifications of $E(\mathbf{\Lambda})$ in Example 4.1 where all covariance terms in $E(\mathbf{\Lambda})$ are negative



4.3 Design robustness when covariances are all negative

Similar simulation studies were conducted when all pairs of random effects are expected to be negatively correlated. Not surprisingly, as shown in Figure 4.2, T_{25}, the design obtained under the negatively correlated $E(\mathbf{\Lambda})$ of the form $\mathbf{I}_p - 0.25\mathbf{J}_p$, is more likely to be ψ^J-efficient than the orthogonal design. On the other hand, D_{05} and D_2 are less likely to be ψ^J-efficient than the orthogonal design. This implies that for the search of ψ^J-efficient designs, the covariance terms in $E(\mathbf{\Lambda})$ should be specified with the anticipated signs.

5 Summary and conclusion

In this paper, we have investigated optimal designs for the joint estimation of the mean and covariance matrix of the random effects in hierarchical linear models under known response error variance. A ψ^J design criterion was specified which requires the integration over the prior distribution of the random effects covariance matrix $\mathbf{\Lambda}$. We showed that level-balanced orthogonal

designs, if they exist, are optimal when the random effects are expected to be independently distributed. However, when the random effects are correlated, nonorthogonal designs tend to be more ψ^J-efficient than orthogonal designs.

The robustness study under different specification of $E(\mathbf{\Lambda})$ showed that, when all pairs of random effects are expected to be positively (negatively) correlated, designs obtained under positively (negatively) correlated $E(\mathbf{\Lambda})$ with moderate and equal correlations are more likely to be ψ^J-efficient than the orthogonal design. Similar results have been found from other studies with different numbers of treatment factors, factor levels and observations. The results imply that, when the signs of the correlations of the random effects are believed to be known but the approximate sizes of the variances and covariances are unknown, $E(\mathbf{\Lambda})$ should be specified with moderate sized correlations with the anticipated signs in the search for ψ^J-efficient designs.

A Proof of Lemma 1

For display clarity, we omit the subscript and use $\mathbf{M}$ to represent $\mathbf{M}(\eta)$. When $\mathbf{M}$ is nonsingular, $\xi = Tr([\mathbf{M}(\mathbf{I} + c\mathbf{M})^{-1}]^2) = Tr(c\mathbf{I} + \mathbf{M}^{-1})^{-2}$. Let $\tilde{\mathbf{M}} = (c\mathbf{I} + \mathbf{M}^{-1})^{-1}$, then for $\mathbf{M}_1 > \mathbf{M}_2$ (i.e., $\mathbf{M}_1 - \mathbf{M}_2$ is positive definite),

$$c\mathbf{I} + \mathbf{M}_1^{-1} < c\mathbf{I} + \mathbf{M}_2^{-1},$$

and since $c\mathbf{I} + \mathbf{M}^{-1}$ is nonsingular,

$$(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} > (c\mathbf{I} + \mathbf{M}_2^{-1})^{-1}, \text{ i.e. } \tilde{\mathbf{M}}_1 > \tilde{\mathbf{M}}_2$$

(Theorem 12.2.14, Graybill, 1983). Now,

$$\begin{aligned}
\xi(\mathbf{M}_1) - \xi(\mathbf{M}_2) &= Tr(\tilde{\mathbf{M}}_1^2) - Tr(\tilde{\mathbf{M}}_2^2) \\
&= Tr\left[(\tilde{\mathbf{M}}_1 - \tilde{\mathbf{M}}_2)(\tilde{\mathbf{M}}_1 + \tilde{\mathbf{M}}_2)\right], \text{ since } \; Tr(\tilde{\mathbf{M}}_1\tilde{\mathbf{M}}_2) = Tr(\tilde{\mathbf{M}}_2\tilde{\mathbf{M}}_1) \\
&= Tr\left[(\tilde{\mathbf{M}}_1 - \tilde{\mathbf{M}}_2)\tilde{\mathbf{M}}_1\right] + Tr\left[(\tilde{\mathbf{M}}_1 - \tilde{\mathbf{M}}_2)\tilde{\mathbf{M}}_2\right] \qquad \text{(A.1)}
\end{aligned}$$

Since $\mathbf{M}$ is positive definite, its eigenvalues $e_i,\ (i = 1, \ldots p)$ are all positive, and the eigenvalues of the symmetric matrix $\tilde{\mathbf{M}}$ which are $(c+1/e_i)^{-1}$ are also all positive. Therefore, $\tilde{\mathbf{M}}$ is positive definite. According to Theorem 12.2.3 in Graybill (1983), for two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ of size $p \times p$, $Tr(\mathbf{AB}) > 0$. If we let $\mathbf{A} = \tilde{\mathbf{M}}_1 - \tilde{\mathbf{M}}_2$, and let $\mathbf{B} = \tilde{\mathbf{M}}_1$ and $\mathbf{B} = \tilde{\mathbf{M}}_2$ respectively for the first term and the second term of (A.1), we get

$$\xi(\mathbf{M}_1) - \xi(\mathbf{M}_2) > 0, \text{ for } \mathbf{M}_1 > \mathbf{M}_2.$$

Therefore, the function ξ is strictly increasing. To prove that ξ is concave, write ξ as

$$\xi = Tr(\tilde{\tilde{\mathbf{M}}}^{-1}), \quad \text{where } \tilde{\tilde{\mathbf{M}}} = (c\mathbf{I} + \mathbf{M}^{-1})^2 = c^2\mathbf{I} + 2c\mathbf{M}^{-1} + \mathbf{M}^{-2}.$$

From A.1 in Silvey (1980), $\mathbf{M}^{-1}$ is convex in $\mathcal{M}$. Replacing $(\mathbf{M}^{+})^{1/2}$ with $\mathbf{M}^{-1}$ in the proof of A.1 in Silvey (1980), it can be easily shown that $\mathbf{M}^{-2}$ is also convex in $\mathcal{M}$. Therefore, $\tilde{\tilde{\mathbf{M}}}$ is convex, and from A.2 in Silvey (1980), $\tilde{\tilde{\mathbf{M}}}^{-1}$ is concave on $\mathcal{M}$. Then since the trace function is a linear increasing function, $\xi = Tr(\tilde{\tilde{\mathbf{M}}}^{-1})$ is also concave. $\square$

B Proof of Theorem 2

Following Silvey (1980), we first obtain the Gâteaux derivative of function ξ:

$$\begin{aligned}
G_\xi\{\mathbf{M}_1, \mathbf{M}_2\} &= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{Tr\Big(c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big)^{-2} - Tr(c\mathbf{I} + \mathbf{M}_1^{-1})^{-2}\Big\} \\
&= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{Tr\Big(\Big[(c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1})^{-1} + (c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big] \\
&\qquad\qquad \Big[(c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1})^{-1} - (cI + \mathbf{M}_1^{-1})^{-1}\Big]\Big)\Big\} \\
&= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{Tr\Big((c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big[(c\mathbf{I} + \mathbf{M}_1^{-1})(c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1})^{-1} + \mathbf{I}\Big] \\
&\qquad\qquad (c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big[(c\mathbf{I} + \mathbf{M}_1^{-1})(c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1})^{-1} - \mathbf{I}\Big]\Big)\Big\} \\
&= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{Tr\Big((c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big[c\mathbf{I} + \mathbf{M}_1^{-1} + c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]
\end{aligned}$$

$$
\begin{aligned}
&\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} \\
&\Big[c\mathbf{I} + \mathbf{M}_1^{-1} - c\mathbf{I} - (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}\Big)\Big\} \\
= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{ & Tr\Big((c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big[c\mathbf{I} + \mathbf{M}_1^{-1} + c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big] \\
&\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} \\
&\mathbf{M}_1^{-1}[\mathbf{I} - (\mathbf{I} + \epsilon\mathbf{M}_2\mathbf{M}_1^{-1})^{-1}]\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}\Big)\Big\} \\
= \lim_{\{\epsilon\to 0^+\}} \frac{1}{\epsilon}\Big\{ & Tr\Big((c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\Big[c\mathbf{I} + \mathbf{M}_1^{-1} + c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big] \\
&\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} \\
&\mathbf{M}_1^{-1}(\mathbf{I} + \epsilon\mathbf{M}_2\mathbf{M}_1^{-1})^{-1}(\epsilon\mathbf{M}_2\mathbf{M}_1^{-1})\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}\Big)\Big\} \\
= \lim_{\{\epsilon\to 0^+\}} \Big\{ & Tr\Big(\mathbf{M}_2\mathbf{M}_1^{-1}\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} \\
&\Big[c\mathbf{I} + \mathbf{M}_1^{-1} + c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1} \\
&(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\mathbf{M}_1^{-1}(\mathbf{I} + \epsilon\mathbf{M}_2\mathbf{M}_1^{-1})^{-1}\Big)\Big\}
\end{aligned}
$$

From Morrison (1990, page 69, Equation 8),

$$
\begin{aligned}
(\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1} &= \mathbf{M}_1^{-1} - \mathbf{M}_1^{-1}\left(\epsilon^{-1}\mathbf{M}_2^{-1} + \mathbf{M}_1^{-1}\right)^{-1}\mathbf{M}_1^{-1} = \mathbf{M}_1^{-1} + O(\epsilon) \\
[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}]^{-1} &= c^{-1}\mathbf{I} - c^{-1}\left(\mathbf{M}_1 + \epsilon\mathbf{M}_2 + c^{-1}\mathbf{I}\right)^{-1}c^{-1} \\
&= c^{-1}\mathbf{I} - c^{-2}\Big[(\mathbf{M}_1 + c^{-1}\mathbf{I})^{-1} + O(\epsilon)\Big] \\
&= c^{-1}\mathbf{I} - c^{-1}(c\mathbf{M}_1 + \mathbf{I})^{-1} + O(\epsilon) \\
&= c^{-1}(c\mathbf{M}_1 + \mathbf{I})^{-1}(c\mathbf{M}_1 + \mathbf{I}) - c^{-1}(c\mathbf{M}_1 + \mathbf{I})^{-1} + O(\epsilon) \\
&= c^{-1}(c\mathbf{M}_1 + \mathbf{I})^{-1}c\mathbf{M}_1 + O(\epsilon) \\
&= (c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} + O(\epsilon)
\end{aligned}
$$

Therefore, The Gâteaux derivative of function ξ is

$$
\begin{aligned}
G_\xi\{\mathbf{M}_1, \mathbf{M}_2\} = \lim_{\{\epsilon\to 0^+\}} \Big\{ & Tr\Big(\mathbf{M}_2\mathbf{M}_1^{-1}\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1}(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} \\
&\Big[c\mathbf{I} + \mathbf{M}_1^{-1} + c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]\Big[c\mathbf{I} + (\mathbf{M}_1 + \epsilon\mathbf{M}_2)^{-1}\Big]^{-1} \\
&(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}\mathbf{M}_1^{-1}(\mathbf{I} + \epsilon\mathbf{M}_2\mathbf{M}_1^{-1})^{-1}\Big)\Big\} \\
= \lim_{\{\epsilon\to 0^+\}} \Big\{ & Tr\Big(\mathbf{M}_2\mathbf{M}_1^{-1}\Big[(c\mathbf{I} + \mathbf{M}_1^{-1})^{-1} + O(\epsilon)\Big](c\mathbf{I} + \mathbf{M}_1^{-1})^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\Big[c\mathbf{I}+\mathbf{M}_1^{-1}+c\mathbf{I}+\mathbf{M}_1^{-1}+O(\epsilon)\Big]\Big[(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}+O(\epsilon)\Big]\\
&(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}\Big[\mathbf{M}_1^{-1}+O(\epsilon)\Big]\Big)\Big\}\\
=&\lim_{\{\epsilon\to 0^+\}}\Big\{Tr\Big(\mathbf{M}_2\mathbf{M}_1^{-1}(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}2(c\mathbf{I}+\mathbf{M}_1^{-1})\\
&(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}(c\mathbf{I}+\mathbf{M}_1^{-1})^{-1}\mathbf{M}_1^{-1}\Big]+O(\epsilon)\Big)\Big\}\\
=&2Tr\Big(\mathbf{M}_2(c\mathbf{M}_1+\mathbf{I})^{-1}\mathbf{M}_1(c\mathbf{M}_1+\mathbf{I})^{-2}\Big)
\end{aligned}
$$

The Fréchet derivative, defined by

$$F_\xi\{\mathbf{M}_1,\mathbf{M}_2\}=G_\xi\{\mathbf{M}_1,\mathbf{M}_2-\mathbf{M}_1\},$$

is therefore

$$
\begin{aligned}
F_\xi\{\mathbf{M}_1,\mathbf{M}_2\}=&2Tr\Big[\mathbf{M}_2(c\mathbf{M}_1+\mathbf{I})^{-1}\mathbf{M}_1(c\mathbf{M}_1+\mathbf{I})^{-2}\Big]\\
&-2Tr\Big[\mathbf{M}_1(c\mathbf{M}_1+\mathbf{I})^{-1}\mathbf{M}_1(c\mathbf{M}_1+\mathbf{I})^{-2}\Big].
\end{aligned}
$$

The Gâteaux derivative is linear in $\mathbf{M}_2$, and since only η for which $\mathbf{M}(\eta)$ is non-singular can be optimal in this case, the Fréchet derivative is differentiable at $\mathbf{M}_1$. Therefore, the necessary and sufficient condition (3.4) for the maximization of ξ follows from Lemma 1 and Theorem 3.7 in Silvey (1980). □

References

Allenby, G. M. and Lenk, P. J. (1994). Modeling household purchase behavior with logistic normal regression, *Journal of American Statistical Association* **89**: 1218–1229.

Arora, N. and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments, *Journal of Consumer Research* **28 (September)**: 273–283.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer.

Bradlow, E. T. and Rao, V. R. (2000). A hierarchical bayes model for assortment choice, *Journal of Marketing Research* **37 (2)**: 259–268.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review, *Statistical Science* **10 (3)**: 273–304.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences, *Journal of Educational and Behavioral Statistics* **20 (2)**: 115–147.

Entholzner, M., Benda, N., Schmelter, T. and Schwabe, R. (2005). A note on designs for estimating population parameters, *Biometrical Letters – Listy Biometryczne* pp. 25–41.

Fedorov, V. V. and Hackl, P. (1997). *Model-oriented design of experiments*, New York: Springer Verlag.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. London: Hodder Arnold.

Graybill, F. A. (1983). *Matrices with Applications in Statistics*, Belmont, California: Wadsworth.

Han, C. and Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with applications to hiv dynamics, *Biometrics* **60**: 25–33.

Kessels, R., Goos, P. and Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments, *Journal of Marketing Research* **43 (3)**: 409–419.

Lenk, P. J., Desarbo, W. S., Green, P. E. and Young, M. R. (1996). Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs, *Marketing Science* **15 (2)**: 173–191.

Liu, Q., Dean, A. M. and Allenby, G. M. (2007). Optimal experimental designs for hyperparameter estimation in hierarchical linear models, `http://www.stat.osu.edu/~amd/dissertations.html`, *Submitted for publication* .

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics*, New York: John Wiley.

Mentré, F., Mallet, A. and Baccar, D. (1997). Optimal design in random-

effects regression models, *Biometrika* **84 (2)**: 429–442.

Montgomery, A. L., Li, S., Srinivasan, K. and Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data, *Marketing Science* **23 (4)**: 579–595.

Morrison, D. F. (1990). *Multivariate Statistical Methods*, New York: McGraw-Hill, Inc.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research, *Journal of Educational Statistics* **18 (4)**: 321–349.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analsis Methods*, Sage Publications.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, John Wiley and Sons, Ltd.

Sándor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers prior beliefs, *Journal of Marketing Research* **38 (4)**: 430–444.

Sándor, Z. and Wedel, M. (2002). Profile construction in experimental choice designs for mixed logit models, *Marketing Science* **21 (4)**: 455–475.

Sándor, Z. and Wedel, M. (2005). Differentiated bayesian conjoint choice designs, *Journal of Marketing Research* **55 (2)**: 210–218.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley, New York.

Silvey, S. D. (1980). *Optimal Design*, Chapman and Hall, London.

Smith, A. and Verdinelli, I. (1980). A note on bayesian designs for inference using a hierarchical linear model, *Biometrika* **67**: 613–619.

Tod, M., Mentré, F., Merlé, Y. and Mallet, A. (1998). Robust optimal design for the estimation of hyperparameters in population pharmacokinetics, *Journal of Pharmacokinetics and Biopharmaceuics* **26**: 689–716.

Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K., Vonesh, E. and Wolfinger, R. (1994). Population pharmacokinetics/pharmacodynamics methodology and applications: A bibliography, *Biometrics* **50**: 566–575.