Estimation of Multivariate Means with Heteroscedastic Errors Using Envelope Models

Zhihua Su and R. Dennis Cook

University of Minnesota

Abstract: In this article, we propose envelope models that accommodate heteroscedastic error structure in the framework of estimating multivariate means for different populations. Envelope models were introduced by Cook et al. (2010) as a parsimonious version of multivariate linear regression, which achieve efficient estimation of the coefficients by linking the mean function and the covariance structure. In the original development, constant covariance structure is assumed. The heteroscedastic envelope models we proposed are more flexible in allowing a more general covariance structure. Their asymptotic variances and Fisher consistency are studied. Simulations and data examples showed that they are more efficient than standard methods of estimating the multivariate means, and also more efficient than the envelope model assuming constant covariance structure.

Key words and phrases: Dimension Reduction, envelope model, Grassmann manifold, reducing subspace.

1. Introduction

The standard model for estimating multivariate means for p populations can be formulated as

$$\mathbf{Y}_{(i)j} = \boldsymbol{\mu} + \boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)j},\ i = 1, \cdots, p,\ j = 1, \cdots, n_{(i)}, \tag{1.1}$$

where $\mathbf{Y}_{(i)j} \in \mathbb{R}^r$ is the jth observation vector in the ith population, $\boldsymbol{\mu} \in \mathbb{R}^r$ is the grand mean over all the observations, $\boldsymbol{\beta}_{(i)} \in \mathbb{R}^r$ is the difference between the mean of the ith population and the grand mean, and the error vector $\boldsymbol{\varepsilon}_{(i)j} \in \mathbb{R}^r$ follows the normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{(i)} > 0$. Throughout this article, subscripts (i) indicate the ith population and subscripts without parentheses are used to number the observations. The sample size from the ith population is $n_{(i)}$, and the total sample size is $n = \sum_{i=1}^{p} n_{(i)}$. As the population means average to the grand mean, we have $\sum_{i=1}^{p} n_{(i)}\boldsymbol{\beta}_{(i)} = 0$. Then model (1.1) will have $pr + pr(r+1)/2$ parameters to estimate. This number grows fast when r increases, making the model potentially inefficient for large r.

Over the years, the multivariate nature of this model has not been used effectively for estimation in standard analyses. For example, if we want to estimate the first element in the $\boldsymbol{\beta}_{(i)}$'s, we can simply take the first element in the $\mathbf{Y}_{(i)j}$'s and do the analysis neglecting the other measurements in $\mathbf{Y}_{(i)j}$'s. In Cook et al. (2010), a new class of models called envelope models was proposed, which connects the mean function and the covariance structure, and as a result, the elements in the response vector are linked, and information in one element will be used in estimating the mean for another element. This connection provides efficiency gains in the estimation of the multivariate means.

The rest of this Introduction is devoted to a brief review of the envelope model, as this is a new area. In Section 2, we introduce a new heteroscedastic envelope model, derive its maximum likelihood estimators (MLE), and study the Fisher consistency of the MLEs. The asymptotic distribution of the MLEs is explored in Section 3. Dimension selection, simulations and an example are discussed in Section 4.

The original development of the envelope model was under the constant covariance assumption, so for now we assume $\mathbf{\Sigma}_{(1)} = \cdots = \mathbf{\Sigma}_{(p)} = \mathbf{\Sigma}_c$, where $\mathbf{\Sigma}_c$ is used to denote this common covariance matrix. Although it differs from the multivariate linear regression framework in Cook et al. (2010), we will introduce the model in the context of (1.1) for consistent flow of the discussion.

In model (1.1), when r is large, some measurements or linear combination of $\mathbf{Y}_{(i)j}$ could distribute the same among all populations, while the other part of $\mathbf{Y}_{(i)j}$ reflects population differences. In other words, there exists a subspace $\mathcal{S} \subseteq \mathbb{R}^r$ so that (i) the distribution of $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ is the same for all i, j, and (ii) with i fixed, $\mathbf{P}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ are independent, where $\mathbf{P}_{(\cdot)}$ is a projection onto the subspace indicated by its argument and $\mathbf{Q}_{(\cdot)} = \mathbf{I} - \mathbf{P}_{(\cdot)}$. Intuitively, $\mathbf{P}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ carries information about the population difference while the distribution of $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ is the same across the populations. We call $\mathbf{P}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ the dynamic and static parts of $\mathbf{Y}_{(i)j}$, because the distribution of $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ is constant (static) across populations, while the distribution of $\mathbf{P}_{\mathcal{S}}\mathbf{Y}_{(i)j}$ changes (dynamic). We will provide more intuition on the dynamic part and static part later in Figure 1.1 when we explain the working mechanism of the envelope model. Conditions (i) and (ii) are equivalent to the following two conditions, (Cook et al. 2010):

$$\mathcal{B} \subseteq \mathcal{S}, \quad \mathbf{\Sigma}_c = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_c\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_c\mathbf{Q}_{\mathcal{S}}, \tag{1.2}$$

where $\mathcal{B} = \text{span}(\boldsymbol{\beta}_{(1)}, \cdots, \boldsymbol{\beta}_{(p)})$, $\text{Var}(\mathbf{P}_{\mathcal{S}}\mathbf{Y}_{(i)j}) = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_c\mathbf{P}_{\mathcal{S}}$ and $\text{Var}(\mathbf{Q}_{\mathcal{S}}\mathbf{Y}_{(i)j}) = \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_c\mathbf{Q}_{\mathcal{S}}$. The equality in (1.2) is a sufficient and necessary condition for $\mathcal{S}$ being a reducing subspace of $\mathbf{\Sigma}_c$, (Conway, 1990), and thus $\mathcal{S}$ is a reducing subspace of $\mathbf{\Sigma}_c$ that contains $\mathcal{B}$. The $\mathbf{\Sigma}_c$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\mathbf{\Sigma}c}(\mathcal{B})$, is defined as the smallest reducing subspace of $\mathbf{\Sigma}_c$ that contains $\mathcal{B}$. The notation is shortened to $\mathcal{E}$ for subscripts. The minimality guarantees that the dynamic part $\mathbf{P}_{\mathcal{E}}\mathbf{Y}_{(i)j}$ carries only the information on population differences, and the static part $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}_{(i)j}$ carries no information on population differences.

With $\mathcal{S} = \mathcal{E}_{\mathbf{\Sigma}c}(\mathcal{B})$, model (1.1) is called the envelope model with (1.2) imposed and called the standard model without (1.2) imposed. The two conditions in (1.2) provide a link between the mean function and the covariance structure, and it is this link that enables the envelope model to achieve efficient estimation of the $\boldsymbol{\beta}_{(i)}$'s. By Theorem 5.1 in Cook et al. (2010), the envelope estimator is always more efficient than or as efficient as the standard estimator. And the efficiency gains can be expected to be substantial when $\|\mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}_c\mathbf{P}_{\mathcal{S}}\| \ll \|\mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}_c\mathbf{Q}_{\mathcal{S}}\|$, where $\|\cdot\|$ is the spectral norm of a matrix.

Figure 1.1 provides a graphical illustration of the working mechanism of the envelope model. Suppose we have two normal populations, represented by the two ellipses in the plot. Take $r = 2$ and label the two elements in a generic response vector $\mathbf{Y}_{(i)}$ vector as $Y_{1(i)}$ and $Y_{2(i)}$. Then standard inference for $\mathrm{E}(Y_{1(1)}) - \mathrm{E}(Y_{1(2)})$, which corresponds to the first element in $\boldsymbol{\beta}_{(1)}$, is based on projecting all the data points onto the Y_1 axis; the projection path is indicated by line "A". We can imagine that the projections for the two populations will have a large part overlapped. But in envelope analysis, only the dynamic part $\mathbf{P}_{\mathcal{E}}\mathbf{Y}_{(i)j}$ reflects population differences, and $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}_{(i)j}$ distributes the same for the two populations. Consequently, inference on the first element in $\boldsymbol{\beta}_{(i)}$ is based on projecting the data first onto the envelope space $\mathcal{E}_{\boldsymbol{\Sigma}c}(\mathcal{B})$ and then onto the Y_1 axis. The projection path is indicated by line "B". We can imagine that the projection of the two populations are well separated and thus inference is more efficient. The efficiency gain is a result of ruling out the variations in the static part $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}_{(i)j}$. In practice, $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}c}(\mathcal{B})$ will have a degree of wobble, which spreads the projections from line "B". The asymptotic variance of $\widehat{\boldsymbol{\beta}}_{(i)}$ takes this into consideration.

Up to now, the issue of heteroscedasticity in envelope models has not been

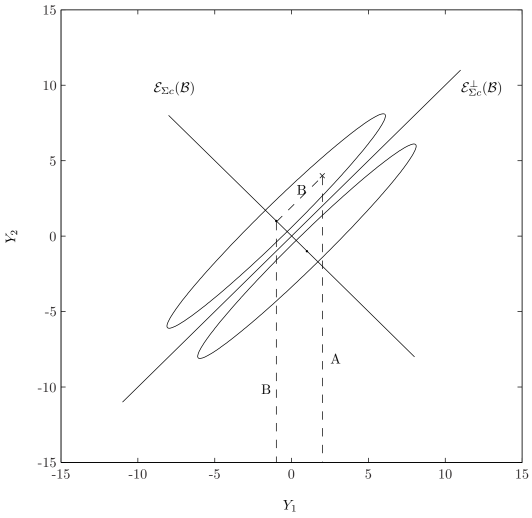

Figure 1.1: Working mechanism of envelope models.

addressed. It was raised in the Discussion of Cook et al. (2010) by Lue, and in the Rejoinder of Cook et al. (2010), the authors considered this to be an important and promising topic for future research. In the next section, we introduce envelope models that accommodate heteroscedastic covariance structure in the framework of estimating multivariate means.

2. Heteroscedastic Envelope Models

2.1. Formulation

No longer assuming $\mathbf{\Sigma}_{(1)} = \cdots = \mathbf{\Sigma}_{(p)}$, we still want to find a subspace $\mathcal{S}$ with the smallest dimension, so that condition (i) and (ii) in Section 1 hold for each population. With heteroscedastic structure, (i) and (ii) expand to the following two conditions:

$$\mathcal{B} \subseteq \mathcal{S}, \;\; \mathbf{\Sigma}_{(i)} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_{(i)}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_{(i)}\mathbf{Q}_{\mathcal{S}}, \; i = 1, \cdots, p. \tag{2.1}$$

The equality in (2.1) indicates that $\mathcal{S}$ is a reducing subspace for all the $\mathbf{\Sigma}_{(i)}$, $i = 1, \cdots, p$. Compared with (1.2), (2.1) suggests a new definition of an envelope that takes a collection of matrices into consideration.

Definition 2.1 *Let $\mathcal{M}$ be a collection of real $p \times p$ symmetric matrices and let $\mathcal{V} \subseteq \mathrm{span}(\mathbf{M})$ for all $\mathbf{M} \in \mathcal{M}$. The $\mathcal{M}$-envelope of $\mathcal{V}$, indicated with $\mathcal{E}_{\mathcal{M}}(\mathcal{V})$, is the intersection of all subspaces that contain $\mathcal{V}$ and that reduce each member of $\mathcal{M}$.*

In our setup, $\mathcal{M} = \{\mathbf{\Sigma}_{(i)} : i = 1, \cdots, p\}$, and $\mathcal{V} = \mathcal{B}$. As the $\mathbf{\Sigma}_{(i)}$ are all positive definite, $\mathrm{span}(\mathbf{\Sigma}_{(i)}) = \mathbb{R}^r$ for all i. The condition $\mathcal{V} \subseteq \mathrm{span}(\mathbf{M})$ is then satisfied by any subspace $\mathcal{V}$ of $\mathbb{R}^r$. The envelope $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ is the subspace with the smallest dimension that contains $\mathcal{B}$ and reduces $\mathcal{M}$, so it is the smallest subspace that satisfies (2.1). The existence of $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ is guaranteed because $\mathcal{E}_{\mathcal{M}}(\mathcal{B}) = \mathbb{R}^r$ satisfies (2.1). Any common eigenspace of the $\mathbf{\Sigma}_{(i)}$'s or the direct sum of the eigenspaces of the $\mathbf{\Sigma}_{(i)}$'s that contains $\mathcal{B}$ satisfies (2.1) (Cook et al, 2010). The envelope $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ is then obtained by taking the intersection of all the subspaces that satisfy (2.1). For example, if $p = 2$, $\mathbf{v}$ is a common eigenvector of $\mathbf{\Sigma}_{(i)}$, $i = 1, 2$, and $\mathcal{B} = \mathrm{span}(\mathbf{v})$, then $\mathrm{span}(\mathbf{v})$ satisfies (2.1). As $\mathrm{span}(\mathbf{v})$ has dimension 1, $\mathcal{E}_{\mathcal{M}}(\mathcal{B}) = \mathrm{span}(\mathbf{v})$.

The envelope $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ can now be used to divide $\mathbf{Y}_{(i),j}$ into its dynamic part $\mathbf{P}_{\mathcal{E}}\mathbf{Y}_{(i),j}$ and its static part $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}_{(i),j}$, where the dynamic part contains information that distinguishs the populations and the static part distributes the same for all populations. From now on, we use the subscript $\mathcal{E}$ for $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$. The working mechanism is similar to that discussed before, but we now have a different covariance structure. Intuition on the heteroscedastic envelope model is provided in the discussion of Figure 3.2 given near the end of Section 3.2 and in Section 4.2.

With $\mathcal{S} = \mathcal{E}_{\mathcal{M}}(\mathcal{B})$, model (1.1) is called the heteroscedastic envelope model with (2.1) imposed. For distinction, the envelope model in Cook et al. (2010) is now called the homoscedastic envelope model. Without (2.1) imposed, model (1.1) is called the heteroscedastic standard model if it allows different covariance structure for different population, and is called the homoscedastic standard model otherwise.

The four models we mentioned in the preceding paragraph are in fact closely related. We use u to denote the dimension of $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$. When $u = r$, $\mathcal{E}_{\mathcal{M}}(\mathcal{B}) = \mathbb{R}^r$. All responses and their linear combinations contain differential population information and there is no static part, so the heteroscedastic envelope model reduces to the heteroscedastic standard model and the homoscedastic envelope model reduces to the homoscedastic standard model (Cook et al., 2010). When we have $\mathbf{\Sigma}_{(1)} = \cdots = \mathbf{\Sigma}_{(p)}$, the heteroscedastic envelope model and the heteroscedastic standard model degenerate to the homoscedastic envelope model and the homoscedastic standard model. When $u = r$ and $\mathbf{\Sigma}_{(1)} = \cdots = \mathbf{\Sigma}_{(p)}$, the four models are the same.

The coordinate form of the heteroscedastic envelope model is similar to (3.2) in Cook et al. (2010), but the error structure now accommodates heteroscedastic cases:

$$\begin{aligned} \mathbf{Y}_{(i)j} &= \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\eta}_{(i)} + \boldsymbol{\varepsilon}_{(i)j}, \\ \mathbf{\Sigma}_{(i)} &= \mathbf{\Sigma}_{1(i)} + \mathbf{\Sigma}_2 = \mathbf{\Gamma}\mathbf{\Omega}_{1(i)}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \end{aligned} \tag{2.2}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ is an orthogonal basis for $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$, and $\mathbf{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ is its completion such that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ is an orthogonal matrix. So we have $\mathbf{\Gamma}\mathbf{\Gamma}^T = \mathbf{P}_{\mathcal{E}}$ and $\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T = \mathbf{Q}_{\mathcal{E}}$. For $i = 1, \cdots, p$, $\boldsymbol{\eta}_{(i)} \in \mathbb{R}^{u\times 1}$ carries the coordinates of $\boldsymbol{\beta}_{(i)}$ with respect to $\mathbf{\Gamma}$, so $\boldsymbol{\beta}_{(i)} = \mathbf{\Gamma}\boldsymbol{\eta}_{(i)}$ and $\sum_{i=1}^{p} n_{(i)}\boldsymbol{\eta}_{(i)} = 0$, $\mathbf{\Omega}_{1(i)} \in \mathbb{R}^{u\times u}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are symmetric matrices that carry the coordinates of $\mathbf{\Sigma}_{(i)}$ with respect to $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$.

The number of parameters in (2.2) is then $u(r-u+p)+pu(u+1)/2+(r-u)(r-u+1)/2$. This parameter counting arises as follows. We need u parameters for each $\boldsymbol{\eta}_{(i)}$, $i = 1, \cdots, p$, but a total of $u(p-1)$ for all of the $\boldsymbol{\eta}_{(i)}$'s as they are linearly dependent, r parameters are needed to specify $\boldsymbol{\mu}$, $u(u+1)/2$ parameters for each $\mathbf{\Omega}_{1(i)}$, $i = 1, \cdots, p$ and $(r-u)(r-u+1)/2$ parameters for $\mathbf{\Omega}_0$. We cannot estimate $\mathbf{\Gamma}$ but only its span, so we are estimating span$(\mathbf{\Gamma})$ on a $r \times u$ Grassmann manifold; therefore, $u(r-u)$ parameters are needed. Compared with the number of parameters in the heteroscedastic standard model as we mentioned at the beginning of Section 1, when $u < r$, the heteroscedastic envelope model has less parameters, which implies a potential for efficiency gains.

2.2. Maximum likelihood estimators

The MLEs of the heteroscedastic envelope model parameters are derived using the coordinate form (2.2). Let $\bar{\mathbf{Y}} = \sum_{i,j} \mathbf{Y}_{(i)j}/n$ be the sample grand mean, and $\bar{\mathbf{Y}}_{(i)} = \sum_j \mathbf{Y}_{(i)j}/n_{(i)}$ be the sample mean for the ith population. We use $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \sum_{i,j}(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}})(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}})^T/n$ for the sample covariance matrix of $\mathbf{Y}$, and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)} = \sum_j(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}}_{(i)})(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}}_{(i)})^T/n_{(i)}$ for the sample covariance matrix of $\mathbf{Y}$ restricted within the ith population, $i = 1, \cdots, p$. Then as shown in Appendix 1, an orthogonal basis $\widehat{\boldsymbol{\Gamma}}$ of the MLE of $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ can be obtained by minimizing the following objective function over the Grassmann manifold $\mathbb{G}^{r \times u}$: $\widehat{\boldsymbol{\Gamma}} = \arg\min_{\mathbf{G}} f_{\mathrm{obj}}(\mathbf{G})$, where

$$f_{\mathrm{obj}}(\mathbf{G}) = n \log |\mathbf{G}^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}^{-1} \mathbf{G}| + \sum_{i=1}^{p} n_{(i)} \log |\mathbf{G}^T \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)} \mathbf{G}|, \tag{2.3}$$

and $\mathbf{G}$ is an $r \times u$ semi-orthogonal matrix. Having $\widehat{\boldsymbol{\Gamma}}$, $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T$ is the projection matrix onto $\mathrm{span}(\widehat{\boldsymbol{\Gamma}})$, and $\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}} = \mathbf{I}_r - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}$. The MLE for the other parameters are listed below:

- $\hat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$;

- $\widehat{\boldsymbol{\beta}}_{(i)} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}(\bar{\mathbf{Y}}_{(i)} - \hat{\boldsymbol{\mu}})$, for $i = 1, \cdots, p$, which is the projection onto the envelope subspace of the difference between the mean for ith population and the grand mean;

- the sample mean for the ith population is $\hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\beta}}_{(i)} = \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\hat{\boldsymbol{\mu}} + \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\bar{\mathbf{Y}}_{(i)}$ for $i = 1, \cdots, p$;
- $\widehat{\boldsymbol{\Gamma}}_0$ is any orthogonal basis of the orthogonal complement of span$(\widehat{\boldsymbol{\Gamma}})$;
- $\hat{\boldsymbol{\eta}}_{(i)} = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\beta}}_{(i)}$, $\widehat{\boldsymbol{\Omega}}_{1(i)} = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}\widehat{\boldsymbol{\Gamma}}$, for $i = 1, \cdots, p$, and $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\widehat{\boldsymbol{\Gamma}}_0$;
- $\widehat{\boldsymbol{\Sigma}}_2 = \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T$, $\widehat{\boldsymbol{\Sigma}}_{1(i)} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}_{1(i)}\widehat{\boldsymbol{\Gamma}}^T$ and $\widehat{\boldsymbol{\Sigma}}_{(i)} = \widehat{\boldsymbol{\Sigma}}_{1(i)} + \widehat{\boldsymbol{\Sigma}}_2$, for $i = 1, \cdots, p$.

2.3. Fisher consistency of the MLEs

As the MLEs are derived under a normality assumption, a natural concern is on their robustness when this assumption does not hold. In this section, we will show that the MLEs are Fisher consistent even the errors are not normally distributed. For the sampling scheme, we assume that the sample proportion for each population is fixed as n increases, in other words, $f_{(i)} = n_{(i)}/n$ is fixed as $n \to \infty$, for $i = 1, \cdots, p$.

Proposition 2.1 *Under the heteroscedastic envelope model (2.2), assume that the errors are independent, but not necessarily normal, and have finite second moments. Then,*

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} &\xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{Y}} = \sum_{i=1}^{p} f_{(i)}\boldsymbol{\Gamma}(\boldsymbol{\Omega}_{1(i)} + \boldsymbol{\eta}_{(i)}\boldsymbol{\eta}_{(i)}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \\
\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)} &\xrightarrow{p} \boldsymbol{\Sigma}_{\mathrm{res}(i)} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,
\end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{Y}}$ and $\mathbf{\Sigma}_{\mathrm{res}(i)}$ are population version of $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{res}(i)}$, for $i = 1, \cdots, p$.

By Proposition 2.1, $f_{\mathrm{obj}}(\mathbf{G})/n$ converges in probability to

$$\tilde{f}_{\mathrm{obj}}(\mathbf{G}) = \log|\mathbf{G}^T \mathbf{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{G}| + \sum_{i=1}^{p} f_{(i)} \log|\mathbf{G}^T \mathbf{\Sigma}_{\mathrm{res}(i)} \mathbf{G}|.$$

Proposition 2.2 *Assume that the conditions in Proposition 2.1 hold, and further assume that the subspace which minimizes $\tilde{f}_{\mathrm{obj}}$ is unique. Then,*

$$\mathbf{\Gamma} = \arg\min_{\mathbf{G}} \tilde{f}_{\mathrm{obj}}(\mathbf{G}),$$

where $\mathbf{\Gamma}$ is any basis matrix for $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$ and $\mathbf{G}$ is a $r \times u$ semi-orthogonal matrix.

Proposition 2.2 indicates that the estimator $\widehat{\mathcal{E}}_{\mathcal{M}}(\mathcal{B})$ is Fisher consistent, which is the basis of the Fisher consistency of the $\widehat{\boldsymbol{\beta}}_{(i)}$'s and $\widehat{\mathbf{\Sigma}}_{(i)}$'s.

Proposition 2.3 *Assume that the conditions in Proposition 2.2 hold, then $\widehat{\boldsymbol{\beta}}_{(i)}$ and $\widehat{\mathbf{\Sigma}}_{(i)}$ are Fisher consistent, for $i = 1, \cdots, p$.*

The proofs of Proposition 2.1, Proposition 2.2 and Proposition 2.3 are in Appendix 2.

3. Asymptotic Distribution

In this section, we study the asymptotic distributions for the $\widehat{\boldsymbol{\beta}}_{(i)}$'s under model (2.2). As the form of the asymptotic variances is too complicated to interpret straightforwardly, we then look into a special case which provides some intuition.

In preparation to state the limiting distribution for the $\widehat{\boldsymbol{\beta}}_{(i)}$'s, we use "vec" as the "vector" operator that rearranges the elements of a matrix into a vector column-wise, and "vech" as the "vector half" operator that extracts the unique elements of a symmetric matrix (Henderson and Searle, 1979). If $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{\mathcal{D}}{\to} N(0, \mathbf{A})$, we write $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\theta}}) = \mathbf{A}$. We define $\mathrm{Bdiag}\{\mathbf{A}_i\}_{i=1}^p$ as a block diagonal matrix with the ith block diagonal as $\mathbf{A}_i$, $i = 1, \cdots, p$, and define the $(p-1) \times 1$ vector $\mathbf{v_f} = (f_{(1)}/f_{(p)}, \cdots, f_{(p-1)}/f_{(p)})^T$. For a population characterizing quantity $\mathbf{A}$, $\mathbf{M}(\mathbf{A})$ is constructed as

$$\mathbf{M}(\mathbf{A}) = f_{(p)}(\mathbf{v_f} \otimes \mathbf{v_f}^T) \otimes \mathbf{A}_{(p)}^{-1} + \mathrm{Bdiag}\{f_{(i)}\mathbf{A}_{(i)}^{-1}\}_{i=1}^{p-1}.$$

We use $\mathbf{B}$ to denote the $r(p-1) \times u(r-u)$ matrix $(\boldsymbol{\eta}_{(1)} \otimes \boldsymbol{\Gamma}_0^T, \cdots, \boldsymbol{\eta}_{(p-1)} \otimes \boldsymbol{\Gamma}_0^T)^T$, use $\mathbf{D}$ to denote the $r \times r$ matrix $\sum_{i=1}^p f_i \boldsymbol{\Sigma}_{(i)}^{-1}$, and $\mathbf{C}$ to denote the $r \times u(p-1)$ matrix $\left(f_1\boldsymbol{\Gamma}(\boldsymbol{\Omega}_{(p)}^{-1} - \boldsymbol{\Omega}_{(1)}^{-1}), \cdots, f_{p-1}\boldsymbol{\Gamma}(\boldsymbol{\Omega}_{(p)}^{-1} - \boldsymbol{\Omega}_{(p-1)}^{-1})\right)$. Then the asymptotic distribution of

$$\hat{h} = \left(\widehat{\boldsymbol{\beta}}_{(1)}^T, \cdots, \widehat{\boldsymbol{\beta}}_{(p-1)}^T\right)^T$$

is given in the following Proposition; justification is given in Appendix 3.

Proposition 3.1 *Under model (2.2), $\sqrt{n}(\hat{h} - h)$ converges in distribution to a $r(p-1)$ dimension multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix*

$$\mathrm{avar}(\sqrt{n}\hat{h}) \quad = \quad (\mathbf{I}_{p-1} \otimes \boldsymbol{\Gamma})\{\mathbf{M}(\boldsymbol{\Omega}_{(1)}) - \mathbf{C}^T\mathbf{D}^{-1}\mathbf{C}\}^{-1}(\mathbf{I}_{p-1} \otimes \boldsymbol{\Gamma})^T + \mathbf{B}^T\mathbf{M}(\boldsymbol{\Sigma}_{(1)})\mathbf{B}.$$

We now look into a special case to gain some insights. Assume that $u = 1$, $p = 2$, $f_{(1)} = f_{(2)} = 1/2$, $\mathbf{\Omega}_{1(i)} = \sigma_i^2 \mathbf{I}_u$, for $i = 1, \cdots, p$ and $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_{r-u}$. Then we can just focus on $\widehat{\boldsymbol{\beta}}_{(1)}$ as $\widehat{\boldsymbol{\beta}}_{(2)} = -\widehat{\boldsymbol{\beta}}_{(1)}$. From Proposition 3.1, we have

$$\mathrm{avar}[\sqrt{n}\widehat{\boldsymbol{\beta}}_{(1)}] = 2^{-1}(\sigma_1^2+\sigma_2^2)\mathbf{\Gamma}\mathbf{\Gamma}^T+\eta_{(1)}^2\left[\sigma_0^{-2}\eta_{(1)}^2 + \sum_{i=1}^{2}\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_i^2}+\frac{\sigma_i^2}{\sigma_0^2}-2\right)\right]^{-1}\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T.$$

The asymptotic variance under the heteroscedastic standard model is

$$\mathrm{avar}[\sqrt{n}\widehat{\boldsymbol{\beta}}_{(1)\mathrm{sm}}] = 2^{-1}(\mathbf{\Sigma}_{(1)} + \mathbf{\Sigma}_{(2)}) = 2^{-1}(\sigma_1^2+\sigma_2^2)\mathbf{\Gamma}\mathbf{\Gamma}^T + \sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T,$$

here we add a subscript "sm" to indicate the heteroscedastic standard model.

Taking an arbitrary linear combination of $\widehat{\boldsymbol{\beta}}_{(1)}$, $l^T\widehat{\boldsymbol{\beta}}_{(1)}$, where $l^T l = 1$, then $\mathrm{avar}[\sqrt{n}l^T\widehat{\boldsymbol{\beta}}_{(1)\mathrm{sm}}]/\mathrm{avar}[\sqrt{n}l^T\widehat{\boldsymbol{\beta}}_{(1)}]$ has the following form:

$$\frac{2^{-1}(\sigma_1^2+\sigma_2^2)l^T\mathbf{\Gamma}\mathbf{\Gamma}^T l + \sigma_0^2 l^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T l}{2^{-1}(\sigma_1^2+\sigma_2^2)l^T\mathbf{\Gamma}\mathbf{\Gamma}^T l + \eta_{(1)}^2\left[\sigma_0^{-2}\eta_{(1)}^2 + \sum_{i=1}^{2}\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_i^2}+\frac{\sigma_i^2}{\sigma_0^2}-2\right)\right]^{-1} l^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T l}.$$

Notice that the numerator is always no less than the denominator, so this ratio will be greater or equal to 1, which means that the heteroscedastic envelope model will be more efficient, or at least equally efficient as the heteroscedastic standard model. If we fix σ_1 and σ_2, and let σ_0 increase, this ratio will diverge to infinity, which means that when the static information accumulates, the advantage of the heteroscedastic envelope model over the heteroscedastic standard model can grow without bound. But if we fix σ_0 and σ_2, and let σ_1 grow, then the ratio will converge to a constant $1+(\sigma_0/\sigma_2)^2 l^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T l$, which means that if we increase the information about the dynamic part, the efficiency gains by enveloping has a finite limit, and this limit depends on the ratio of σ_0/σ_2.

These conclusions can also be demonstrated in Figure 3.2. The background of the plots is similar to that in Figure 1.1, we just added heteroscedatic structure to the covariance matrix. In the left panel, σ_0 is greater than both of σ_1 and σ_2. Here we may image that by enveloping, efficiency will be gained as the projections of the two population following line “B” will be better separated than the projections following line “A”. This is because we have considerable variation in the direction of $\mathcal{E}_{\mathcal{M}}^{\perp}(\mathcal{B})$, which, in envelope analysis, will be taken away. In the right panel, σ_0 is smaller than both of σ_1 and σ_2, and we can imagine that the performance of the heteroscedastic envelope model and the heteroscedastic standard model will be very similar, as it is shown in the plot that for a fixed data point, the projection following line “A” differs little from the projection following line “B”. The reason is that the data does not contain much static information, so enveloping makes little difference.

4. Simulations and Data Example

4.1. Dimension selection and computing

In this section, we will introduce information criteria and likelihood ratio testing (LRT) for the selection of u, the dimension of $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$. Both of the methods work reasonably well in our numerical experiments.

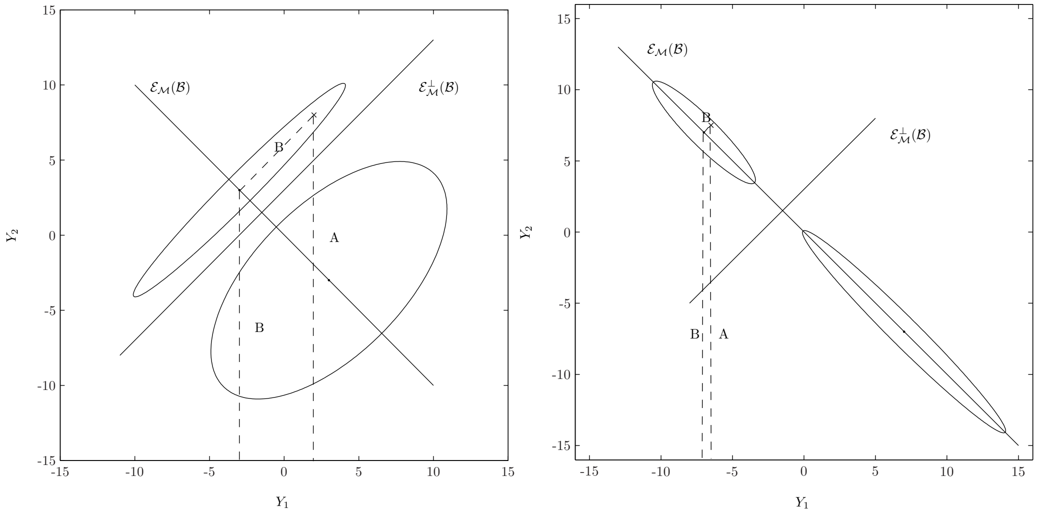

Figure 3.2: Demonstration of efficiency gains by envelopping.

The two most commonly used information criteria are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). For AIC, with a fixed dimension u, $u = 0, \cdots, r$, AIC $= 2N(u) - 2L(u)$, where $N(u) = u(r - u + p) + pu(u+1)/2 + (r-u)(r-u+1)/2$ is the number of parameters in the model and $L(u)$ is the log likelihood function, which has the form

$$L(u) = -\frac{nr}{2}(1+\log 2\pi) - \frac{n}{2}\log|\widehat{\mathbf{\Gamma}}^T\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}^{-1}\widehat{\mathbf{\Gamma}}| - \frac{n}{2}\log|\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}| - \frac{1}{2}\sum_{i=1}^{p} n_{(i)}\log|\widehat{\mathbf{\Gamma}}^T\widehat{\mathbf{\Sigma}}_{\mathrm{res}(i)}\widehat{\mathbf{\Gamma}}|.$$

We compute AIC for all possible values of u and select the value that minimizes AIC. For BIC, with a fixed dimension u, $u = 0, \cdots, r$, BIC $= \log(n)N(u) - 2L(u)$. Again, we search through all possible values and select u at the value that minimizes BIC.

LRT is performed as a sequential testing of hypotheses, starting from $u = 0$ at a prechosen common significance level α and picking u to be the first hypothesized value that is not rejected. For testing the hypothesis $u = u_0$, the test statistics is $\Lambda(u_0) = 2[L(r) - L(u_0)]$, and the reference distribution is chi-squared with degrees of freedom $N(r) - N(u_0)$. The test of $u = 0$ is the same as the likelihood ratio test that the populations means are equal or, equivalently, that $\boldsymbol{\beta}_{(1)} = \ldots = \boldsymbol{\beta}_{(p-1)} = 0$.

To compare the performance of model selection criteria, we set a simulation with the same settings as the upper left panel of Figure 4.3, but we used different sample sizes and different u's. When $u = 3$, we need 67 parameters for heteroscedastic envelope model; when $u = 6$, we need 88 parameters, and when $u = 9$, we need 118 parameters. So we used 80, 160 and 320 to represent small, moderate and large sample sizes. With each u and n combination, we simulated 100 datasets and compared the frequency at which the criteria selected the correct u. The results are reported in Table 1. LRT1 represents the likelihood ratio testing procedure introduced above with $\alpha = 0.05$. LRT2 represents a single test on $H_0 : u = u_0$, LRT2 alone is not used for selecting u, but it provides intuition on the performance of LRT. From Table 4.1, we notice that the LRT is most stable with small sample size, but asymptotically it makes error with the rate α. BIC is consistent, but it is sometimes slow to respond to sample size (Cook and Forzani, 2009, Figure 3). AIC tends to overestimate u, and it works better for larger u as shown in Table 4.1.

Table 4.1: The number of times out of 100 replications that the criteria selected u correctly.

(u, n)	(3, 80)	(3, 160)	(3, 320)	(6, 80)	(6, 160)	(6, 320)	(9, 80)	(9, 160)	(9, 320)
AIC	15	33	23	49	71	66	86	93	100
BIC	83	100	99	82	100	100	34	95	100
LRT1	85	94	96	79	99	96	84	92	95
LRT2	88	94	97	80	98	98	88	89	96

The numerical Grassmann optimization of (2.3) can be performed by using the MATLAB package *sg-min 2.4.1* by Lippert (*http://www-math.mit.edu/ lippert/sgmin.html*). It uses the analytical first derivative and numerical second derivative of the objective function and offers several methods including Newton–Raphson to perform the optimization. We find it very stable.

4.2. Simulations

In this section, we demonstrate the performance of the heteroscedastic envelope model, with comparison to the heterscedastic standard model and the homoscedastic standard model. To connect with the discussion of the special case at the end of Section 3, the data were generated from two normal populations following model (2.2), with $r = 10$, $u = 1$ and $u = 2$, $\mathbf{\Omega}_{1(i)} = \sigma_i^2 \mathbf{I}_u$, for $i = 1, 2$, and $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_{r-u}$. The matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ was obtained by orthogonalizing an $r \times r$ matrix of random uniform $(0, 1)$ variables, and the elements in $\boldsymbol{\eta}_{(1)}$ were sampled from a standard normal population. We sampled equal numbers of observations for each population, and $\boldsymbol{\eta}_{(2)} = -\boldsymbol{\eta}_{(1)}$. The sample size n was fixed at 100, 200, 300, 500, 800 and 1200 and, for each sample size, 200 replications were performed to compute the actual estimation standard deviations for elements in $\widehat{\boldsymbol{\beta}}_{(1)}$. The estimation standard deviations for $\widehat{\boldsymbol{\beta}}_{(2)}$ will be the same as $\widehat{\boldsymbol{\beta}}_{(2)} = -\widehat{\boldsymbol{\beta}}_{(1)}$. Bootstrap standard deviations were obtained by computing the standard deviations for 200 bootstrap samples, as a way to estimate the actual estimation standard deviations for $\widehat{\boldsymbol{\beta}}_{(1)}$. The results are shown in Figure 4.3. No computational problems arose as long as $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}$ are a positive definite, $i = 1, \ldots, p$. Under that condition (2.3) will not have any points at which a determinant is 0.

In all four panels, the results for homoscedastic standard mode were almost the same as the heteroscedastic standard model and their lines overlapped with each other. For that reason the results for the homoscedastic standard model are not shown. In the two right panels, the asymptotic standard deviation for homoscedastic envelope model and heteroscedastic envelope model are quite close, so it is difficult to see the line for homoscedastic envelope model because of the chosen line types. The upper left panel is under the simulation case when $\sigma_1 < \sigma_0$, $\sigma_2 < \sigma_0$ and $u = 1$. For better visibility, we cut the vertical axis at 0.45, while

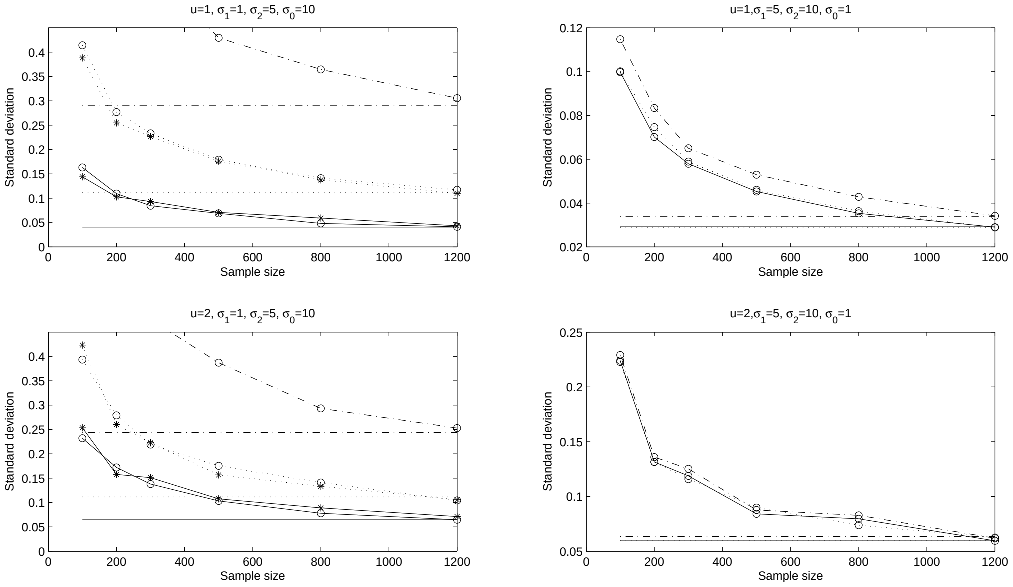

Figure 4.3: Estimated standard deviations for a randomly picked element in $\widehat{\boldsymbol{\beta}}_{(1)}$. Lines — mark the heteroscedastic envelope model, lines $\cdots\cdots$ mark the homoscedastic envelope model, and lines $- \cdot -$ mark the standard models. Lines with $\circ$ mark the sample standard deviations of the method indicated by the line type, the line with $*$ marks the bootstrap standard deviations of heteroscesdastic envelope model, and lines without $\circ$ or $*$ mark the asymptotic standard deviations.

the line for the standard model reaches as high as 0.96 for $n = 100$. We notice that the heteroscedastic envelope model is much more efficient than the standard models even with relatively small sample size and the reason can be explained by the left panel of Figure 3.2. The heteroscedastic envelope model is also more efficient than the homoscedastic envelope model by more accurately capturing the error structure. It is also indicated in the plot that the bootstrap standard deviations estimate the actual estimation standard deviations well. In the upper right panel of Figure 4.3, we omitted the results from bootstrap as the lines almost overlap. From this plot, it is hard to tell the difference between the estimation standard deviations for the homoscedastic and heteroscedastic standard model as the two lines almost overlap with each other, but the heteroscedastic envelope model is more efficient than both of them for all the sample sizes. And the magnitude of the difference between the actual standard deviations is almost the same as the difference between the asymptotic estimation standard deviations. This matches our discussion at the end of Section 3 and of the right panel in Figure 3.2, that the heteroscedastic envelope model does not achieve much reduction when $\sigma_0 < \sigma_1$ and $\sigma_0 < \sigma_2$, as the majority of the variation comes from the dynamic part of $\mathbf{Y}$. This also agrees the discussion in the Rejoinder of Cook et al. (2010), that we will achieve more reduction when $\|\mathbf{\Sigma}_2\| \gg \|\mathbf{\Sigma}_{(1)}\|$. The lower panels have the same simulation settings as the upper panels, but

$u = 2$. The results are similar to the upper panels. We expect that the difference between the heteroscedastic envelope model and the standard models for $u = 2$ is smaller than that for $u = 1$, because the dynamic part has a larger dimension and there is less space for efficiency gains in the first place. And this is confirmed in the plots. Not shown here, in our simulation with $u = 3$, the difference is smaller, in the $\sigma_0 < \sigma_1$ and $\sigma_0 < \sigma_2$ case, the lines that mark actual standard deviations overlap with each other.

The results with more than two populations, $p > 2$, are qualitatively similar to those for $p = 2$. An example with $p = 3$ is provided in the Section 4.3.

To test the sensitivity of the MLEs to non-normal errors, we did a simulation using the same setup as that in the left panel of Figure 4.3, but we used a centered t distribution with degrees of freedom 6, a centered uniform $(0, 1)$ distribution and a chi-squared distribution with degrees of freedom 4 to represent distributions with heavier tails, shorter tails, and skewness. The results are shown in Figure 4.4, as the results from different error types are so close, we did not mark the curves for the different distributions. We concluded that moderate departures from normality, as reflected by the distributions used on our simiulations, do not materially affect the performance of the heteroscedastic envelope. This agrees with our discussion in Section 2.3, that the MLEs are Fisher consistent regardless of the error distribution.

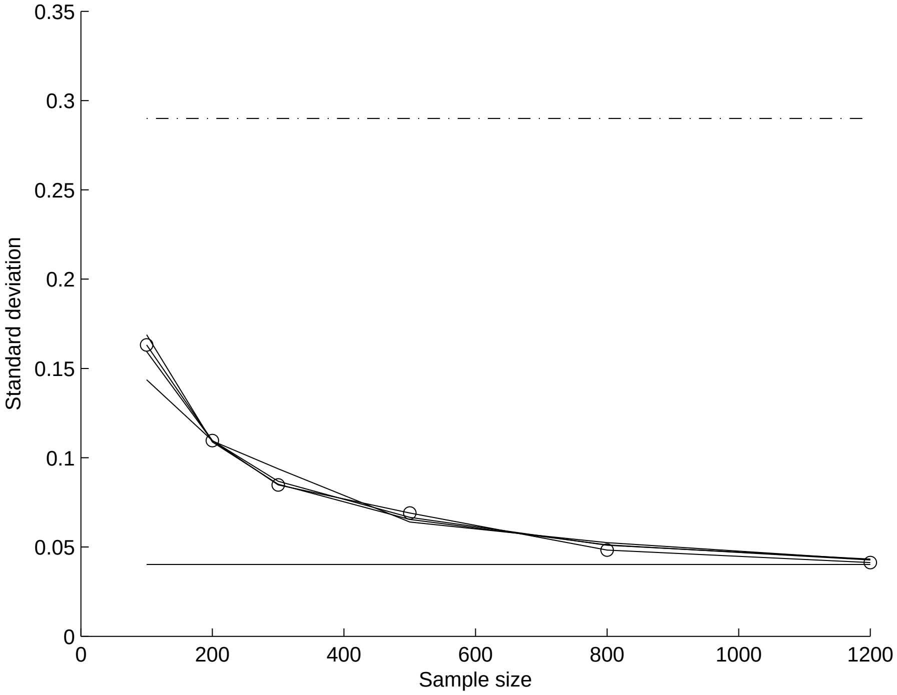

Figure 4.4: The line types for the horizontal lines are the same as in Figure 4.3. The solid line with circles represents the results of normal errors, the other three lines around it represent the results from other three errors distributions.

4.3. Two Data Examples

The athletes data (Cook, 1998) contains measurements of plasma ferritin concentration and white cell counts from 102 male and 100 female athletes collected at the Australian Institute of Sport. So we have two populations: male and female, and $r = 2$. Box's M tests (Johnson and Wichern, 2007) was used to test homogeneity of covariance matrices, with $H_0 : \mathbf{\Sigma}_{(1)} = \cdots = \mathbf{\Sigma}_{(p)}$, and H_a : not H_0. The test yielded a p-value of $2.3e - 06$, which indicates that the covariance structure is different for male and female athletes. We then fitted a heteroscedastic envelope model and $u = 1$, was inferred by AIC, BIC and LRT with $\alpha = 0.01$. The ratios of the estimation standard deviations for elements in $\widehat{\boldsymbol{\beta}}_{(1)}$ are 1.02 and 2.37 for homoscedastic standard model versus heteroscedastic envelope model, and 1.00 and 2.32 for heteroscedastic standard model versus heteroscedastic envelope model. To achieve such efficiency gains in a standard analysis, we may need to increase the original sample size by a factor of 5. The standard deviation ratios for the homoscedastic envelope model versus the heteroscedastic envelope model are 1.02 and 1.02.

Water striders are insects that live on the surface of ponds or streams. They can be easily identified because of their ability to walk on water. Like other insects, water striders have six legs and two antennae. Before the adult stage, the water strider grows through five stages of nymphal forms, called instars, at the end of which they shed their skins, also their skeletons. This water strider dataset contains eight measures of characteristics – the lengths of fomora and tibiae of the middle and hind legs and the lengths of four antennal segments – for three water strider species L. dissortis, L. rufoscutellatus and L. esakii, with 90 samples for each species. This is part of a larger dataset analyzed by Klingenberg and Spence (1993) to study heteroschrony, and they found "a remarkable variety of heterochronic changes among different species" using principal component analysis of the eight characteristics. We consider species differences by comparing the mean of the characteristics. To avoid the effect of female and male, we only look at the data from the first three instars, when sex is hard to determine, leaving us 30 samples for each species. Box's M tests gave a p-value of less than 0.001, indicating heteroscedastic error structure. For the dimension of the heteroscedastic envelope model, LRT inferred $u = 6$ while AIC suggested $u = 5$ and BIC suggested $u = 4$. We will take $u = 6$ since LRT is more stable with small sample size as mention in Section 4.1. The ratios of the standard deviations from the homoscedastic standard model versus the heteroscedastic envelope model for the elements in the $\widehat{\boldsymbol{\beta}}_{(i)}$'s fell between 5.11 and 16.77, with an average of 9.95. A comparison with the heteroscedastic standard model produced similar results: the standard deviation ratios of the heteroscedastic standard model versus the heteroscedastic envelope model ranged from 4.92 to 16.21, with an average of

9.58. To achieve such efficiency gains in the standard analysis, we may need more than 280×30 samples. Comparing with the homoscedastic envelope model of $u = 6$, the standard deviation ratios of from the homoscedastic envelope model versus those from the heteroscedastic envelope model ranged from 4.81 to 15.95, with an average of 9.48. Therefore, considering the heteroscedastic nature of the covariance matrix does bring us additional efficiency gains. However, in practice, the inferred dimensions of the homoscedastic and heteroscedastic envelope models may not agree. In this example, LRT suggested $u = 4$ for homoscedastic envelope model. But the standard deviation ratios does not change much, they fell between 4.79 and 15.95 with an average of 9.47.

5. Discussion

When there are multiple populations, we do not need to envelope on all of the $\boldsymbol{\beta}_{(i)}$'s if our interest is in just a few of them. For example, suppose we have three populations and three characteristics in $\mathbf{Y}$. Suppose also that two of the populations are placed in the Y_1Y_2 plane as the left panel of Figure 3.2, and the elliptical contour for the third population is in a different plane with neither its major axis nor the minor axis aligning with the envelope for the Y_1Y_2 plane. Then we will have $\mathcal{E}_{\mathcal{M}}(\mathcal{B}) = \mathbb{R}^r$ if we envelop on all the $\boldsymbol{\beta}_{(i)}$'s, and no gains are offered. But if we are interested in a contrast between $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$, and just envelop on $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$, according to the discussion in Section 3, we will have significant gains. This idea is parallel to the partial envelope idea in Su and Cook (2010).

Acknowledgment

The work is supported in part by a Doctoral Dissertation Fellowship from the Graduate School, University of Minnesota, and by National Science Foundation Grant DMS-1007547.

Appendix 1 Derivation of the MLEs for the Heteroscedastic Envelope Model

The derivation will be easier if we change the parameterization in (2.2) to the following form (Cook and Forzani, 2009b):

$$\begin{aligned}\mathbf{Y}_{(i)j} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\bar{\boldsymbol{\Omega}}_1\boldsymbol{\nu}_{(i)} + \boldsymbol{\varepsilon}_{(i)j},\\ \boldsymbol{\Sigma}_{(i)} &= \boldsymbol{\Sigma}_{1(i)} + \boldsymbol{\Sigma}_2 = \boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,\end{aligned}$$

where $\bar{\boldsymbol{\Omega}}_1 = \sum_{i=1}^p n_{(i)}\boldsymbol{\Omega}_{1(i)}/n$, $\boldsymbol{\eta}_{(i)} = \bar{\boldsymbol{\Omega}}_1\boldsymbol{\nu}_{(i)}$ and $\sum_{i=1}^p n_{(i)}\boldsymbol{\nu}_{(i)} = 0$.

During the derivation, we use ˆ over a quantity both for intra-derivation steps and final estimators. The log likelihood function L based on observation $\mathbf{Y}_{(i)j}$'s,

$i = 1, \cdots, p$, $j = 1, \cdots, n_{(i)}$ is

$$\begin{aligned} L &= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \log|\boldsymbol{\Omega}_{1(i)}| \\ &- \frac{1}{2}\sum_{i=1}^{p} n_{(i)}[\boldsymbol{\Gamma}^T(\bar{\mathbf{Y}}_{(i)} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\bar{\boldsymbol{\Omega}}_1\boldsymbol{\nu}_{(i)})]^T\boldsymbol{\Omega}_{1(i)}^{-1}[\boldsymbol{\Gamma}^T(\bar{\mathbf{Y}}_{(i)} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\bar{\boldsymbol{\Omega}}_1\boldsymbol{\nu}_{(i)})] \\ &- \frac{1}{2}\sum_{i=1}^{p} n_{(i)}[\boldsymbol{\Gamma}_0^T(\bar{\mathbf{Y}}_{(i)} - \boldsymbol{\mu})]^T\boldsymbol{\Omega}_0^{-1}[\boldsymbol{\Gamma}_0^T(\bar{\mathbf{Y}}_{(i)} - \boldsymbol{\mu})] \\ &- \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}^{-1}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}) - \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \operatorname{tr}(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}). \end{aligned}$$

Only the fourth term above involves $\boldsymbol{\nu}_{(i)}$'s, which have the constraint $\sum_{i=1}^{p} n_{(i)}\boldsymbol{\nu}_{(i)} = 0$, so we apply the Lagrange multiplier, set the derivative at zero, and get

$$\hat{\boldsymbol{\nu}}_{(i)} = \bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\Gamma}^T(\bar{\mathbf{Y}}_{(i)} - \boldsymbol{\mu}),$$

for $i = 1, \cdots, p$. Substitute them back, and maximize over $\boldsymbol{\mu}$, we have $\hat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$, where $\bar{\mathbf{Y}} = \sum_{i,j}\mathbf{Y}_{(i)j}/n$. Substitute this also into the log likelihood function, we have

$$\begin{aligned} L &= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \log|\boldsymbol{\Omega}_{1(i)}| \\ &- \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}^{-1}\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}) - \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \operatorname{tr}(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}(i)}), \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}(i)} = \sum_{j=1}^{n_i}(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}})(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}})^T$.

Now if we fix $\boldsymbol{\Gamma}$, by Lemma 4.3 in Cook et al. (2010), the maximum value of the log likelihood function function is

$$L = -\frac{nr}{2}(1 + \log 2\pi) - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}| - \frac{1}{2}\sum_{i=1}^{p} n_{(i)} \log|\boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}(i)}\boldsymbol{\Gamma}|,$$

so the objective function to minimize over the $r \times u$ Grassmann manifold is

$$f_{\text{obj}}(\mathbf{G}) = n \log |\mathbf{G}^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}^{-1} \mathbf{G}| + \sum_{i=1}^{p} n_{(i)} \log |\mathbf{G}^T \widehat{\boldsymbol{\Sigma}}_{\text{res}(i)} \mathbf{G}|.$$

Appendix 2 Proofs of Proposition 2.1, Proposition 2.2 and Proposition 2.3

Proof of Proposition 2.1

Since the errors are independent and have finite second moments, and also because that the number of populations is finite, we have $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{Y}}$. We use $\mathcal{I}$ for population indices, then

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{Y}} &= \text{var}(\mathbf{Y}) \\
&= \text{E}(\text{var}(\mathbf{Y}|\mathcal{I} = i)) + \text{var}(\text{E}(\mathbf{X}|\mathcal{I} = i)) \\
&= \sum_{i=1}^{p} f_{(i)}(\boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) + \sum_{i=1}^{p} f_{(i)}\boldsymbol{\Gamma}\boldsymbol{\eta}_{(i)}\boldsymbol{\eta}_{(i)}^T\boldsymbol{\Gamma}^T \\
&= \sum_{i=1}^{p} f_{(i)}\boldsymbol{\Gamma}(\boldsymbol{\Omega}_{1(i)} + \boldsymbol{\eta}_{(i)}\boldsymbol{\eta}_{(i)}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.
\end{aligned}$$

Since $\widehat{\boldsymbol{\Sigma}}_{\text{res}(i)} = \sum_{j=1}^{n_{(i)}}(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}}_{(i)})(\mathbf{Y}_{(i)j} - \bar{\mathbf{Y}}_{(i)})^T/n_{(i)}$ and $\mathbf{Y}_{(i)j} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}_{(i)} + \boldsymbol{\varepsilon}_{(i)j}$,

$$\widehat{\boldsymbol{\Sigma}}_{\text{res}(i)} = \sum_{j=1}^{n_{(i)}}(\boldsymbol{\varepsilon}_{(i)j} - \bar{\boldsymbol{\varepsilon}}_{(i)})(\boldsymbol{\varepsilon}_{(i)j} - \bar{\boldsymbol{\varepsilon}}_{(i)})^T/n_{(i)} \xrightarrow{p} \boldsymbol{\Gamma}\boldsymbol{\Omega}_{1(i)}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T = \boldsymbol{\Sigma}_{\text{res}(i)}.$$

Proof of Proposition 2.2

Let $\mathbf{G}_0 \in \mathbb{R}^{r \times r-u}$ be the completion of $\mathbf{G}$, so that $(\mathbf{G}, \mathbf{G}_0) \in \mathbb{R}^{r \times r}$ is an orthogonal matrix. Let $\mathbf{H} = \mathbf{G}_0^T\boldsymbol{\Gamma}(\sum_{i=1}^p f_{(i)}\boldsymbol{\eta}_{(i)}\boldsymbol{\eta}_{(i)}^T)^{1/2}$, then

$$\begin{aligned}
\tilde{f}_{\text{obj}} &= \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{G}_0| + \sum_{i=1}^p f_{(i)}\log|\mathbf{G}^T\boldsymbol{\Sigma}_{\text{res}(i)}\mathbf{G}| \\
&= \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\text{res}}\mathbf{G}_0| + \log|\mathbf{I}_{r-u} + \mathbf{H}^T(\mathbf{G}_0^T\boldsymbol{\Sigma}_{\text{res}}\mathbf{G}_0)^{-1}\mathbf{H}| + \sum_{i=1}^p f_{(i)}\log|\mathbf{G}^T\boldsymbol{\Sigma}_{\text{res}(i)}\mathbf{G}| \\
&\geq \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\text{res}}\mathbf{G}_0| + \sum_{i=1}^p f_{(i)}\log|\mathbf{G}^T\boldsymbol{\Sigma}_{\text{res}(i)}\mathbf{G}| \\
&\geq \sum_{i=1}^p f_{(i)}\log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\text{res}(i)}\mathbf{G}_0| + \sum_{i=1}^p f_{(i)}\log|\mathbf{G}^T\boldsymbol{\Sigma}_{\text{res}(i)}\mathbf{G}| \\
&\geq \sum_{i=1}^p f_{(i)}\log|\boldsymbol{\Sigma}_{\text{res}(i)}| \\
&= \sum_{i=1}^p f_{(i)}\log|\boldsymbol{\Omega}_{1(i)}||\boldsymbol{\Omega}_0|.
\end{aligned}$$

When $\mathbf{G}$ spans $\mathcal{E}_{\mathcal{M}}(\mathcal{B})$, the three inequality will hold simultaneously, and the second inequality will hold only when $\text{span}(\mathbf{G}) = \mathcal{E}_{\mathcal{M}}(\mathcal{B})$. So we have $\boldsymbol{\Gamma} = \arg\min_{\mathbf{G}} \tilde{f}_{\text{obj}}(\mathbf{G})$.

Proof of Proposition 2.3

The Fisher consistency of $\hat{\boldsymbol{\eta}}_{(i)}$, $\widehat{\boldsymbol{\Omega}}_{(i)}$ and $\widehat{\boldsymbol{\Omega}}_0$ follows from the theory of MLE, and Proposition 2.2 gives the Fisher consistency of $\widehat{\boldsymbol{\Gamma}}$. Then as $\widehat{\boldsymbol{\beta}}_{(i)}$ and $\widehat{\boldsymbol{\Sigma}}_{(i)}$ are simple functions of $\hat{\boldsymbol{\eta}}_{(i)}$, $\widehat{\boldsymbol{\Omega}}_{(i)}$ and $\widehat{\boldsymbol{\Omega}}_0$: $\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\eta}}_{(i)}$ and $\widehat{\boldsymbol{\Sigma}}_{(i)} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}_{1(i)}\widehat{\boldsymbol{\Gamma}}^T + \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T$, $\widehat{\boldsymbol{\beta}}_{(i)}$ and $\widehat{\boldsymbol{\Sigma}}_{(i)}$ are Fisher consistent, for $i = 1, \cdots, p$.

Appendix 3 Proof of Proposition 3.1

As there is overparameterization in $\boldsymbol{\Gamma}$, the asymptotic distribution can be derived by using Proposition 4.1 in Shapiro (1986). The parameters are

$$\boldsymbol{\phi} = \left(\boldsymbol{\eta}_{(1)}^T, \cdots, \boldsymbol{\eta}_{(p-1)}^T, \mathrm{vec}^T(\boldsymbol{\Gamma}), \boldsymbol{\mu}^T, \mathrm{vech}^T(\boldsymbol{\Omega}_{1(1)}), \cdots, \mathrm{vech}^T(\boldsymbol{\Omega}_{1(p)}), \mathrm{vech}^T(\boldsymbol{\Omega}_0)\right)^T,$$

and the functions to estimate is

$$g = \left(\boldsymbol{\beta}_{(1)}^T, \cdots, \boldsymbol{\beta}_{(p-1)}^T, \boldsymbol{\mu}^T, \mathrm{vech}^T(\boldsymbol{\Sigma}_{(1)}), \cdots, \mathrm{vech}^T(\boldsymbol{\Sigma}_{(p)})\right)^T,$$

we then have $\sqrt{n}(\hat{g}-g) \overset{\mathcal{D}}{\rightarrow} N(\mathbf{0}, \mathbf{V_0})$, where $\mathbf{V}_0 = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T$, $\mathbf{J}$ is the Fisher information under the heterpscedastic standard model, and $\mathbf{H} = (\partial \mathbf{h}_i/\partial \boldsymbol{\phi}_j^T)_{i,j}$ is the gradient matrix. The asymptotic variance for $\hat{h}$ corresponds to the upper left $r(p-1) \times r(p-1)$ block of $\mathbf{V}_0$. After some lengthly but straightforward matrix algebra, we get the form of $\mathrm{avar}(\sqrt{n}\hat{h})$ as displayed in Proposition 3.1.

References

Conway, J. (1990). *A Course in Functional Analysis.* New York: Springer.

Cook, R. D. (1998). *Regression Graphics: Ideas for studying Regressions Through Graphics.* Wiley, New York.

Cook, R. D., Li, B. and Chiaromente, F. (2010). Envelope Models for Parsimonious and Efficient Multivariate Linear Regression (with discussion). *Statist. Sinica* **20**, 927-1010.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104**, 197–208.

Henderson, H. V. and Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canad. J. Statist.* **7**, 65-81.

Johnson, R. A. and Wichern, D. A. (1998). *Applied Multivariate Statistical Analysis*. Sixth Edition. Pearson Prentice Hall.

Klingenberg, C. R. and Spence, J. R. (1993). Heterochrony and Allometry: Lessons from the Water Strider Genus Limnoporus. *Evolution* **47**, 1834-1853.

Shapiro, A. (1986). Asymptotic Theory of Overparameterized Structural Models. *J. Amer. Statist. Assoc.* **81**, 142-149.

Su, Z. and Cook, R. D. (2010). Partial Envelopes for Efficient Estimation in Multivariate Linear Regression. *Biometrika*, to appear.

University of Minnesota

E-mail: suzhihua@stat.umn.edu

University of Minnesota

E-mail: dennis@stat.umn.edu