Scandinavian Journal of Statistics Theory and Applications

Envelopes for multivariate linear regression with linearly constrained coefficients

Journal:	*Scandinavian Journal of Statistics*
Manuscript ID	SJS-22-011.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Cook, Dennis; University of Minnesota Twin Cities, School of Statistics Forzani, Liliana; IMAL, Liu, Lan; University of Minnesota at Twin Cities, Statistics

Envelopes for multivariate linear regression with linearly constrained coefficients

R. Dennis Cook,* Liliana Forzani† and Lan Liu‡

November 14, 2022

Abstract

A constrained multivariate linear model is a multivariate linear model with the columns of its coefficient matrix constrained to lie in a known subspace. This class of models includes those typically used to study growth curves and longitudinal data. Envelope methods have been proposed to improve the estimation efficiency in unconstrained multivariate linear models, but have not yet been developed for constrained models. We pursue that development in this article. We first compare the standard envelope estimator with the standard estimator arising from a constrained multivariate model in terms of bias and efficiency. To further improve efficiency, we propose a novel envelope estimator based on a constrained multivariate model. We show the advantage of our proposals by simulations and by studying the probiotic capacity to reduced Salmonella infection.

*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: dennis@stat.umn.edu).

†Liliana Forzani is Professor, Facultad de Ingeniería Química, UNL. Researcher of CONICET, Santa Fe, Argentina (E-mail: liliana.forzani@gmail.com).

‡Lan Liu is Associate Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: liu1815@gmail.com).

Key Words: Growth curves, envelope models, repeated measures

1 Introduction

Consider the multivariate linear regression model

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{1}$$

with stochastic response $\mathbf{Y}_i \in \mathbb{R}^r$, non-stochastic predictors $\mathbf{X}_i \in \mathbb{R}^p$, $\boldsymbol{\beta}_0 \in \mathbb{R}^r$, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ and error vectors $\boldsymbol{\varepsilon}_i$ independent copies of $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$. The predictors are naturally non-stochastic when they are selected by design. When the predictors are sampled, we condition on them at the outset and so treating them as non-stochastic because they are ancillary under model (1). Model (1) is unconstrained in the sense that each response is allowed a separate linear regression: the maximum likelihood estimator (MLE) of the j-th row of $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is the same as the estimator of the coefficients from the linear regression of the j-th response on $\mathbf{X}$. In many applications, particularly analyses of growth curves and longitudinal data, we may have information that $\mathrm{span}(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is contained in a known subspace $\mathcal{U}$ with basis matrix $\mathbf{U} \in \mathbb{R}^{r\times k}$. The classic dental data (Potthoff and Roy, 1964; Lee and Geisser, 1975; Rao, 1987; Lee, 1988) is an example of such a case.

Example 1 *A study of dental growth measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14. The goal was to study the growth measurement as a function of time and sex. We revisited this example using the methodology presented in this paper in Supplement Section 5.*

Let Y_{ik} denote the continuous measure of distance for child i at age t_k, for $t_k = 8, 10, 12, 14$, and let $\mathbf{X}_i$ denote the gender indicator for child i (1 for boy and 0 for girl). After graphical inspection, many researchers treated the population means for distance as linear in time for each gender. Following that, a mixed effects repeated measure model is $Y_{ik} = \alpha_{00} + b_{0i} + \alpha_{01}\mathbf{X}_i + (\alpha_{10} + b_{1i} + \alpha_{11}\mathbf{X}_i)t_k + \varepsilon^*_{ik}$, where $\boldsymbol{\varepsilon}^*_i = (\varepsilon^*_{i1}, \ldots, \varepsilon^*_{ir})^T \overset{\text{i.i.d.}}{\sim} N(0, \boldsymbol{\Sigma}^*)$, b_{0i} and b_{1i} denote the random intercept and slope, $(b_{0i}, b_{1i}) \overset{\text{i.i.d.}}{\sim} N(0, \mathbf{D})$, where $\mathbf{D} \in \mathbb{S}^{2\times 2}$. We rewrite this model as

$$\mathbf{Y}_i = \mathbf{U}\boldsymbol{\alpha}_0 + \mathbf{U}\boldsymbol{\alpha}\mathbf{X}_i + \epsilon_i \tag{2}$$

with $\mathbf{U} := (\mathbf{1}, \mathbf{t})$ with $\mathbf{t} = (8, 10, 12, 14)^T$, $\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10})$ $\boldsymbol{\alpha} = (\alpha_{01}, \alpha_{11})$, $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}^*_i + b_{0i}1_{r\times 1} + b_{1i}\mathbf{t}$ and $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} N(0, \boldsymbol{\Sigma})$. Applying the same ideas to just $\boldsymbol{\beta}$ in (1), so $\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{U}$ without requiring that $\text{span}(\boldsymbol{\beta}_0) \subseteq \mathcal{U}$, leads to the model

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\alpha}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \; i = 1, \ldots, n. \tag{3}$$

Let $\mathcal{B} = \text{span}(\boldsymbol{\beta})$. If we set $\mathbf{U} = 1_r$, so in model (3) $\boldsymbol{\alpha}$ is a row vector of length p, then the mean functions for the individual responses are parallel. Although motivated in the context of the dental data, we use models (2) and (3) as general forms that can be adapted to different applications by varying the choice of $\mathbf{U}$, referring to them as constrained multivariate linear models. Cooper and Evans (2002) used a version of model (2) with $\mathbf{U}$ reflecting charge balance constraints on chemical constituents of water samples.

Constrained models occur in various areas including growth curve and longitudinal studies where the elements of $\mathbf{Y}_i$ are repeated observations on the i-th experimental unit over time. It is common in such settings to model the rows of $\mathbf{U}$ as a user-specified vector-valued function $\mathbf{u}(t) \in \mathbb{R}^k$ of time t, the i-th row of $\mathbf{U}$ then being $\mathbf{u}^T(t_i)$. Polynomial bases $\mathbf{u}^T(t) = (1, t, t^2, \ldots, t^{k-1})$ are prevalent, particularly in the foundational work of Potthoff and Roy (1964), Rao (1965), Grizzle and Allen (1969) and others, but splines (Nummi and Koskela, 2008) or other basis constructions (Izenman and Williams, 1989) could be used as well. In longitudinal studies, model (2) might be used when modeling profiles, while model (3) could be used when modeling just profile differences. For instance, if $\mathbf{X} = 0, 1$ is a population indicator then under model (2) the mean profiles are modeled as $\mathbf{U}\boldsymbol{\alpha}_0$ and $\mathbf{U}(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha})$, while under model (3) the profile means are $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\alpha}$. It is known in the literature that constrained models gain efficiency in the estimators compared with model (1), provided that $\mathbf{U}$ is correctly specified. However, it may be very difficult to correctly specify $\mathbf{U}$ in some applications, as in the following study:

Example 2 *Kenward (1987). An experiment was carried out to compare two treatments for the control of gut worm in cattle. Each treatments was randomly assigned to 30 cows whose weights were measured at* 2*,* 4*,* $6, \ldots, 18$ *and* 19 *weeks after treatment. The goal of the experiment was to see if a differential treatment effect could be detected and, if so, the time point when the difference was first manifested.*

The constrained models (2) and (3) require that we select $\mathcal{U}$. Lacking prior knowledge, it is natural to inspect plots of the average weight by time, as shown in Figure 1. It seems clear from the figure that it would be difficult to model the treatment profiles, particularly their two crossing points, without running into problems of over fitting. Envelopes provide a way to model data like this without specifying a subspace $\mathcal{U}$.

Envelope methodology is based on a relatively new paradigm for dimension reduction that,

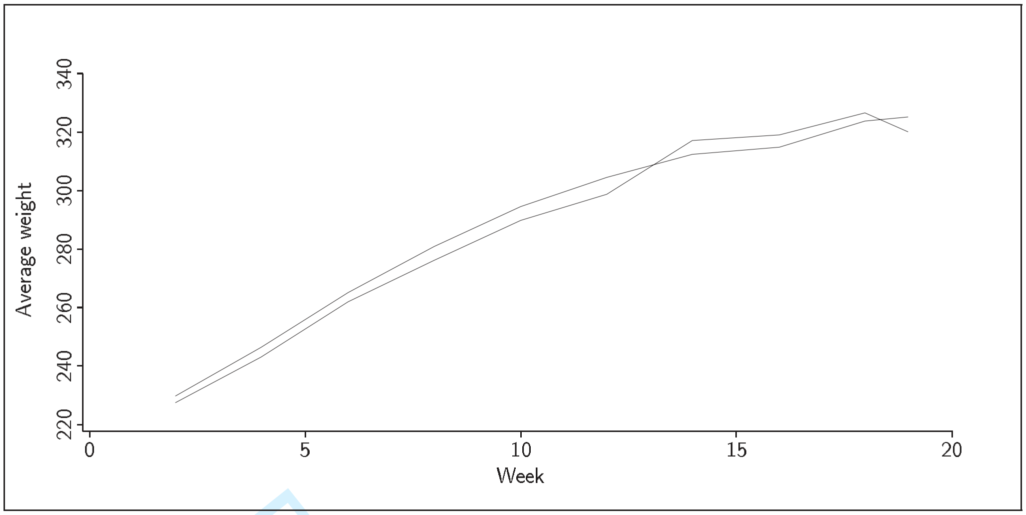


Figure 1: Cattle data: Average weight by treatment and time.

when applied in the context of model (1), has some similarity with constrained multivariate models. Briefly, envelopes produce a re-parameterization of model (1) in terms of a basis $\mathbf{\Gamma} \in \mathbb{R}^{r \times u}$ for the smallest reducing subspace of $\mathbf{\Sigma}$ that contains $\mathcal{B}$. Like the constrained model, envelopes produce an upper bound for $\mathcal{B}$, $\mathcal{B} \subseteq \mathrm{span}(\mathbf{\Gamma})$, but unlike the constrained model, the bound is unknown and must be estimated. Also, unlike the constrained model, $\mathbf{\Gamma}^T\mathbf{Y}$ contains the totality of $\mathbf{Y}$ that is affected by changing $\mathbf{X}$. Since $\mathcal{B} \subseteq \mathrm{span}(\mathbf{\Gamma})$, we have $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$. Model (1) can be re-paremeterized to give its envelope counterpart. For $i = 1, \dots, n$,

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \tag{4}$$

where $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r \times r}$, orthogonal, $\mathbf{\Omega} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma} > 0$ and $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0 > 0$. Envelopes are reviewed in more detail in Section 2.2.

Comparing (2)–(3) with (4), both express $\boldsymbol{\beta}$ as a basis times a coordinate matrix: $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ in (2)–(3) and $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$ in (4). However, as mentioned previously, $\mathbf{\Gamma}$ is estimated but $\mathbf{U}$ is assumed known. Envelopes were first proposed by Cook et al. (2007) to facilitate dimension reduction and later were shown by Cook et al. (2010) to have the potential massive efficiency gains relative to the standard MLE of $\boldsymbol{\beta}$, and for these gains to be passed on to other tasks such as prediction. There are now a number of extensions and applications of this basic envelope methodology, each demonstrating the potential for substantial efficiency gains (Su and Cook, 2011; Cook and Zhang, 2015a,b; Forzani and Su, 2021; Su et al., 2016; Li and Zhang, 2017; Rekabdarkolaee et al., 2020). Studies over the past several years have demonstrated repeatedly that sometimes the efficiency gains of the envelope methods relative to standard methods amount to increasing the sample size many times over. See Cook (2018) for a review and additional extensions of envelope methodology.

The choice between a constrained model, (2) or (3), and the envelope model (4) hinges on the ability to correctly specify an upper bound $\mathcal{U}$ for $\mathrm{span}(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ or $\mathcal{B}$. In Section 2 we obtain the MLE estimators and the asymptotic variances for the parameters of model (2) and show that if we have a correct parsimonious basis $\mathbf{U}$ then the constrained models are more efficient. But if bias is present or if we use a correct but excessive $\mathbf{U}$, then the envelope model (4) can be much more efficient. To the best of our knowledge, no such a comparison has been investigated theoretically or empirically in the literature despite the similarity between both models. Although considerable methodology has been developed for the envelope version (4) of the unconstrained model (1), there are apparently no envelope counterparts available for the class of models represented by (2) and (3) when a correct parsimonious $\mathbf{U}$ is available. In Section 3 we adapt the present envelope paradigm to model versions for (2) and (3) to achieve efficiency gains over those models. Simulations to support our finding are given in Section 4 and in Section 5 we compare our methodology with others in an example. We conclude the paper with a discussion section. Proofs for all propositions, two extra data set comparison and discussions of related issues are available in a Supplement to this article.

Notational conventions. Given a sample $(\mathbf{a}_i, \mathbf{b}_i), i = 1, \ldots, n$, let $\mathbf{T}_{\mathbf{a},\mathbf{b}} = n^{-1}\sum_{i=1}^n \mathbf{a}_i\mathbf{b}_i^T$ denote the matrix of raw second moments, and let $\mathbf{T}_{\mathbf{a}} = n^{-1}\sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^T$. For raw second moments involving $\mathbf{Y}_S$ and $\mathbf{Y}_D$ defined bellow we use S and D as subscripts. We use a subscript 1 in residuals computed from a model containing a vector of intercepts. The absence of a 1 indicates no intercept was included. For instance, $\mathbf{R}_{\mathbf{a}|\mathbf{b}}$ means the residuals from the regression of $\mathbf{a}$ on $\mathbf{b}$ without an intercept vector, $\mathbf{a}_i = \boldsymbol{\beta}\mathbf{b}_i + \mathbf{e}$, while $\mathbf{R}_{\mathbf{a}|(1,\mathbf{b})}$ means those with an intercept vector, $\mathbf{a}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{b}_i + \mathbf{e}_i$. Similarly, $\mathbf{R}_{D|S}$ means a residual from the regression of $\mathbf{Y}_D$ and $\mathbf{Y}_S$ without an intercept, and $\mathbf{R}_{D|(1,S)}$ with an intercept.

Sample variances are written as $\mathbf{S}_{\mathbf{a}} = n^{-1}\sum_{i=1}^n (\mathbf{a}_i - \bar{\mathbf{a}})(\mathbf{a}_i - \bar{\mathbf{a}})^T$ and sample covariance matrices are written as $\mathbf{S}_{\mathbf{a},\mathbf{b}} = n^{-1}\sum_{i=1}^n (\mathbf{a}_i - \bar{\mathbf{a}})(\mathbf{b}_i - \bar{\mathbf{b}})^T$. For variances and covariances involving $\mathbf{Y}_D$ and $\mathbf{Y}_S$ we again use D and S as subscripts, e.g. $\mathbf{S}_D = n^{-1}\sum_{i=1}^n (\mathbf{Y}_{Di} - \bar{\mathbf{Y}}_D)(\mathbf{Y}_{Di} - \bar{\mathbf{Y}}_D)^T$. $\mathbf{S}_{\mathbf{a}|\mathbf{b}}$ denotes the covariance matrix of the residuals from fit of the model $\mathbf{a}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{b}_i + \mathbf{e}_i$, which always includes an intercept. That is, $\mathbf{S}_{\mathbf{a}|\mathbf{b}} = n^{-1}\sum_{i=1}^n \mathbf{R}_{\mathbf{a}|(1,\mathbf{b}),i}\mathbf{R}_{\mathbf{a}|(1,\mathbf{b}),i}^T$. Similarly, $\mathbf{S}_{D|S} = \sum_{i=1}^n \mathbf{R}_{D|(1,S),i}\mathbf{R}_{D|(1,S),i}^T$.

We use $\mathrm{span}(\mathbf{A})$ to denote the subspace spanned by the columns of the matrix $\mathbf{A}$. The projection onto $\mathcal{S} = \mathrm{span}(\mathbf{A})$ will be denoted using either the subspace itself $\mathbf{P}_{\mathcal{S}}$ or its basis $\mathbf{P}_{\mathbf{A}}$. Projections onto an orthogonal complement will be denoted similarly using $\mathbf{Q}_{(\cdot)} = \mathbf{I} - \mathbf{P}_{(\cdot)}$. For a subspace $\mathcal{S}$ and conformable matrix $\mathbf{B}$, $\mathbf{B}\mathcal{S} = \{\mathbf{B}S \mid S \in \mathcal{S}\}$. If an estimator $\mathbf{a} \in \mathbb{R}^r$ of $\boldsymbol{\alpha} \in \mathbb{R}^r$ has the property that $\sqrt{n}(\mathbf{a} - \boldsymbol{\alpha})$ is asymptotically normal with mean 0 and variance $\mathbf{A}$, we write $\mathrm{avar}(\sqrt{n}\mathbf{a}) = \mathbf{A}$ to denote its asymptotic variance.

2 Comparison of the envelope and constrained estimators

Models (2)–(3) and (4) are similar in the sense that $\boldsymbol{\beta}$ is represented as a basis times a coordinate matrix, $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ in (2)–(3) and $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ in (4). It might be thought that (2) and (3) would yield better estimators because $\mathbf{U}$ is known while $\boldsymbol{\Gamma}$ is not, but that turns out not to be true in general. This is in part because we may have $\mathcal{B} \not\subseteq \mathcal{U}$, which raises the issue of bias as discussed in Section 2.3, and in part because the envelope model capitalizes automatically on the structure in $\boldsymbol{\Sigma}$, which can improve efficiency as discussed in Section 2.4. Our general conclusion is that, in practice, it may be necessary to compare their fits before selecting an estimator and that the envelope estimator may have a clear advantage when there is uncertainty in the choice of $\mathcal{U}$, as illustrated in Figure 1.

Developments under models (2) and (3) are very similar since they differ only on how the intercept is handled. In the remainder of this article we focus on model (2) and comment from time to time on modifications necessary for model (3).

2.1 Maximum likelihood estimators for constrained models

Our treatment of maximum likelihood estimation from (2) is based on linearly transforming $\mathbf{Y}$. Let $\mathbf{U}_0$ be a semi-orthogonal basis matrix for $\mathcal{U}^{\perp}$, and let $\mathbf{W} = (\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}, \mathbf{U}_0) := (\mathbf{W}_1, \mathbf{W}_2)$. Then the transformed model becomes

$$\mathbf{W}^T\mathbf{Y}_i := \begin{pmatrix} \mathbf{Y}_{Di} \\ \mathbf{Y}_{Si} \end{pmatrix} = \begin{pmatrix} (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}_i \\ \mathbf{U}_0^T\mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}\mathbf{X}_i \\ 0 \end{pmatrix} + \mathbf{W}^T\boldsymbol{\varepsilon}_i, \ i = 1, \dots, n, \quad (5)$$

where $\mathbf{Y}_{Di} \in \mathbb{R}^k$ and $\mathbf{Y}_{Si} \in \mathbb{R}^{r-k}$ with k the number of columns of $\mathbf{U}$. The transformed variance can be represented block-wise as $\boldsymbol{\Sigma}_{\mathbf{W}} := \mathrm{var}(\mathbf{W}^T\boldsymbol{\varepsilon}) = (\mathbf{W}_i^T\boldsymbol{\Sigma}\mathbf{W}_j)$, $i, j = 1, 2$, where $\boldsymbol{\Sigma}$ is as defined for model (2). The mean $E(\mathbf{Y}_D \mid \mathbf{X})$ depends non-trivially on $\mathbf{X}$ and thus, as indicated by the subscript D, we think of $\mathbf{Y}_D$ as providing direct information about the regression. On the other hand, $E(\mathbf{Y}_S \mid \mathbf{X}) = 0$ and thus $\mathbf{Y}_S$ provides no direct information but may provide useful subordinate information by virtue of its association with $\mathbf{Y}_D$.

To find the MLEs from model (5), we write the full log likelihood as the sum of the log likelihoods for the marginal model for $\mathbf{Y}_S \mid \mathbf{X}$ and the conditional model for $\mathbf{Y}_D \mid (\mathbf{X}, \mathbf{Y}_S)$:

$$\mathbf{Y}_{Si} \mid \mathbf{X} = \mathbf{e}_{Si} \tag{6}$$

$$\mathbf{Y}_{Di} \mid (\mathbf{X}_i, \mathbf{Y}_{Si}) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}\mathbf{X}_i + \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} + \mathbf{e}_{D|Si}, \tag{7}$$

where $\boldsymbol{\phi}_{D|S} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}_0(\mathbf{U}_0^T\boldsymbol{\Sigma}\mathbf{U}_0)^{-1} \in \mathbb{R}^{k\times(r-k)}$, $\mathbf{e}_{D|S} = \mathbf{W}_1^T\boldsymbol{\varepsilon}$, $\mathbf{e}_S = \mathbf{W}_2^T\boldsymbol{\varepsilon}$. The variances of the errors are $\boldsymbol{\Sigma}_S := \mathrm{var}(\mathbf{e}_S) = \mathbf{U}_0^T\boldsymbol{\Sigma}\mathbf{U}_0$ and $\boldsymbol{\Sigma}_{D|S} := \mathrm{var}(\mathbf{e}_{D|S}) = (\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U})^{-1}$. The number of free real parameters in this conditional model is $N_{\mathrm{cm}}(k) = k(p+1) + r(r+1)/2$. The subscript 'cm' is used to indicate estimators arising from the conditional model (7). The MLE and its asymptotic variance for (2) are

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}} = \mathbf{S}_{D,\mathbf{R}_{\mathbf{X}|(1,S)}}\mathbf{S}_{\mathbf{X}|S}^{-1} = (\mathbf{S}_{D,\mathbf{X}} - \mathbf{S}_{D,S}\mathbf{S}_S^{-1}\mathbf{S}_{S,\mathbf{X}})\mathbf{S}_{\mathbf{X}|S}^{-1} \tag{8}$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{cm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}} = \mathbf{U}\mathbf{S}_{D,\mathbf{R}_{\mathbf{X}|(1,S)}}\mathbf{S}_{\mathbf{X}|S}^{-1} = \mathbf{U}(\mathbf{S}_{D,\mathbf{X}} - \mathbf{S}_{D,S}\mathbf{S}_S^{-1}\mathbf{S}_{S,\mathbf{X}})\mathbf{S}_{\mathbf{X}|S}^{-1} \tag{9}$$

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{D|S} \tag{10}$$

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{cm}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{U}\boldsymbol{\Sigma}_{D|S}\mathbf{U}^T, \tag{11}$$

The estimation for model (3) requires just a few modifications of the procedure for model (2). All modifications stem from the presence of an intercept vector in model (6), which becomes

$\mathbf{Y}_S = \mathbf{W}_2^T\boldsymbol{\beta}_0 + \mathbf{e}_S$. The variance $\boldsymbol{\Sigma}_S$ is estimated as $\widehat{\boldsymbol{\Sigma}}_S = \mathbf{S}_S$ with corresponding changes in the estimator of $\boldsymbol{\Sigma}$, and the estimator of the intercept $\mathbf{W}_2^T\boldsymbol{\beta}_0$ is just $\bar{\mathbf{Y}}_S$. The intercept in (7) is redefined as $\boldsymbol{\alpha}_0 = \mathbf{W}_1^T\boldsymbol{\beta}_0 - \boldsymbol{\phi}_{D|S}\mathbf{W}_2^T\boldsymbol{\beta}_0$. The MLE of $\boldsymbol{\beta}_0$ in model (3) can be constructed in a straightforward way from the estimators of $\boldsymbol{\alpha}_0$, $\mathbf{W}_2^T\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}_{D|S}$. The number of real parameters in (6) becomes $N_{\text{cm}} + r - k$. The estimators of the parameters in (7) are unchanged. In particular, $\widehat{\boldsymbol{\alpha}}_{\text{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ along with their asymptotic variances are the same under models (2) and (3), although different $\mathbf{U}$'s might be used in their construction.

2.2 Envelope estimator stemming from Model (1)

Consider a subspace $\mathcal{S} \subseteq \mathbb{R}^r$ that satisfies the two conditions (i) $\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ and (ii) $\mathbf{P}_{\mathcal{S}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{Y} \mid \mathbf{X}$. Condition (i) insures that the marginal distribution of $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ does not depend on $\mathbf{X}$, while statement (ii) insures that, given $\mathbf{X}$, $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ cannot provide material information via an association with $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$. Together these conditions imply that the impact of $\mathbf{X}$ on the distribution of $\mathbf{Y}$ is concentrated solely in $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$. One motivation underlying envelopes is then to characterize linear combinations $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ that are unaffected by changes in $\mathbf{X}$ and that produce gains in estimative and predictive efficiency.

In terms of model (1), condition (i) holds if and only if $\mathcal{B} \subseteq \mathcal{S}$ and condition (ii) holds if and only if $\mathcal{S}$ is a reducing subspace of $\boldsymbol{\Sigma}$; that is, $\mathcal{S}$ must decompose $\boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}}$. The intersection of all subspaces with these properties is by construction the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\mathcal{B}$, which is called the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$ and is represented as $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ (Cook et al., 2010). These consequences of conditions (i) and (ii) can be incorporated into model (1) by using a basis, leading to model (4). Let $u \in \{0, 1, \ldots, r\}$ denote the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The number of free real parameters is $N_{\text{em}} = r + pu + r(r+1)/2$. The subscript 'em' is used to indicate selected quantities arising from this envelope model. The goal here is still to estimate $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$. Cook et al. (2010) derived the maximum likelihood envelope estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ along with their asymptotic variances. They showed that substantial efficiency gains in estimation of $\boldsymbol{\beta}$ are possible under this model, particularly when a norm of $\mathrm{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \boldsymbol{\Omega}_0$ is considerably larger than the same norm of $\mathrm{var}(\boldsymbol{\Gamma}^T\mathbf{Y}) = \boldsymbol{\Omega}$.

Given the envelope dimension u, Cook et al. (2010) proved that the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ of $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ from envelope model (4) has asymptotic variance given by

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{em}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T), \tag{12}$$

where for a $\mathbf{C} \in \mathbb{R}^{p \times p}$, $\mathbf{M}(\mathbf{C}) := \boldsymbol{\eta}\mathbf{C}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}$ and $\dagger$ denotes the Moore-Penrose inverse. Cook et al. (2010) showed that $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{em}})) \leq \mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{um}}))$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ is the MLE under the uncontrained model (1). In consequence, estimators from the envelope model (4) are always superior to those from the unconstrained multivariate model (1). Cook et al. (2010) also showed that the envelope estimator is $\sqrt{n}$-consistent even when the normality assumption is violated as long as the data has finite fourth moments.

2.3 Potential bias in $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$

Assuming that $\mathcal{B} \subseteq \mathcal{U}$, $\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ are unbiased estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. However, if $\mathcal{B} \not\subseteq \mathcal{U}$ then both $\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ are biased, which could materially affect the estimators: $E(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}) = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\beta}$ and $E(\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}) = \mathbf{P}_{\mathbf{U}}\boldsymbol{\beta}$. Consequently, the bias in $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ is $\boldsymbol{\beta} - \mathbf{P}_{\mathbf{U}}\boldsymbol{\beta} = \mathbf{Q}_{\mathbf{U}}\boldsymbol{\beta}$. A nonzero bias must necessarily dominate the mean squared error asymptotically and so could limit the utility of $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$. Simulation results that illustrate the potential bias effects are discussed in Section

4.2. We assume that $\mathcal{B} \subseteq \mathcal{U}$ for the remainder of this article except for where otherwise indicated.

2.4 Comparison of asymptotic variances of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$

We now compare the asymptotic variances of the envelope and constrained estimators of $\boldsymbol{\beta}$, (12) and (11). Depending on the dimensions involved, the relationship between $\mathcal{U}$ and the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and other factors, the difference between the asymptotic covariance matrices for the estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ from these two models can be positive definite, negative definite or indefinite. Since all comparisons are in terms of $\boldsymbol{\beta}$'s, we assume without loss of generality that $\mathbf{U}$ is a semi-orthogonal matrix. Also, since $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ is the same under models (2) and (3) we do not distinguish between these two models in this section.

2.4.1 $\mathcal{B} \subseteq \mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$

Assuming that $\mathcal{U}$ is correct so that $\mathcal{B} \subseteq \mathcal{U}$ and $\mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can simplify the variance comparison:

Proposition 2.1 *If* $\mathcal{B} \subseteq \mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, *then* $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{cm}})) \leq \mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{em}}))$.

In consequence, under this hypothesis, the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ is superior to the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$. However, this comparison may be seen as loaded in favor of $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ since the constrained estimator uses the additional knowledge that $\mathcal{B} \subseteq \mathcal{U}$ and the envelope estimator does not. Additionally, neither estimator makes use of the proposition's hypothesis. The next proposition provides help in assessing the impact of the hypothesis on the underlying structure by connecting it with $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$, the $\boldsymbol{\Sigma}$-envelope of $\mathcal{U}$.

Proposition 2.2 *Assume that* $\mathcal{B} \subseteq \mathcal{U}$. *Then*

1. $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$,

2. *$\mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ if and only if $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$,*

3. *If $\mathrm{rank}(\boldsymbol{\alpha}) = k$ then $\mathcal{B} = \mathcal{U}$ and $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$.*

This proposition says essentially that if $\mathcal{B} \subseteq \mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ we can start with model (1) and parameterize in terms of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$ rather than $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. A key distinction here is that $\mathcal{U}$ is known while $\mathcal{B}$ is not. In consequence, we expect less estimative variation when parameterizing (1) in terms of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$ instead of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Since $\mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$ we can construct a semi-orthogonal basis for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$ as $\boldsymbol{\Gamma} = (\mathbf{U}, \boldsymbol{\Gamma}_2)$ with $\mathbf{U}_0 = (\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0)$ and, recognizing that $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, we get a new model

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{U}\boldsymbol{\alpha}_0 + \mathbf{U}\boldsymbol{\alpha}\mathbf{X}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{13}$$

Consider estimating $\boldsymbol{\alpha}$ from this model using the steps sketched in Section 2.1, and partition $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij})$ to conform to the partition of $\boldsymbol{\Gamma} = (\mathbf{U}, \boldsymbol{\Gamma}_2)$. The envelope structure of (13) induces a special structure on the reduced model that corresponds to (6)–(7): $\boldsymbol{\Sigma}_S = \mathrm{bdiag}(\boldsymbol{\Omega}_{22}, \boldsymbol{\Omega}_0)$ is block diagonal, $\boldsymbol{\Sigma}_{D|S} = \boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$ and $\boldsymbol{\phi}_{D|S} = (\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}, 0)$. It can now be shown that the estimators of $\boldsymbol{\alpha}$ from the constrained model (6)–(7) and from (13) have the same asymptotic variance. In other words, if we neglect the hypothesized condition that $\mathcal{U} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ then the constrained estimator is better, but if we formulate the envelope model making use of that condition then the constrained and envelope estimators are asymptotically equivalent.

Rao (1967) posited a simple structure for the analysis of balanced growth curve data (See also Geisser, 1970; Lee and Geisser, 1975; Geisser, 1981; Lee, 1988; Pan and Fang, 2002). In our context, Rao's structure is obtained by assuming that $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U}) = \mathcal{U}$, which corresponds to model

(13) with $\mathbf{\Gamma} = \mathbf{U}$, which seems too specialized to warrant further attention. Additional discussion of Rao's structure is available in Supplement Section 10.

2.4.2 $\mathcal{U} \supseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$

Assuming that $\mathcal{U} \supseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is another way to simplify the variance comparison. Let $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be an orthogonal matrix. Since $\mathcal{U} \supseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, we can construct semi-orthogonal bases $\mathbf{U} = (\mathbf{\Gamma}, \mathbf{\Gamma}_{01})$ and $\mathbf{\Gamma}_0 = (\mathbf{\Gamma}_{01}, \mathbf{\Gamma}_{02})$. Partition $\mathbf{\Omega}_0 = (\mathbf{\Omega}_{0,ij})$ to correspond to the partitioning of $\mathbf{\Gamma}_0$. Then

Proposition 2.3 *Assume that $\mathcal{U} \supseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and let $\mathbf{c} \in \mathbb{R}^r$. Then*

1. *If $\mathbf{c} \in \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ then* $\mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}) = \mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{em}})$.

2. *If $\mathbf{c} \in \mathrm{span}(\mathbf{\Gamma}_{02})$ then* $\mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}) \leq \mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{em}})$.

3. *If $\mathbf{c} \in \mathrm{span}(\mathbf{\Gamma}_{01})$, $\mathrm{rank}(\mathbf{M}(\mathbf{\Sigma}_{\mathbf{X}})) = \mathrm{rank}(\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \mathbf{\Omega}_0^{-1})$ and $\mathbf{\Omega}_{12} = 0$ then* $\mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}) \geq \mathrm{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{em}})$.

The main takeaway of this lemma is that the difference between the asymptotic covariance matrices for the estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ can be positive semi-definite or negative semi-definite, depending on the characteristics of the problem.

Although the above derivation is under two simple cases where $\mathcal{U}$ and the envelope space are nested, the conclusion actually holds for the general case: if we have a correct parsimoniously parameterized constrained model then the envelope model (4) is less efficient, but if the basis $\mathbf{U}$ in the constrained model is incorrect or excessively parameterized, then envelopes can be much more efficient. This motivated us to incorporate envelopes into the constrained model so that we can further improve efficiency if constraints are reasonably well modeled for the data

3 Envelopes in constrained models

In this section, we consider two different ways of imposing envelopes in a constrained model when $\mathcal{B} \subseteq \mathcal{U}$. As previously dane, we focus on envelope estimators in the constrained model (2) and later describe the modifications necessary for model (3). In Section 3.1 we describe the envelope estimation of $\boldsymbol{\alpha}$ when there is an application-grounded basis $\mathbf{U}$ that is key to interpretation and inference. In Section 3.2 we address envelope estimation of $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$. Here the choice of basis $\mathbf{U}$ has no effect on the MLE of $\boldsymbol{\beta}$ under the constrained models (2), but it does affect the envelope estimator of $\boldsymbol{\beta}$. Basis selection is also addressed in Section 3.2.

3.1 Enveloping $\boldsymbol{\alpha}$

Estimation of $\boldsymbol{\alpha}$ will be of interest when it is desirable to interpret $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ in terms of its coordinates $\boldsymbol{\alpha}$ relative to the known application-grounded basis $\mathbf{U}$. Let $\mathcal{A} = \mathrm{span}(\alpha)$. The envelope estimator of $\boldsymbol{\alpha}$ in model (5) can be found by first transforming (5) into (6)–(7) and then parameterizing (7) in terms of a semi-orthogonal basis matrix $\boldsymbol{\phi} \in \mathbb{R}^{k\times u}$ for $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$, the $\boldsymbol{\Sigma}_{D|S}$-envelope of $\mathcal{A}$ with dimension $u \leq k$. Since $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{D|S}$ is in the form of a Kronecker product that allows separation of row and column effects of $\boldsymbol{\alpha}$, this structure follows also from the theory of Cook and Zhang (2015a,b) for matrix-valued envelope estimators based on envelopes of the form $\mathbb{R}^p \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$, where $\oplus$ denotes the direct sum.

Let $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ be an unconstrained matrix giving the coordinates of $\boldsymbol{\alpha}$ in terms of a semi-orthogonal basis matrix $\boldsymbol{\phi}$, so $\boldsymbol{\alpha} = \boldsymbol{\phi}\boldsymbol{\eta}$, and let $(\boldsymbol{\phi}, \boldsymbol{\phi}_0) \in \mathbb{R}^{k\times k}$ be an orthogonal matrix. Then the envelope version of model (6)–(7) is

$$\begin{aligned} \mathbf{Y}_{Si} &= \mathbf{e}_{Si} \\ \mathbf{Y}_{Di} \mid (\mathbf{X}_i, \mathbf{Y}_{Si}) &= \boldsymbol{\alpha}_0 + \boldsymbol{\phi}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} + \mathbf{e}_{D|Si}, \\ \boldsymbol{\Sigma}_{D|S} &= \boldsymbol{\phi}\boldsymbol{\Omega}\boldsymbol{\phi}^T + \boldsymbol{\phi}_0\boldsymbol{\Omega}_0\boldsymbol{\phi}_0^T, \end{aligned} \tag{14}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(k-u)\times(k-u)}$ are positive definite matrices. Part of this model can be seen as a version of the partial envelope model (Su and Cook, 2011)

The total real parameters in model (14) is $N_{\mathrm{ecm}}(u) = k + pu + r(r+1)/2$, which reduces to that given previously for model (6)–(7) when $u = k$. The subscript ecm is used to indicate selected key quantities that arise from enveloping $\mathcal{A}$ in the constrained model (2). A basis $\widehat{\boldsymbol{\phi}}$ for the MLE $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$ of $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$ is constructed as

$$\widehat{\boldsymbol{\phi}} = \arg\min_{\mathbf{G}} \log|\mathbf{G}^T\mathbf{S}_{D|(\mathbf{X},S)}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{D|S}^{-1}\mathbf{G}|, \tag{15}$$

where the minimum is computed over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{k\times u}$ with $u \leq k$. The fully maximized log likelihood is

$$\hat{L}_u = c - \frac{n}{2}\left\{\log|\mathbf{T}_S| + \log|\mathbf{S}_{D|S}| + \log|\widehat{\boldsymbol{\phi}}^T\mathbf{S}_{D|(\mathbf{X},S)}\widehat{\boldsymbol{\phi}}| + \log|\widehat{\boldsymbol{\phi}}^T\mathbf{S}_{D|S}^{-1}\widehat{\boldsymbol{\phi}}|\right\}. \tag{16}$$

where $c = n\log|\mathbf{W}| - (nr/2)(1+\log(2\pi))$ with the $\log|\mathbf{W}|$ term corresponding to the Jacobian transformation back to the scale of $\mathbf{Y}$.

Once $\widehat{\boldsymbol{\phi}}$ is obtained we get the following envelope estimators for constrained model (2). Specifically, we have

- $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$, $\widehat{\boldsymbol{\alpha}}_{\text{ecm}} = \mathbf{P}_{\widehat{\boldsymbol{\Phi}}}\widehat{\boldsymbol{\alpha}}_{\text{cm}} = \widehat{\boldsymbol{\phi}}\widehat{\boldsymbol{\eta}}$, and $\widehat{\boldsymbol{\alpha}}_0 = \bar{\mathbf{Y}}_D - \widehat{\boldsymbol{\alpha}}_{\text{ecm}}\bar{\mathbf{X}} - \widehat{\boldsymbol{\phi}}_{D|S}\bar{\mathbf{Y}}_S$.
- $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\phi}}^T\widehat{\boldsymbol{\alpha}}_{\text{cm}}$, $\widehat{\boldsymbol{\phi}}_{D|S} = \mathbf{S}_{D,S}\mathbf{S}_S^{-1} - \widehat{\boldsymbol{\alpha}}_{\text{ecm}}\mathbf{S}_{\mathbf{X},S}\mathbf{S}_S^{-1}$, and $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$
- $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\phi}}^T\mathbf{S}_{D|(\mathbf{X},S)}\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\phi}}_0^T\mathbf{S}_{D|S}\widehat{\boldsymbol{\phi}}_0$,
- $\widehat{\boldsymbol{\Sigma}}_{D|S} = \widehat{\boldsymbol{\phi}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\phi}}^T + \widehat{\boldsymbol{\phi}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\phi}}_0^T$ and $\widehat{\boldsymbol{\Sigma}}_S = \mathbf{T}_S$.

The variances $\boldsymbol{\Sigma}_{\mathbf{W}}$ and $\boldsymbol{\Sigma}$ can be estimated as indicated in Section 2.1. The variances $\boldsymbol{\Sigma}_{\mathbf{W}}$ and $\boldsymbol{\Sigma}$ can be estimated as indicated in Section 2.1. The asymptotic variances for $\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}}$ can be deduced from recognizing that in our application $\mathbf{Y}_S$ is random, $\mathbf{X}$ is fixed, and the distribution of $\mathbf{Y}_S|\mathbf{X}$ is the same as that of the marginal of $\mathbf{Y}_S$:

$$\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{ecm}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\phi}\boldsymbol{\Omega}\boldsymbol{\phi}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\phi}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\phi}_0^T) \tag{17}$$

$$\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ecm}})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{U}\left[\boldsymbol{\phi}\boldsymbol{\Omega}\boldsymbol{\phi}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\phi}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\phi}_0^T)\right]\mathbf{U}^T \tag{18}$$

We have $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{ecm}})) \le \text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{cm}}))$ and $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ecm}})) \le \text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{cm}}))$ being equal when $u = k$, so using an envelope in the constrained model always improves estimation asymptotically.

Because $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A}) \subseteq \mathbb{R}^k$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) \subseteq \mathbb{R}^r$ and $k \le r$, it is reasonable to expect that $\dim\{\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})\} \le \dim\{\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})\}$, as we have estimated in many examples. However, this relationship between the envelope dimension is not guaranteed in general. The following proposition gives sufficient conditions to bound $\dim\{\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})\}$.

Proposition 3.1 *Assume that* $\mathbf{U} = (\mathbf{\Gamma G}, \mathbf{\Gamma}_0\mathbf{G}_0)$*, where the* $\mathbf{\Gamma}$*'s are as defined for model (4), and that* $\mathbf{G} \in \mathbb{R}^{u\times u_1}$ *and* $\mathbf{G}_0 \in \mathbb{R}^{(r-u)\times(k-u_1)}$ *both have full column rank, so that* $u_1 \leq u$*. Then* $\dim\{\mathcal{E}_{\mathbf{\Sigma}_{D|S}}(\mathcal{A})\} \leq u_1 \leq \dim\{\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})\}$.

We can assess the model fitting of (14) using BIC, assuming that the error terms follows a normal distribution. That is, we can compare the constrained envelope model with alternative models by inspecting whether $-2\hat{L}_u + N_{ecm}(u)log(n)$ is small. By comparing the BICs of the constrained model with different dimensions u, we can also select the dimension that has the best fit. More about estimating the envelope dimension is given in Supplement 9.

3.2 Enveloping $\boldsymbol{\beta}$

Estimation of $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ will be of interest in applications where prediction is important or where $\mathbf{U}$ is selected based on convenience, say, rather than on criteria that facilitate understanding and inference. For instance, if $\mathbf{X}$ serves to indicate different treatments then plots of the columns of $\boldsymbol{\beta}$ versus time give a visual comparisons of the treatment profiles. The choice of $\mathcal{U}$ is of course relevant to estimation of $\boldsymbol{\beta}$, but a basis $\mathbf{U}$ is not uniquely determined. While this flexibility has no effect on the MLE of $\boldsymbol{\beta}$ under the constrained model (2), it does affect the envelope estimator of $\boldsymbol{\beta}$. This raises the issue of selecting a good basis for the purpose of estimating $\boldsymbol{\beta}$ via envelopes.

Consider re-parameterizing $\mathbf{U}$ as $\mathbf{U}\mathbf{V}^{-1}$ and $\boldsymbol{\alpha}$ as $\mathbf{V}\boldsymbol{\alpha}$ for some positive definite matrix $\mathbf{V} \in \mathbb{R}^{k\times k}$, giving $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha} = (\mathbf{U}\mathbf{V}^{-1})(\mathbf{V}\boldsymbol{\alpha})$. We could use either $\mathcal{E}_{\mathbf{\Sigma}_{D|S}}(\mathcal{A})$ or $\mathcal{E}_{\mathbf{V}\mathbf{\Sigma}_{D|S}\mathbf{V}^T}(\mathbf{V}\mathcal{A})$ to estimate $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$ or, in terms of re-parameterized coordinates $\mathbf{V}\boldsymbol{\alpha}$, as $\widehat{\boldsymbol{\beta}}_{\text{ecm},\mathbf{V}} = \mathbf{U}\mathbf{V}^{-1}\widehat{(\mathbf{V}\boldsymbol{\alpha})}_{\text{ecm}}$. In general $\widehat{\boldsymbol{\beta}}_{\text{ecm}} \neq \widehat{\boldsymbol{\beta}}_{\text{ecm},\mathbf{V}}$ and we cannot tell which estimator is better. In this section, we show that the envelope estimator of $\boldsymbol{\beta}$ is invariant under orthogonal re-parameterization, so that we only need to consider diagonal re-parameterization: $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha} = (\mathbf{U}\mathbf{\Lambda}^{-1})(\mathbf{\Lambda}\boldsymbol{\alpha})$, where

$\mathbf{\Lambda}$ is a diagonal matrix with positive diagonal elements. In growth curve or longitudinal analyses, the columns of $\mathbf{U}$ may correspond to different powers of time, and then it seems natural to consider rescaling to bring the columns of $\mathbf{U}$ closer to the same scale.

In Supplement Section 3.1 we provided technical tools for demonstrating that the maximum likelihood envelope estimator of $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$, when $\mathbf{U}$ is semi-orthogonal, is simply $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$.

Thus, to consider the constrained model envelope under a linear transformation of $\mathbf{U}$, it suffices to consider a re-scaling transformation. That is, we consider $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha} = (\mathbf{U}\mathbf{\Lambda}^{-1})(\mathbf{\Lambda}\boldsymbol{\alpha})$, where $\mathbf{\Lambda} = \mathrm{diag}(1, \lambda_2, \ldots, \lambda_k)$. The first diagonal element of $\mathbf{\Lambda}$ is 1 to ensure identifiability. We follow the general logic of Cook and Su (2013) in their development of a scaled version of model (2).

Without loss of generality, we cast our discussion of scaling in the context of the conditional model (7). We assume that there is a scaling of the response $\mathbf{Y}_D$ so that the scaled response $\mathbf{\Lambda}\mathbf{Y}_D$ follows an envelope model in $\mathbf{\Lambda}\boldsymbol{\alpha}$ with the envelope $\mathcal{E}_{\mathbf{\Lambda}\mathbf{\Sigma}_{D|S}\mathbf{\Lambda}}(\mathbf{\Lambda}\mathcal{A})$ having dimension v and semi-orthogonal basis matrix $\mathbf{\Theta} \in \mathbb{R}^{k\times v}$. Let $(\mathbf{\Theta}, \mathbf{\Theta}_0)$ denote an orthogonal matrix. Then we can parametrize $\mathbf{\Lambda}\boldsymbol{\alpha} = \mathbf{\Theta}\boldsymbol{\eta}$ and $\mathbf{\Lambda}\mathbf{\Sigma}_{D|S}\mathbf{\Lambda} = \mathbf{\Theta}\mathbf{\Omega}\mathbf{\Theta}^T + \mathbf{\Theta}_0\mathbf{\Omega}\mathbf{\Theta}_0^T$; equivalently, this setup can also be viewed as a rescaling $\mathbf{U} \mapsto \mathbf{U}\mathbf{\Lambda}^{-1}$ of $\mathbf{U}$, since $\mathbf{\Lambda}\mathbf{Y}_D = \mathbf{\Lambda}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y} = (\mathbf{\Lambda}^{-1}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{-1})^{-1}\mathbf{\Lambda}^{-1}\mathbf{U}^T\mathbf{Y}$. Since $\mathbf{\Lambda}\mathbf{Y}_D$ is unobserved, we now transform back to the original scale for analysis, leading to the marginal model $\mathbf{Y}_{Si} \mid \mathbf{X} = \mathbf{e}_{Si}$ and conditional model

$$\begin{aligned} \mathbf{Y}_{Di} \mid (\mathbf{X}_i, \mathbf{Y}_{Si}) &= \boldsymbol{\alpha}_0 + \mathbf{\Lambda}^{-1}\mathbf{\Theta}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} + \mathbf{e}_{D|Si}, \\ \mathbf{\Sigma}_{D|S} &= \mathbf{\Lambda}^{-1}(\mathbf{\Theta}\mathbf{\Omega}\mathbf{\Theta}^T + \mathbf{\Theta}_0\mathbf{\Omega}_0\mathbf{\Theta}_0^T)\mathbf{\Lambda}^{-1}. \end{aligned} \tag{19}$$

The total real parameters in this scaled envelope model is $N_{\text{secm}}(v) = 2k - 1 + pv + r(r+1)/2$, where the subscript secm is used to indicate quantities arising from the scaled envelope version of the conditional model. For identifiability we typically need $N_{\text{secm}}(v) \leq N_{\text{cm}}$ or $p(k-v) \geq k-1$. The goal now is to estimate $\boldsymbol{\alpha}_0$, the coefficient matrix $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Theta}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{D|S}$, which requires the estimation of several constituent parameters. We presented in Supplement Section 3.2 the maximum likelihood estimators under this model and prove that the asymptotic variance of the estimator $\boldsymbol{\beta}_{\text{secm}}$ of $\boldsymbol{\beta}$ is $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{secm}})) = (\mathbf{I}_p \otimes \mathbf{U})\mathbf{V}_{\text{secm}}(\mathbf{I}_p \otimes \mathbf{U}^T)$ and therefore it is never less efficient than $\widehat{\boldsymbol{\beta}}_{\text{cm}}$.

3.3 Estimation under model (3)

The modifications necessary to adapt the results in Sections 3.1–3.2 for model (3) all stem from the new model for the subordinate response, $\mathbf{Y}_S = \mathbf{W}_2^T\boldsymbol{\beta}_0 + \mathbf{e}_S$, and the new definitions of $\boldsymbol{\alpha}_0 = \mathbf{W}_1^T\boldsymbol{\beta}_0 - \boldsymbol{\phi}_{D|S}\mathbf{W}_2^T\boldsymbol{\beta}_0$ for models (14) and (19). This implies that $\mathbf{T}_S$ is replaced by $\mathbf{S}_S$ throughout, including log likelihoods (16) and (13) and that the estimator of $\boldsymbol{\beta}_0$ can be constructed as indicated near the end of Section 2.1. There is no change in the objective functions (15) and (12), and consequently no change in the envelope estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

4 Simulations

4.1 Efficiency Comparison

We first evaluate the efficiency of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{em}}$, the constrained estimator $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ and the constrained envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{ecm}}$ using simulations in two scenarios. We also include the unconstrained estimator $\widehat{\boldsymbol{\beta}}_{\text{um}}$ as a reference. In the first scenario considered the eigenvalue corresponding to the material part is small relative to the immaterial part and the dimension of $\mathbf{U}$ is large; therefore the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{em}}$ is expected to have substantial efficiency gain. In the second scenario the eigenvalue of the immaterial part is small relative to the one of the material part and the envelope estimator is not expected to have substantial efficiency gain.

4.1.1 Scenario 1

The simulation for Scenario 1 is carried out in the following steps:

Step 1. We first generated a sample of size $n = 200$. For each individual i, we generated $p = 8$ predictors $\mathbf{X}_i$ from a multivariate normal distribution with mean 0 and variance $\mathbf{C}\mathbf{C}^T$, where each element in $\mathbf{C}$ is identically and independently distributed with a standard normal distribution $N(0, 1)$.

Step 2. Set $r = 20$, $u = 6$, $q = 15$, $q_1 = 4$ and $q_2 = q - q_1$. Set $\boldsymbol{\Omega} = \text{bdiag}(0.5\mathbf{I}_{u-q_1}, 1.5\mathbf{I}_{q_1})$ and $\boldsymbol{\Omega}_0 = 50\mathbf{I}_{r-u}$. Set $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) = \mathbf{O}$ and let $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\mathbf{O}$ is an orthogonal matrix obtained by singular value decomposition of a randomly generated matrix. Set $\boldsymbol{\eta} = \mathbf{K}_1\mathbf{K}_2$, where $\mathbf{K}_1 \in \mathbb{R}^{u\times q_1}$, $\mathbf{K}_2 \in \mathbb{R}^{q_1\times p}$ and each element in $\mathbf{K}_1$ and $\mathbf{K}_2$ is identically and independently generated from $N(0, 1)$. Let $\mathbf{U} = (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\boldsymbol{\phi}$, where $\boldsymbol{\phi} = \text{bdiag}\{\mathbf{M}^{\mathbf{U}}, \mathbf{M}_0^{\mathbf{U}}\}$, $\mathbf{M}^{\mathbf{U}} = \mathbf{K}_1$ and $\mathbf{M}_0^{\mathbf{U}} = (\mathbf{I}_{q_2}, \mathbf{0}_{q_2\times(r-u-q_2)})^T$, given $\mathbf{U} \in \mathbb{R}^{r\times q}$. Set $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$; notice that it also satisfies $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = (\mathbf{K}_2^T, \mathbf{0}_{p\times q_2})^T$.

Step 3. For each individual i, generated $\mathbf{Y}_i$ identically and independently from a normal distribution $N(\boldsymbol{\beta}\mathbf{X}_i, \boldsymbol{\Sigma})$.

Step 4. Calculate $\widehat{\boldsymbol{\beta}}_{\text{um}}$, $\widehat{\boldsymbol{\beta}}_{\text{em}}$, $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}}$, where $\mathbf{U}$ is correctly specified when calculating $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}}$.

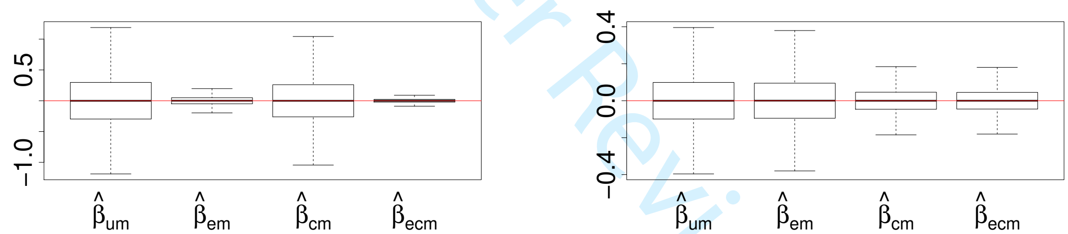
Step 5. Repeat Steps 3–4 1000 times.

We also carried out simulations with a smaller sample size of $n = 80$, with results similar to those presented below. They are presented in Supplement Section 4.1

From the choice of $\boldsymbol{\eta}$ in Step 2 we have colrank$(\boldsymbol{\eta}) = q_1$, and span$(\boldsymbol{\beta})$ is strictly contained in both span$(\boldsymbol{\Gamma})$ and span$(\mathbf{U})$ since the dimension of span$(\boldsymbol{\beta})$ is $q_1 = 4$ which is smaller than $\min(u, q) = 6$. Specifically, we have span$(\boldsymbol{\beta})$ = span$(\boldsymbol{\Gamma}^{\mathbf{U}})$ = span$(\boldsymbol{\Gamma}) \cap$ span$(\mathbf{U})$. As mentioned before, $\boldsymbol{\alpha} = (\mathbf{K}_2^T, 0_{p\times q_2})^T$ in this example. Therefore, we also have a non-trivial constrained envelope of dimension 4: Under the conditions of Scenario 1, we have that $\boldsymbol{\Sigma}_{\mathbf{D}|\mathbf{S}}^{-1} = \boldsymbol{\phi}^T \text{bdiag}(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Omega}_0^{-1})\boldsymbol{\phi} = \text{bdiag}(\mathbf{K}_1^T \boldsymbol{\Omega}^{-1}\mathbf{K}_1, 50^{-1}\mathbf{I}_{q_2})$. As a consequence, the envelope of $\boldsymbol{\alpha}$ with respect to $\boldsymbol{\Sigma}_{\mathbf{D}|\mathbf{S}}$ is $(\tilde{\mathbf{K}}_2^T, \mathbf{0}_{q_2\times(r-u-q_2)})^T$ with $\tilde{\mathbf{K}}_2 \in \mathbb{R}^{q_1\times q_1}$ such that $\mathbf{K}_2 = \tilde{\mathbf{K}}_2\mathbf{J}$ with $\mathbf{J} \in \mathbb{R}^{q_1\times p}$ and $\tilde{K}_2$ orthogonal. Such a decomposition of $\mathbf{K}_2$ is possible since $\mathbf{K}_2$ as rank q_1. In the 1000 simulations repetitions, the envelope dimension was always correctly estimated as 6 using BIC. The dimension of the constrained envelope estimator was correctly estimated as 4 for 989 times and 5 for 11 times. The empirical results for $\widehat{\boldsymbol{\beta}}_{\text{um}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\text{em}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\text{cm}} - \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}} - \boldsymbol{\beta}$ are shown in Figure 2a, where all the elements of $\boldsymbol{\beta}$ are plotted in the same boxplot as if they are from the same population and the outliers are suppressed for a cleaner representation. Since $\mathbf{U}$ is correctly specified, $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}}$ are asymptotically unbiased estimators, as are $\widehat{\boldsymbol{\beta}}_{\text{um}}$ and $\widehat{\boldsymbol{\beta}}_{\text{em}}$. Hence, the boxplots of the four estimators are all centered at 0. In Step 2, the larger eigenvalues of $\boldsymbol{\Sigma}$ are contained in $\boldsymbol{\Omega}_0$ rather than $\boldsymbol{\Omega}$. That is, the variability of the immaterial part is bigger than that of the material part. Additionally, the column space of $\mathbf{U}$ is very conservatively specified as $q = 15$, which is much bigger than the dimension of q_1 = colrank$(\boldsymbol{\beta}) = 4$, and the span$(\mathbf{U})$ contains 11 eigenvectors corresponds to large eigenvalues (i.e., 50 in this simulation). Hence, this scenario favors of the envelope estimator in terms of the efficiency: the envelope estimator is the most efficient estimator among the three estimators, while $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ is also more efficient than the saturated estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$.

The average estimated asymptotic variances were close to the theoretical asymptotic variances calculated using the true parameter values for all three estimators. The mean of the empirical asymptotic variances across all the elements in four estimators are 51.66 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ and 39.48 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ but is only 1.12 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and 0.37 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. That is, in this setting, the envelope estimator is about 40 times more efficient that the saturated estimator and the unconstrained estimator and the constrained envelope estimator is about 100 times more efficient than those two. A comparison between the envelope and constrained envelope estimator demonstrates the advantage of leveraging prior information in terms of achieving better efficiency.

Figure 2: Box plot of $\widehat{\boldsymbol{\beta}}_{\mathrm{um}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{em}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}} - \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}} - \boldsymbol{\beta}$ in two scenarios in 1000 simulations.



(a) Scenario 1 (b) Scenario 2

4.1.2 Scenario 2

To carry out simulations in Scenario 2, we modify some of the parameters in Step 2. In the new Scenario 2, $q = 6$ and we redefine the eigenvalues of $\boldsymbol{\Sigma}$ by setting $\boldsymbol{\Omega} = \mathrm{bdiag}(50\mathbf{I}_{u-q_1}, 0.5\mathbf{I}_{q_1})$ and $\boldsymbol{\Omega}_0 = 0.5\mathbf{I}_{r-u}$. In this scenario, the larger eigenvalues of $\boldsymbol{\Sigma}$ are associated with $\boldsymbol{\Omega}$. Now the dimension of $\mathbf{U}$ is 6 and therefore is only 2 dimension larger than the dimension of $\boldsymbol{\beta}$.

Since the envelope is also of dimension 6 and needs to be estimated, the envelope method is at a disadvantage in terms of the efficiency as compared with $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$. We have $\boldsymbol{\alpha} = (\mathbf{K}_2^T, \mathbf{0}_{p\times q_2})^T$ and

$\boldsymbol{\Sigma}_{\mathbf{D}|\mathbf{S}} = \mathrm{bdiag}(\mathbf{K}_1^T\boldsymbol{\Omega}^{-1}\mathbf{K}_1, (0.5)^{-1}\mathbf{I}_{q_2})$ and the dimension of the constraint envelope is still 4. In the 1000 simulations repetitions, the envelope dimension and the constrained envelope dimension are always correctly estimated as 6 and 4, respectively. The empirical biases of the estimators are shown in Figure 2b. Again, all four estimators are centered around 0, indicating the asymptotic unbiasedness. As expected, the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ are the most efficient among the four estimators, while the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ is still more efficient than $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$.

The average estimated asymptotic variance of the three estimators were all close to their theoretical values. The average empirical variances of all the elements in three estimators are 8.26 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, 7.87 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, 1.50 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and 1.55 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. That is, in this setting, the estimators using a correctly specified $\mathbf{U}$ are on average about 4 times of more efficient that the saturated estimator and the envelope estimator, but the envelope constraint estimator does not provide additional advantages over the constrained estimator. We also carried out simulations with a smaller sample size of $n = 80$, with results similar to those presented here. The details are presented in Supplement Section 4.2

4.2 Potential Bias of the constrained estimator

We conducted a small simulation generating data from envelope model (4), to further illustrate potential bias effects. The sample size is again $n = 200$, and the parameters to generate the data are chosen as in Scenario 1, only changing the definition of $\mathbf{U}$, which now is $\mathbf{U} = (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\mathbf{A}_k$ with $\mathbf{A}_k = (\mathbf{I}_k, 0)^T$, $k = 1, \dots, r$. For $k < u$, $\mathcal{B} \not\subseteq \mathcal{U}$ and so both $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ are biased. But for $k \geq u$, $\mathcal{B} \subseteq \mathcal{U}$ and there is no bias in $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. Again, the dimension of the constrained envelope remains at 4 when $k \geq u$.

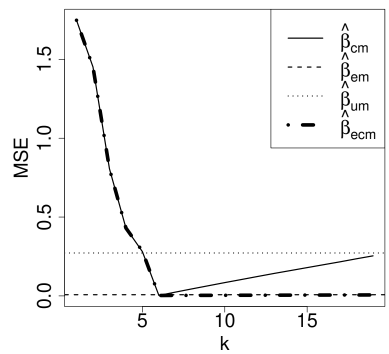


Figure 3: Illustration of potential bias in the constrained estimator (3) under Scenario 1, where $k = \dim(\mathcal{U})$, $\mathcal{U} = \mathrm{span}\{(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)(\mathbf{I}_k, 0)^T\}$ and MSE denotes the average element-wise squared error for the indicated estimators.

We generated response vectors according to model (4) using normal errors, and fitted the resulting data to obtain the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$. We used the same data to construct the unconstrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and the constrained envelope $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ with different selections for $\mathbf{U} = (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\mathbf{A}_k$ where $\mathbf{A}_k = (\mathbf{I}_k, 0)^T$, $k = 1, \ldots, r$. For $k < u$, $\mathcal{B} \not\subseteq \mathcal{U}$ and so both $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ are biased, but for $k \geq u$, $\mathcal{B} \subseteq \mathcal{U}$ and there is no bias in $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. Actually, the dimension of the constrained envelope remains at 4 when $k \geq u$. We summarized the bias by computing the mean squared error over all elements β_{ij} of $\boldsymbol{\beta}$: $\mathrm{MSE} = (rp)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{p}(\widehat{\boldsymbol{\beta}}_{(\cdot),ij} - \beta_{ij})^2$ for the three four estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. Shown in Figure 3 are plots of the MSE averaged over 1000 replications of this scheme for Scenario 1, each replication starting with the generation of the response vectors. The constant MSE for $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ was 7.4×10^{-2} and that for unconstrained model was about 36 times greater at 0.27. The MSE for both the constrained estimator and the constrained envelope estimator decreased monotonically from its maximum value of about 1.75 at $k = 1$ to around 2.5×10^{-3}, at $k = u = 6$, the constrained estimator increased monotonically to 0.25 at $k = 20$, while the constrained envelope estimator remains about the same. This is because the constraint envelope is adapted to the data and does not loose much efficiency even if

$\mathbf{U}$ is large as it was shown in Section 4. This suggests that it may be a good practice to specify a conservative $\mathbf{U}$ and apply the constrained envelope to gain more efficiency so that we can enjoy the benefit of prior information but do not suffer from large bias. The corresponding plot for Scenario 2 is similar and therefore is not presented here. It seems clear that the bias in the constrained and constrained envelope estimators can be substantial until we achieve $\mathcal{B} \subseteq \mathcal{U}$.

4.3 A more general case

The previous simulations were conducted so that both the constrained model and the envelope model can hold under the data generating mechanism. Here, we consider a general case where $\mathbf{U}$ is arbitrarily generated but correctly specified. Because $\mathbf{U}$ is correctly specified but the envelope model no longer holds, we only compare the constrained model and the constrained envelope model. We carried out the simulations similar to those in Section 4.1, replacing Steps 2–4 with:

Step 2*. Set $r = 20$, $u^* = 3$, $q = 15$. Set $\mathbf{\Omega}^* = 0.5\mathbf{I}_{u^*}$ and $\mathbf{\Omega}_0 = 50\mathbf{I}_{q-u^*}$. Set $(\mathbf{\Gamma}^*, \mathbf{\Gamma}_0^*) = \mathbf{I}$ and let $\mathrm{var}(\boldsymbol{\varepsilon}_{\mathbf{D}|\mathbf{S}}) = \mathbf{\Sigma}_{\mathbf{D}|\mathbf{S}} = \mathbf{\Gamma}^*\mathbf{\Omega}^*\mathbf{\Gamma}^{*T} + \mathbf{\Gamma}_0^*\mathbf{\Omega}_0^*\mathbf{\Gamma}_0^{*T}$. Generate $\boldsymbol{\eta}^* \in \mathbb{R}^{u^* \times p}$ and $\mathbf{U}$, where each element in $\boldsymbol{\eta}^*$ and $\mathbf{U}$ is identically and independently generated from $N(0, 1)$. Set $\boldsymbol{\alpha}^* = \mathbf{\Gamma}^*\boldsymbol{\eta}^*$ and $\boldsymbol{\beta}^* = \mathbf{U}\boldsymbol{\alpha}^*$.

Step 3*. For each i, generate $\mathbf{Y}_{\mathbf{S}i}$ identically and independently from normal distribution $N(0, \mathbf{I}_{r-q})$. Generate $\boldsymbol{\phi} \in \mathbb{R}^{q \times (r-q)}$, where each element is generated identically and independently from standard normal. Generate $\mathbf{Y}_{\mathbf{D}i}$ from the distribution $N(\boldsymbol{\alpha}^*\mathbf{Z}_i + \boldsymbol{\phi}\mathbf{Y}_{\mathbf{S}i}, \mathbf{\Sigma}_{\mathbf{D}|\mathbf{S}})$

Step 4*. Calculate $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$., where $\mathbf{U}$ is correctly specified for both estimators.

The average MSE of $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ was 0.12 and 0.03. The Monte Carlo mean variances over all the elements were 24.86 and 6.81 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{cm}}$ and $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{ecm}}$, demonstrating the efficiency of the additional envelope structure over the $\widehat{\boldsymbol{\beta}}_{\text{cm}}$ estimator.

5 Application: Postbiotics study

The aim of the posbiotics study (Dunand et al., 2019) was to determine the protective capacity against Salmonella infection in mice of the cell-free fraction (postbiotic) of fermented milk, produced at laboratory and industrial level. The capacity of the postbiotics produced by pH-controlled fermentation to stimulate the production of secretory IgA in feces and to protect mice against Salmonella infection was evaluated. There were 3 study groups with seven mice per group: (i) a control group (C) where mice received the unfermented milk supernatant, (ii) a F36 group (F36) where mice received the cell-free supernatant obtained by DSM-100H fermentation in 10% (w/v) skim milk produced in the laboratory, and (iii) a F36D group (F36D) where mice received the product F36 diluted 1/10 in tap water. Feces samples (approximately 50 mg per mouse) were collected once a week for 6 weeks and the concentration of secretory IgA (S-IgA) was determinate by ELISA. The response was the IgA measures over the 6 weeks period and the predictors the group indicators. The research question was whether there were differences in the IgA measures among the treatment groups. We present the average response by group over the weeks in Figure 4. We set the control group as the baseline and therefore $\boldsymbol{\beta} \in \mathbb{R}^{6\times 2}$. We calculate all estimators based on various envelopes on model (3) because we were interested in profile contrasts rather than modeling profiles. We use $\mathbf{U}^T(t) = (1, t/6, (t/6)^2, \cos(2\pi t/6), \sin(2\pi t/6))$, where $t = 1, \ldots, 6$ are the weeks where the measures were taking. The unconstrained estimator $\widehat{\boldsymbol{\beta}}_{\text{um}}$ was considered in Dunand et al. (2019) and it did not show a difference between treatment groups even when

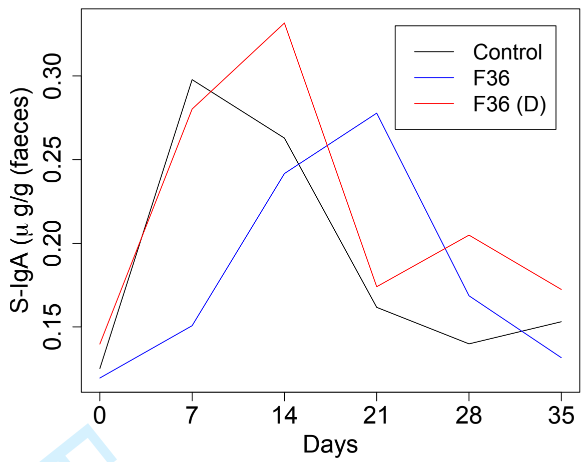


Figure 4: Average of IgA by group over time in the Posbiotics Study data

exploratory differences can be seen (Figure 4).

Table 1 shows the BIC, envelope dimension and MSE of the estimators. We listed the maximum envelope dimension for the two non-envelope methods as their estimated envelope dimensions. The unconstrained estimator performs the worst and the scaled constrained envelope estimator performs the best in terms of both the BIC and the efficiency.

Table 1: Envelope dimension, BIC, BIC order, and MSE for the Postbiotics Study

Estimator	Dimension	BIC	BIC order	MSE
$\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$	6	-133.90	6	0.15
$\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$	1	-163.52	2	0.13
$\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$	2	-144.37	5	0.15
$\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$	1	-160.76	3	0.14
$\widehat{\boldsymbol{\beta}}_{\mathrm{secm}}$	1	-251.48	1	0.13

To answer the researcher question, we look the p-values of the $\widehat{\boldsymbol{\beta}}$ components. From Table 2 we can see that the unconstrained estimator does not reveal any difference, which aligns with the findings in Dunand et al. (2019). None of the estimators demonstrate any evidence of difference between F36D group and the control group at any time. On the other hand, $\widehat{\boldsymbol{\beta}}_{\mathrm{secm}}$ reveals a significance difference between the control and F36 groups in all followup weeks. The p-values for such a comparison of $\widehat{\boldsymbol{\beta}}_{\rm em}$ are only significant in week 3. Other estimators also fail to find all followup weeks significant between F36 and control groups, e.g., the scaled envelope is not significant in week 5 and 6, and constrained envelope is significant only in week 2. The variance gains for the

Table 2: The p-values for coefficients for $\widehat{\boldsymbol{\beta}}_{\rm um}$, $\widehat{\boldsymbol{\beta}}_{\rm em}$ and $\widehat{\boldsymbol{\beta}}_{\rm secm}$

week	F36 vs control			F36 D vs control		
	$\widehat{\boldsymbol{\beta}}_{\rm um}$	$\widehat{\boldsymbol{\beta}}_{\rm em}$	$\widehat{\boldsymbol{\beta}}_{\rm secm}$	$\widehat{\boldsymbol{\beta}}_{\rm um}$	$\widehat{\boldsymbol{\beta}}_{\rm em}$	$\widehat{\boldsymbol{\beta}}_{\rm secm}$
1	0.91	0.07	0.13	0.77	0.27	0.30
2	0.09	0.10	0.01	0.83	0.28	0.21
3	0.83	0.01	0.01	0.48	0.20	0.22
4	0.26	0.06	0.02	0.90	0.23	0.22
5	0.55	0.05	0.00	0.16	0.20	0.20
6	0.57	0.63	0.01	0.59	0.64	0.21

scale version of the constrained envelope model over the unconstrained model (and therefore the p-values) are reflected by the eigenvalue 1×10^{-4} of $\widehat{\boldsymbol{\Omega}}$ and the four eigenvalues of $\widehat{\boldsymbol{\Omega}}_0$ which are 23.06, 13.67, 0.41 and 0.22. The reason for the envelope estimator to be not as significant when comparing F36 and control groups is that there is not as big a discrepancy between the eigenvalues of $\widehat{\boldsymbol{\Omega}}$ (2×10^{-3}) and those of $\widehat{\boldsymbol{\Omega}}_0$ $(0.02, 0.04, 0.03, 0.01, 4 \times 10^{-3})$.

6 Discussion

In this paper, we first compared the envelope model with the commonly used linear constraint model in terms of both the potential bias and efficiency. We then proposed a constrained envelope model for studying growth curve and longitudinal data when a well-grounded linear constraint is available. We recommend using the constrained envelope model with a relatively conservative $\mathbf{U}$ so that it is likely to contain the space of interest and to achieve efficiency gain. Extensions to unbalanced data and random effects models are designated for future research.

The primary computational step for all of the envelope methods described herein involves finding $\widehat{\mathbf{G}} = \arg\min_{\mathbf{G}\in\mathcal{G}} \log|\mathbf{G}^T\mathbf{M}_1\mathbf{G}| + \log|\mathbf{G}^T\mathbf{M}_2\mathbf{G}|$ over a class $\mathcal{G}$ of semi-orthogonal matrices, where the inner product matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ depend on the application. The R package Renvlp by M. Lee and Z. Su contains a routine for minimizing objective functions of this form. Computations are straightforward once $\widehat{\mathbf{G}}$ has been found. Renvlp also implements specialized methodology for data analysis under envelope model (4) and the partial envelope model. The associated routines can be modified for the models described herein. The codes to reproduce the examples and simulations from this paper can be found at `https://github.com/lanliu1815/constrained_env`.

7 Supplementary Material

Discussion of certain well-established aspects of envelope methodology is available in Supplement Section 1–3. We include an additional simulation with smaller sample size in Section 4. We revisited the Dental data using the methodology presented in this paper in Section 5, and we studied the China Health and Nutrition Survey data set in Section 6. Enveloping for $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ jointly is discuss in Section 7. Non-normality and the bootstrap are discussed in Section 8 and methods for selecting the envelope dimension are reviewed in Section 9. Finally a brief discussions of envelopes and Rao's simple structure is in Section 10.

References

Cook, R. D. (2018). An Introduction to Envelopes. Wiley, Hoboken, NJ.

Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. Biometrika, 94(3):569–584.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica, 20(3):927–960.

Cook, R. D. and Su, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. Biometrika, 100(4):939–954.

Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods. Journal of the American Statistical Association, 110(510):599–611.

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. Technometrics, 57(1):11–25.

Cooper, D. M. and Evans, C. D. (2002). Constrained multivariate trend analysis applied to water quality variables. Environmentrics, 13:42–53.

Dunand, E., Burns, P., Binetti, A., Bergamini, C., Peralta, G., Forzani, L., Reinheimer, J., and Vinderola, G. (2019). Postbiotics produced at laboratory and industrial level as potential functional food ingredients with the capacity to protect mice against salmonella infection. J Appl Microbiol., to appear, 127(1):219–229.

Forzani, L. and Su, Z. (2021). Envelopes for elliptical multivariate linear regression. Statistica Sinica, 31:301–332.

Geisser, S. (1970). Bayesian analysis of growth curves. Sankhya, Ser. A, 32(1):53–64.

Geisser, S. (1981). Sample reuse procedures for prediction of the unobserved portion of a partially observed vector. Biometrika, 68(1):243–250.

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves. Biometrics, 25(2):357–381.

Izenman, A. J. and Williams, J. S. (1989). A class of linear spectral models and analysis for the study of longitudinal data. Biometrics, 45(3):831–849.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. ournal of the Royal Statistical Society C, 36(3):296–308.

Lee, J. C. (1988). Prediction and estmation of growth curves with special covariance structures. Journal of the American Statistical Association, 83(402):432–440.

Lee, J. C. and Geisser, S. (1975). Applications of growth curve prediction. Sankhyā: The Indian Journal of Statistics, Series A, 37(2):239–256.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. Journal of the American Statistical Association, 112(519):1131–1146.

Nummi, T. and Koskela, L. (2008). Analysis of growth curve data by using cubic smoothing splines. Journal of Applied Statistics, 35(6):681–691.

Pan, J.-X. and Fang, K.-T. (2002). Growth Curve Models and Statistical Diagnostics. Springer, New York.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51(3):313–326.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 52(3):447–458.

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In LeCam, L. M. and Neyman, J., editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 355–372. Berkeley: University of California Press.

Rao, C. R. (1987). Prediction of future observations in growth curve models. Statistical Science, 2(4):434–441.

Rekabdarkolaee, H. M., Wang, Q., Naji, Z., and Fluentes, M. (2020). New parsimonious multivariate spatial model: Spatial envelope. Statistica Sinica, 30:1583–1604. http://www3.stat.sinica.edu.tw/statistica/j30n3/j30n320/j30n320.html.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. Biometrika, 98(1):133–146.

Su, Z., Zhu, G., Chen, X., and Yang, Y. (2016). Sparse envelope model: estimation and response variable selection in multivariate linear regression. Biometrika, 103(3):579–593.

Supplement to "Envelopes for multivariate linear regression with linearly constrained coefficients"

R. Dennis Cook,* Liliana Forzani† and Lan Liu‡

November 14, 2022

1 Supplementary material for Section 2.1

We will derive here the formula for $\mathbf{\Sigma}_{D|S}$, the maximum likelihood estimator for $\boldsymbol{\alpha}_{\rm cm}$ and $\mathbf{\Sigma}_{D|S}$, the asymptotic variance for $\widehat{\boldsymbol{\alpha}}_{\rm cm}$ and the estimator of $\mathbf{\Sigma}$ from the constraint model.

Derivation of $\mathbf{\Sigma}_{D|S}$. Direct calculation gives

$$\mathbf{\Sigma}_{D|S} = (\mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{\Sigma}\mathbf{U} - \mathbf{U}^T\mathbf{\Sigma}\mathbf{U}_0(\mathbf{U}_0^T\mathbf{\Sigma}\mathbf{U}_0)^{-1}\mathbf{U}_0^T\mathbf{\Sigma}\mathbf{U})(\mathbf{U}^T\mathbf{U})^{-1}.$$

*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: dennis@stat.umn.edu).

†Liliana Forzani is Professor, Facultad de Ingeniería Química, UNL. Researcher of CONICET, Santa Fe, Argentina (Email: liliana.forzani@gmail.com).

‡Lan Liu is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: liu1815@gmail.com).

The result follows by multiplying the identity

$$\mathbf{P}_{\boldsymbol{\Sigma}^{-1/2}\mathbf{W}_1} + \mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{W}_2} = \mathbf{I}_r$$

on left and right by $\mathbf{W}_1^T\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{\Sigma}^{1/2}\mathbf{W}_1$, and then rearranging terms. (See also Cook and Forzani, 2008, eq. (A1)).

Derivation of $\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$. First construct a version of (7) is construct so that its predictors are orthogonal: Let $\boldsymbol{\alpha}_0^* = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}\bar{\mathbf{X}} + \boldsymbol{\phi}_{D|S}\bar{\mathbf{Y}}_S$ and $\boldsymbol{\phi}_{D|S}^* = \boldsymbol{\phi}_{D|S} + \boldsymbol{\alpha}\mathbf{S}_{\mathbf{X},S}\mathbf{S}_S^{-1}$. Then

$$\mathbf{Y}_{Di} \mid (\mathbf{X}_i, \mathbf{Y}_{Si}) \quad = \quad \boldsymbol{\alpha}_0^* + \boldsymbol{\alpha}\mathbf{R}_{\mathbf{X}|(1,S)i} + \boldsymbol{\phi}_{D|S}^*(\mathbf{Y}_{Si} - \bar{\mathbf{Y}}_S) + \mathbf{e}_{D|Si}, \tag{1}$$

where for clarity $\mathbf{R}_{\mathbf{X}|(1,S)i} = \mathbf{X}_i - \bar{\mathbf{X}} - \mathbf{S}_{\mathbf{X},S}\mathbf{S}_S^{-1}(\mathbf{Y}_{Si} - \bar{\mathbf{Y}}_S)$. The three addends on the right side are orthogonal and so the three terms can be fitted separately. The starred parameters are not of direct interest, but $\boldsymbol{\alpha}$ is the same as in the original model.

Derivation of $\widehat{\boldsymbol{\Sigma}}_{D|S}$. It can be express as

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{D|S} &= \mathbf{S}_{D|(\mathbf{X},S)} = \mathbf{S}_D - \mathbf{S}_{D,\mathbf{R}_{\mathbf{X}|(1,S)}}\mathbf{S}_{\mathbf{X}|S}^{-1}\mathbf{S}_{D,\mathbf{R}_{\mathbf{X}|(1,S)}}^T - \mathbf{S}_{D,S}\mathbf{S}_S^{-1}\mathbf{S}_{D,S}^T \\ &= \mathbf{S}_{D|S} - \widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}\mathbf{S}_{\mathbf{X}|S}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}^T.\end{aligned}$$

Derivation of $\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}))$. First, suppressing notation for the conditioning on $\mathbf{X}$,

$$\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}})) = E\{\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}) \mid \mathbf{Y}_{S1}, \ldots, \mathbf{Y}_{Sn})\} + \mathrm{var}\{E(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}) \mid \mathbf{Y}_{S1}, \ldots, \mathbf{Y}_{Sn})\}$$

The second addend on the right side is 0 since $E(\boldsymbol{\alpha}_{\text{cm}}) = \boldsymbol{\alpha}$. For notational convenience, let $\mathbf{R}_i = \mathbf{R}_{\mathbf{X}|(1,S)i}$ and recall that $\sum_{i=1}^n \mathbf{R}_i = 0$ and that the $\mathbf{X}$'s are non-stochastic. To evaluate the first term we have from (9),

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{\text{cm}} &= n^{-1}\sum_{i=1}^n (\mathbf{Y}_{Di} - \bar{\mathbf{Y}}_D)\mathbf{R}_i^T \mathbf{S}_{\mathbf{X}|S}^{-1} = n^{-1}\sum_{i=1}^n \mathbf{Y}_{Di}\mathbf{R}_i^T \mathbf{S}_{\mathbf{X}|S}^{-1} \\
\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{cm}}) &= n^{-1}\sum_{i=1}^n (\mathbf{S}_{\mathbf{X}|S}^{-1}\mathbf{R}_i \otimes \mathbf{I}_k)\mathbf{Y}_{D_i} \\
\text{var}(\text{vec}(\widehat{\boldsymbol{\alpha}}_{\text{cm}}) \mid \mathbf{Y}_{S1}, \ldots, \mathbf{Y}_{Sn}) &= n^{-2}\sum_{i=1}^n (\mathbf{S}_{\mathbf{X}|S}^{-1}\mathbf{R}_i \otimes \mathbf{I}_k)\boldsymbol{\Sigma}_{D|S}(\mathbf{R}_i^T \mathbf{S}_{\mathbf{X}|S}^{-1} \otimes \mathbf{I}_k) \\
&= n^{-2}\sum_{i=1}^n (\mathbf{S}_{\mathbf{X}|S}^{-1}\mathbf{R}_i\mathbf{R}_i^T \mathbf{S}_{\mathbf{X}|S}^{-1}) \otimes \boldsymbol{\Sigma}_{D|S} \\
&= n^{-1}\mathbf{S}_{\mathbf{X}|S}^{-1} \otimes \boldsymbol{\Sigma}_{D|S}
\end{aligned}$$

Taking the expectation with respect to $\mathbf{Y}_S \mid \mathbf{X}$ gives the finite sample result. The asymptotic variance follows by direct calculation from the Fisher information, recalling that the distribution of $\mathbf{Y}_S$ does not depend on $\mathbf{X}$. It can be similarly computed from the finite sample variance.

Estimator of Σ from the constrained model Starting from the main paper,we have

$$\widehat{\boldsymbol{\Sigma}}_{\text{cm}} = \mathbf{W}^{-T}\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}}\mathbf{W}^{-1} = \mathbf{W}^{-T}\begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_D & \widehat{\boldsymbol{\Sigma}}_{D,S} \\ \widehat{\boldsymbol{\Sigma}}_{S,D} & \widehat{\boldsymbol{\Sigma}}_S \end{pmatrix}\mathbf{W}^{-1}, \tag{2}$$

2 Proofs from Section 2.4

Proof of Proposition 2.1 Comparing (12) and (11), it is sufficient to show that

$$\mathbf{U}(\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U})^{-1}\mathbf{U}^T \leq \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T,$$

where $\mathbf{\Gamma}$ is a semi-orthogonal basis matrices for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and without loss of generality we take $\mathbf{U}$ to be a semi-orthogonal basis matrix for $\mathcal{U}$. Since by hypothesis $\mathcal{U} \subseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ there is a semi-orthogonal matrix $\mathbf{H} \in \mathbb{R}^{p \times k}$ so that $\mathbf{U} = \mathbf{\Gamma}\mathbf{H}$. Let $(\mathbf{H}, \mathbf{H}_0)$ be an orthognal matrix and recall that $\mathbf{\Gamma}^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma} = \mathbf{\Omega}^{-1}$. Then

$$\mathbf{U}(\mathbf{U}^T\mathbf{\Sigma}^{-1}\mathbf{U})^{-1}\mathbf{U}^T = \mathbf{\Gamma}\mathbf{H}(\mathbf{H}^T\mathbf{\Omega}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{\Gamma}^T \leq \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T,$$

where the inequality follows from the identity $\mathbf{P}_{\mathbf{\Omega}^{-1/2}\mathbf{H}} + \mathbf{P}_{\mathbf{\Omega}^{1/2}\mathbf{H}_0} = \mathbf{I}_u$. □

Proof of Proposition 2.2 Recall that $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$. (1) follows immediately since $\mathrm{span}(\mathbf{U}\boldsymbol{\alpha}) \subseteq \mathrm{span}(\mathbf{U})$.

For (2) assume that $\mathrm{span}(\mathbf{U}) \subseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathrm{span}(\mathbf{U}\boldsymbol{\alpha}))$. Then $\mathcal{E}_{\mathbf{\Sigma}}(\mathrm{span}(\mathbf{U}\boldsymbol{\alpha}))$ is a reducing subspace of $\mathbf{\Sigma}$ that contains $\mathrm{span}(\mathbf{U})$. Using (1) and the fact that $\mathcal{E}_{\mathbf{\Sigma}}(\mathrm{span}(\mathbf{U}))$ is the intersection of all reducing subspaces of $\mathbf{\Sigma}$ that contains $\mathrm{span}(\mathbf{U})$, it follows that $\mathcal{E}_{\mathbf{\Sigma}}(\mathrm{span}(\mathbf{U}\boldsymbol{\alpha})) = \mathcal{E}_{\mathbf{\Sigma}}(\mathrm{span}(\mathbf{U}))$. The reverse implication is immediate. Part (3) is also immediate. □

Preparatory lemma The following lemma, which follows immediately from Milliken and Akdeniz (1977), will be used in the justification of Proposition 2.3.

Lemma 2.1 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be two real positive semi-definite* $r \times r$ *matrices with* $\mathbf{U} \geq \mathbf{V}$ *and* $\mathrm{rank}(\mathbf{U}) = \mathrm{rank}(\mathbf{V})$. *Then (a)* $\{\mathbf{a} \in \mathbb{R}^r \mid \mathbf{U}\mathbf{a} = 0\} = \{\mathbf{a} \in \mathbb{R}^r \mid \mathbf{V}\mathbf{a} = 0\}$, *(b)* $\mathrm{span}(\mathbf{U}) = \mathrm{span}(\mathbf{V})$ *and (c)* $\mathbf{U}^\dagger \leq \mathbf{V}^\dagger$, *where* $\dagger$ *denotes the Moore-Penrose inverse.*

Proof of Proposition 2.3 Let

$$\begin{aligned}
\mathbf{D}(\mathbf{c}) &= \operatorname{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\text{cm}}) - \operatorname{avar}(\sqrt{n}\mathbf{c}^T\widehat{\boldsymbol{\beta}}_{\text{em}}) \\
&= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{c}^T(\mathbf{U}(\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U})^{-1}\mathbf{U}^T - \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T)\mathbf{c} - (\boldsymbol{\eta}^T \otimes \mathbf{c}^T\boldsymbol{\Gamma}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T\mathbf{c})
\end{aligned}$$

We first simplify terms in $\mathbf{D}(\mathbf{c})$. Let $\boldsymbol{\Omega}_{0,1|2} = \boldsymbol{\Omega}_{0,11} - \boldsymbol{\Omega}_{0,12}\boldsymbol{\Omega}_{0,22}^{-1}\boldsymbol{\Omega}_{0,21}$, and recall that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. Then

$$\begin{aligned}
\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U} &= \operatorname{bdiag}(\boldsymbol{\Omega}^{-1}, (\boldsymbol{\Omega}_0^{-1})_{11}) \\
\mathbf{U}(\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U})^{-1}\mathbf{U}^T &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{0,1|2}\boldsymbol{\Gamma}_{01}^T.
\end{aligned}$$

Consequently,

$$\mathbf{D}(\mathbf{c}) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{c}^T\boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{0,1|2}\boldsymbol{\Gamma}_{01}^T\mathbf{c} - (\boldsymbol{\eta}^T \otimes \mathbf{c}^T\boldsymbol{\Gamma}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T\mathbf{c}).$$

It follows that $\mathbf{D}(\mathbf{c}) = 0$ for all $\mathbf{c} \in \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and $\mathbf{D}(\mathbf{c}) \leq 0$ for all $\mathbf{c} \in \operatorname{span}(\boldsymbol{\Gamma}_{02})$. This established parts 1 and 2 of the lemma.

Since $\boldsymbol{\Omega}_{0,12} = 0$ by hypothesis, we justify part 3 by first replacing $\boldsymbol{\Omega}_{0,1|2}$ with $\boldsymbol{\Omega}_{0,11}$ in $\mathbf{D}(\mathbf{c})$ to get

$$\mathbf{D}(\mathbf{c}) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{c}^T\boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{0,11}\boldsymbol{\Gamma}_{01}^T\mathbf{c} - (\boldsymbol{\eta}^T \otimes \mathbf{c}^T\boldsymbol{\Gamma}_0)\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T\mathbf{c}).$$

Next, since $\mathbf{M}(\boldsymbol{\Sigma}_{\mathbf{X}}) \geq \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1}$ and by hypothesis $\operatorname{rank}(\mathbf{M}(\boldsymbol{\Sigma}_{\mathbf{X}})) = \operatorname{rank}(\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1})$, 55 we have from Lemma 2.1 that $\mathbf{M}^{\dagger}(\boldsymbol{\Sigma}_{\mathbf{X}}) \leq (\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T)^{\dagger} \otimes \boldsymbol{\Omega}_0$. Substituting into $\mathbf{D}(\mathbf{c})$ we get

$$\begin{aligned} \mathbf{D}(\mathbf{c}) &\geq \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{c}^T\boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{0,11}\boldsymbol{\Gamma}_{01}^T\mathbf{c} - (\boldsymbol{\eta}^T \otimes \mathbf{c}^T\boldsymbol{\Gamma}_0)\{(\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T)^{\dagger} \otimes \boldsymbol{\Omega}_0\}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T\mathbf{c}) \\ &= (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \boldsymbol{\eta}^T(\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T)^{\dagger}\boldsymbol{\eta}) \otimes \mathbf{c}^T\boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{0,11}\boldsymbol{\Gamma}_{01}^T\mathbf{c} \geq 0. \end{aligned}$$

56 3 Proofs from Section 3

57 **Derivation of the MLE for model (14).** In this case estimation of $\boldsymbol{\alpha}$ corresponds to the partial
58 envelope model and so the estimators for most of the parameters can be taken directly from Su
59 and Cook. (2011). To be self-inclusive, we give a sketch of the derivation here. The likelihood
60 function for model (14) is

$$\begin{aligned} L &= -\frac{n}{2}\log|\boldsymbol{\Sigma}_S| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}(\mathbf{Y}_{Si}^T\boldsymbol{\Sigma}_S^{-1}\mathbf{Y}_{Si}) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{D|S}| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}[(\mathbf{Y}_{Di} - \boldsymbol{\alpha}_0 - \boldsymbol{\phi}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si})^T\boldsymbol{\Sigma}_{D|S}^{-1}(\mathbf{Y}_{Di} - \boldsymbol{\alpha}_0 - \boldsymbol{\phi}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si})] \quad (3) \end{aligned}$$

61 The MLE of $\boldsymbol{\Sigma}_S$ is $\widehat{\boldsymbol{\Sigma}}_S = \mathbf{T}_S$. To get the estimators of all other quantities we can always write

$$\begin{aligned} &\mathbf{Y}_{Di} - \boldsymbol{\alpha}_0 - \boldsymbol{\phi}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} \\ =\ &\mathbf{Y}_{Di} - \{\boldsymbol{\alpha}_0 + \boldsymbol{\phi}_{D|S}\bar{\mathbf{Y}}_S + \boldsymbol{\phi}\boldsymbol{\eta}\bar{\mathbf{X}}\} - \boldsymbol{\phi}\boldsymbol{\eta}(\mathbf{X}_i - \bar{\mathbf{X}}) - \boldsymbol{\phi}_{D|S}(\mathbf{Y}_{Si} - \bar{\mathbf{Y}}_S) \\ :=\ &\mathbf{Y}_{Di} - \boldsymbol{\alpha}_{0c} - \boldsymbol{\phi}\boldsymbol{\eta}(\mathbf{X}_i - \bar{\mathbf{X}}) - \boldsymbol{\phi}_{D|S}(\mathbf{Y}_{Si} - \bar{\mathbf{Y}}_S), \end{aligned}$$

62 where $\boldsymbol{\alpha}_{0c} = \boldsymbol{\alpha}_0 + \boldsymbol{\phi}_{D|S}\bar{\mathbf{Y}}_S + \boldsymbol{\phi}\boldsymbol{\eta}\bar{\mathbf{X}}$ denotes the intercept in the model with centered predictors.
63 The predictors $\mathbf{X}_i - \bar{\mathbf{X}}$ and $\mathbf{Y}_{Si} - \bar{\mathbf{Y}}_S$ in this model are now centered although their coefficients

are the same as those in the uncentered version. The intercept vector has changed but this is of no consequence since we are not estimating $\boldsymbol{\alpha}_0$. Consequently, we can use the centered version to derive the estimators of interest.

The centered model leading immediately to $\widehat{\boldsymbol{\alpha}}_{0c} = \bar{\mathbf{Y}}_D$. Holding all other parameters fixed, the value of $\boldsymbol{\phi}_{D|S}$ that maximizes L is

$$\boldsymbol{\phi}_{D|S} = \mathbf{S}_{DS}\mathbf{S}_S^{-1} - \boldsymbol{\phi\eta}\mathbf{S}_{\mathbf{X}S}\mathbf{S}_S^{-1}. \quad (4)$$

Let $\mathbf{R}_{D|(1,S)} = \mathbf{Y}_D - \bar{\mathbf{Y}}_D - \mathbf{S}_{DS}\mathbf{S}_S^{-1}(\mathbf{Y}_S - \bar{\mathbf{Y}}_S)$ denote a typical residual vector from the regression of $\mathbf{Y}_D$ on $\mathbf{Y}_S$ including an intercept, and let $\mathbf{R}_{\mathbf{X}|(1,S)} = \mathbf{X} - \bar{\mathbf{X}} - \mathbf{S}_{\mathbf{X}S}\mathbf{S}_S^{-1}(\mathbf{Y}_S - \bar{\mathbf{Y}}_S)$ denote a typical residual from the regression of $\mathbf{X}$ on $\mathbf{Y}_S$ including an intercept.

Then, evaluating at (4) we have

$$\mathbf{Y}_{Di} - \bar{\mathbf{Y}}_D - \boldsymbol{\phi\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} = \mathbf{R}_{D|(1,S)} - \boldsymbol{\phi\eta}\mathbf{R}_{\mathbf{X}|(1,S)}$$

and the first partially maximized log likelihood becomes

$$\begin{aligned} L_1 = &-\frac{n}{2}\log|\mathbf{T}_S| - \frac{np}{2} - \frac{n}{2}\log|\boldsymbol{\Sigma}_{D|S}| \\ &-\frac{1}{2}\sum_{i=1}^{n}[(\mathbf{R}_{D|(1,S)} - \boldsymbol{\phi\eta}\mathbf{R}_{\mathbf{X}|(1,S)})^T\boldsymbol{\Sigma}_{D|S}^{-1}(\mathbf{R}_{D|(1,S)} - \boldsymbol{\phi\eta}\mathbf{R}_{\mathbf{X}|(1,S)})] \\ = &-\frac{n}{2}\log|\widehat{\boldsymbol{\Sigma}}_S| - \frac{np}{2} - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| \\ &-\frac{1}{2}\sum_{i=1}^{n}[\mathbf{R}_{D|(1,S)} - \boldsymbol{\phi\eta}\mathbf{R}_{\mathbf{X}|(1,S)})^T(\boldsymbol{\phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\phi}^T + \boldsymbol{\phi}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\phi}_0^T)(\mathbf{R}_{D|(1,S)} - \boldsymbol{\phi\eta}\mathbf{R}_{\mathbf{X}|(1,S)})], \quad (5) \end{aligned}$$

where the i subscripts on $\mathbf{R}_{D|(1,S)}$ and $\mathbf{R}_{\mathbf{X}|(1,S)}$ have been suppressed. The value of $\boldsymbol{\eta}$ that maximizes L_1 is

$$\boldsymbol{\eta} = \boldsymbol{\phi}^T \mathbf{S}_{\mathbf{R}_{D|(1,S)}, \mathbf{R}_{\mathbf{X}|(1,S)}} \mathbf{S}^{-1}_{\mathbf{R}_{\mathbf{X}|(1,S)}} = \boldsymbol{\phi}^T \mathbf{B}_{\mathbf{R}_{D|(1,S)}|\mathbf{R}_{\mathbf{X}|(1,S)}},$$

where $\mathbf{B}_{\mathbf{R}_{D|(1,S)}|\mathbf{R}_{\mathbf{X}|(1,S)}} \in \mathbb{R}^{p_D}$ is the OLS coefficient vector from the regression of $\mathbf{R}_{D|(1,S)}$ on $\mathbf{R}_{\mathbf{X}|(1,S)}$. These coefficients can also be interpreted as the OLS coefficients of $\mathbf{X}$ from the regression of $\mathbf{Y}_D$ on $(\mathbf{X}, \mathbf{Y}_S)$ including an intercept. That is, $\mathbf{B}_{\mathbf{R}_{D|(1,S)}|\mathbf{R}_{\mathbf{X}|(1,S)}} = \tilde{\boldsymbol{\alpha}}$. Substituting this into L_1 we get the following partially maximized log likelihood

$$\begin{aligned} L_2 &= -\frac{n}{2}\log|\widehat{\boldsymbol{\Sigma}}_S| - \frac{np}{2} - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}[(\mathbf{R}_{D|(1,S)} - \tilde{\boldsymbol{\alpha}}\mathbf{R}_{\mathbf{X}|(1,S)})^T \boldsymbol{\phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\phi}^T(\mathbf{R}_{D|(1,S)} - \tilde{\boldsymbol{\alpha}}\mathbf{R}_{\mathbf{X}|(1,S)})] \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{R}^T_{D|(1,S)}\boldsymbol{\phi}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\phi}_0^T\mathbf{R}_{D|(1,S)}). \end{aligned} \tag{6}$$

But $\mathbf{R}_{D|(1,S)} - \tilde{\boldsymbol{\alpha}}\mathbf{R}_{\mathbf{X}|(1,S)} = \mathbf{R}_{D|(1,\mathbf{X},S)}$, the residual vectors from the regression of $\mathbf{Y}_D$ on $(\mathbf{X}, \mathbf{Y}_S)$ including an intercept. We have

$$\begin{aligned} L_2 &= -\frac{n}{2}\log|\mathbf{T}_S| - \frac{np}{2} - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}[\mathbf{R}^T_{D|(1,\mathbf{X},S)}\boldsymbol{\phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\phi}^T\mathbf{R}_{D|(1,\mathbf{X},S)}] \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{R}^T_{D|(1,S)}\boldsymbol{\phi}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\phi}_0^T\mathbf{R}_{D|(1,S)}). \end{aligned} \tag{7}$$

Taking $\boldsymbol{\phi}$ fixed,

$$\begin{aligned} \boldsymbol{\Omega} &= \boldsymbol{\phi}^T\mathbf{S}_{D|(\mathbf{X},S)}\boldsymbol{\phi} \\ \boldsymbol{\Omega}_0 &= \boldsymbol{\phi}_0^T\mathbf{S}_{D|S}\boldsymbol{\phi}_0 \end{aligned}$$

and, letting u denote the dimension of the envelope, the following partially maximized log likelihood becomes

$$L_3 = -\frac{n}{2}\log|\mathbf{T}_S| - \frac{np}{2} - \frac{n}{2}\log|\boldsymbol{\phi}^T\mathbf{S}_{D|(\mathbf{X},S)}\boldsymbol{\phi}| - \frac{n}{2}\log|\boldsymbol{\phi}_0^T\mathbf{S}_{D|S}\boldsymbol{\phi}_0| \quad (8)$$

From what ot follows that the MLE of ϕ can be found as

$$\widehat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \log|\boldsymbol{\phi}^T\mathbf{S}_{D|(\mathbf{X},S)}\boldsymbol{\phi}| + \log|\boldsymbol{\phi}^T\mathbf{S}_{D|S}^{-1}\boldsymbol{\phi}|, \quad (9)$$

where for clarity $\mathbf{S}_{D|S}$ is the sample residual covariance matrix of the regression of $\mathbf{Y}_D$ on $\mathbf{Y}_S$ with an intercept, and $\mathbf{S}_{D|(\mathbf{X},S)}$ is the sample residual covariance matrix of the regression of $\mathbf{Y}_D$ on $(\mathbf{X}, \mathbf{Y}_S)$ with an intercept.

Once we get $\widehat{\boldsymbol{\phi}}$, we get the estimators for the rest of the parameters.

Proof of the Asymptotic distribution of $\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\mathrm{ecm}})$ **and** $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}})$ Let us call the parameters of model (2) as $h = (\boldsymbol{\alpha}_0, \mathrm{vech}(\boldsymbol{\alpha}_{\mathrm{cm}}), \boldsymbol{\phi}_{D|S}, \mathrm{vech}(\boldsymbol{\Sigma}_{D|S}), \mathrm{vech}(\boldsymbol{\Sigma}_S))$, the asymptitic distribution of the MLE estimator $\hat{h}$ is the inverse of its Fisher information matrix that can be obtained using straightforward computing of the second derivative of the log-likelihood as

$$J = \begin{pmatrix} \mathbf{\Sigma}_{D|S}^{-1} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{\Sigma}_S \otimes \mathbf{\Sigma}_{D|S}^{-1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{\Sigma}_X \otimes \mathbf{\Sigma}_{D|S}^{-1} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}E_k^T(\mathbf{\Sigma}_{D|S}^{-1} \otimes \mathbf{\Sigma}_{D|S}^{-1})E_k & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}E_{r-k}^T(\mathbf{\Sigma}_S^{-1} \otimes \mathbf{\Sigma}_S^{-1})E_{r-k} \end{pmatrix} \quad (10)$$

where E_r is the expansion matrix that satisfy $\text{vec}(\Sigma) = E_r \,\text{vech}(\Sigma)$ for Σ a symmetric matrix of dimension r.

Since the envelope model (14) is over-parameterized, we will apply Proposition 4.1 from Shapiro (1986) to prove the asymptotic distribution (17) as in Cook et al. (2015) and Cook et al. (2010). To apply Proposition 4.1 of Shapiro (1986), we will check the assumptions first. Along the discussion, we will match Shapiro's notations in our context. Let us call $F(\hat{h}) = \hat{L}_u(\hat{h}) - \hat{L}_u(h)$ where $\hat{L}$ is the likelihood function (16). Then F satisfies the four conditions for F in Section 3 in Shapiro (1986). The function g defined by Shapiro in (2.1) is the function

$$h = g(\psi) = \begin{pmatrix} \text{vec}(\boldsymbol{\alpha}_0) \\ \text{vec}(\boldsymbol{\alpha}) \\ \text{vec}(\boldsymbol{\phi}_{D|S}) \\ \text{vech}(\mathbf{\Sigma}_{D|S}) \\ \text{vech}(\mathbf{\Sigma}_S) \end{pmatrix} = \begin{pmatrix} \text{vec}(\boldsymbol{\alpha}_0) \\ \text{vec}(\boldsymbol{\phi}\boldsymbol{\eta}) \\ \text{vec}(\boldsymbol{\phi}_{D|S}) \\ \text{vech}(\boldsymbol{\phi}\mathbf{\Omega}\boldsymbol{\phi}^T + \boldsymbol{\phi}_0\mathbf{\Omega}_0\boldsymbol{\phi}_0^T) \\ \text{vech}(\mathbf{\Sigma}_S) \end{pmatrix}$$

with $\psi^T = (\text{vec}^T(\boldsymbol{\alpha}_0), \text{vec}^T(\boldsymbol{\eta}), \text{vec}^T(\boldsymbol{\phi}), \text{vec}^T(\boldsymbol{\phi}_{D|S}),\ \text{vech}^T(\mathbf{\Sigma}_{D|S}),\ \text{vech}^T(\mathbf{\Sigma}_S))$ denote the parameters in the envelope model (14). It is obvious that g is twice continuous differentiable. Therefore all the assumptions of Shapiro's Proposition 4.1 are satisfied, and we can get the asymptotic distribution of the estimators $\hat{h}(\hat{\psi})$ from model (14) as

$$H(H^T J H)^{\dagger} H^T,$$

where J is the Fisher information under model (2), and H is the gradient matrix, which equals to $\partial h / \partial^T \psi$. Computing H we have

$$H = \begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\phi} & (\boldsymbol{\eta}^T \otimes \mathbf{I}) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I} & 0 & 0 & 0 \\ 0 & 0 & 2C_k(\boldsymbol{\phi}\boldsymbol{\Omega} \otimes \mathbf{I} - \boldsymbol{\phi} \otimes \boldsymbol{\phi}_0 \boldsymbol{\Omega}_0 \boldsymbol{\phi}_0^T) & 0 & C_k(\boldsymbol{\phi} \otimes \boldsymbol{\phi})E_u & C_k(\boldsymbol{\phi}_0 \otimes \boldsymbol{\phi}_0)E_{k-u} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{I} \end{pmatrix}. \tag{11}$$

where C_r is the contraction matrix that satisfy $\operatorname{vech}(\Sigma) = C_r \operatorname{vec}(\Sigma)$ for Σ a symmetric matrix of dimension r.

Following similar calculations than in Su and Cook (2011, Prop. 1) we get (17) and since $H(H^T J H)^{\dagger} H^T \leq J^{-1}$ we get the efficiency of the estimator $\operatorname{vec}(\widehat{\boldsymbol{\alpha}}_{\text{ecm}}))$ under model (14) compare with the estimator of $\operatorname{vec}(\widehat{\boldsymbol{\alpha}}_{\text{cm}})$ under model (2) when $u < k$ being the same when $u = k$. The asymptotic distribution of $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\text{ecm}}))$ follows since $\boldsymbol{\beta}_{\text{ecm}} = \mathbf{U}\boldsymbol{\alpha}_{\text{ecm}}$ and as a consequence the asymptotic efficiency of $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}})$ compare with $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{cm}})$

Proof of Proposition 3.1 The proof consists of showing that (1) $\boldsymbol{\alpha} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma}\boldsymbol{\eta}$, (2) $\mathrm{rank}(\mathbf{U}^T\boldsymbol{\Gamma}) = u_1$, (3) $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma}) = \mathrm{span}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma})$, and (4) $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$ is a reducing subspace of $\boldsymbol{\Sigma}_{D|S}$ that contains $\mathcal{A} = \mathrm{span}(\boldsymbol{\alpha})$. It follows from these statements that $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A}) \subseteq \mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$. Since $\dim(\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})) = u_1$, the conclusion will follow: $\dim\{\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})\} \leq u_1 \leq u = \dim\{\mathcal{E}_{\boldsymbol{\Sigma}}(\mathrm{span}(\boldsymbol{\beta}))\}$. It remains then to show (1)–(4).

(1) Since $\mathbf{U}$ has full column rank and $\boldsymbol{\Gamma}\boldsymbol{\eta} = \mathbf{U}\boldsymbol{\alpha}$ it follows immediately that $\boldsymbol{\alpha} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma}\boldsymbol{\eta}$.

(2) Direct multiplication gives $\mathbf{U}^T\boldsymbol{\Gamma} = (\mathbf{G}, 0)^T$. The conclusion follows since $\mathbf{G}^T \in \mathbb{R}^{u_1 \times u}$ has rank u_1.

(3) Since $(\mathbf{U}^T\mathbf{U})$ is full rank, $\mathrm{rank}(\mathbf{U}^T\boldsymbol{\Gamma}) = \mathrm{rank}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma})$. It is then sufficient to show that $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma}) \subseteq \mathrm{span}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma})$. For an arbitrary $\boldsymbol{\gamma} \in \mathbb{R}^u$, let $\mathbf{g} = (\mathbf{U}^T\boldsymbol{\Gamma})\boldsymbol{\gamma}$ and $\mathbf{h} = \mathbf{G}\mathbf{G}^T\boldsymbol{\gamma}$. Then by direct multiplication we have $\mathbf{g} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma}\mathbf{h}$. Consequently we have that every vector $\mathbf{g} \in \mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$ is also in $\mathrm{span}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Gamma})$.

(4) That $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$ contains $\mathcal{A}$ follows immediately from (1) and (3). It remains to show that $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$ reduces $\boldsymbol{\Sigma}_{D|S}$ or, equivalently, that it reduces

$$
\begin{aligned}
\boldsymbol{\Sigma}_{D|S}^{-1} &= \mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U} \\
&= \mathbf{U}^T\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{U} + \mathbf{U}^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_0^T\mathbf{U}.
\end{aligned}
$$

Since $\mathbf{U}^T\boldsymbol{\Gamma}_0 = (0, \mathbf{G}_0)^T$, we have immediately that $\mathbf{U}^T\boldsymbol{\Gamma}$ and $\mathbf{U}^T\boldsymbol{\Gamma}_0$ are orthogonal. Let $\mathbf{C}$ be a semi-orthogonal basis matrix for $\mathrm{span}(\mathbf{U}^T\boldsymbol{\Gamma})$ and let $\mathbf{C}_0$ be a semi-orthogonal basis matrix for

$\text{span}(\mathbf{U}^T\mathbf{\Gamma}_0)$. Then there exists matrices $\mathbf{A}$ and $\mathbf{A}_0$ so that

$$\begin{aligned}\mathbf{\Sigma}_{D|S}^{-1}\mathbf{C} &= \mathbf{U}^T\mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^T\mathbf{U}\mathbf{C} = \mathbf{C}\mathbf{A} \\ \mathbf{\Sigma}_{D|S}^{-1}\mathbf{C}_0 &= \mathbf{U}^T\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{U}\mathbf{C}_0 = \mathbf{C}_0\mathbf{A}_0.\end{aligned}$$

The conclusion follows from Cook (2018, Lemma A.1). □

3.1 Maximum likelihood estimator when U is semi-orthogonal

The following two propositions provide technical tools for demonstrating that the maximum likelihood envelope estimator of $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ is simply $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$ when $\mathbf{U}$ is semi-orthogonal.

Proposition 3.1 *(a) Let $\mathcal{S} \subseteq \mathbb{R}^k$ be a reducing subspace of the symmetric matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$, and let $\mathbf{V} \in \mathbb{R}^{p\times k}$ be a semi-orthogonal matrix. Then $\mathbf{V}\mathcal{S}$ is a reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$. (b) Let $\mathcal{D} \in \mathbb{R}^p$ be a reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$. Then $\mathbf{V}^T\mathcal{D}$ is a reducing subspace of $\mathbf{M}$.*

Proof of Proposition 3.1 Since $\mathbf{M}$ is symmetric it is sufficient to show for conclusion (a) that $\mathbf{V}\mathcal{S}$ is an invariant subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$. That is, by definition we must show that $\mathbf{V}\mathbf{M}\mathbf{V}^T(\mathbf{V}\mathcal{S}) \subseteq \mathbf{V}\mathcal{S}$. But this follows immediately from the condition that $\mathcal{S}$ reduces $\mathbf{M}$: $\mathbf{M}\mathcal{S} \subseteq \mathcal{S}$. Conclusion (b) follows similarly since $\mathbf{V}\mathbf{M}\mathbf{V}^T\mathcal{D} \subseteq \mathcal{D}$ and multiplying both sides by $\mathbf{V}^T$ gives the desired conclusion. □

Proposition 3.2 *Let $\mathcal{E}_{\mathbf{M}}(\mathcal{S}) \subseteq \mathbb{R}^k$ be the smallest reducing subspace of the symmetric matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ that contains $\mathcal{S} \subseteq \mathbb{R}^k$, and let $\mathbf{V} \in \mathbb{R}^{p\times k}$ be a semi-orthogonal matrix. Then $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$*

is the smallest reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$ *that contains* $\mathbf{V}\mathcal{S}$*; that is,* $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S}) = \mathcal{E}_{\mathbf{V}\mathbf{M}\mathbf{V}^T}(\mathbf{V}\mathcal{S})$.

Proof of Proposition 3.2 We need to show that (a) $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is a reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$, (b) that $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ contains $\mathbf{V}\mathcal{S}$ and (c) that $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is minimal. It follows immediately from Proposition 3.1 that $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is a reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$, so (a) is satisfied. Since $\mathcal{S} \subseteq \mathcal{E}_{\mathbf{M}}(\mathcal{S})$ it follows immediately that $\mathbf{V}\mathcal{S} \subseteq \mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, so (b) holds.

To show (c) we use the fact that the smallest reducing subspace of $\mathbf{V}\mathbf{M}\mathbf{V}^T$ that contains $\mathbf{V}\mathcal{S}$ is the intersection of all such subspaces. Consequently, if $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is not minimal then there is a subspace $\mathcal{D}$ that satisfies (a) and (b) while being a proper subset of $\mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, $\mathcal{D} \subset \mathbf{V}\mathcal{E}_{\mathbf{M}}(\mathcal{S})$. But we know from Proposition 3.1 that $\mathbf{V}^T\mathcal{D}$ is a reducing subspace of $\mathbf{M}$ and from part (b) it contains $\mathcal{S}$. However, $\mathbf{V}^T\mathcal{D} \subset \mathcal{E}_{\mathbf{M}}(\mathcal{S})$ which contradicts the minimality of $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$. □

These two propositions show that the results of Section 3.1 can be used directly to get the envelope estimator of $\mathbf{U}\boldsymbol{\alpha}$ when $\mathbf{U}$ is semi-orthognonal. The standard MLE of $\mathbf{U}\boldsymbol{\alpha}$ is just $\mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{cm}}$ with asymptotic covariance matrix $\mathbf{U}\boldsymbol{\Sigma}_{D|S}\mathbf{U}^T$. In consequence, following the rationale at the beginning of Section 3.1, we seek the MLE of $\mathcal{E}_{\mathbf{U}\boldsymbol{\Sigma}_{D|S}\mathbf{U}^T}(\mathbf{U}\mathcal{A})$, which by Proposition 3.2 is equal to $\mathbf{U}\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$. From Proposition 3.2, the MLE of $\mathcal{E}_{\mathbf{U}\boldsymbol{\Sigma}_{D|S}\mathbf{U}^T}(\mathbf{U}\mathcal{A})$ is $\mathbf{U}\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$, which implies that the envelope estimator of $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ is $\widehat{\boldsymbol{\beta}}_{\text{ecm}} = \mathbf{U}\widehat{\boldsymbol{\alpha}}_{\text{ecm}}$ with asymptotic variance $\mathbf{U}\text{avar}(\sqrt{n}\widehat{\boldsymbol{\alpha}}_{\text{ecm}})\mathbf{U}^T$. Propositions 3.1 and 3.2 also suggest how to proceed when re-prameterizing as $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha} = (\mathbf{U}\mathbf{O}^T)(\mathbf{O}\boldsymbol{\alpha})$, where $\mathbf{O}$ is an orthognal matrix and $\mathbf{U}$ is not necessarily orthogonal. In that case the envelope estimator of $\mathbf{O}\boldsymbol{\alpha}$ is simply $\mathbf{O}\boldsymbol{\alpha}_{\text{ecm}}$, and so the envelope estimator of $\boldsymbol{\beta}$ is invariant under orthogonal re-paramterization of the kind used here.

3.2 Maximum likelihood estimators and asymptotic distribution under Model (19)

Given the envelope dimension u, the log likelihood based on $\mathbf{Y}_D$ is

$$\begin{aligned} L_u &= -(nk/2)\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{D|S}| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_{Di} - \boldsymbol{\alpha}_0 - \boldsymbol{\Lambda}^{-1}\boldsymbol{\Theta}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si})^T\boldsymbol{\Sigma}_{D|S}^{-1}(\mathbf{Y}_{Di} - \boldsymbol{\alpha}_0 - \boldsymbol{\Lambda}^{-1}\boldsymbol{\Theta}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si}), \end{aligned}$$

where

$$\begin{aligned} -\frac{n}{2}\log|\boldsymbol{\Sigma}_{D|S}| &= n\log|\boldsymbol{\Lambda}| - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| \\ \boldsymbol{\Sigma}_{D|S}^{-1} &= \boldsymbol{\Lambda}(\boldsymbol{\Theta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Theta}^T + \boldsymbol{\Theta}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Theta}_0^T)\boldsymbol{\Lambda}. \end{aligned}$$

Let $\boldsymbol{\nu}_0 = \Lambda\boldsymbol{\alpha}_0$ and $\boldsymbol{\nu} = \Lambda\boldsymbol{\phi}_{D|S}$. Substituting these various quantities we get

$$\begin{aligned} L_u = &-(nk/2)\log(2\pi) + n\log|\boldsymbol{\Lambda}| - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| \\ &-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\Lambda}\mathbf{Y}_{Di} - \boldsymbol{\nu}_0 - \boldsymbol{\Theta}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\nu}\mathbf{Y}_{Si})^T(\boldsymbol{\Theta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Theta}^T + \boldsymbol{\Theta}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Theta}_0^T)(\boldsymbol{\Lambda}\mathbf{Y}_{Di} - \boldsymbol{\nu}_0 - \boldsymbol{\Theta}\boldsymbol{\eta}\mathbf{X}_i - \boldsymbol{\nu}\mathbf{Y}_{Si}), \end{aligned}$$

Aside from the addend $n\log|\boldsymbol{\Lambda}|$, this log likelihood has the same form as that associated with model (14) after replacing the response $\mathbf{Y}_D$ with the transformed response $\boldsymbol{\Lambda}\mathbf{Y}_D$. This enables us to adapt the log likelihood and estimators listed in Section 3.1 for the present setting.

After maximizing the log likelihood over all parameters except $(\boldsymbol{\Lambda}, \boldsymbol{\Theta})$ we have

$$(\widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{\Theta}}) = \arg\min_{\mathbf{A},\mathbf{G}} \log|\mathbf{G}^T\mathbf{A}\mathbf{S}_{D|(\mathbf{X},S)}\mathbf{A}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{A}^{-1}\mathbf{S}_{D|S}^{-1}\mathbf{A}^{-1}\mathbf{G}|, \tag{12}$$

where the minimum is computed over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{k\times v}$ and diagonal matrices $\mathbf{A} = \mathrm{diag}(1, a_2, \ldots, a_k)$. Aside from the inner product matrices $\mathbf{S}_{D|(\mathbf{X},S)}$ and $\mathbf{S}_{D|S}^{-1}$ this is the same as the objective function that Cook and Su (2013) derived for response scaling prior to using model (4), which allowed us to adapt their optimization algorithm to handle (12).

Having determined the MLEs $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{\Theta}}$, the remaining parameter estimators are

- $\widehat{\boldsymbol{\beta}}_{\mathrm{secm}} = \mathbf{U}\widehat{\mathbf{\Lambda}}^{-1}\mathbf{P}_{\widehat{\mathbf{\Theta}}}\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$, $\widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{\Lambda}}^{-1}\mathbf{P}_{\widehat{\mathbf{\Theta}}}\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$, $\widehat{\boldsymbol{\alpha}}_0 = \bar{\mathbf{Y}}_D - \widehat{\boldsymbol{\beta}}_{\mathrm{secm}}\bar{\mathbf{X}} - \widehat{\boldsymbol{\phi}}_{D|S}\bar{\mathbf{Y}}_S$
- $\widehat{\boldsymbol{\eta}} = \widehat{\mathbf{\Theta}}^T\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$, $\widehat{\boldsymbol{\phi}}_{D|S} = (\mathbf{S}_{D,S} - \widehat{\boldsymbol{\alpha}}\mathbf{S}_{\mathbf{X},S})\mathbf{S}_S^{-1}$
- $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Theta}}^T\widehat{\mathbf{\Lambda}}\mathbf{S}_{D|(\mathbf{X},S)}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Theta}}$, $\widehat{\mathbf{\Omega}}_0 = \widehat{\mathbf{\Theta}}_0^T\widehat{\mathbf{\Lambda}}\mathbf{S}_{D|S}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Theta}}_0$
- $\widehat{\mathbf{\Sigma}}_{D|S} = \widehat{\mathbf{\Lambda}}^{-1}(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{\Theta}}^T + \widehat{\mathbf{\Theta}}_0\widehat{\mathbf{\Omega}}_0\widehat{\mathbf{\Theta}}_0^T)\widehat{\mathbf{\Lambda}}^{-1}$, $\widehat{\mathbf{\Sigma}}_S = \mathbf{T}_S$.

The variances $\mathbf{\Sigma}_{\mathbf{W}}$ and $\mathbf{\Sigma}$ can be estimated as indicated in Section 2.1.

This representation of the scaled envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{secm}}$ shows the construction process. First the direct-information response is transformed to $\widehat{\mathbf{\Lambda}}\mathbf{Y}_D$. The constrained estimator $\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$ and the envelope estimator $\mathbf{P}_{\widehat{\mathbf{\Theta}}}\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$ are then determined in the transformed scale. Next, the estimator is transformed back to the original scale by multiplying by $\widehat{\mathbf{\Lambda}}^{-1}$ to get $\widehat{\mathbf{\Lambda}}^{-1}\mathbf{P}_{\widehat{\mathbf{\Theta}}}\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{\alpha}}_{\mathrm{cm}}$, which is the estimator of $\boldsymbol{\alpha}$ in the original scale. Finally, the estimator in the original scale is multiplied by $\mathbf{U}$ to give the scaled envelope estimator of $\boldsymbol{\beta}$. In effect, $\widehat{\mathbf{\Lambda}}$ is a similarity transformation to represent $\mathbf{P}_{\widehat{\mathbf{\Theta}}}$ in the original coordinate system as $\widehat{\mathbf{\Lambda}}^{-1}\mathbf{P}_{\widehat{\mathbf{\Theta}}}\widehat{\mathbf{\Lambda}}$.

The fully maximized log likelihood is

$$\hat{L}_u = c - \frac{n}{2}\left\{\log|\mathbf{T}_S| + \log|\mathbf{S}_{D|S}| + \log|\widehat{\mathbf{\Theta}}^T\widehat{\mathbf{\Lambda}}\mathbf{S}_{D|(\mathbf{X},S)}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Theta}}| + \log|\widehat{\mathbf{\Theta}}^T\widehat{\mathbf{\Lambda}}^{-1}\mathbf{S}_{D|S}^{-1}\widehat{\mathbf{\Lambda}}^{-1}\widehat{\mathbf{\Theta}}|\right\}, \quad (13)$$

where $c = n \log |\mathbf{W}| - (nr/2)(1 + \log(2\pi))$. To describe the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{secm}}$, let $\mathbf{V}_{\text{secm}}$ denote the upper $pk \times pk$ diagonal block of the asymptotic variance $\mathbf{V}$ given by Proposition 2 from Cook and Su (2013) with $\boldsymbol{\Sigma}$ replaced by $\boldsymbol{\Sigma}_{D|S}$, $\boldsymbol{\Gamma}$ by $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}_0$ by $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}^{-1}$. Additionally, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ in the Cook-Su notation are the same as the corresponding quantities in the decomposition of $\boldsymbol{\Sigma}_{D|S}$ for model (19). For the asymptotic variance of the estimators, we need to recognize that the log-likelihood function (12) correspond to the log-likelihood function from Su and Cook (2013) where $\mathbf{S}_{D|S}$ correspond to their $\widetilde{\Sigma}_y$. Then $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{secm}})) = (\mathbf{I}_p \otimes \mathbf{U})\mathbf{V}_{\text{secm}}(\mathbf{I}_p \otimes \mathbf{U}^T)$ is an estimator that is never less efficient than $\widehat{\boldsymbol{\beta}}_{\text{cm}}$.

4 Additional simulations

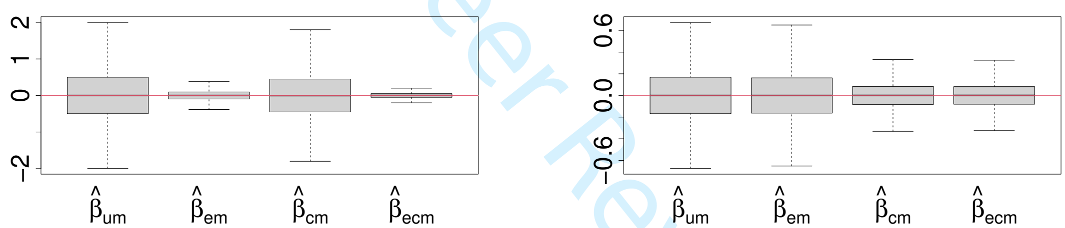
We carry out an additional set of simulations under the settings described in Section 4.1 with a small sample size $n = 80$. The overall performance of the four estimators are similar to that with a larger sample size shown in the main text.

4.1 Scenario 1

In the 1000 simulations repetitions, the envelope dimension was correctly estimated as 6 for 329 times, under estimated as 4 for 132 times and 5 for 398 times, overly estimated as 7 for 131 times and 8 for 10 times using BIC. The dimension of the constrained envelope estimator was correctly estimated as 4 for 787 times, overly estimated as 5 for 207 times, and 6 for 6 times. The empirical results for $\widehat{\boldsymbol{\beta}}_{\text{um}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\text{em}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\text{cm}} - \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\text{ecm}} - \boldsymbol{\beta}$ are shown in Figure 1a, where all the elements of $\boldsymbol{\beta}$ are plotted in the same boxplot as if they are from the same population and the outliers are suppressed for a cleaner representation.

The average estimated asymptotic variances were close to the theoretical asymptotic variances calculated using the true parameter values for all three estimators. The mean of the empirical asymptotic variances across all the elements in four estimators are 54.94 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ and 40.07 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ but is only 1.59 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and 1.16 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. That is, in this setting, the envelope estimator and the constrained envelope estimator are about 40 times more efficient that the saturated estimator and the unconstrained estimator. A comparison between the envelope and constrained envelope estimator demonstrates the advantage of leveraging prior information in terms of achieving better efficiency.

Figure 1: Box plot of $\widehat{\boldsymbol{\beta}}_{\mathrm{um}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{em}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}} - \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}} - \boldsymbol{\beta}$ in two scenarios with a relatively small sample size $n = 80$ in 1000 simulations.



(a) Scenario 1 (b) Scenario 2

4.2 Scenario 2

In the 1000 simulations repetitions, the envelope dimension is correctly estimated as 6 for 675 times, and overly estimated as 7 for 281 times and 8 for 44 times. The constrained envelope dimension is correctly estimated as 4 for 994 times and overly estimated as 5 for 6 times. The empirical biases of the estimators are shown in Figure 2b. Again, all four estimators are centered around 0, indicating the asymptotic unbiasedness. As expected, the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ are the most efficient among the four estimators, while the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ is still more efficient than $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$.

The average estimated asymptotic variance of the three estimators were all close to their theoretical values. The average empirical variances of all the elements in three estimators are 8.76 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, 8.36 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, 1.37 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and 1.60 for $\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. That is, in this setting, the estimators using a correctly specified $\mathbf{U}$ are on average about 4 times of more efficient that the saturated estimator and the envelope estimator, but the envelope constraint estimator does not provide additional advantages over the constrained estimator.

5 Dental data revisited

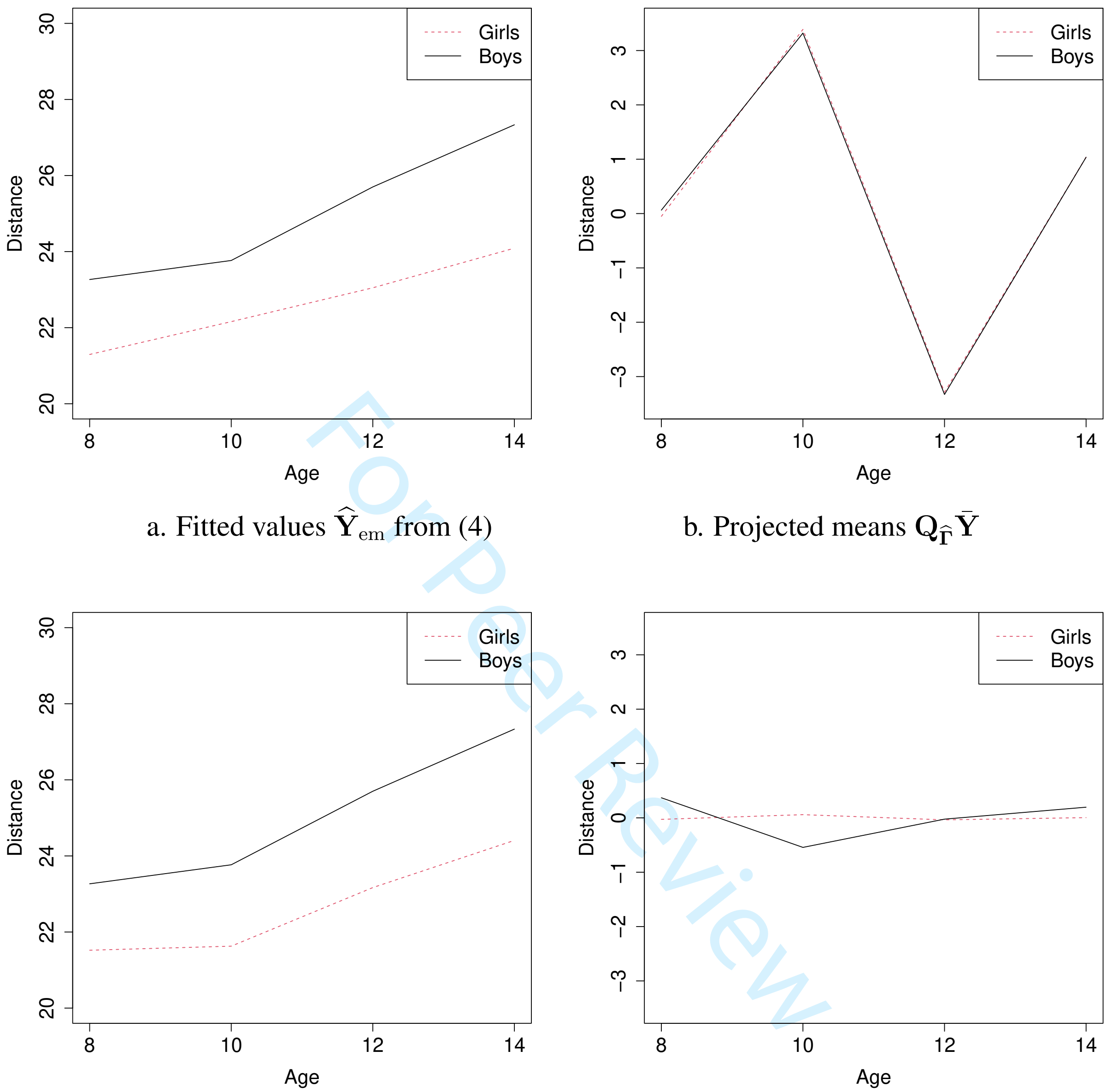
The dental data consists of measurements of the distance (mm) from the center of the pituitary to the pterygomaxillary fissure for each of 11 girls and 16 boys at ages 8, 10, 12, and 14 years (t). Since their introduction by Potthoff and Roy (1964), these data have been used frequently to illustrate the analysis of longitudinal data. We respect that tradition in this section. We removed the outlying and influential male case described by Pan and Fang (2002) prior to application of the methods discussed herein. We set as a goal to characterize the differences between boys and girls rather than to profile modeling and so we contrasted the behavior of estimators from the unconstrained model (1), the envelope model (4), the constrained model (3), and the envelope version of model (3) discussed in Section 3.1.

Consistent with the literature, we fitted constrained model (3) and its envelope counterpart with the rows of $\mathbf{U}$ being $\mathbf{U}^T(t) = (1, t)$. The estimated dimension of the envelope for model (4) was $u = 2$, and thus it was inferred that only two linear combinations of the response vectors are needed to fully characterize the differences between boys and girls. The estimated dimension of the envelope for the constrained envelope model (14) was $u = 1$. Table 1 shows the estimated asymptotic variances, determined by the plug-in method, for the four estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$. The unconstrained model has the worst estimated performance, followed by the regular envelope model and the constrained model. The enveloping in the constrained model has the best estimated performance. We would need to increase the sample size by about 2.5 times for the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ to have the performance estimated for the enveloped version $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ with the current sample size. The relatively bad performance of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ can be

Table 1: Estimated asymptotic variances $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{(\cdot)})$ of the four elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ from the unconstrained model (1), $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ from the envelope model (4), $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ from the constrained model (3) and $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ from the envelope version of constrained model (3).

	Age			
$\widehat{\boldsymbol{\beta}}_{(\cdot)}$	8	10	12	14
$\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$	15.53	16.41	25.42	18.95
$\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$	15.29	13.56	22.79	18.73
$\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$	13.97	13.57	15.00	18.27
$\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$	5.88	9.16	13.16	17.89

traced back to the estimated eigen-structure of $\boldsymbol{\Sigma}$. The eigenvalues of $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Omega}}_0$ were $(14.61, 1.10)$ and $(2.20, 0.70)$. Envelopes offer relatively little gain when most of the variation in the response is associated material information, as is the case here. On the other hand, the eigenvalues of $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Omega}}_0$ arising from enveloping in the constrained model were 0.02 and 8.31. In this case most of the variation in the direct response $\mathbf{Y}_D$ is associated with immaterial information, the general setting when envelopes perform well. Figure 2a gives a profile plot of the fitted vectors from envelope model (4). The implied fit is quite good and close to the profile plot of the raw mean vectors shown in Supplement Figure 3a (profile plots of residuals are also shown in Figure 3). Under envelope theory, the distribution of $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y}$ should be independent of the predictor values, in this case, sex.



a. Fitted values $\widehat{\mathbf{Y}}_{\rm em}$ from (4)

b. Projected means $\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\bar{\mathbf{Y}}$

c. Fitted values $\widehat{\mathbf{Y}}_{\rm cm}$ from (3)

d. Projected means $\mathbf{Q}_{\mathbf{U}}\bar{\mathbf{Y}}$

Figure 2: Profile plots by sex of (a) the fitted vectors from envelope model (4), (b) means projected onto $\mathrm{span}^{\perp}(\widehat{\boldsymbol{\Gamma}})$, (c) the fitted vectors from the constrained model (3) and (d) means projected onto $\mathcal{U}^{\perp}$. The vertical axis for each plot is the distance for the plotted vectors.

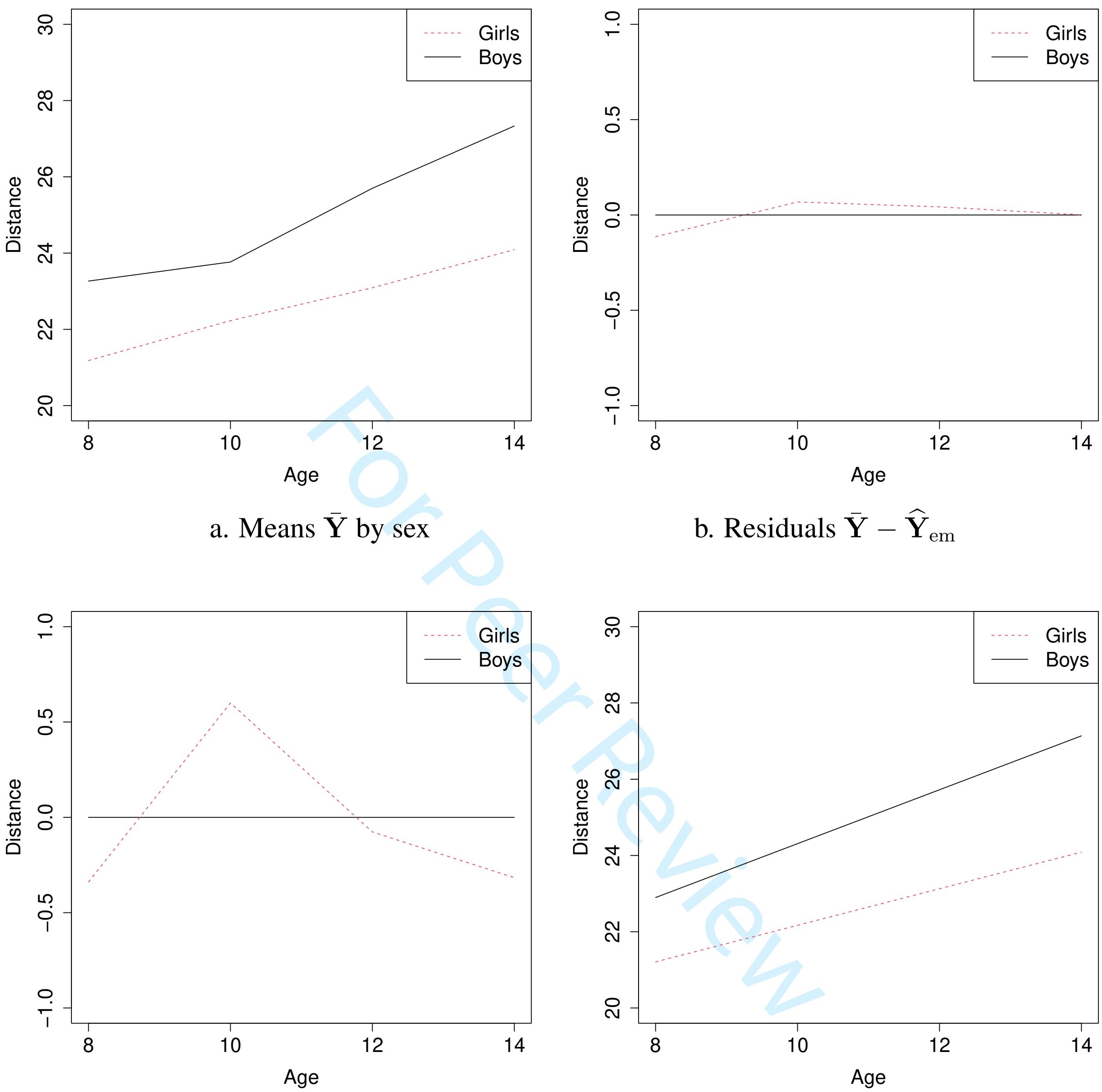
The profile plot of $\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\bar{\mathbf{Y}}$ by sex shown in Figure 2b reflects this property. Figures 2cd show the corresponding plots from the fit of the constrained model (3). The fit of the constrained model altered the shape of the profile for girls so that it more closely matches that for boys, which was not done by the fit of the envelope model. This type of conformity is an intrinsic property of constrained model (3).

If there is uncertainty about the containment $\mathcal{B} \subseteq \mathcal{U}$ needed for the constrained model then it may be desirable to base an analysis on the envelope model (4). Otherwise, the results in the last two rows of Table 1 indicate that enveloping in the constrained model (3) is the best option from among those considered.

We also applied the scaled envelope estimator discussed in Section 3.2. The asymptotic variances of the elements of the corresponding estimator of $\boldsymbol{\beta}$ did not differ materially from those shown in Table 1 for $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$: Scaling offered no gains in this example. This was as expected since good scale estimation generally requires large sample size.

5.1 Additional plots for the dental data

Figure 3 gives additional plots for the dental data. Figure 3a is a profile plot of the raw mean vectors by sex, where there is a slight upward bend in the line for girls at age 10. This bend accounts for the residual pattern from the fit of constrained model (3) shown in Figure 3c. The bend is reduced in the plot of the residual vectors $\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathrm{em}}$ from envelope model (3) shown in Figure 3b. A comparison of the residual plots in Figures 3bc indicates that the envelope model fits the raw means noticeably better than the constrained model. For contrast, Figure 2d is a plot of the fitted vectors from model (2). In that model we have $\mathrm{span}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \subseteq \mathcal{U}$ and consequently the profile plots of the fitted vectors are linear.



a. Means $\bar{\mathbf{Y}}$ by sex

b. Residuals $\bar{\mathbf{Y}} - \widehat{\mathbf{Y}}_{\mathrm{em}}$

c. Residuals $\bar{\mathbf{Y}} - \widehat{\mathbf{Y}}_{\mathrm{cm}}$ from (3)

b. Fitted vectors $\widehat{\mathbf{Y}}_{\mathrm{cm}}$ from (2)

Figure 3: Dental data: Profile plots by sex of (a) the raw mean vectors, (b) Residual vectors from the fit of envelope model (4),(c) Residual vectors from the fit of constrained model (2) and (d) Fitted vectors $\widehat{\mathbf{Y}}_{\mathrm{cm}}$ from the fit of constrained model (3).

6 The China Health and Nutrition Survey

The China Health and Nutrition Survey (CHNS) was designed to evaluate the effects of the health, nutrition and family planning policies on the health and nutritional status of its population (Popkin et al., 2009). The survey used a multistage, random cluster process to draw samples of households in 15 provinces and municipal cities that vary substantially in geography, economic development, public resources, and health indicators. In total, 9 surveys were carried out between 1989 and 2011. We included in our analysis only the 1209 individuals that participated in all of the 9 surveys, giving a total of $9 \times 1209 = 10,881$ records. Five individuals were deleted for having unreasonable changes in weight or height. For instance, one individual had a height of 65 cm in the seventh survey but a height of 160 cm in all other surveys. The baseline predictors we considered include age at the first survey, binary indicators for gender and region (urban or rural), and a six-level indicator for highest education levels obtained at the first survey. About 98.2% of the individuals in the analysis were over 21 years old. Age at first survey, gender and region were fully observed but there were 28 individuals with missing education levels at baseline. We imputed the missing values with the education level collected at the next available visit. The response was the change in BMI from baseline at the 8 followup surveys. In the $10,881$ records, there was a total of 371 values of either missing height or weight information needed to calculate BMI. We assumed that height and weight were missing at random and imputed them by carrying the last observation forward.

We compared the estimated asymptotic variances of the unconstrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ from model (3) using $\mathbf{U}^T = (1, t, t^2)$, where t is the time in years from baseline. We also included the envelope version of the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$, the scaled envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{sem}}$ from Cook and Su (2013) and its constrained version $\widehat{\boldsymbol{\beta}}_{\mathrm{secm}}$ corresponding to model (3). We used version (3) of the constrained model because we were interested in profile contrasts rather than modeling profiles per se.

Since $\widehat{\boldsymbol{\beta}}_{(\cdot)} \in \mathbb{R}^{8\times 8}$, we first report in columns 4–9 of Table 2 various location statistics computed over the estimated variances of the individual elements in $\widehat{\boldsymbol{\beta}}_{(\cdot)}$. Using these summary statistics as the basis for comparison, we see that the estimators fall into two clear groups. The unconstrained estimator does the worst, followed closely by the envelope estimator and the constrained estimator. Our assessment based on just the variance summary statistics and taking computational difficulty into account leads us to prefer the envelope constrained estimator $\widehat{\boldsymbol{\beta}}_{\rm ecm}$. The model order determined by BIC given in the third column of Table 2 tells a similar story. Based on the actual BIC values, the unconstrained estimator in the first row appears clearly inferior to the others, while the scaled constrained envelope model in the last row is clearly the best. The remaining models are relatively difficult to distinguish. We next give a few additional details.

Table 2: BIC order, minimum, maximum, mean and quartiles Q_1–Q_3 of the estimated asymptotic variances of the elements in $\widehat{\boldsymbol{\beta}}_{(\cdot)}$ for the CHNS study

Estimator	Dimension	BIC order	Min	Q_1	Q_2	Mean	Q_3	Max
$\widehat{\boldsymbol{\beta}}_{\rm um}$	8	6	0.03	0.07	0.12	0.12	0.15	0.20
$\widehat{\boldsymbol{\beta}}_{\rm em}$	2	5	0.02	0.05	0.11	0.11	0.15	0.21
$\widehat{\boldsymbol{\beta}}_{\rm cm}$	3	4	0.03	0.05	0.10	0.10	0.13	0.19
$\widehat{\boldsymbol{\beta}}_{\rm ecm}$	1	3	0.00	0.00	0.02	0.05	0.06	0.17
$\widehat{\boldsymbol{\beta}}_{\rm sem}$	1	2	0.00	0.03	0.03	0.04	0.04	0.08
$\widehat{\boldsymbol{\beta}}_{\rm secm}$	1	1	0.00	0.00	0.02	0.05	0.06	0.17

The estimated dimensions of the various envelopes based using BIC are listed in the second column of Table 2. We listed the maximum envelope dimension for the two non-envelope methods. The variance gains for the envelope model over the unconstrained model shown in Table 2 are reflected by the two eigenvalues $(17.44, 15.90)$ of $\widehat{\boldsymbol{\Omega}}$ and the six eigenvalues of $\widehat{\boldsymbol{\Omega}}_0$ which ranged between 1.11 and 1.62. Turning to the envelope version of the constrained model (3), the estimated dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathcal{A})$ using BIC was 1. The variance gain over the unconstrained model shown in Table 2 is again reflected by the value of $\widehat{\boldsymbol{\Omega}} = 3 \times 10^{-5}$ and the two eigenvalues of $\widehat{\boldsymbol{\Omega}}_0$, 0.16 and 2.74. As with the regular envelope model, most of the variability lies in the immaterial part of the response.

The point estimates and standard errors of the considered estimators are given in Tables 3-8. Out of 8 predictors across 8 time points (64 variables in total), the unconstrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$, the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, the constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ and the scaled envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{sem}}$ find 53, 41, 51, and 56 variables significant, while both the envelope constrained estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ and the scaled version $\widehat{\boldsymbol{\beta}}_{\mathrm{secm}}$ find all 64 variables significant.

Table 3: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\mathrm{um}}$ for CHNS study, where the rows correspond to surveys, columns are predictors and 0.00 means < 0.01

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	0.09[0.05]	0.22[0.05]	0.25[0.05]	0.21[0.05]	0.17[0.05]	0.07[0.05]	0.21[0.05]	0.16[0.04]
2	0.15[0.07]	0.17[0.06]	0.25[0.07]	0.09[0.07]	0.14[0.07]	0.06[0.07]	0.19[0.07]	0.15[0.05]
3	-0.16[0.08]	-0.04[0.08]	-0.09[0.08]	-0.01[0.08]	-0.04[0.08]	-0.21[0.08]	-0.10[0.08]	-0.09[0.06]
4	-0.27[0.10]	-0.26[0.09]	-0.21[0.10]	-0.23[0.10]	-0.28[0.10]	-0.49[0.10]	-0.31[0.10]	-0.30[0.07]
5	-0.51[0.11]	-0.37[0.10]	-0.42[0.11]	-0.44[0.11]	-0.50[0.11]	-0.60[0.11]	-0.44[0.11]	-0.48[0.08]
6	-0.64[0.11]	-0.61[0.11]	-0.58[0.11]	-0.71[0.11]	-0.69[0.11]	-0.81[0.11]	-0.63[0.11]	-0.68[0.08]
7	-0.91[0.13]	-0.73[0.12]	-0.77[0.13]	-0.92[0.12]	-0.80[0.12]	-0.99[0.12]	-0.91[0.12]	-0.88[0.09]
8	-1.00[0.13]	-0.87[0.12]	-0.86[0.13]	-1.10[0.13]	-0.94[0.13]	-1.11[0.13]	-1.07[0.13]	-0.99[0.10]

Table 4: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ for CHNS study, where the rows correspond to surveys, columns are predictors

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	0.09[0.05]	0.13[0.05]	0.14[0.05]	0.15[0.05]	0.11[0.05]	0.03[0.05]	0.14[0.05]	0.09[0.04]
2	0.05[0.07]	0.10[0.06]	0.12[0.07]	0.12[0.07]	0.07[0.07]	-0.03[0.07]	0.10[0.07]	0.04[0.05]
3	-0.04[0.08]	0.04[0.08]	0.06[0.08]	0.04[0.08]	0.00[0.08]	-0.14[0.08]	0.02[0.08]	-0.04[0.06]
4	-0.20[0.10]	-0.11[0.09]	-0.09[0.10]	-0.13[0.10]	-0.16[0.10]	-0.32[0.10]	-0.15[0.10]	-0.21[0.07]
5	-0.51[0.11]	-0.40[0.10]	-0.40 [0.11]	-0.49[0.11]	-0.47[0.11]	-0.64[0.11]	-0.49[0.11]	-0.52[0.08]
6	-0.63[0.11]	-0.52[0.11]	-0.52[0.11]	-0.64[0.11]	-0.59[0.11]	-0.75[0.11]	-0.63[0.11]	-0.65[0.08]
7	-0.85[0.13]	-0.73[0.12]	-0.74[0.13]	-0.89[0.12]	-0.81[0.12]	-0.97[0.12]	-0.87[0.12]	-0.87[0.09]
8	-1.01[0.13]	-0.89[0.12]	-0.91[0.13]	-1.09[0.13]	-0.97[0.13]	-1.12[0.13]	-1.06[0.13]	-1.03[0.10]

Table 5: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\mathrm{cm}}$ for CHNS study, where the rows correspond to surveys, columns are predictors

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	0.12[0.06]	0.23[0.06]	0.27[0.06]	0.20[0.06]	0.19[0.06]	0.09[0.06]	0.22[0.06]	0.18[0.05]
2	0.05[0.06]	0.14[0.06]	0.17[0.06]	0.13[0.06]	0.10[0.06]	-0.01[0.06]	0.13[0.06]	0.09[0.05]
3	-0.11[0.08]	-0.06[0.07]	-0.04[0.08]	-0.05[0.08]	-0.10[0.08]	-0.23[0.08]	-0.07[0.08]	-0.10[0.06]
4	-0.27[0.09]	-0.22[0.08]	-0.21[0.09]	-0.22[0.09]	-0.26[0.09]	-0.41[0.09]	-0.25[0.09]	-0.27[0.07]
5	-0.51[0.10]	-0.45[0.09]	-0.45[0.10]	-0.50[0.10]	-0.49[0.10]	-0.66[0.10]	-0.52[0.10]	-0.52[0.08]
6	-0.65[0.11]	-0.57[0.10]	-0.58[0.10]	-0.65[0.10]	-0.62[0.10]	-0.79[0.10]	-0.67[0.10]	-0.65[0.08]
7	-0.88[0.11]	-0.76[0.11]	-0.77[0.11]	-0.92[0.11]	-0.82[0.11]	-1.00[0.11]	-0.91[0.11]	-0.87[0.09]
8	-1.05[0.13]	-0.90[0.12]	-0.90[0.13]	-1.12[0.12]	-0.96[0.12]	-1.15[0.12]	-1.09[0.12]	-1.03[0.10]

Table 6: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\mathrm{ecm}}$ for CHNS study, where the rows correspond to surveys, columns are predictors and 0.00 means < 0.01

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]
2	-0.04[0.00]	-0.03[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]
3	-0.15[0.02]	-0.14[0.01]	-0.14[0.02]	-0.17[0.02]	-0.14[0.02]	-0.15[0.02]	-0.16[0.02]	-0.15[0.01]
4	-0.28[0.03]	-0.26[0.03]	-0.27[0.03]	-0.32[0.03]	-0.27[0.03]	-0.29[0.03]	-0.31[0.03]	-0.28[0.02]
5	-0.52[0.06]	-0.48[0.05]	-0.50[0.06]	-0.59[0.05]	-0.50[0.05]	-0.54[0.05]	-0.57[0.05]	-0.52[0.04]
6	-0.67[0.07]	-0.62[0.07]	-0.64[0.07]	-0.76[0.07]	-0.64[0.07]	-0.69[0.07]	-0.74[0.07]	-0.67[0.05]
7	-0.93[0.10]	-0.86[0.09]	-0.88[0.10]	-1.05[0.10]	-0.89[0.10]	-0.96[0.10]	-1.02[0.10]	-0.93[0.07]
8	-1.13[0.12]	-1.04[0.11]	-1.07[0.12]	-1.27[0.12]	-1.07[0.12]	-1.16[0.12]	-1.23[0.12]	-1.13[0.09]

Table 7: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\mathrm{sem}}$ for CHNS study, where the rows correspond to surveys, columns are predictors

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	0.32[0.05]	0.34[0.05]	0.35[0.05]	0.39[0.05]	0.33[0.05]	0.33[0.05]	0.39[0.05]	0.32[0.04]
2	0.30[0.05]	0.32[0.05]	0.33[0.06]	0.37[0.06]	0.31[0.05]	0.31[0.05]	0.37[0.06]	0.31[0.05]
3	0.20[0.05]	0.21[0.06]	0.22[0.06]	0.24[0.06]	0.20[0.05]	0.20[0.05]	0.24[0.06]	0.20[0.05]
4	0.09[0.05]	0.10[0.06]	0.10[0.06]	0.11[0.06]	0.09[0.05]	0.09[0.05]	0.11[0.06]	0.09[0.05]
5	0.06[0.02]	0.06[0.02]	0.07[0.02]	0.07[0.02]	0.06[0.02]	0.06[0.02]	0.07[0.02]	0.06[0.02]
6	-0.19[0.04]	-0.20[0.04]	-0.21[0.05]	-0.23[0.05]	-0.20[0.04]	-0.20[0.04]	-0.23[0.05]	-0.19[0.04]
7	-0.37[0.06]	-0.39[0.06]	-0.40[0.06]	-0.45[0.07]	-0.37[0.06]	-0.37[0.06]	-0.44[0.07]	-0.37[0.05]
8	-0.52[0.08]	-0.56[0.08]	-0.57[0.08]	-0.64[0.08]	-0.53[0.08]	-0.53[0.08]	-0.63[0.08]	-0.53[0.07]

Table 8: Point estimates ($\times 10$), standard error in backets ($\times 10$) of $\widehat{\boldsymbol{\beta}}_{\rm secm}$ for CHNS study, where the rows correspond to surveys, columns are predictors and 0.00 means < 0.01

S	Gender	edu1b	edu2b	edu3b	edu4b	edu5b	Urban	ageb
1	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]	-0.01[0.00]
2	-0.04[0.00]	-0.03[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]	-0.04[0.00]
3	-0.15[0.02]	-0.14[0.01]	-0.14[0.02]	-0.17[0.02]	-0.14[0.02]	-0.15[0.02]	-0.16[0.02]	-0.15[0.01]
4	-0.28[0.03]	-0.26[0.03]	-0.27[0.03]	-0.32[0.03]	-0.27[0.03]	-0.29[0.03]	-0.31[0.03]	-0.28[0.02]
5	-0.52[0.06]	-0.48[0.05]	-0.50[0.06]	-0.59[0.05]	-0.50[0.05]	-0.54[0.05]	-0.57[0.05]	-0.52[0.04]
6	-0.67[0.07]	-0.62[0.07]	-0.64[0.07]	-0.76[0.07]	-0.64[0.07]	-0.69[0.07]	-0.74[0.07]	-0.67[0.05]
7	-0.93[0.10]	-0.86[0.09]	-0.88[0.10]	-1.05[0.10]	-0.89[0.10]	-0.96[0.10]	-1.02[0.10]	-0.93[0.07]
8	-1.13[0.12]	-1.04[0.11]	-1.07[0.12]	-1.27[0.12]	-1.07[0.12]	-1.16[0.12]	-1.23[0.12]	-1.13[0.09]

7 Enveloping $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$

Our focus has so far been on estimation of $\boldsymbol{\beta}$, either in the unconstrained model (1), the constrained model with $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$, the envelope model (4) with $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, the envelope constrained model with $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\phi}\boldsymbol{\eta}$ or in one of the scaled version. This emphasis on $\boldsymbol{\beta}$ reflects an interest in profile contrasts rather than on the profiles themselves. When $\mathbf{U}$ is selected to model profiles rather than profile contrasts, as in model (2), the intercept vector $\boldsymbol{\alpha}_0$ may be of interest because it represents coordinates of the profile when $\mathbf{X} = 0$. In this section we again consider the model (2), but now we pursue envelop estimation of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}$ simultaneously, which may be appropriate when profiles are important.

The model decomposition (6)–(7) still holds so only (7) is required to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$, although both are again required for the full likelihood function and asymptotic variances. The standard estimator $(\widehat{\boldsymbol{\alpha}}_{0,\rm cm}, \widehat{\boldsymbol{\alpha}}_{\rm cm})$ is asymptotically normal with variance $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{0,\rm cm}, \widehat{\boldsymbol{\alpha}}_{\rm cm})) =$

345 $\boldsymbol{\tau}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{D|S}$ where

$$\boldsymbol{\tau}_{\mathbf{X}} = \lim_{n\to\infty} \begin{pmatrix} 1 & \bar{\mathbf{X}}^T \\ \bar{\mathbf{X}} & \mathbf{T}_{\mathbf{X}} \end{pmatrix}.$$

346 Turning to envelopes and writing $\mathbf{Z} = (1, \mathbf{X}^T)^T$, model (7) can be rewritten as

$$\begin{aligned} \mathbf{Y}_{Di} \mid \mathbf{Y}_{Si} &= (\boldsymbol{\alpha}_0, \boldsymbol{\alpha})\mathbf{Z}_i + \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} + \mathbf{e}_{D|Si} \\ &= \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{Z}_i + \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si} + \mathbf{e}_{D|Si} \qquad (14) \\ \boldsymbol{\Sigma}_{D|S} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}\boldsymbol{\Gamma}_0^T, \end{aligned}$$

347 where $\boldsymbol{\Gamma}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathrm{span}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}))$. The number of real parameters
348 in this model is $u(p+1) + r(r+1)/2$. Since we are enveloping only on $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ this is again in
349 the form of a partial envelope, but now there is no intercept term. The likelihood function for
350 parameters $(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\eta}, \boldsymbol{\phi}_{D|S}, \boldsymbol{\Sigma}_S)$ is

$$\begin{aligned} L_u &= -nr(1+\log(2\pi))/2 - \frac{n}{2}\log|\boldsymbol{\Sigma}_S| - \frac{1}{2}\sum_{i=1}^n \mathbf{Y}_{Si}^T\boldsymbol{\Sigma}_S^{-1}\mathbf{Y}_{Si} - \frac{n}{2}\log|\boldsymbol{\Sigma}_{D|S}| \\ &\quad -\frac{1}{2}\sum_{i=1}^n [(\mathbf{Y}_{Di} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{Z}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si})^T\boldsymbol{\Sigma}_{D|S}^{-1}(\mathbf{Y}_{Di} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{Z}_i - \boldsymbol{\phi}_{D|S}\mathbf{Y}_{Si})] \end{aligned}$$

351 The maximum likelihood estimator of $\boldsymbol{\Sigma}_S$ is $\mathbf{T}_S$, which is the same as for the other models we
352 have considered. Holding all other parameters fixed, the value of $\boldsymbol{\phi}_{D|S}$ that maximizes L is

$$\boldsymbol{\phi}_{D|S} = \mathbf{T}_{D,S}\mathbf{T}_S^{-1} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{T}_{\mathbf{Z},S}\mathbf{T}_S^{-1}.$$

353 Substituting this into the log likelihood we find its maximized over $\boldsymbol{\eta}$ by $\boldsymbol{\Gamma}^T$ times the coefficient

matrix from the ordinary least squares regression of $\mathbf{Y}_D$ on $(\mathbf{Z}, \mathbf{Y}_S)$, $\boldsymbol{\eta} = \boldsymbol{\Gamma}^T \mathbf{S}_{D,(\mathbf{Z},S)} \mathbf{S}^{-1}_{(\mathbf{Z},S)}$ Continuing to maximize the resulting partially maximized log likelihoods, we have $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T \mathbf{S}_{D|(\mathbf{Z},S)} \boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T \mathbf{T}_{\mathbf{R}_{D|S}} \boldsymbol{\Gamma}_0$, which results in the log likelihood maximized over all parameters except $\boldsymbol{\Gamma}$:

$$L_u(\boldsymbol{\Gamma}) = -\frac{n}{2} \left\{ r(1 + \log(2\pi)) + \log|\widehat{\boldsymbol{\Sigma}}_S| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{D|(\mathbf{Z},S)} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T \mathbf{T}_{\mathbf{R}_{D|S}} \boldsymbol{\Gamma}_0| \right\}.$$

The maximum likelihood estimator of an envelope basis can thus be represented as

$$\widehat{\boldsymbol{\Gamma}} = \arg\min_{\mathbf{G}} \log|\mathbf{G}^T \mathbf{S}_{D|(\mathbf{Z},S)} \mathbf{G}| + \log|\mathbf{G}^T \mathbf{T}^{-1}_{\mathbf{R}_{D|S}} \mathbf{G}|,$$

where the minimum is computed over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{k \times u}$. The fully maximized log likelihood is then

$$\hat{L}_u = c - \frac{n}{2} \left\{ \log|\widehat{\boldsymbol{\Sigma}}_S| + \log|\mathbf{T}_{\mathbf{R}_{D|S}}| + \log|\widehat{\boldsymbol{\Gamma}}^T \mathbf{S}_{D|(\mathbf{Z},S)} \widehat{\boldsymbol{\Gamma}}| + \log|\widehat{\boldsymbol{\Gamma}}^T \mathbf{T}^{-1}_{\mathbf{R}_{D|S}} \widehat{\boldsymbol{\Gamma}}| \right\},$$

where $c = n \log|\mathbf{W}| - (nr/2)(1 + \log(2\pi))$.

Finally, the maximum likelihood estimator of a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{D|S}}(\mathrm{span}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}))$ can be represented as $\widehat{\boldsymbol{\Gamma}} = \arg\min_{\mathbf{G}} \log|\mathbf{G}^T \mathbf{S}_{D|(\mathbf{Z},S)} \mathbf{G}| + \log|\mathbf{G}^T \mathbf{T}^{-1}_{\mathbf{R}_{D|S}} \mathbf{G}|$, where the minimum is computed over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{k \times u}$. The envelope estimator of $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ is then $(\widehat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\alpha}}) = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}(\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}})$ where $(\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}})$ is the ordinary least squares estimators for the coefficient of the predictor $(1, \mathbf{X})$ in the regression of $\mathbf{Y}_D$ onto $(1, \mathbf{X}, \mathbf{Y}_S)$. The estimators of $\boldsymbol{\phi}_{D|S}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Sigma}_{D|S}$ can then be constructed by substituting $\widehat{\boldsymbol{\Gamma}}$ into previously given expressions for them. The asymptotic variance of $\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\alpha}})$ is $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\alpha}})) = \boldsymbol{\tau}_{\mathbf{Z}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\tau}_{\mathbf{Z}})(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0)$.

8 Non-normality and the bootstrap

The methods presented herein are all based on maximum likelihood estimators, assuming normal errors. When the errors are non-normal with finite fourth moments, all estimators are still root-n consistent and asymptotically normal (see, for example, Cook and Zhang, 2015, Section 5.1), but the asymptotic variances based on the Fisher information under normality may no longer be accurate. In such cases the residual bootstrap can be used for variances and inference. For illustration, we next describe how to use the bootstrap to estimate $\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}))$, the variance of the envelope estimator of $\boldsymbol{\alpha}$ as described in Section 3.1. The procedure is similar for other settings.

Let $\mathcal{R} = \{\mathbf{R}_1, \ldots, \mathbf{R}_n\}$ denote the collection of residual vectors $\mathbf{R}_i$ from the standard fit of model (2) and recall that the estimators for the corresponding envelope model are denoted with "hats", as described in Section 3.1. Then a bootstrap sample $\{\mathbf{R}_1^*, \ldots, \mathbf{R}_n^*\}$ from $\mathcal{R}$ is used to generate a bootstrap sample $\{\mathbf{Y}_i^*\}$ of the responses as follows,

$$\mathbf{Y}_i^* = \mathbf{U}\widehat{\boldsymbol{\alpha}}_0 + \mathbf{U}\widehat{\boldsymbol{\alpha}}\mathbf{X}_i + \mathbf{R}_i^*,\ i = 1, \ldots, n.$$

The resulting bootstrap data $\{\mathbf{Y}_i^*, \mathbf{X}_i,\ i = 1, \ldots, n\}$ are then used to construct the first bootstrap estimator $\widehat{\boldsymbol{\alpha}}_1^*$ of $\boldsymbol{\alpha}$, employing the value of u used in the construction of $\widehat{\boldsymbol{\alpha}}$ along with methods described in Section 3.1. Repeating this process B times gives bootstrap estimates $\widehat{\boldsymbol{\alpha}}_j^*$, $j = 1, \ldots, B$. The sample variance $\mathbf{S}_{\mathrm{vec}(\widehat{\boldsymbol{\alpha}}^*)}$ is then a bootstrap estimator of $\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\alpha}}))$. Background on using the residual bootstrap with envelope models is available from Cook (2018, Section 11.1).

9 Estimating the envelope dimension

Methods for estimating the envelope dimension u include likelihood ratio testing, an information criterion or cross validation (Cook et al., 2010). A review and expanded discussion of estimation of u is available from Cook (2018).

The likelihood ratio for testing an envelope model with dimension $u < k$ against the model with $u = k$ can be cast as a test of the hypothesis $u = u_0$ versus the alternative $u = k$. The likelihood ratio statistic for this hypothesis is $\Lambda(u_0) = 2(\hat{L}_k - \hat{L}_{u_0})$, where $\hat{L}_a$ is the maximized envelope log likelihood for the envelope model in question with $u = a$. Under the null hypothesis this statistic is distributed asymptotically as a chi-squared random variable with $(k - u_0)$ degrees of freedom, the number of real parameters for the standard model ($u = k$) minus that for the envelope model with $u = u_0$, The likelihood ratio test statistic $\Lambda(u_0)$ can be used sequentially to estimate u: Starting with $u_0 = 0$, test the hypothesis $u = u_0$ against $u = k$ at a selected level α. If the hypothesis is rejected, increment u_0 by 1 and test again. The estimate $\hat{u}$ of u is the first hypothesized value that is not rejected.

The envelope dimension can also be selected by using an information criterion:

$$\widehat{u} = \arg\min_u \{-2\hat{L}_u + h(n)N(u)\}, \tag{15}$$

where $N(u)$ is the number of real parameters in the envelope model with envelope dimension u, and $h(n) = \log n$ for BIC and $h(n) = 2$ for AIC. Theoretical results (Su and Cook, 2013, Prop. 4) supported by simulations indicate that AIC tends to overestimate u. BIC will select the correct u with probability tending to 1 as $n \to \infty$ (Yang, 2005), but it can be slow to respond in small samples. Selection by likelihood ratio testing can perform well depending on the sample size, but asymptotically it makes an error with rate α. It may be useful to use all three methods in applications, giving a preference to BIC and LRT if there is disagreement, or using the largest estimate of u in cases where it is desirable to be conservative. It is also possible to avoid the selection of u by using model averaging to combine the envelope estimators over all possible values of u (Eck and Cook, 2017).

10 Envelopes and Rao's simple structure

In this section we contrast envelopes and Rao's structure. Since Rao's structure is typically employed the context of prediction we assume that $\mathbf{U}$ is semi-orthogonal to ease exposition. In the context of model (2), Rao's simple structure is

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T + \mathbf{U}_0\boldsymbol{\Delta}_0\mathbf{U}_0^T, \tag{16}$$

where $\boldsymbol{\Delta} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{\Delta}_0 \in \mathbb{R}^{(r-k)\times(r-k)}$ are positive definite and $(\mathbf{U}, \mathbf{U}_0)$ is orthogonal, as defined previously. It follows from this structure that the eigenvectors of $\boldsymbol{\Sigma}$ must be in either $\mathcal{U}$ or $\mathcal{U}^\perp$, and that $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U}) = \mathcal{U}$. This simplifies the analysis considerably since it implies that $\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}_0 = 0$ and thus that $\mathbf{Y}_S \perp\!\!\!\perp \mathbf{Y}_D \mid \mathbf{X}$, $\boldsymbol{\phi}_{D|S} = 0$, $\boldsymbol{\Sigma}_{D|S} = \boldsymbol{\Delta}$, $\boldsymbol{\Sigma}_S = \boldsymbol{\Delta}_0$ and that analysis can be based on the unconstrained model $\mathbf{Y}_{Di} \mid \mathbf{X}_i = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}\mathbf{X}_i + \mathbf{e}_{D|Si}$, which becomes the basis for an analysis based on the envelope $\mathcal{E}_{\boldsymbol{\Delta}}(\mathcal{A})$, recalling that $\mathcal{A} = \mathrm{span}(\boldsymbol{\alpha})$.

In an envelope analysis based on a model like (14), the envelope dimension u is in effect a model-selection parameter that typically needs to be inferred from the data (see Section 9). Assuming that an originating model like (2) holds, the only remaining model selection issue is the choice of u. If it is concluded that $u = k$, so the model defaults to the original growth curve model, then envelopes offer no gain. If it is concluded that $u < k$ then there is a proper envelope model and some perhaps substantial efficiency gains can be expected. In this sense an envelope analysis is adaptive through the choice of u. In contrast, Rao's simple structure is non-adaptive because it relies on the strong assumption that $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U}) = \mathcal{U}$. One possible generalization of Rao's structure is to base analysis on $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{U})$ without requiring that it equal $\mathcal{U}$.

Rao's approach was to impose a structure on $\boldsymbol{\Sigma}$ via (16), while in the envelope approach we use an adaptive structure on $\boldsymbol{\Sigma}_{D|S}$. It seems most informative to compare these structures on the $\mathbf{W}$ scale via the precision matrix, $\boldsymbol{\Sigma}_{\mathbf{W}}^{-1}$. Under Rao's structure,

$$\boldsymbol{\Sigma}_{\mathbf{W}}^{-1} = \begin{pmatrix} \boldsymbol{\Delta}^{-1} & 0 \\ 0 & \boldsymbol{\Delta}_0^{-1} \end{pmatrix},$$

while under the envelope model (14)

$$\boldsymbol{\Sigma}_{\mathbf{W}}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{D|S}^{-1} & -\boldsymbol{\Sigma}_{D|S}^{-1}\boldsymbol{\phi}_{D|S} \\ -\boldsymbol{\phi}_{D|S}^{T}\boldsymbol{\Sigma}_{D|S}^{-1} & \boldsymbol{\Sigma}_{S}^{-1} + \boldsymbol{\phi}_{D|S}^{T}\boldsymbol{\Sigma}_{D|S}^{-1}\boldsymbol{\phi}_{D|S} \end{pmatrix}.$$

It seems clear from these representations that the envelope structure is much less restrictive, which is reflected also by the parameter counts in the $\boldsymbol{\Sigma}_{\mathbf{W}}$'s. The number of free real parameters in Rao's $\boldsymbol{\Sigma}_{\mathbf{W}}$ is $r(r+1)/2 - k(r-k)$, while that for the envelope model is $r(r+1)/2$, the difference being reflected by the absence of $\boldsymbol{\phi}_{D|S}$ in Rao's structure.

References

Cook, R. D. (2018), An Introduction to Envelopes, Wiley, Hoboken, NJ.

Cook, R. D. and Forzani, L. (2008), 'Covariance reducing models: An alternative to spectral modelling of covariance matrices', Biometrika **95**(4), 799–812.
URL: *https://doi.org/10.1093/biomet/asn052*

Cook, R. D., Forzani, L. and Zhang, X. (2015), 'Envelopes and reduced-rank regression', Biometrika **102**(2), 439–456.
URL: *https://doi.org/10.1093/biomet/asv001*

Cook, R. D., Li, B. and Chiaromonte, F. (2010), 'Envelope models for parsimonious and efficient multivariate linear regression', Statistica Sinica **20**(3), 927–960.

Cook, R. D. and Su, Z. (2013), 'Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression', Biometrika **100**(4), 939–954.

Cook, R. D. and Zhang, X. (2015), 'Simultaneous envelopes for multivariate linear regression', Technometrics **57**(1), 11–25.

Eck, D. J. and Cook, R. D. (2017), 'Weighted envelope estimation to handle variability in model selection.', Biometrica **104**(3), 743–749.
URL: *https://doi.org/10.1093/biomet/asx035*

Milliken, G. A. and Akdeniz, F. (1977), 'A theorem on the difference of the generalized inverse of two nonnegative matrices', Communications in Statistics – Theory and Methods **6**, 73–79.

Pan, J.-X. and Fang, K.-T. (2002), Growth Curve Models and Statistical Diagnostics, Springer, New York.

Popkin, B., Du, S., Zhai, F. and Zhang, B. (2009), 'Cohort profile: The china health and nutrition surveymonitoring and understanding socio-economic and health change in china, 1989–2011', International Journal of Epidemiology **39**, 1435–1440.

460 Potthoff, R. F. and Roy, S. N. (1964), 'A generalized multivariate analysis of variance model useful especially for growth curve problems', Biometrika **51**(3), 313–326.
URL: *https://www.jstor.org/stable/2334137*

463 Shapiro, A. (1986), 'Asymptotic theory of overparameterized structural models', J. Am. Statist. Assoc. **81**, 142–149.

464 Su, Z. and Cook, R. D. (2011), 'Partial envelopes for efficient estimation in multivariate linear regression', Biometrika **98**(1), 133–146.
URL: *http://dx.doi.org/10.1093/biomet/asq063*

467 Su, Z. and Cook, R. D. (2013), 'Estimation of multivariate means with heteroscedastic errors using envelope models', Statistica Sinica **23**(1), 213–230.

469 Yang, Y. (2005), 'Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation', Biometrika **92**(4), 937–950.