A Note on Fast Envelope Estimation

R. Dennis Cook,[*] Liliana Forzani[†] and Zhihua Su[‡]

May 7, 2016

Abstract

We propose a new algorithm for envelope estimation, along with a new $\sqrt{n}$-consistent method for computing starting values. The new algorithm, which does not require optimization over a Grassmannian, is shown by simulation to be much faster and typically more accurate than the best existing algorithm proposed by Cook and Zhang [7].

Key Words: Envelopes; Grassmann manifold; reducing subspaces.

1. Introduction

The goal of envelope methods is to increase efficiency in multivariate parameter estimation and prediction by exploiting variation in the data that is effectively immaterial to the goals

[*] 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, Universidad Nacional del Litoral and Instituto Matemática Aplicada Litoral, CONICET-UNL, Santa Fe, Argentina (Email: liliana.forzani@gmail.com).

[‡] Zhihua Su is Assistant Professor, Department of Statistics, University of Florida, Gainsville, FL 32611-8545 (E-mail: zhihuasu@stat.ufl.edu).

of the analysis. Envelopes achieve efficiency gains by basing estimation on the variation that is material to those goals, while simultaneously excluding that which is immaterial. It now seems evident that immaterial variation is often present in multivariate analyses and that the estimative improvement afforded by envelopes can be quite substantial when the immaterial variation is large, sometimes equivalent to taking thousands of additional observations.

Algorithms for envelope estimation require optimization of a non-convex objective function over a Grassmannian, which can be quite slow in all but small or modest sized problems, possibly taking hours or even days to complete an analysis of a sizable problem. Local optima are another complication that may increase the difficulty of the computations and the analysis generally. Until recently, envelope methods were available only in Matlab, as these computing issues hindered implementation in R.

In this article we propose new easily computed $\sqrt{n}$-consistent starting values and a novel non-Grassmann algorithm for optimization of the most common envelope objective function. These computing tools are much faster than current algorithms in sizable problems and can be implemented straightforwardly in R. The new starting values have proven quite effective and can be used as fast standalone estimators in exploratory analyses. An R package that implements the algorithm was developed and is available at `http://www.stat.ufl.edu/~zhihuasu/Renvlp`.

In the remainder of this introduction we review envelopes and describe the computing issues in more detail. We let $\mathbf{P}_{(\cdot)}$ denote a projection with $\mathbf{Q}_{(\cdot)} = \mathbf{I} - \mathbf{P}_{(\cdot)}$, let $\mathbb{R}^{r\times c}$ be the set of all real $r \times c$ matrices, and let $\mathbb{S}^{k\times k}$ be the set of all real and symmetric $k \times k$ matrices.

If $\mathbf{M} \in \mathbb{R}^{r\times c}$, then $\mathrm{span}(\mathbf{M}) \subseteq \mathbb{R}^r$ is the subspace spanned by columns of $\mathbf{M}$. vec is the vectorization operator that stacks the columns of a matrix. A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.

1.1. Review of envelopes

Envelopes were originally proposed and developed by Cook et al. [2, 3] in the context of multivariate linear regression,

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

where $\boldsymbol{\varepsilon}_i \in \mathbb{R}^r$ is a normal error vector with mean 0, variance $\boldsymbol{\Sigma} > 0$ and is independent of $\mathbf{X}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix in which we are primarily interested. Immaterial variation can occur in $\mathbf{Y}$ or $\mathbf{X}$ or both. Cook et al. [3] operationalized the idea of immaterial variation in the response vector by asking if there are linear combinations of $\mathbf{Y}$ whose distribution is invariant to changes in $\mathbf{X}$. Specifically, let $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ denote the projection onto a subspace $\mathcal{E} \subseteq \mathbb{R}^r$ with the properties (1) the distribution of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid \mathbf{X}$ does not depend on the value of the non-stochastic predictor $\mathbf{X}$ and (2) $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is independent of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ given $\mathbf{X}$. These conditions imply that the distribution of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ is not affected by $\mathbf{X}$ marginally or through an association with $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$. Consequently, changes in the predictor affect the distribution of $\mathbf{Y}$ only via $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ and so we refer to $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ informally as the material part of $\mathbf{Y}$ and to $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ as the immaterial part of $\mathbf{Y}$.

Conditions (1) and (2) hold if and only if (a) $\mathcal{B} := \mathrm{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}$ (so $\mathcal{E}$ *envelopes* $\mathcal{B}$) and (b) $\mathcal{E}$ reduces $\boldsymbol{\Sigma}$. The $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, denoted $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, is defined formally as the intersection of all reducing subspaces of $\boldsymbol{\Sigma}$ that contain $\mathcal{B}$. Let $u = \dim\{\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})\}$ and let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix with $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ and $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. This leads directly to the envelope version of model (1),

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \text{ with } \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\top, \; i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ gives the coordinates of $\boldsymbol{\beta}$ relative to basis $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega} \in \mathbb{S}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{(r-u)\times(r-u)}$ are positive definite matrices. While $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ depend on the basis $\boldsymbol{\Gamma}$ selected to represent $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, the parameters of interest $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ depend only on $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and not on the basis. All parameters in (2) can be estimated by maximizing its likelihood with the envelope dimension u determined by using standard methods like likelihood ratio testing, information criteria, cross-validation or a hold-out sample, as described by Cook et al. [3]. The envelope estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is just the projection of the ordinary least squares estimator $\mathbf{B}$ of $\boldsymbol{\beta}$ onto the estimated envelope, $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\mathcal{E}}}\mathbf{B}$, and $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta})\}$ is asymptotically normal with mean 0 and covariance matrix given by Cook et al. [3], where u is assumed to be known. An introductory example of response envelopes is available in Cook and Zhang [5].

Similar reasoning leads to partial envelopes for use when only selected columns of $\boldsymbol{\beta}$ are of interest (Su and Cook [10]), to predictor envelopes allowing for immaterial variation in $\mathbf{X}$ (Cook et al. [1]), to predictor-response envelopes allowing simultaneously for immaterial variation in $\mathbf{X}$ and $\mathbf{Y}$ (Cook and Zhang [6]) and to heteroscedastic envelopes for comparing the means of multivariate populations with unequal covariance matrices (Su and Cook [11]).

Cook and Zhang [5] extended envelopes beyond multivariate linear models by proposing the following estimative construct for vector-valued parameters. Let $\widetilde{\boldsymbol{\theta}}$ denote an estimator of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ based on a sample of size n and assume, as is often the case, that $\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\mathbf{V}(\boldsymbol{\theta}) > 0$ as $n \to \infty$. To accommodate the presence of nuisance parameters, decompose $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\psi}^\top, \boldsymbol{\phi}^\top)^\top$, where $\boldsymbol{\phi} \in \mathbb{R}^p$, $p \leq m$, is the parameter vector of interest and $\boldsymbol{\psi} \in \mathbb{R}^{m-p}$ is the nuisance parameter vector. The asymptotic covariance matrix of $\widetilde{\boldsymbol{\phi}}$ is represented as $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$, which is the $p \times p$ lower right block of $\mathbf{V}(\boldsymbol{\theta})$. Then Cook and Zhang [5] defined the envelope for improving $\widetilde{\boldsymbol{\phi}}$ as the smallest reducing subspace of $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$ that contains $\mathrm{span}(\boldsymbol{\phi})$, $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})}\{\mathrm{span}(\boldsymbol{\phi})\} \subseteq \mathbb{R}^p$. This definition links the envelope to a particular pre-specified method of estimation through the covariance matrix $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$, while normal-theory maximum likelihood is the only method of estimation allowed by the previous approaches. The goal of an envelope is to improve that pre-specified estimator, perhaps a maximum likelihood, least squares or robust estimator. Second, the matrix to be reduced – here $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$ – is dictated by the method of estimation. Third, the matrix to be reduced can now depend on the parameter being estimated, in addition to perhaps other parameters. Cook and Zhang [5] sketched application details for generalized linear models, weighted least squares, Cox regression and described an extension to matrix-valued parameters.

1.2. Computational issues

The approaches reviewed in the last section all require estimation of an envelope, now represented generically as $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$, the smallest reducing subspace of $\mathbf{M} \in \mathbb{S}^{r \times r}$ that contains $\mathcal{U} \subseteq \mathbb{R}^r$, where $\mathbf{M} > 0$. Let $u = \dim\{\mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$, let $\mathbf{\Gamma} \in \mathbb{R}^{r \times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$, let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be an orthogonal matrix, let $\widehat{\mathbf{M}}$ be a $\sqrt{n}$-consistent estimator of $\mathbf{M}$, and let $\widehat{\mathbf{U}}$ be a positive semi-definite $\sqrt{n}$-consistent estimator of a basis matrix $\mathbf{U}$ for $\mathcal{U}$. With u specified, the most common objective function used for envelope estimation is

$$L_u(\mathbf{G}) = \ln|\mathbf{G}^\top \widehat{\mathbf{M}} \mathbf{G}| + \ln|\mathbf{G}^\top (\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1} \mathbf{G}|, \tag{3}$$

and the envelope is estimated as $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}) = \operatorname{span}\{\arg\min L_u(\mathbf{G})\}$, where the minimum is taken over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{r \times u}$. Objective function (3) corresponds to maximum likelihood estimation under normality for many envelopes, including those associated with (1). Otherwise it provides a $\sqrt{n}$-consistent estimator of the projection onto $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ provided $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent (Cook and Zhang [7], who also provided additional background on $L_u(\mathbf{G})$).

In the case of response envelopes reviewed in Section 1.1, $\widehat{\mathbf{M}}$ is the covariance matrix of the residuals from the ordinary least squares fit of (1), denoted $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ is marginal sample covariance matrix of $\mathbf{Y}$, denoted $\mathbf{S}_{\mathbf{Y}}$. The envelope estimator $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\mathcal{E}}}\mathbf{B}$ is the maximum likelihood estimator if the errors are normal. If the errors are not normal but have finite fourth moments then $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal. In the general context of Cook and Zhang [5], also reviewed in Section 1.1, $\widehat{\mathbf{M}}$ is set to a

$\sqrt{n}$-consistent estimator of $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$ and $\widehat{\mathbf{U}} = \widetilde{\boldsymbol{\phi}}\widetilde{\boldsymbol{\phi}}^{\top}$.

For any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$, $L_u(\mathbf{G}) = L_u(\mathbf{GO})$, so $L_u(\mathbf{G})$ depends only on $\mathrm{span}(\mathbf{G})$ and not on a particular basis. Thus the optimization problem is over a Grassmannian (See Edelman et al. [8] for background on optimization over Grassmann manifolds.). Since it takes $u(r-u)$ real numbers to specify $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ uniquely, Grassmann optimization is usually computationally straightforward when $u(r-u)$ is not too large, but it can be very slow when $u(r-u)$ is large. Also, since $L_u(\mathbf{G})$ is non-convex, the solution returned may correspond to a local rather than global minimum, particularly when the signal is small relative to the noise.

It is important that we have a fast and reliable method of determining $\arg\min L_u(\mathbf{G})$ because we may need to repeat that operation hundreds or even thousands of times in an analysis. An information criterion like AIC or BIC is often used to select a suitable value for u, and this requires that we find $\arg\min L_u(\mathbf{G})$ for $u = 0, \ldots, r$. Predictive cross validation might also be used to select u, again requiring many optimizations of $L_u(\mathbf{G})$; repeating five fold cross validation with 50 random partitions require in total $250 \times r$ optimizations. Asymptotic standard errors are available for many normal models, but we may wish to use a few hundred bootstrap samples to determine standard errors when normality is in doubt or when we wish to check the accuracy of the asymptotic approximations. And may more bootstrap samples may be required if we want accurate inference statements. In some analyses we may wish to fit a few model variations, again multiplying the computation time. In cases like those discussed at the end of Section 1.1, $\mathbf{M} = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$, which may depend on unknown parameters, necessitating another level of iteration for the best results (See Cook and Zhang [5] for further discussion of this point.) In short, a seemingly small savings in computation time for one optimization of $L_u(\mathbf{G})$ can translate into massive savings over the course of an analysis. Additionally, the choice of starting value for $\mathbf{G}$ can be crucial since the objective function is non-convex. Converging to a local minimum can negate the advantages of maximum likelihood estimation, for example. Trying several different starting values is not really an effective method since it again multiplies the total computation time and in our experience is not likely to result in the global optimum.

Cook, Su and Yang ([4]; https://github.com/emeryyi/envlp) developed a fairly comprehensive Matlab toolbox *envlp* for envelope estimation based on Lippert's *sg_min* program for optimization over Stiefel and Grassmann manifolds (http://web.mit.edu/$\sim$ ripper/www/software/). This is a very effective toolbox for small to moderate sized analyses, but otherwise is susceptible to all of the issues mentioned previously. Cook and Zhang [7] replaced $L_u(\mathbf{G})$ with a sequential *1D algorithm* that can be computationally much faster than *sg_min* and is less dependent on good starting values. Nevertheless, it is still susceptible to the problems described previously, although less so than methods based on *sg_min*. Additionally, since it does not provide $\arg\min L_u(\mathbf{G})$, it loses the advantages of that accrue with maximum likelihood estimation when normality is a reasonable assumption. For instance, information criteria like AIC and BIC are no longer available straightforwardly, and likelihood ratio testing is problematic and thus dimension selection must typically be guided by cross validation.

In this paper we propose an iterative non-Grassmann method to compute $\arg\min L_u(\mathbf{G})$ that is faster and more reliable that existing methods in large analyses and otherwise performs about the same. It depends crucially on new effective $\sqrt{n}$-consistent starting values that can also be used as standalone estimators. We restrict our comparisons to the 1D algorithm, since Cook and Zhang [7] have demonstrated its superiority over direct optimization methods based on *sg_min*.

The new starting values are developed in Section 2 and the new algorithm, which relies the new starting values, is described in Section 3. Supporting simulation results are given in Section 4 and contrasts on real data are given in Section 5. Proofs are given in an appendix.

2. Starting values

In this section we describe how to choose the u columns of the starting value for $\mathbf{G}$ from the eigenvectors of $\widehat{\mathbf{M}}$ or $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. To gain intuition about the approach, consider the following population representations. Since $\mathcal{U} \subseteq \mathcal{E}_{\mathbf{M}}(\mathcal{U})$, we have $\mathbf{U} = \boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^{\top}$ for some positive semi-definite $\mathbf{V} \in \mathbb{S}^{u \times u}$. Similarly, $\mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\top}$ and $(\mathbf{M} + \mathbf{U})^{-1} = \boldsymbol{\Gamma}(\boldsymbol{\Omega} + \mathbf{V})^{-1}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^{\top}$. For the starting values selected from the eigenvectors of $\widehat{\mathbf{M}}$ to work well, the eigenvalues of $\boldsymbol{\Omega}$ need to be well distinguished from those of $\boldsymbol{\Omega}_0$. If some of the eigenvalues of $\boldsymbol{\Omega}$ are close to a subset of the eigenvaues of $\boldsymbol{\Omega}_0$ then in samples the corresponding eigenspaces will likely be confused when attempting to minimize $L_u(\mathbf{G})$. In other words, we may well pick vectors near $\mathrm{span}(\boldsymbol{\Gamma}_0)$ instead of eigenvectors near $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\mathbf{M}}(\mathcal{U})$. In such cases we may obtain a better starting value by choosing from the eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ rather than the eigenvectors of $\widehat{\mathbf{M}}$. The same argument applies to choosing the starting values from the eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$:

the eigenvalues of $\boldsymbol{\Omega} + \mathbf{V}$ need to be well distinguished from those of $\boldsymbol{\Omega}_0$. If some of the eigenvalues of $\boldsymbol{\Omega} + \mathbf{V}$ are close to a subset of the eigenvalues of $\boldsymbol{\Omega}_0$ then in samples the corresponding eigenspaces will again likely be confused. In such cases we may obtain better starting values by starting with the eigenvectors of $\widehat{\mathbf{M}}$ rather than the eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. The general conclusion from this discussion is that for effective starting values we will need to consider both $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. Scaling will also be an issue, as discussed later in this section, leading to four potential starting values. The actual starting value used is the one that minimizes $L_u(\mathbf{G})$.

We make use of the following result.

Proposition 2.1 *Let* $(\mathbf{G}, \mathbf{G}_0) \in \mathbb{R}^{r\times r}$ *be an orthogonal matrix with* $\mathbf{G} \in \mathbb{R}^{r\times u}$ *and let* $\mathbf{M} \in \mathbb{S}^{r\times r}$ *be a positive definite matrix. Then* $\ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}_0^\top\mathbf{M}\mathbf{G}_0|$ *and* $\ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}^\top\mathbf{M}^{-1}\mathbf{G}|$ *are both minimized globally when the columns of* $\mathbf{G}$ *span any* u *dimensional reducing subspace of* $\mathbf{M}$.

In the next section we describe how to select starting values from the eigenvectors of $\widehat{\mathbf{M}}$.

2.1. Choosing the starting value from the eigenvectors of $\widehat{\mathbf{M}}$

Define $J_1(\mathbf{G}) = \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0|$, $J_2(\mathbf{G}) = \ln|\mathbf{I}_{r-u} + \mathbf{G}_0^\top\widehat{\mathbf{U}}_{\mathbf{M}}\mathbf{G}_0|$ and $J(\mathbf{G}) = J_1(\mathbf{G}) + J_2(\mathbf{G})$, where $\widehat{\mathbf{U}}_{\mathbf{M}} = \widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{M}}^{-1/2}$ is a standardized version of $\widehat{\mathbf{U}}$. Assume for convenience that the eigenvalues of $\widehat{\mathbf{M}}$ are unique, which will typically hold with probability 1, and let $\mathcal{V}_u$ be the collection of all subsets of u eigenvectors of $\widehat{\mathbf{M}}$. Then

Proposition 2.2 $\arg\min_{\mathbf{G}\in\mathcal{V}_u} L_u(\mathbf{G}) = \arg\min_{\mathbf{G}\in\mathcal{V}_u} J(\mathbf{G})$.

Consequently, instead of $L_u(\mathbf{G})$ we can work with the more amenable objective function $J(\mathbf{G}) = J_1(\mathbf{G}) + J_2(\mathbf{G})$ when restricting starting values to the eigenvectors of $\widehat{\mathbf{M}}$. It follows from Proposition 2.1 that $J_1(\mathbf{G})$ is minimized when the columns of $\mathbf{G}$ are any u eigenvectors of $\widehat{\mathbf{M}}$. Restricting $\mathbf{G} \in \mathcal{V}_u$, we next need to find $\arg\min_{\mathbf{G}\in\mathcal{V}_u} J_2(\mathbf{G})$. This does not have a closed-form solution and evaluating at all r-choose-u elements of $\mathcal{V}_u$ will be effectively impossible when r is large. For these reasons we replace the $\ln$-determinant in $J_2(\mathbf{G})$ with the trace and minimize $\operatorname{tr}(\mathbf{I}_{r-u} + \mathbf{G}_0^\top \widehat{\mathbf{U}}_{\mathbf{M}} \mathbf{G}_0)$, which is equivalent to maximizing

$$K_{\mathbf{M}}(\mathbf{G}) := \operatorname{tr}(\mathbf{G}^\top \widehat{\mathbf{U}}_{\mathbf{M}} \mathbf{G}) = \sum_{i=1}^{u} \mathbf{g}_i^\top \widehat{\mathbf{U}}_{\mathbf{M}} \mathbf{g}_i,$$

where $\mathbf{g}_i$ is the i-th selected eigenvector of $\widehat{\mathbf{M}}$ (the i-th column of $\mathbf{G}$). Computation is now easy, since we just select the u eigenvectors of $\widehat{\mathbf{M}}$ that maximize $\mathbf{g}_i^\top \widehat{\mathbf{U}}_{\mathbf{M}} \mathbf{g}_i$.

Applying this in response envelopes, let $\mathbf{S}_{\mathbf{X}}$ denote the marginal sample covariance matrix of the predictors. Then $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, $\widehat{\mathbf{U}} = \mathbf{B}\mathbf{S}_{\mathbf{X}}\mathbf{B}^\top$, $\widehat{\mathbf{U}}_{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}\mathbf{B}^\top\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}$, and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}^{1/2}$ is a standardized version of the ordinary least squares estimator $\mathbf{B}$ of $\boldsymbol{\beta}$.

2.2. Choosing the starting value from the eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$

Define $J_1^*(\mathbf{G}) = \ln|\mathbf{G}^\top(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\mathbf{G}| + \ln|\mathbf{G}^\top(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{G}|$, $J_2^*(\mathbf{G}) = \ln|\mathbf{I}_u - \mathbf{G}^\top\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{G}|$ and $J^*(\mathbf{G}) = J_1^*(\mathbf{G}) + J_2^*(\mathbf{G})$, where $\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}} = (\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1/2}\widehat{\mathbf{U}}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1/2}$ is another standardized version of $\widehat{\mathbf{U}}$. Let $\mathcal{V}_u^*$ be the collection of all subsets of u eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. Then

Proposition 2.3 $\arg\min_{\mathbf{G}\in\mathcal{V}_u^*} L_u(\mathbf{G}) = \arg\min_{\mathbf{G}\in\mathcal{V}_u^*} J^*(\mathbf{G})$.

Consequently, instead of $L_u(\mathbf{G})$ we can again work with a more amenable objective function, this time $J^*(\mathbf{G}) = J_1^*(\mathbf{G}) + J_2^*(\mathbf{G})$. It follows from Proposition 2.1 that $J_1^*(\mathbf{G})$ is minimized when the columns of $\mathbf{G}$ are any u eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. Restricting $\mathbf{G} \in \mathcal{V}_u^*$, we next need to find $\arg\min_{\mathbf{G}\in\mathcal{V}_u} J_2^*(\mathbf{G})$. Again, this does not have a closed-form solution and evaluating at all r-choose-u elements of $\mathcal{V}_u^*$ will be effectively impossible when r is large. For these reasons we again replace the $\ln$-determinant with the trace and minimize $\mathrm{tr}(\mathbf{I}_u - \mathbf{G}^\top\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{G})$, which is equivalent to maximizing

$$K_{\mathbf{M}+\mathbf{U}}(\mathbf{G}) := \mathrm{tr}(\mathbf{G}^\top\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{G}) = \sum_{i=1}^{u} \mathbf{g}_i^\top\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{g}_i,$$

where g_i is the i-th selected eigenvector of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ (the i-th column of $\mathbf{G}$). Computation is again easy, since we just select the u eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ that maximize $\mathrm{g}_i^\top\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{g}_i$. This is exactly the same as the previous case, except the standardization of $\widehat{\mathbf{U}}$ is with $(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1/2}$ instead of $\widehat{\mathbf{M}}^{-1/2}$.

Applying this in response envelopes, $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, $\widehat{\mathbf{U}} = \mathbf{B}\mathbf{S}_{\mathbf{X}}\mathbf{B}^\top$, $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$, $\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}\mathbf{B}^\top\mathbf{S}_{\mathbf{Y}}^{-1/2}$ and $\mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}^{1/2}$ is another standardized matrix of ordinary least squares regression coefficients as before.

2.3. Scaling and consistency

The standardized forms $\widehat{\mathbf{U}}_{\mathbf{M}}$ and $\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}$ are important when the scales involved in $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}+\widehat{\mathbf{U}}$ are very different. This can perhaps be appreciated readily in the context of response envelopes, where $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$. In this case the standardization will be important and effective if the scales of the elements of $\mathbf{Y}$ are very different. However, the standardization will be effectively unnecessary when the scales are similar. In the case of response envelopes this means that the scales of the elements of $\mathbf{Y}$ are the same or similar.

Depending on the scales involved, standardization can also be counterproductive when the sample size is not large enough to give sufficiently accurate estimates of $\mathbf{M}$ and $\mathbf{U}$. In such cases, we abandon the standardization and use either $K^*_{\mathbf{M}}(\mathbf{G}) = \sum_{i=1}^{u} \mathbf{g}_i^\top \widehat{\mathbf{U}} \mathbf{g}_i$ or $K^*_{\mathbf{M}+\mathbf{U}}(\mathbf{G}) = \sum_{i=1}^{u} \mathbf{g}_i^\top \widehat{\mathbf{U}} \mathbf{g}_i$ as the objective function. The only difference between these is that $K^*_{\mathbf{M}}(\mathbf{G})$ confines $\mathbf{G}$ to the eigenvectors of $\widehat{\mathbf{M}}$, while $K^*_{\mathbf{M}+\mathbf{U}}(\mathbf{G})$ confines $\mathbf{G}$ to the eigenvectors of $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. We now have four possible starting values from which to choose, corresponding to the arguments that minimize $K_{\mathbf{M}}(\mathbf{G})$, $K^*_{\mathbf{M}}(\mathbf{G})$, $K_{\mathbf{M}+\mathbf{U}}(\mathbf{G})$, and $K^*_{\mathbf{M}+\mathbf{U}}(\mathbf{G})$. The value $\mathbf{G}_{\text{start}}$ chosen to start the algorithm described in Section 3 is the one that minimizes $L_u(\mathbf{G})$. The following proposition summarizes an asymptotic property of this starting value.

Proposition 2.4 *Let* $\mathbf{P}_{\text{start}}$ *denote the projection onto* $\text{span}(\mathbf{G}_{\text{start}})$. *Then with known* u, $\mathbf{P}_{\text{start}}$ *is a* $\sqrt{n}$*-consistent estimator of the projection onto* $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$.

3. New iterative algorithm

In this section we describe a re-parameterized version of $L_u(\mathbf{G})$ that does not require optimization over a Grassmannian. The new parameterization requires first selecting u rows of $\mathbf{G} \in \mathbb{R}^{r \times u}$ and then constraining the matrix $\mathbf{G}_1 \in \mathbb{R}^{u \times u}$ formed with these rows to be non-singular. Without loss of generality, assume that $\mathbf{G}_1$ is constructed from the first u rows of $\mathbf{G}$ which we can then partition as

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_u \\ \mathbf{A} \end{pmatrix} \mathbf{G}_1 = \mathbf{C}_{\mathbf{A}} \mathbf{G}_1,$$

where $\mathbf{A} = \mathbf{G}_2 \mathbf{G}_1^{-1} \in \mathbb{R}^{(r-u)\times u}$ is an unconstrained matrix and $\mathbf{C}_{\mathbf{A}} = (\mathbf{I}_u, \mathbf{A}^\top)^\top$. Since $\mathbf{G}^\top \mathbf{G} = \mathbf{I}_u$ and $\mathbf{G}_1$ is non-singular, $\mathbf{G}_1 \mathbf{G}_1^\top = (\mathbf{C}_{\mathbf{A}}^\top \mathbf{C}_{\mathbf{A}})^{-1}$. Using these relationships, $L_u(\mathbf{G})$ can be re-parameterized as a function of only $\mathbf{A}$:

$$L_u(\mathbf{A}) = -2 \ln |\mathbf{C}_{\mathbf{A}}^\top \mathbf{C}_{\mathbf{A}}| + \ln \left| \mathbf{C}_{\mathbf{A}}^\top \widehat{\mathbf{M}} \mathbf{C}_{\mathbf{A}} \right| + \ln \left| \mathbf{C}_{\mathbf{A}}^\top (\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1} \mathbf{C}_{\mathbf{A}} \right| .$$

With this objective function minimization over $\mathbf{A}$ is unconstrained. The number of real parameters $u(r-u)$ comprising $\mathbf{A}$ is the same as the number of reals needed to specify uniquely a u-dimensional subspace of $\mathbb{R}^r$; that is, a single element in the Grassmannian.

If $u(r-u)$ is not too large, $L_u(\mathbf{A})$ might be minimized directly by using standard optimization software and the starting values described in Section 2. In other cases minimization can be carried out by minimizing iteratively over the rows of $\mathbf{A}$. Suppose that we wish to minimize over the last row $\mathbf{a}^\top$ of $\mathbf{A}$. Partition

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{a}^\top \end{pmatrix}, \ \mathbf{C}_{\mathbf{A}} = \begin{pmatrix} \mathbf{C}_{\mathbf{A}_1} \\ \mathbf{a}^\top \end{pmatrix}, \ \widehat{\mathbf{M}} = \begin{pmatrix} \widehat{\mathbf{M}}_{11} & \widehat{\mathbf{M}}_{12} \\ \widehat{\mathbf{M}}_{21} & \widehat{M}_{22} \end{pmatrix}, \ (\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1} = \begin{pmatrix} \widehat{\mathbf{V}}_{11} & \widehat{\mathbf{V}}_{12} \\ \widehat{\mathbf{V}}_{21} & \widehat{V}_{22} \end{pmatrix}.$$

Then after a little algebra, the objective function for minimizing over $\mathbf{a}^\top$ with $\mathbf{A}_1$ held fixed can be written up to terms that do not depend on $\mathbf{a}$ as

$$\begin{aligned} L_u(\mathbf{a} \mid \mathbf{A}_1) &= -2\ln\left\{1+\mathbf{a}^\top(\mathbf{C}_{\mathbf{A}_1}^\top\mathbf{C}_{\mathbf{A}_1})^{-1}\mathbf{a}\right\} \\ &\quad + \ln\left\{1+\widehat{M}_{22}(\mathbf{a}+\widehat{M}_{22}^{-1}\mathbf{C}_{\mathbf{A}_1}^\top\widehat{\mathbf{M}}_{12})^\top\mathbf{W}_1^{-1}(\mathbf{a}+\widehat{M}_{22}^{-1}\mathbf{C}_{\mathbf{A}_1}^\top\widehat{\mathbf{M}}_{12})\right\} \\ &\quad + \ln\left\{1+\widehat{V}_{22}(\mathbf{a}+\widehat{V}_{22}^{-1}\mathbf{C}_{\mathbf{A}_1}^\top\widehat{\mathbf{V}}_{12})^\top\mathbf{W}_2^{-1}(\mathbf{a}+\widehat{V}_{22}^{-1}\mathbf{C}_{\mathbf{A}_1}^\top\widehat{\mathbf{V}}_{12})\right\}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{W}_1 &= \mathbf{C}_{\mathbf{A}_1}^\top\left(\widehat{\mathbf{M}}_{11}-\widehat{M}_{22}^{-1}\widehat{\mathbf{M}}_{12}\widehat{\mathbf{M}}_{21}\right)\mathbf{C}_{\mathbf{A}_1} \\ \mathbf{W}_2 &= \mathbf{C}_{\mathbf{A}_1}^\top\left(\widehat{\mathbf{V}}_{11}-\widehat{V}_{22}^{-1}\widehat{\mathbf{V}}_{12}\widehat{\mathbf{V}}_{21}\right)\mathbf{C}_{\mathbf{A}_1}. \end{aligned}$$

The objective function $L_u(\mathbf{a} \mid \mathbf{A}_1)$, which depends only on logarithms of quadratics in $\mathbf{a}$, can now be minimized using any suitable off-the-shelf algorithm. Iteration then cycles over rows of $\mathbf{A}$ until a convergence criterion is met.

This algorithm requires the starting value $\mathbf{G}_{\text{start}}$ described in Section 2. Prior to application of the algorithm we must identify u rows of $\mathbf{G}_{\text{start}}$ and then constrain the matrix $\mathbf{G}_{\text{start},u}$ formed from those u rows to be non-singular. This implies that the matrix formed from the corresponding rows of a basis matrix for $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ should also be non-singular. This can be achieved asymptotically at rate $\sqrt{n}$ by first applying Gaussian elimination with partial pivoting to $\mathbf{G}_{\text{start}}$. The u rows of $\mathbf{G}_{\text{start}}$ identified during this process then form $\mathbf{G}_{\text{start},u}$.

Proposition 3.1 *Assume that the eigenvalues of* $\mathbf{M}$ *and* $\mathbf{M}+\mathbf{U}$ *are distinct. Then the* $u \times u$

submatrix of $\mathbf{G}_{\text{start}}$ *that consists of the* u *rows selected by Gaussian elimination converges to a non-singular matrix with rate* $\sqrt{n}$*.*

This proposition shows that asymptotically Gaussian elimination produces a non-singular submatrix. The condition that the eigenvalues of $\mathbf{M}$ and $\mathbf{M}+\mathbf{U}$ be distinct is mainly for clarity of exposition and is not necessary. The proof given in the appendix demonstrates a more complete result. Let $\mathbf{\Gamma}_{\text{start}}$ denote the population version of $\mathbf{G}_{\text{start}}$, and let $\mathbf{\Gamma}_{\text{start},u} \in \mathbb{R}^{u\times u}$ consist of the u rows of $\mathbf{\Gamma}_{\text{start}}$ formed by applying Gaussian elimination to $\mathbf{\Gamma}_{\text{start}}$. Then $\mathbf{\Gamma}_{\text{start}}$ is a basis matrix for $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ and $\mathbf{G}_{\text{start},u}$ converges to $\mathbf{\Gamma}_{\text{start},u}$ at rate $\sqrt{n}$.

The new algorithm estimates a basis $\mathbf{\Gamma}$ row by row, while the 1D algorithm optimizes column by column. When u is small, the 1D algorithm tends to be a bit more efficient as it optimizes one column at a time and it needs only one pass through those columns. When u is larger, the new algorithm dominates, and sometimes substantially. In each estimation, the 1D algorithm uses conjugate gradient with Polak-Ribiere updates while our algorithm uses Newton updates.

4. Simulations

4.1. Starting values

The first series of simulations was designed to illustrate why it is important to consider the eigenvalues of both $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}+\widehat{\mathbf{U}}$. All simulations are for response envelopes reviewed in Section 1.1, model (2). The results displayed in the tables of this section are the average over 50 replications in each simulation scenario. The angle $\angle\{\text{span}(\mathbf{A}_1), \text{span}(\mathbf{A}_2)\}$ between the subspaces spanned by columns of the semi-orthogonal basis matrices $\mathbf{A}_1 \in \mathbb{R}^{r\times u}$ and $\mathbf{A}_2 \in \mathbb{R}^{r\times u}$ was computed in degrees as the arc cosine of the smallest absolute singular value of $\mathbf{A}_1^\top \mathbf{A}_2$, and $\widehat{\boldsymbol{\beta}}_{\text{start}} = \mathbf{P}_{\text{start}}\mathbf{B}$, where $\mathbf{P}_{\text{start}}$ is as defined in Proposition 2.4. The starting value is still denoted as $\mathbf{G}_{\text{start}}$ but its definition depends on the simulation. $\widehat{\boldsymbol{\Gamma}} = \arg\min L_u(\mathbf{G})$ was obtained from the new algorithm described in Section 3 using the simulation-specific starting value $\mathbf{G}_{\text{start}}$, and $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}) = \operatorname{span}(\widehat{\boldsymbol{\Gamma}})$.

Scenario I. This simulation was designed to illustrate a regression in which the eigenvalues of $\boldsymbol{\Sigma}$ are close and the signal is strong. We generated the data with $p = r = 100$, $n = 500$ and $u = 20$, taking $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ to be diagonal matrices with diagonal elements generated as independent uniform $(49, 51)$ variates. Elements in $\boldsymbol{\eta}$ were independent uniform $(0, 10)$ variates, $\mathbf{X}$ followed a multivariate normal distribution with mean 0 and covariance matrix $400\mathbf{I}_p$, and the elements of $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ were obtained by standardizing a matrix of independent uniform $(0, 1)$ variates. In this scenario, the eigenvalues of $\boldsymbol{\Sigma}$ are close to each other, but we have a strong signal arising from the distribution of $\mathbf{X}$. Starting values based on the eigenvectors of $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ were expected to perform poorly, while starting values based on $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$ were expected to perform well, as conjectured at the start of Section 2 and confirmed by the results in Table 1.

The overarching conclusion from Table 1 is that the starting values from $\mathbf{S}_{\mathbf{Y}}$ did very well, whether $\widehat{\mathbf{U}}$ was standardized or not, while the starting values from $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ were effectively equivalent to choosing a 20-dimensional subspace at random. Additionally, iteration from the starting value produced essentially no change in the angle, the value of the objec-

	Standardized $\widehat{\mathbf{U}}$		Unstandardized $\widehat{\mathbf{U}}$	
Summary statistic	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\vert\mathbf{X}} = \widehat{\mathbf{M}}$	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\vert\mathbf{X}} = \widehat{\mathbf{M}}$
$\angle\{\mathrm{span}(\mathbf{G}_{\mathrm{start}}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	0.58	89.05	0.58	88.98
$\angle\{\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	0.58	88.58	0.58	88.74
$L_u(\mathbf{G}_{\mathrm{start}})$	-182.10	-13.18	-182.10	-9.94
$L_u(\widehat{\mathbf{\Gamma}})$	-182.10	-21.95	-182.10	-20.01
$\Vert\widehat{\boldsymbol{\beta}}_{\mathrm{start}} - \boldsymbol{\beta}\Vert_2$	0.27	149.58	0.27	136.51
$\Vert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\Vert_2$	0.27	113.02	0.27	101.67

Table 1: Results for Scenario I. The starting value $\mathbf{G}_{\mathrm{start}}$ was constructed from the eigenvectors of the matrices indicated by the headings for columns 2-5.

tive function or the envelope estimator of $\boldsymbol{\beta}$.

Scenario II. We generated data with $p = r = 100$, $n = 500$ and $u = 5$, taking $\mathbf{\Omega} = \mathbf{I}_u$ and $\mathbf{\Omega}_0 = 100\mathbf{I}_{r-u}$. Elements in $\boldsymbol{\eta}$ were independent uniform $(0, 10)$ variates, $\mathbf{X}$ followed multivariate normal distribution with mean 0 and covariance matrix $25\mathbf{I}_p$, and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ was obtained by standardizing an $r \times r$ matrix of independent uniform $(0, 1)$ variates. Since the eigenvalues in $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ are very different and the signal is modest, the results in Table 2 show as expected from the argument given in Section 2 that the starting values based on $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ did much better than those based on $\mathbf{S}_{\mathbf{Y}}$. As in Scenario I, the starting value did very well. Iteration improved the starting value a small amount and scaling had no notable affect.

Scenario III. The intent of this simulation is to demonstrate the importance of scaling $\widehat{\mathbf{U}}$. We generated data with $p = r = 30$, $n = 200$ and $u = 5$, taking $\mathbf{\Omega}$ to be a diagonal matrix with diagonal elements $1.5^1, \ldots, 1.5^u$ and $\mathbf{\Omega}_0$ to be a diagonal matrix with diagonal elements 1.5^{u+1}, ..., 1.5^r. Elements in $\boldsymbol{\eta}$ were generated as independent uniform $(0, 10)$

	Standardized $\widehat{\mathbf{U}}$		Unstandardized $\widehat{\mathbf{U}}$	
Summary statistic	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}} = \widehat{\mathbf{M}}$	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}} = \widehat{\mathbf{M}}$
$\angle\{\mathrm{span}(\mathbf{G}_{\mathrm{start}}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	45.88	3.87	45.88	3.87
$\angle\{\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	36.55	3.78	36.55	3.78
$L_u(\mathbf{G}_{\mathrm{start}})$	-16.19	-30.88	-16.19	-30.88
$L_u(\widehat{\boldsymbol{\Gamma}})$	-20.74	-30.95	-20.74	-30.95
$\|\widehat{\boldsymbol{\beta}}_{\mathrm{start}} - \boldsymbol{\beta}\|_2$	1.93	0.66	1.93	0.66
$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	1.64	0.57	1.64	0.57

Table 2: Results for scenario II. The starting value $\mathbf{G}_{\mathrm{start}}$ was constructed from the eigenvectors of the matrices indicated by the headings for columns 2-5.

variates, $\mathbf{X}$ followed the multivariate normal distribution with mean 0 and covariance matrix $100\mathbf{I}_p$, and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) = \mathbf{I}_r$. We see from the results of Table 3 that standardization performed well and that now iteration improved the starting value considerably. Here and in all other results of this section, the smallest value of $L_u(\mathbf{G}_{\mathrm{start}})$ produced best results.

	Standardized $\widehat{\mathbf{U}}$		Unstandardized $\widehat{\mathbf{U}}$	
Summary statistic	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}} = \widehat{\mathbf{M}}$	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}} = \widehat{\mathbf{M}}$
$\angle\{\mathrm{span}(\mathbf{G}_{\mathrm{start}}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	48.63	16.72	89.35	33.31
$\angle\{\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	17.92	1.54	89.34	22.77
$L_u(\mathbf{G}_{\mathrm{start}})$	-13.43	-35.75	-12.69	-34.09
$L_u(\widehat{\boldsymbol{\Gamma}})$	-32.48	-46.93	-23.26	-44.84
$\|\widehat{\boldsymbol{\beta}}_{\mathrm{start}} - \boldsymbol{\beta}\|_2$	11.82	8.56	20.32	11.13
$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	4.37	0.72	20.17	5.39

Table 3: Results for scenario III. The starting value $\mathbf{G}_{\mathrm{start}}$ was constructed from the eigenvectors of the matrices indicated by the headings for columns 2-5.

Scenario IV. For this simulation we kept the same settings as Scenario III, except that diagonal elements of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ were 1.05^1, …, 1.05^u and 1.05^{u+1}, …, 1.05^r, and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was generated by standardizing a matrix of uniform $(0, 1)$ random variables. In this setup heteroscedasticity across the elements is reduced substantially from that in Scenario III. As indicated in Table 4, the standardization no longer provides much improvement. Also, since the eigenvalues of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ are similar, $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ again does not work well.

	Standardized $\widehat{\mathbf{U}}$		Unstandardized $\widehat{\mathbf{U}}$	
Summary statistic	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\vert\mathbf{X}} = \widehat{\mathbf{M}}$	$\mathbf{S}_{\mathbf{Y}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}\vert\mathbf{X}} = \widehat{\mathbf{M}}$
$\angle\{\mathrm{span}(\mathbf{G}_{\mathrm{start}}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	0.30	79.57	0.30	80.66
$\angle\{\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U}), \mathcal{E}_{\mathbf{M}}(\mathcal{U})\}$	0.30	73.40	0.30	75.58
$L_u(\mathbf{G}_{\mathrm{start}})$	-53.54	-7.92	-53.54	-7.27
$L_u(\widehat{\mathbf{\Gamma}})$	-53.54	-13.40	-53.54	-12.56
$\Vert\widehat{\boldsymbol{\beta}}_{\mathrm{start}} - \boldsymbol{\beta}\Vert_2$	0.08	33.36	0.08	31.59
$\Vert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\Vert_2$	0.08	25.04	0.08	22.61

Table 4: Results for scenario IV. The starting value $\mathbf{G}_{\mathrm{start}}$ was constructed from the eigenvectors of the matrices indicated by the headings for columns 2-5.

4.2. Comparisons with the 1D algorithm

In this section we give three different simulation scenarios based on response envelopes for comparing the new non-Grassmann algorithm with the 1D algorithm. In all scenarios $p = 100$, $\boldsymbol{\alpha} = 0$, orthogonal bases $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ were obtained by normalizing an $r \times r$ matrix of independent uniform $(0, 1)$ variates, the elements in $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ were generated as independent uniform $(0, 10)$ variates, and $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$. The predictors $\mathbf{X}$ were generated as independent normal random vectors with mean 0 and variance $400\mathbf{I}_r$. We varied u from 1 to 90 and recorded and computing times and the angles between the true and estimated subspaces.

The 1D algorithm was implemented in R for all simulations reported in this and the next section. Using efficient programming tools in R, it is now much faster than its Matlab version, which produced the results in Cook and Zhang [7]. To insure a fair comparison, we used the default convergence criterion in R for optimizations within both the 1D algorithm and the new algorithm. The angle between subspaces was computed as described previously. In all case the results tabled are the averages over 50 replications. We use $\widehat{\mathbf{\Gamma}}_{\mathrm{1D}}$ to denote the basis generated by the the 1D algorithm.

Scenario V. In this scenario we set $r = 100$ and $n = 250$. To reflect multivariate regressions with large immaterial variation, so envelopes give large gains, we generated the error covariance matrix as $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^\top + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^\top$, where $\mathbf{\Omega} = \mathbf{A}\mathbf{A}^\top$, $\mathbf{\Omega}_0 = \mathbf{C}\mathbf{C}^\top$, the elements in $\mathbf{A}$ were generated as independent standard normal variates and elements in $\mathbf{C}$ were generated as independent normal $(0, 5^2)$ variates. The results are shown in Table 5.

The 1D algorithm tends to perform a bit better on accuracy (Table 5) for small values of u, while performing poorly for large values of u. The same phenomenon occurs in terms of time: the 1D algorithm tends to be a bit faster for small values of u, but otherwise can take much longer than the new non-Grassmann algorithm. The relatively small times for the new algorithm at $u = 5, 10, 20, 60$ occurred because in those cases the starting value was quite good and little iteration was required. The same qualitative differences hold when considering the norm between the estimated coefficient matrix and the true value from the simulation. Note also that the angle for the starting value by itself was often smaller than that for the 1D algorithm.

Scenario VI. We again set $r = 100$ and $n = 250$. To reflect multivariate regressions with small immaterial variation, so envelopes give worthwhile but relatively modest gains,

(A) Angle	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\mathrm{span}(\mathbf{G}_{\mathrm{start}})$	$\mathrm{span}(\widehat{\mathbf{\Gamma}}_{\mathrm{1D}})$
$u=1$	0.92	1.68	0.64
$u=5$	3.56	3.60	1.81
$u=10$	4.67	4.73	4.60
$u=20$	5.83	5.84	42.77
$u=30$	4.84	6.07	12.37
$u=40$	5.59	7.39	6.24
$u=50$	6.81	7.62	39.57
$u=60$	8.48	8.49	70.37
$u=80$	7.61	10.01	25.51
$u=90$	7.15	12.04	21.02

(B) Time	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\mathrm{span}(\mathbf{G}_{\mathrm{start}})$	$\mathrm{span}(\widehat{\mathbf{\Gamma}}_{\mathrm{1D}})$
$u=1$	2.30	0.03	0.23
$u=5$	0.19	0.03	1.45
$u=10$	0.37	0.03	2.71
$u=20$	0.34	0.03	5.16
$u=30$	7.49	0.04	6.23
$u=40$	7.58	0.04	7.30
$u=50$	5.53	0.05	9.18
$u=60$	0.97	0.05	10.59
$u=80$	2.21	0.07	11.07
$u=90$	1.55	0.08	10.40

Table 5: Scenario V: (A) Angle between $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ and the indicated subspace. (B) Computing time in seconds for the indicated subspace. $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$, $\mathrm{span}(\mathbf{G}_{\mathrm{start}})$ and $\mathrm{span}(\widehat{\mathbf{\Gamma}}_{\mathrm{1D}})$ denote the estimated subspaces by the new non Grassmann algorithm, the starting values described in Section 2 and the 1D algorithm.

we generated the error covariance matrix as $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^{\top} + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^{\top}$, where $\mathbf{\Omega} = \mathbf{A}\mathbf{A}^{\top}$, $\mathbf{\Omega}_0 = \mathbf{C}\mathbf{C}^{\top}$, the elements in $\mathbf{A}$ were generated as independent normal $(0, 5^2)$ variates variates and elements in $\mathbf{C}$ were generated as independent standard normal variates. The results shown in Table 6 broadly parallel those in Table 5 for Scenario V, but now the performance of the new algorithm is stronger, both in terms of accuracy and time.

(A) Angle	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\mathrm{span}(\mathbf{G}_{\mathrm{start}})$	$\mathrm{span}(\widehat{\boldsymbol{\Gamma}}_{\mathrm{1D}})$
$u = 1$	0.32	0.32	0.33
$u = 5$	0.75	0.75	0.70
$u = 10$	0.89	0.89	2.94
$u = 20$	1.13	1.14	21.00
$u = 30$	1.24	1.24	10.73
$u = 40$	1.36	1.36	12.68
$u = 50$	1.40	1.40	16.97
$u = 60$	1.45	1.45	31.20
$u = 80$	1.64	1.64	6.67
$u = 90$	1.14	1.14	4.10
(B) Time	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\mathrm{span}(\mathbf{G}_{\mathrm{start}})$	$\mathrm{span}(\widehat{\boldsymbol{\Gamma}}_{\mathrm{1D}})$
$u = 1$	0.08	0.04	0.30
$u = 5$	0.10	0.03	1.13
$u = 10$	0.18	0.03	2.52
$u = 20$	0.29	0.04	3.82
$u = 30$	0.42	0.04	6.42
$u = 40$	0.71	0.04	7.71
$u = 50$	0.38	0.05	9.74
$u = 60$	0.31	0.06	10.62
$u = 80$	0.61	0.07	11.69
$u = 90$	0.21	0.09	11.00

Table 6: Scenario VI: (A) Angle between $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$ and the indicated subspace. (B) Computing time in seconds for the indicated subspace. $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$, $\mathrm{span}(\mathbf{G}_{\mathrm{start}})$ and $\mathrm{span}(\widehat{\boldsymbol{\Gamma}}_{\mathrm{1D}})$ denote the estimated subspaces by the new non Grassmann algorithm, the starting values described in Section 2 and the 1D algorithm.

Scenario VII. This scenario was designed to emphasize the time differences between the 1D algorithm and the non Grassmann algorithm. We set $n = 500$ and varied r from 150 to 350. The error covariance matrix was constructed as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\top}$, where $\boldsymbol{\Omega} = \mathbf{I}$, $\boldsymbol{\Omega}_0 = 25\mathbf{I}$. The estimative performance of the two algorithms was essentially the same in this scenario, with the angles between the estimated subspaces and the envelope varying between about 0.3 degrees for $(u, r) = (1, 150)$ and 10 degrees for $(u, r) = (90, 350)$. However, as shown in Table 7 the 1D algorithm can take considerably longer than the non Grassmann algorithm. To emphasize the differences, the 1D algorithm with $r = 350$ would take about 2.5 hours to estimate the envelope for each u between 1 and 90, while the non Grassmann algorithm would take only about 0.15 hours. In practice we would normally need to estimate the envelope for each u between 1 and 350, leading to much longer computing times.

	$r = 150$		$r = 250$		$r = 350$	
	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\text{span}(\widehat{\mathbf{\Gamma}}_{1\text{D}})$	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\text{span}(\widehat{\mathbf{\Gamma}}_{1\text{D}})$	$\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$	$\text{span}(\widehat{\mathbf{\Gamma}}_{1\text{D}})$
$u = 1$	0.16	0.18	0.45	0.64	0.96	1.65
$u = 5$	0.21	0.85	0.54	3.30	1.08	8.23
$u = 10$	0.28	2.26	0.64	8.31	1.22	20.89
$u = 20$	0.42	6.16	0.94	23.4	1.64	51.61
$u = 30$	0.62	9.56	1.33	37.00	2.18	81.46
$u = 40$	0.86	12.94	1.83	48.45	2.86	110.06
$u = 50$	1.09	16.03	2.38	59.20	3.73	135.05
$u = 60$	1.40	18.62	3.13	68.23	4.85	157.65
$u = 80$	2.08	22.50	9.77	87.08	15.07	196.84
$u = 90$	2.50	23.91	11.74	91.20	27.97	212.87

Table 7: Scenario VII. Computing time in seconds for the indicated subspace. $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathcal{U})$ and $\text{span}(\widehat{\mathbf{\Gamma}}_{1\text{D}})$ denote the subspaces by the new non Grassmann algorithm and the 1D algorithm.

5. Contrasts on real data

In this section we compare the computing time for the new non Grassmann algorithm and the 1D algorithm to select an envelope dimension by minimizing prediction errors from five fold cross validation, the method typically used in conjunction with the 1D algorithm. The time reported is, for each u, the total optimization time over 250 optimizations comprised of 50 replications of five fold cross validation.

5.1. Alzheimer data

The Alzheimer data contains volumes of $r = 93$ regions of the brain from each of 749 Alzheimer patients (Zhu et al. [12]). We used gender, age, the logarithm of intracerebroventricular volume, and interactions involving gender as predictors, so $p = 5$. After taking the logarithms of all brain volumes, we fitted the response envelope model using both the new algorithm and the 1D algorithm. There was little to distinguish the methods based on predictive performance, but the time differences are clear, as displayed in Figure 1. As we observed in the simulations, the times for the two algorithms are close for relatively small values of u and diverge for larger values of u. The total optimization time over all $250 \times r = 23,250$ optimizations was about 22 hours for the new algorithm and 60 hours for the 1D algorithm. The overall computation time is relatively large because the signal in the data is somewhat weak.

5.2. Glass data

Our algorithm is applicable in many envelope contexts other than response envelopes. We used predictor envelopes (Cook et al. [1]) for this illustration.

The dataset contains measurements of the chemical composition and electron-probe-X-ray microanalysis for 180 archeological glass vessels from 15th to 17th century excavated

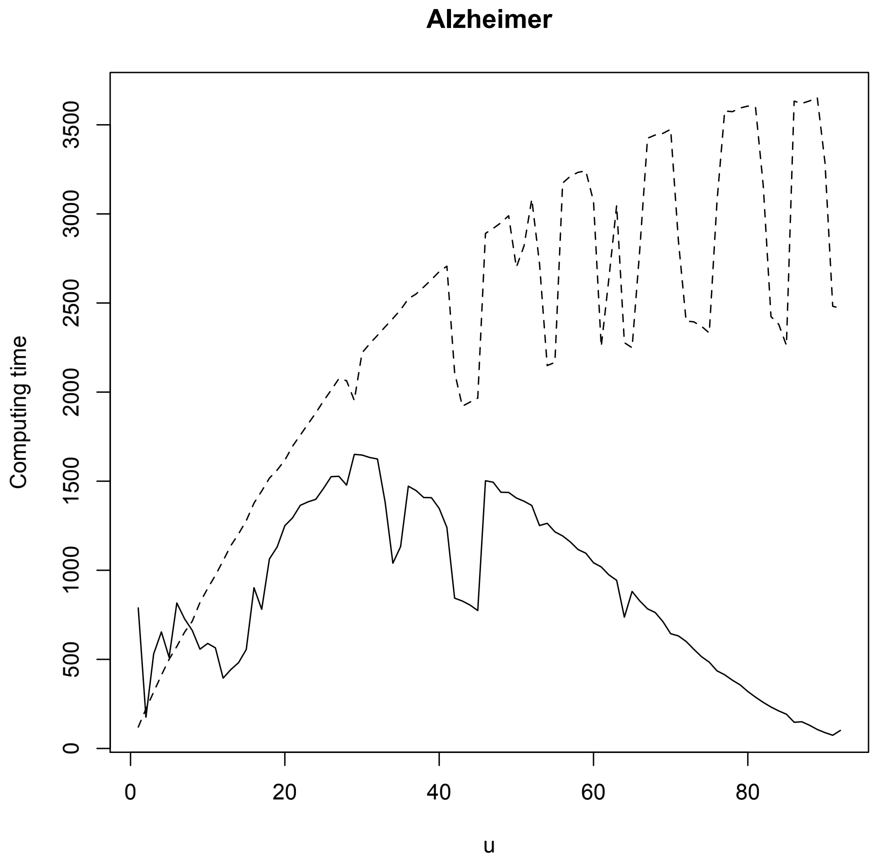

Figure 1: Alzheimer data: for each u the vertical axis is the total optimization time over 250 optimizations comprised of 50 replications of five fold cross validation. The solid line marks the new non Grassmann algorithm and the dashed line marks 1D algorithm.

in Antwerp, Belgium. For each vessel, a spectrum on a set of equispaced frequencies between 1 and 1920 is measured. Since the values below 100 and above 400 are almost null, following Kudraszow and Maronna [9], we chose 13 equispaced frequencies between 100 and 400 as predictors. The response variable is the amount of sulfur trioxide. For each $u = 1, \dots, 13$, we ran the 1D algorithm and the new algorithm, recoding the prediction error from 50 replications of five fold cross validation and the average computing time for these 250 optimizations. The new algorithm gave a four percent improvement in prediction error over the 1D algorithm at $u = 3$, which was best for both methods. As in the Alzheimer data, there were clear differences in computing time, as shown in Figure 2. The total time for computing all u was 86 seconds for the new algorithm and 541 seconds for the 1D algorithm.

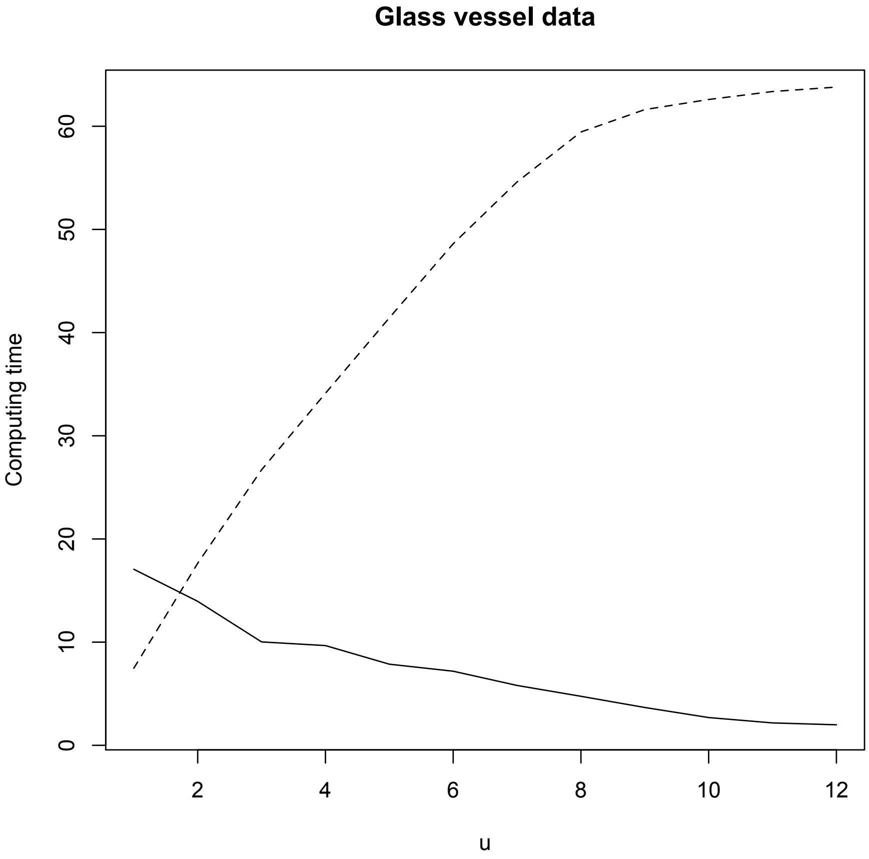

Figure 2: The solid line marks the new non Grassmann algorithm and the dashed line marks 1D algorithm.

Acknowledgements

The authors thank Hongtu Zhu for providing the Alzheimer data and Nadia L. Kudraszow for providing the glass data. The authors also thank the Editor and two Associate Editors for an incredibly fast review, faster than we have ever experienced before. Research for this article was supported in part by grant DMS-1407460 from the National Science Foundation.

References

[1] Cook, R.D., Helland, I. S. and Su, Z. (2013), Envelopes and partial least squares regression. *Journal of the Royal Statistical Society B* **75**, 851–877.

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

[3] Cook, R.D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, **20**, 927–1010.

[4] Cook, R. D., Su, Z. and Yang, Y. (2014). A MATLAB toolbox for computing envelope estimators in multivariate analyses. *Journal of Statistical Software* **62**, http://www.jstatsoft.org/v62/i08/paper.

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

[6] Cook, R.D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics* **57**, 11–25.

[7] Cook, R.D. and Zhang, X. (2015c). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, to appear (http://arxiv.org/pdf/1403.4138.pdf)

[8] Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications* **20**, 303 – 353.

[9] Kudraszow, N. L. and Maronna, R. A. (2011). Estimates of MM type for the multivariate linear model. *Journal of Multivariate Analysis*, **102** 1280–1292.

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

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

[12] Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. (2014). Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers. *Journal of the American Statistical Association*, **109**, 977–990.

Appendix

A. Proof of Proposition 2.1

Let $(\mathbf{G}, \mathbf{G}_0) \in \mathbb{R}^{r\times r}$ be a column partitioned orthogonal matrix and let $\mathbf{M} \in \mathbb{S}^{r\times r}$ be positive definite. The conclusion that $\ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}_0\mathbf{M}\mathbf{G}_0|$ is minimized when $\mathrm{span}(\mathbf{G})$ is any u-dimensional reducing subspace of $\mathbf{M}$ will follow by showing that $|\mathbf{M}| \leq |\mathbf{G}^\top\mathbf{M}\mathbf{G}| \times |\mathbf{G}_0^\top\mathbf{M}\mathbf{G}_0|$ with equality if and only if $\mathrm{span}(\mathbf{G})$ reduces $\mathbf{M}$.

$$\begin{aligned}
|\mathbf{M}| &= |(\mathbf{G},\mathbf{G}_0)^\top\mathbf{M}(\mathbf{G},\mathbf{G}_0)| = \begin{vmatrix} \mathbf{G}^\top\mathbf{M}\mathbf{G} & \mathbf{G}^\top\mathbf{M}\mathbf{G}_0 \\ \mathbf{G}_0^\top\mathbf{M}\mathbf{G} & \mathbf{G}_0^\top\mathbf{M}\mathbf{G}_0 \end{vmatrix} \\
&= |\mathbf{G}^\top\mathbf{M}\mathbf{G}| \times |\mathbf{G}_0^\top\mathbf{M}\mathbf{G}_0 - \mathbf{G}_0^\top\mathbf{M}\mathbf{G}(\mathbf{G}^\top\mathbf{M}\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{M}\mathbf{G}_0| \\
&\leq |\mathbf{G}^\top\mathbf{M}\mathbf{G}| \times |\mathbf{G}_0^\top\mathbf{M}\mathbf{G}_0|,
\end{aligned}$$

with equality if and only if $\mathbf{G}_0^\top\mathbf{M}\mathbf{G} = 0$, which is equivalent to requiring that $\mathrm{span}(\mathbf{G})$ reduce $\mathbf{M}$.

The conclusion that $\ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}\mathbf{M}^{-1}\mathbf{G}|$ is also minimized when $\mathrm{span}(\mathbf{G})$ is any u-dimensional reducing subspace of $\mathbf{M}$ follows because

$$\ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}\mathbf{M}^{-1}\mathbf{G}| = \ln|\mathbf{G}^\top\mathbf{M}\mathbf{G}| + \ln|\mathbf{G}_0\mathbf{M}\mathbf{G}_0| - \ln|\mathbf{M}|.$$

B. Proof of Proposition 2.2

Recall that $J_1(\mathbf{G}) = \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0|$, $J_2(\mathbf{G}) = \ln|\mathbf{I}_{r-u} + \mathbf{G}_0^\top\widehat{\mathbf{U}}_{\mathbf{M}}\mathbf{G}_0|$ and $J(\mathbf{G}) = J_1(\mathbf{G}) + J_2(\mathbf{G})$, where $\widehat{\mathbf{U}}_{\mathbf{M}} = \widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{M}}^{-1/2}$ is a standardized version of $\widehat{\mathbf{U}}$. Then from Proposition 2.1, an argument minimizes $L_u(\mathbf{G})$ if and only if it minimizes

$$\begin{aligned}
f(\mathbf{G}) &= \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})\mathbf{G}_0| \\
&= \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top(\widehat{\mathbf{M}} + \widehat{\mathbf{u}}\widehat{\mathbf{u}}^\top)\mathbf{G}_0| \\
&= \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0| + \ln|\mathbf{I}_k + \widehat{\mathbf{u}}^\top\mathbf{G}_0(\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1}\mathbf{G}_0^\top\widehat{\mathbf{u}}| \\
&= \ln|\mathbf{G}^\top\widehat{\mathbf{M}}\mathbf{G}| + \ln|\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0| \\
&\quad + \ln|\mathbf{I}_{r-u} + (\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}\mathbf{G}_0^\top\widehat{\mathbf{u}}\widehat{\mathbf{u}}^\top\mathbf{G}_0(\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}| \\
&= J_1(\mathbf{G}) + f_2(\mathbf{G}),
\end{aligned}$$

where f_2 is defined implicitly and $\widehat{\mathbf{U}} = \widehat{\mathbf{u}}\widehat{\mathbf{u}}^\top$ is a decomposition of $\widehat{\mathbf{U}}$ with $\widehat{\mathbf{u}} \in \mathbb{R}^{r\times k}$. To see that $f_2 = J_2$ over $\mathcal{V}_u$ we have

$$\begin{aligned}
f_2(\mathbf{G}) &= \ln|\mathbf{I}_{r-u} + (\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}\mathbf{G}_0^\top\widehat{\mathbf{U}}\mathbf{G}_0(\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}| \\
&= \ln|\mathbf{I}_{r-u} + (\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}\mathbf{G}_0^\top\widehat{\mathbf{M}}^{1/2}(\widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{M}}^{-1/2})\widehat{\mathbf{M}}^{1/2}\mathbf{G}_0(\mathbf{G}_0^\top\widehat{\mathbf{M}}\mathbf{G}_0)^{-1/2}| \\
&= \ln|\mathbf{I}_{r-u} + \mathbf{G}_0^\top(\widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{M}}^{-1/2})\mathbf{G}_0| \\
&= \ln|\mathbf{I}_{r-u} + \mathbf{G}_0^\top\widehat{\mathbf{U}}_{\mathbf{M}}\mathbf{G}_0| \\
&= J_2(\mathbf{G}),
\end{aligned}$$

where the third equality follows because $\mathbf{G}_0 \in \mathcal{V}_u$ reduces $\widehat{\mathbf{M}}$.

C. Proof of Proposition 2.3

Let $\widehat{\mathbf{W}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ for notational convenience and start with the objective function

$$\begin{aligned}
L_u(\mathbf{G}) &= \ln|\mathbf{G}^\top \widehat{\mathbf{M}} \mathbf{G}| + \ln|\mathbf{G}^\top (\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1} \mathbf{G}| \\
&= \ln|\mathbf{G}^\top \widehat{\mathbf{M}} \mathbf{G}| + \ln|\mathbf{G}^\top \widehat{\mathbf{W}}^{-1} \mathbf{G}| \\
&= \ln|\mathbf{G}^\top (\widehat{\mathbf{M}} + \widehat{\mathbf{U}}) \mathbf{G} - \mathbf{G}^\top \widehat{\mathbf{U}} \mathbf{G}| + \ln|\mathbf{G}^\top \widehat{\mathbf{W}}^{-1} \mathbf{G}| \\
&= \ln|\mathbf{G}^\top \widehat{\mathbf{W}} \mathbf{G} - \mathbf{G}^\top \widehat{\mathbf{u}} \widehat{\mathbf{u}}^\top \mathbf{G}| + \ln|\mathbf{G}^\top \widehat{\mathbf{W}}^{-1} \mathbf{G}| \\
&= \ln|\mathbf{G}^\top \widehat{\mathbf{W}} \mathbf{G}| + \ln|\mathbf{I}_k - \widehat{\mathbf{u}}^\top \mathbf{G} (\mathbf{G}^\top \widehat{\mathbf{W}} \mathbf{G})^{-1} \mathbf{G}^\top \widehat{\mathbf{u}}| + \ln|\mathbf{G}^\top \widehat{\mathbf{W}}^{-1} \mathbf{G}|,
\end{aligned}$$

where $\widehat{\mathbf{u}}$ is as defined in the proof of Proposition 2.2. The sum of the first and last terms on the right side of this representation is always non-negative and equals 0, its minimum value, when the columns of $\mathbf{G}$ span any reducing subspace of $\widehat{\mathbf{W}} = \widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. Restricting $\mathbf{G}$ in this way,

$$\begin{aligned}
\widehat{\mathbf{u}}^\top \mathbf{G} (\mathbf{G}^\top \widehat{\mathbf{W}} \mathbf{G})^{-1} \mathbf{G}^\top \widehat{\mathbf{u}} &= \widehat{\mathbf{u}}^\top \widehat{\mathbf{W}}^{-1/2} \widehat{\mathbf{W}}^{1/2} \mathbf{G} (\mathbf{G}^\top \widehat{\mathbf{W}} \mathbf{G})^{-1} \mathbf{G}^\top \widehat{\mathbf{W}}^{1/2} \widehat{\mathbf{W}}^{-1/2} \widehat{\mathbf{u}} \\
&= \widehat{\mathbf{u}}^\top \widehat{\mathbf{W}}^{-1/2} \mathbf{G} \mathbf{G}^\top \widehat{\mathbf{W}}^{-1/2} \widehat{\mathbf{u}},
\end{aligned}$$

and the middle term of $L(\mathbf{G})$ reduces to

$$\begin{aligned}\ln|\mathbf{I}-\widehat{\mathbf{u}}^{\top}\mathbf{G}\{\mathbf{G}^{\top}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\mathbf{G}\}^{-1}\mathbf{G}^{\top}\widehat{\mathbf{u}}| &= \ln|\mathbf{I}_k-\widehat{\mathbf{u}}^{\top}\widehat{\mathbf{W}}^{-1/2}\mathbf{G}\mathbf{G}^{\top}\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{u}}| \\ &= \ln|\mathbf{I}_u-\mathbf{G}^{\top}\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{u}}\widehat{\mathbf{u}}^{\top}\widehat{\mathbf{W}}^{-1/2}\mathbf{G}| \\ &= \ln|\mathbf{I}_u-\mathbf{G}^{\top}\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}\mathbf{G}|,\end{aligned}$$

where $\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}} = \widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{u}}\widehat{\mathbf{u}}^{\top}\widehat{\mathbf{W}}^{-1/2} = \widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{W}}^{-1/2}$ is $\widehat{\mathbf{U}}$ standardized by $\widehat{\mathbf{W}}^{-1/2} = (\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1/2}$.

D. Proof of Proposition 2.4

We demonstrate the result in detail for $K_{\mathbf{M}}$. The corresponding result for the other three K functions follows similarly.

Recall that $K_{\mathbf{M}}(\mathbf{G}) = \sum_{i=1}^{u}(\mathbf{g}_i^{\top}\widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{U}}\widehat{\mathbf{M}}^{-1/2}\mathbf{g}_i)$ where $\mathbf{g}_i$ is an eigenvector of $\widehat{\mathbf{M}}$. The population version of this objective function is

$$\tilde{K}_{\mathbf{M}}(\tilde{\mathbf{G}}) = \sum_{i=1}^{u}\tilde{\mathbf{g}}_i^{\top}\mathbf{M}^{-1/2}\mathbf{U}\mathbf{M}^{-1/2}\tilde{\mathbf{g}}_i$$

where $\tilde{\mathbf{g}}$ is an eigenvector of $\mathbf{M}$ and $\tilde{\mathbf{G}} = (\tilde{\mathbf{g}}_1,\ldots,\tilde{\mathbf{g}}_u)$. We next show that

$$\mathrm{span}\left\{\arg\max \tilde{K}_{\mathbf{M}}(\tilde{\mathbf{G}})\right\} = \mathcal{E}_{\mathbf{M}}(\mathcal{U}).$$

Consider a generic envelope $\mathcal{E}_{\mathbf{A}}(\mathcal{S})$, where $\mathbf{A} > 0$ with eigenspaces $\mathcal{A}_i$, $i = 1,\ldots,q$.

Cook et al. [3] show that this envelope can be characterizes as $\mathcal{E}_{\mathbf{A}}(\mathcal{S}) = \sum_{i=1}^{q} \mathbf{P}_{\mathcal{A}_i}\mathcal{S}$. As a consequence there are $u = \dim\{\mathcal{E}_{\mathbf{A}}(\mathcal{S})\}$ orthogonal eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_u$ of $\mathbf{A}$ so that

$$\mathcal{E}_{\mathbf{A}}(\mathcal{S}) = \sum_{i=1}^{u} \mathbf{P}_{\mathbf{a}_i}\mathcal{S} = \operatorname{span}\left(\sum_{i=1}^{u} \mathbf{P}_{\mathbf{a}_i}\mathbf{s}\mathbf{s}^\top \mathbf{P}_{\mathbf{a}_i}\right)$$

where s is a basis matrix for $\mathcal{S}$. By definition of $\mathcal{E}_{\mathbf{A}}(\mathcal{S})$, there exists exactly u eigenvectors of $\mathbf{A}$ that are not orthogonal to $\mathcal{S}$ and these eigenvectors are $\mathbf{a}_1, \ldots, \mathbf{a}_u$. Consequently, we must have

$$\mathcal{E}_{\mathbf{A}}(\mathcal{S}) = \operatorname{span}\left\{\arg\max\ \operatorname{tr}\left(\sum_{i=1}^{u} \mathbf{P}_{\mathbf{v}_i}\mathbf{s}\mathbf{s}^\top \mathbf{P}_{\mathbf{v}_i}\right)\right\} = \operatorname{span}\left(\arg\max \sum_{i=1}^{u} \mathbf{v}_i^\top \mathbf{s}\mathbf{s}^\top \mathbf{v}_i\right),$$

where the maximum is taken over the eigenvectors $\mathbf{v}_i$ of $\mathbf{A}$. Equality holds since the maximum must select u eigenvectors of $\mathbf{A}$ that are not orthogonal to $\mathrm{ss}^\top$.

Comparing this general argument with $\tilde{K}_{\mathbf{M}}(\mathbf{G})$ we see that $\arg\max \tilde{K}_{\mathbf{M}}$ will select u eigenvectors of $\mathbf{M}$ that are not orthogonal to $\mathbf{M}^{-1/2}\mathbf{U}\mathbf{M}^{-1/2}$ and consequently

$$\operatorname{span}\left\{\arg\max \tilde{K}_{\mathbf{M}}(\tilde{\mathbf{G}})\right\} = \mathcal{E}_{\mathbf{M}}\{\operatorname{span}(\mathbf{M}^{-1/2}\mathbf{U}\mathbf{M}^{-1/2})\} = \mathcal{E}_{\mathbf{M}}(\mathcal{U}),$$

where the final equality follows from Cook et al. ([3], Prop. 2.4).

The $\sqrt{n}$ consistency now follows straightforwardly since the matrices involved in the determination of the four potential starting values – $\widehat{\mathbf{M}}$, $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$, $\widehat{\mathbf{U}}_{\mathbf{M}}$ and $\widehat{\mathbf{U}}_{\mathbf{M}+\mathbf{U}}$ – are all $\sqrt{n}$-consistent estimators of their corresponding population versions.

E. Proof of Proposition 3.1

Let $\mathbf{\Gamma}_{\text{start}} \in \mathbb{R}^{r\times u}$ denote the population counterpart of $\mathbf{G}_{\text{start}}$. Based on previous discussion, the columns of $\mathbf{\Gamma}_{\text{start}}$ are eigenvectors of $\mathbf{M}$ or $\mathbf{M}+\mathbf{U}$. Since $\text{rank}(\mathbf{\Gamma}_{\text{start}}) = u$ we can find u linearly independent rows of $\mathbf{\Gamma}_{\text{start}}$ and, letting $\mathbf{\Gamma}_u$ denote the $u \times u$ matrix forms by these u rows, we get $|\mathbf{\Gamma}_u| \neq 0$. Now, let $\mathbf{G}_u$ denote the submatrix of $\mathbf{G}_{\text{start}}$ forms by these same u rows. It follows straightforwardly in the manner of Proposition 2.4 that $\mathbf{G}_u$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Gamma}_u$. Since the determinant is a continuous function this implies that for n sufficiently large $|\mathbf{G}_u| \neq 0$ with a specified high probability. As a consequence, for n sufficiently large, $\text{rank}(\mathbf{G}_{\text{start}}) = u$ with arbitrarily high probability.

Perform Gaussian elimination with partial pivoting on $\mathbf{G}_{\text{start}}$ and denote the resulting $u \times u$ submatrix by $\mathbf{G}_{\text{start},u}$. From the preceding discussion, $\mathbf{G}_{\text{start},u}$ is nonsingular with high probability for sufficiently large n. Also, perform Gaussian elimination with partial pivoting to $\mathbf{\Gamma}_{\text{start}}$ and denote the resulting nonsingular $u \times u$ submatrix by $\mathbf{\Gamma}_{\text{start},u}$. The proposition is then established if $\mathbf{G}_{\text{start},u}$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Gamma}_{\text{start},u}$.

First we assume that the pivot elements for $\mathbf{\Gamma}_{\text{start}}$ are unique and occur in rows r_i, $i = 1, \ldots, u$. In the first step of Gaussian elimination, for an arbitrary $\epsilon > 0$, we can find an N_1 such that when $n > N_1$, the corresponding element in row r_1 of $\mathbf{G}_{\text{start}}$ is the one having the largest absolute value with probability at least $1 - \epsilon$. In other words, row r_1 will be selected in $\mathbf{G}_{\text{start}}$ with probability at least $1 - \epsilon$. We call the resulting matrices $\mathbf{\Gamma}_{\text{start},1} \in \mathbb{R}^{r\times u}$ and $\mathbf{G}_{\text{start},1} \in \mathbb{R}^{r\times u}$. As Gaussian elimination involves only simple arithmetic operations, $\mathbf{G}_{\text{start},1}$ converges to $\mathbf{\Gamma}_{\text{start},1}$ at rate $\sqrt{n}$. Now, for the second step in Gaussian elimination, we do partial pivoting in the second columns of $\mathbf{G}_{\text{start},1}$ and $\mathbf{\Gamma}_{\text{start},1}$. Then, for an arbitrary $\epsilon > 0$, we can find an $N_2 > N_1$ such that when $n > N_2$, the elements chosen for $\mathbf{G}_{\text{start},1}$ and $\mathbf{\Gamma}_{\text{start},1}$ will be the same with probability at least $(1-\epsilon)$.

Continuing this process, for $n > N_u$, rows $r_1, \ldots, r_u$ in $\mathbf{G}_{\text{start}}$ are selected with probability at least $(1-\epsilon)^u$. Let $\|\cdot\|$ denote some matrix norm. As $\mathbf{G}_{\text{start}}$ converges to $\mathbf{\Gamma}_{\text{start}}$ with rate $\sqrt{n}$, we have $\|\mathbf{G}_{\text{start},u} - \mathbf{\Gamma}_{\text{start},u}\| = O_p(n^{-1/2})$ and consequently for any $\epsilon > 0$ there exists $K > 0$ and N_0 so that for all $n > N_0$,

$$\text{pr}\left(\sqrt{n}\|\mathbf{G}_{\text{start},u} - \mathbf{\Gamma}_{\text{start},u}\| > K \,\middle|\, \text{rows } r_1, \ldots r_u \text{ are selected}\right) < \epsilon.$$

Then with $n > \max(N_0, N_u)$,

$$\begin{aligned} & \text{pr}\left(\sqrt{n}\|\mathbf{G}_{\text{start},u} - \mathbf{\Gamma}_{\text{start},u}\| > K\right) \\ < \; & \text{pr}\left(\sqrt{n}\|\mathbf{G}_{\text{start},u} - \mathbf{\Gamma}_{\text{start},u}\| > K \,\middle|\, \text{rows } r_1, \ldots r_u \text{ are selected}\right) * \text{pr}(\text{rows } r_1, \ldots r_u \text{ are selected}) \\ & + \text{pr}(\text{not all rows } r_1, \ldots r_u \text{ are selected}) \\ < \; & \epsilon + [1 - (1-\epsilon)^u]. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary and $\epsilon + \{1 - (1-\epsilon)^u\}$ tends to 0 as ϵ tends to 0, $\mathbf{G}_{\text{start},u}$ converges to $\mathbf{\Gamma}_{\text{start},u}$ at rate $\sqrt{n}$.

To deal with non-unique pivot elements, assume that there are ties in one column. When we perform Gaussian elimination with partial pivoting on $\mathbf{\Gamma}_{\text{start}}$ in the step with k ties, we can choose whichever of the tied elements, resulting in all the cases in non-singular matrices. We call the resulting matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$. When Gaussian elimination was perform with partial pivoting on $\mathbf{G}_{\text{start}}$, using the preceding reasoning, there will be probability at least $(1-\epsilon)^u/k$ that we pick the rows in $\mathbf{A}_i$, $i = 1, \ldots, k$. Then $\widehat{\mathbf{A}}$ converges to $\mathbf{A}_1, \mathbf{A}_2, \ldots$ or $\mathbf{A}_k$ with rate $\sqrt{n}$, so $\widehat{\mathbf{A}}$ converges to a non-singular matrix with rate $\sqrt{n}$. If we have ties in more than one step we divide further probabilities, since the number of the steps and u are fixed the proof flows similarly.