Envelopes and partial least squares regression[*]

R. D. Cook, University of Minnesota[†]

I. S. Helland, University of Oslo and

Z. Su, University of Florida

December 17, 2012

Abstract

We build connections between envelopes, a recently proposed context for efficient estimation in multivariate statistics, and multivariate partial least squares (PLS) regression. In particular, we establish an envelope as the nucleus of both univariate and multivariate PLS, which opens the door to pursuing the same goals as PLS but using different envelope estimators. It is argued that a likelihood-based envelope estimator is less sensitive to the number of PLS components selected and that it outperforms PLS in prediction and estimation.

1 Introduction

Prediction of a univariate or multivariate response $\mathbf{y} \in \mathbb{R}^r$ from multivariate data $\mathbf{x} \in \mathbb{R}^p$ is at the core of applied statistics, and many different predictive methods have been developed in response to numerous diverse settings encountered across the applied sciences. In this article we address the predictive culture in chemometrics, where partial least squares (PLS) is the dominant method. For chemometricians, who have been mainly responsible for the development of PLS, empirical prediction is a main issue. They tend not to address population PLS models or regression coefficients, but directly the predictions resulting from PLS algorithms. This custom of forgoing population considerations is at odds with statistical tradition. While PLS is known and increasingly used within

[*] AMS 1991 subject classification: Primary 62J05 Secondary 62H12. Keywords and phrases: Dimension reduction; envelopes; envelope models; maximum likelihood estimation; partial least squares; PLS; SIMPLS.

[†] Corresponding author: School of Statistics, 313 Ford Hall, 224 Church St SE, University of Minnesota, Minneapolis, MN 55455. email: dennis@stat.umn.edu

the statistics community, it is perhaps still not widely accepted here because it is based on sample algorithms that have not been cast into a conventional Fisherian framework of well-defined population parameters. But see Helland (1990), where a population model was defined for PLS in the case $r = 1$. This population model was further discussed by Næs and Helland (1993) and by Helland (2001), and a first attempt at maximum likelihood estimation was given by Helland (1992). Martens and Næs (1989) is a classical reference for PLS within the chemometrics community. Frank and Friedman (1993) gave an informative discussion of PLS from various statistical views.

The overarching goal of this article is to show that there is a very close connection between PLS and the recently developed envelopes of Cook, Li and Chiaromonte (2007, 2010). In particular, we show that PLS depends fundamentally on an envelope at the population level and that this envelope can be used as a well-defined parameter that characterizes PLS. The establishment of an envelope as the nucleus of PLS then opens the door to pursuing the same goals as PLS but using different and perhaps better envelope estimators.

While PLS is an integral part of the chemometrics culture, envelope methodology is new and not yet widely recognized in statistics. As shown in past studies (Cook, et al., 2010; Su and Cook, 2011, 2012, 2013) envelope methodology has the potential to achieve substantial efficiency gains in a variety of multivariate problems and thus also has the potential to become wonted methodology. The efficiency gains afforded by envelopes are achieved by a targeted form of dimension reduction that can effectively separate information that is material for the goals at hand from that which is immaterial. The particular advances described in this article can be viewed as another instance of the utility of envelopes in understanding and improving statistical methodology, specifically PLS. As in past studies, it will be seen that the most advantageous envelope methods require numerical optimization over a Grassmann manifold, which is non-standard in statistics but commonplace in other disciplines. A MATLAB toolbox that implements past methods, in addition to the methods described in this article, is available at http://code.google.com/p/envlp/.

We begin in Section 2 by briefly reviewing the relevant algebraic basis for envelopes and establishing the context for our exposition. Since much more is known about univariate PLS ($r = 1$) than multivariate PLS ($r > 1$), we first develop a connection between univariate PLS and envelopes in Section 3, relying primarily on the work of Helland (1988, 1990) for PLS. We present new results in Section 4 to show the role of envelopes in multivariate PLS as implemented in the SIMPLS algorithm (de Jong, 1993). We also discuss two alternative envelope estimators, one based on a multivariate Krylov matrix and one originating from a likelihood-based objective function. It is argued that the likelihood-based estimator will typically provide better predictions than traditional PLS methods. Numerical illustrations are given in Section 5 and a concluding discussion is given in Section 6. Proofs are available in Appendix A. We use group theory in Appendix B to provide a characterization of regressions in which PLS may be most appropriate. To aid intuition and understanding, a little background on how Grassmann optimization algorithms are constructed is provided in Appendix C.

Our exposition makes use of the following notation and conventions. We use $\mathbb{R}^{a\times b}$ to denote the space of real $a \times b$ matrices. $\mathrm{span}(\mathbf{R})$ denotes the subspace spanned by the columns of the matrix $\mathbf{R} \in \mathbb{R}^{a\times b}$. A matrix $\mathbf{R} \in \mathbb{R}^{a\times b}$ with $a > b$ is called semi-orthogonal if its columns are orthogonal and have norm 1, so that $\mathbf{R}^T\mathbf{R} = \mathbf{I}$. For a subspace $\mathcal{R} \subseteq \mathbb{R}^p$ and a matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ we let $\mathbf{M}\mathcal{R}$ denote the space of all vectors $\mathbf{M}\mathbf{x}$ as $\mathbf{x}$ runs through $\mathcal{R}$. The projection onto the subspace $\mathcal{R}$ in the inner product $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{M}\mathbf{y}$ determined by $\mathbf{M}$ is represented as $\mathbf{P}_{\mathcal{R}(\mathbf{M})}$, so that $\mathbf{P}_{\mathcal{R}(\mathbf{M})}\mathbf{z} = \mathbf{R}(\mathbf{R}^T\mathbf{M}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{M}\mathbf{z}$ when $\mathcal{R} = \mathrm{span}(\mathbf{R})$ and the inverse exists. We let $\mathbf{Q}_{\mathcal{R}(\mathbf{M})} = \mathbf{I} - \mathbf{P}_{\mathcal{R}(\mathbf{M})}$. The second subscript '$(\mathbf{M})$' will be suppressed when employing the usual inner product, $\mathbf{M} = \mathbf{I}$, so that $\mathbf{P}_{\mathcal{R}}\mathbf{z} = \mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{z}$ when $\mathbf{R}$ is a basis matrix for $\mathcal{R}$. The orthogonal complement $\mathcal{R}^\perp$ of a subspace $\mathcal{R}$ is with respect to the usual inner product, unless explicitly stated otherwise. We have $\mathbf{P}_{\mathcal{R}^\perp} = \mathbf{Q}_{\mathcal{R}}$. The subspace sum $\mathcal{R}_1 \oplus \mathcal{R}_2$ is the space of all sums $\mathbf{x}_1 + \mathbf{x}_2$ where $\mathbf{x}_1 \in \mathcal{R}_1$ and $\mathbf{x}_2 \in \mathcal{R}_2$.

2 Envelopes

Throughout this article we consider the multivariate linear model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\beta}^T(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^r$, $\boldsymbol{\mu} \in \mathbb{R}^r$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ is non-zero, and the random predictor vector $\mathbf{x}$ has mean $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{x}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$. Independently, $\boldsymbol{\varepsilon}$ is distributed with mean $\mathbf{0}$ and constant covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}$; for emphasis we write this as $\sigma^2_{y|\mathbf{x}}$ when $\mathbf{y}$ is univariate. The data $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, is assumed to consist of independent and identically distributed copies of $(\mathbf{y}, \mathbf{x})$ with finite fourth moments.

Cook, Li and Chiaromonte (2010; hereinafter CLC) introduced the novel idea of an envelope for parsimonious parameterizations of multivariate statistical problems. An expository introduction to the underlying structure of an envelope was given by Su and Cook (2011). Concentrating on reduction in the $\mathbf{y}$-dimension and assuming normal errors and non-stochastic predictors $\mathbf{x}$, CLC demonstrated that the envelope estimator of the coefficient matrix $\boldsymbol{\beta}$ in the multivariate regression model (1) has the potential to produce truly massive gains in efficiency relative to the standard estimator. In contrast, we here consider regressions in which $\mathbf{x}$ is random, focus on reduction in the $\mathbf{x}$-dimension, and do not necessarily assume normal errors.

2.1 Introduction to envelopes for predictor reduction

Our goal, like the usual goal in chemometric applications, is to predict $\mathbf{y} \in \mathbb{R}^r$ from multivariate data $\mathbf{x} \in \mathbb{R}^p$. We make no distributional assumptions, but assume that moments and hence correlations exist. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$ so that (i) $\mathbf{Q}_{\mathcal{S}}\mathbf{x}$ is uncorrelated with $\mathbf{P}_{\mathcal{S}}\mathbf{x}$. While such a subspace may be chosen in many ways, we focus on situations in which it is desirable to base predictions on $\mathbf{P}_{\mathcal{S}}\mathbf{x}$ alone by requiring also that (ii) $\mathbf{y}$ be uncorrelated with $\mathbf{Q}_{\mathcal{S}}\mathbf{x}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{x}$. For any $\mathcal{S}$ with properties (i) and (ii), we say that $\mathbf{Q}_{\mathcal{S}}\mathbf{x}$ is linearly immaterial to the regression since $\mathbf{Q}_{\mathcal{S}}\mathbf{x}$ depends linearly on neither $\mathbf{P}_{\mathcal{S}}\mathbf{x}$ nor $\mathbf{y}$. Consequently, $\mathbf{P}_{\mathcal{S}}\mathbf{x}$ must carry all of the information that is linearly material to the regression; that is, all of the information that is available about $\boldsymbol{\beta}$ from $\mathbf{x}$.

The following proposition connects the statistical conditions (i) and (ii) with equivalent algebraic conditions that lead to the notion of envelopes. These conditions are restated in the proposition for ease of reference.

Proposition 2.1 *Assuming model (1), assertion (i)* $\operatorname{corr}(\mathbf{P}_{\mathcal{S}}\mathbf{x}, \mathbf{Q}_{\mathcal{S}}\mathbf{x}) = 0$ *is equivalent to the algebraic condition (a) both* $\boldsymbol{\Sigma}_{\mathbf{x}}\mathcal{S} \subseteq \mathcal{S}$ *and* $\boldsymbol{\Sigma}_{\mathbf{x}}\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$*. When (a) holds, we say that* $\mathcal{S}$ *is a* reducing subspace *of* $\boldsymbol{\Sigma}_{\mathbf{x}}$*. Assertion (ii)* $\operatorname{corr}(\mathbf{y}, \mathbf{Q}_{\mathcal{S}}\mathbf{x} \mid \mathbf{P}_{\mathcal{S}}\mathbf{x}) = 0$ *is equivalent to the algebraic condition (b)* $\operatorname{span}(\boldsymbol{\beta}) \subseteq \mathcal{S}$*.*

Finally, we want the dimension of $\mathcal{S}$ to be as small as possible. The smallest $\mathcal{S}$ satisfying (a) and (b) is called the $\boldsymbol{\Sigma}_{\mathbf{x}}$-*envelope* of $\operatorname{span}(\boldsymbol{\beta})$ and denoted as $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\operatorname{span}(\boldsymbol{\beta}))$. The equivalence with assertions (i) and (ii) will later be related to connections with PLS. We let $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$ and use $\mathcal{E}$ as shorthand for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ in subscripts. We consider the implications of this structure for prediction in Section 2.3, after reviewing the algebraic basis for envelopes in Section 2.2.

2.2 Review of envelopes

Here the definitions will be restated in complete generality. Our intent is to provide just enough background from CLC to allow us to later develop firm connections with PLS. Many additional results on envelopes were given by CLC.

Definition 2.1 *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 both $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$ and $\mathbf{M}\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$ we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

This definition of a reducing subspace is standard in linear algebra and functional analysis (cf. Conway, 1990, p. 38), but its notion of reduction is not compatible with how it is usually understood in statistics. Nevertheless, it is the foundation for the next definition which is directly relevant to the methodology discussed here.

Definition 2.2 *(CLC) Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ and let $\mathcal{B} \subseteq \mathrm{span}(\mathbf{M})$. Then the $\mathbf{M}$-envelope of $\mathcal{B}$ – denoted by $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ – is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.*

In applications $\mathcal{B}$ is typically the span of a regression vector or matrix and $\mathbf{M}$ will be a covariance matrix. In this article we will often have $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$ and $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{x}}$. The intersection of two reducing subspaces is always a reducing subspace. This together with the weak condition $\mathcal{B} \subseteq \mathrm{span}(\mathbf{M})$ ensures that the $\mathbf{M}$-envelope always exists.

These definitions yield three important consequences that relate reducing subspaces and envelopes to the eigenstructure of the reduced matrix; distinct eigenvalues are not required.

Proposition 2.2 *(CLC)*

(a) $\mathcal{R}$ reduces $\mathbf{M} \in \mathbb{R}^{p\times p}$ if and only if $\mathbf{M} = \mathbf{M}_{\mathcal{R}} + \mathbf{M}_{\mathcal{R}^{\perp}}$, where $\mathbf{M}_{\mathcal{R}} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}}$ and $\mathbf{M}_{\mathcal{R}^{\perp}} = \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$.

(b) If $\mathbf{M} \in \mathbb{R}^{p\times p}$ is symmetric, then $\mathbf{M}$ has a spectral decomposition with eigenvectors only in $\mathcal{R}$ or in $\mathcal{R}^{\perp}$ if and only if $\mathcal{R}$ reduces $\mathbf{M}$.

(c) If $\mathbf{M} \in \mathbb{R}^{p \times p}$ is symmetric with $q \leq p$ eigenspaces, then the $\mathbf{M}$-envelope of $\mathcal{B} \subseteq \operatorname{span}(\mathbf{M})$ can be characterized by $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \oplus_{i=1}^{q} \mathbf{P}_i \mathcal{B}$, where $\mathbf{P}_i$ is the projection onto the i-th eigenspace of $\mathbf{M}$.

Consequence (a) in this proposition shows how the mathematical notion of reduction in Definition 2.1 is linked to the task of algebraically reducing a matrix to the sum of two orthogonal matrices. When applied to a covariance matrix, this type of reduction (decomposition), along with the usual form of statistical dimension reduction, plays a key role in the development of envelope methods.

Envelopes are quite versatile and can be adapted to any multivariate setting that involves a non-negative definite symmetric matrix $\mathbf{M}$ and a location matrix $\boldsymbol{\beta}$. Candidates for $\mathbf{M}$ include $\operatorname{var}(\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{x}}$, $\operatorname{var}(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{y}}$ and the error covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}$. Models like (1) allow for crisp development of envelope methodology by permitting, for example, a likelihood analysis when the errors are normal and reduction of $\mathbf{y}$ is sought (CLC). However, the concept of an envelope as represented in Definition 2.2 is not model-based and can be useful in studies involving only moments, as is the case here.

2.3 Overview of predictor reduction via envelopes

As mentioned previously, CLC studied reduction in the $\mathbf{y}$-dimension, while here we are concerned with reduction in the $\mathbf{x}$-dimension. This distinction means that operationally we work with the column space $\mathcal{B} = \operatorname{span}(\boldsymbol{\beta}) \subseteq \mathbb{R}^p$, while CLC largely worked with the row space $\mathcal{B}' = \operatorname{span}(\boldsymbol{\beta}^T) \subseteq \mathbb{R}^r$. Additionally, CLC assumed normal errors and relied on various conditional independence conditions for motivation. Here we use correlation rather than independence for the underlying rationale.

Let $m = \dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}))$, let $\boldsymbol{\Sigma}_{\mathbf{xy}} = \operatorname{cov}(\mathbf{x}, \mathbf{y})$ when $r > 1$, let $\boldsymbol{\sigma}_{\mathbf{x}y} = \operatorname{cov}(\mathbf{x}, y)$ when $r = 1$ and let $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times m}$ denote a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$. If a basis $\boldsymbol{\Gamma}$ was known then we could reduce the predictors $\mathbf{x} \rightarrow \boldsymbol{\Gamma}^T\mathbf{x}$ and base prediction on the reduced linear model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\alpha}^T \{\boldsymbol{\Gamma}^T(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\alpha} = \operatorname{var}^{-1}(\boldsymbol{\Gamma}^T\mathbf{x})\operatorname{cov}(\boldsymbol{\Gamma}^T\mathbf{x}, \mathbf{y}) = (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{xy}} \in \mathbb{R}^{m \times r}$. The *envelope coefficient* *matrix* of $\mathbf{x}$ in (2) is simply

$$\boldsymbol{\beta}_{\mathcal{E}} \equiv \boldsymbol{\Gamma}\boldsymbol{\alpha} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}} = \mathbf{P}_{\mathcal{E}(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\beta} = \boldsymbol{\beta}, \tag{3}$$

where the third equality follows because $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}}$ and the last equality follows because $\mathcal{B} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ by construction. It follows from conditions (i) and (ii) stated in Section 2.1 that there is no loss of focus on $\boldsymbol{\beta}$ when using model (2) instead of (1).

The envelope coefficient matrix does not depend on the particular basis $\boldsymbol{\Gamma}$ chosen for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ since $\boldsymbol{\beta}_{\mathcal{E}}$ is unchanged by replacing $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}\mathbf{O}$ for any conforming orthogonal matrix $\mathbf{O}$. Consequently, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is the essential parameter with corresponding parameter space being the set of all m-dimensional subspaces of $\mathbb{R}^p$. This set is called a Grassmann manifold or Grassmannian, which we denote by $\mathcal{G}_{(m,p)}$, and $m(p-m)$ real numbers are required to uniquely identify a single subspace in $\mathcal{G}_{(m,p)}$. Background on Grassmann optimization is available from Edelman et al. (1998) and Liu et al. (2004). See Appendix C for a cursory introduction.

Let $\mathbf{S}_{\mathbf{x}}$, $\mathbf{S}_{\mathbf{x}\mathbf{y}}$ and $\mathbf{s}_{\mathbf{x}y}$ denote the sample versions of $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}}$ and $\boldsymbol{\sigma}_{\mathbf{x}y}$. There are now two estimators of $\boldsymbol{\beta}$ to consider: the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{S}_{\mathbf{x}}^{-1}\mathbf{S}_{\mathbf{x}\mathbf{y}}$ from model (1) and, assuming that an estimator $\widehat{\boldsymbol{\Gamma}}$ of a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is available, the envelope estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{x}}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{x}\mathbf{y}}$ from model (2), which is just the estimator $\widehat{\boldsymbol{\alpha}}$ from the OLS fit of $\mathbf{y}$ on $\widehat{\boldsymbol{\Gamma}}^T\mathbf{x}$ left multiplied by $\widehat{\boldsymbol{\Gamma}}$. Several different methods for estimating a basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ are discussed in later sections. For instance, SIMPLS uses a sequential estimator, as discussed in Section 4.3, while the likelihood-based estimator of Section 4.5 requires optimization over a Grassmann manifold.

Let $\mathbf{x}_N$ denote a new independent observation on $\mathbf{x}$, let $\mathbf{z}_N = \mathbf{x}_N - \boldsymbol{\mu}_{\mathbf{x}}$, and let $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\alpha}}$ denote the envelope estimator of $\boldsymbol{\beta}$ when a basis $\boldsymbol{\Gamma}$ is known. The following proposition provides intuition about regressions in which prediction of $\mathbf{y}$ at $\mathbf{z}_N$ via (2) might be superior to those from (1).

Proposition 2.3 *If the regression is univariate with $\mathbf{x} \sim N_p(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})$, $n > p + 2$ and a known semi-orthogonal basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times m}$ for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, then*

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T\mathbf{z}_N) = \frac{n-m-2}{n-p-2}\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}^T\mathbf{z}_N) + \frac{\sigma_{y|\mathbf{x}}^2}{n-p-2}\mathbf{z}_N^T\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\mathbf{z}_N, \tag{4}$$

where the variances are computed over all of the data (both y and $\mathbf{x}$), $\mathbf{\Gamma}_0 \in \mathbb{R}^{p\times(p-m)}$ is a semi-orthogonal basis matrix for $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\mathbf{z}_N$ is held fixed.

We see from this proposition that only the part of $\mathbf{z}_N$ that lies in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is relevant for prediction in the reduced model (2). If $\mathbf{z}_N \in \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ then $\mathrm{var}(\widehat{\boldsymbol{\beta}}^T_{\mathrm{OLS}}\mathbf{z}_N) = (n-p-2)^{-1}(n-m-2)\mathrm{var}(\widehat{\boldsymbol{\beta}}^T_{\mathbf{\Gamma}}\mathbf{z}_N)$. The gain implied in this case depends on the relationships between n, m and p. If p is close to n, in the extreme $p = n-3$, and m is small then the reduction in predictive variance could be substantial. On the other hand, when fitting only the full model, predictions depend on the whole of $\mathbf{z}_N$ and in particular on the part of $\mathbf{z}_N$ that lies in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ via the second term on the right hand side of (4). Here we find a connection between collinearity and $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$. If the predictors are collinear, so $\mathbf{\Sigma}_{\mathbf{x}}$ has some small eigenvalues, and if some of the eigenvectors corresponding to those small eigenvalues fall in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ then the second term on the right side of (4) could be large and the predictive gain realized by using the reduced model could again be substantial.

In practice the envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ will be unknown and thus will need to be estimated. The variability in the estimate of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ will mitigate the predictive gains discussed above, but we have found in simulations that (4) gives a useful qualitative feeling for the advantages of pursuing predictor reduction via envelopes. In general, the bias contribution to the prediction error should also be taken into account. The mean square error of any $\widehat{\boldsymbol{\beta}}^T_{*}\mathbf{z}_N$ as a predictor of y is the sum of three contributions: The conditional variance of y, which cannot be altered, the squared bias of $\widehat{\boldsymbol{\beta}}^T_{*}\mathbf{z}_N$ and its variance. It will follow from later results here that the squared bias of the envelope estimator proposed here is of smaller order than its variance as $n \rightarrow \infty$. In the remainder of this article we discuss how $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is estimated by using PLS and other methods.

3 Univariate partial least squares

In this section we first review relevant aspects of univariate PLS, relying primarily on Helland (1988, 1990), and then turn to its connection with envelopes, showing that PLS provides a root-n consistent estimator of a basis of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$. Like our commentary on envelopes in Section 2, our review of univariate PLS is intended to give just enough background to allow the development of connections with envelopes.

3.1 Review of univariate PLS

The population PLS algorithm (Helland, 1990) may be described as follows: Take $\mathbf{e}_0 = \mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}$, $f_0 = y - \mu$, and for $a = 1, 2, ..., A \le p$ compute successively: $\mathbf{w}_a = \mathrm{cov}(\mathbf{e}_{a-1}, f_{a-1})$, $t_a = \mathbf{w}_a^T \mathbf{e}_{a-1}$, $\mathbf{p}_a = \mathrm{cov}(\mathbf{e}_{a-1}, t_a)/\mathrm{var}(t_a)$, $q_a = \mathrm{cov}(f_{a-1}, t_a)/\mathrm{var}(t_a)$, $\mathbf{e}_a = \mathbf{e}_{a-1} - \mathbf{p}_a t_a$, and $f_a = f_{a-1} - q_a t_a$, continuing until $A = p$ or $\mathbf{w}_A = 0$. After A steps we get the representations

$$\mathbf{x} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{p}_1 t_1 + ... + \mathbf{p}_A t_A + \mathbf{e}_A, \;\; y = \mu + q_1 t_1 + ... + q_A t_A + f_A \tag{5}$$

with the corresponding PLS population prediction

$$y_{A,\mathrm{PLS}} = \mu + q_1 t_1 + ... + q_A t_A = \mu + \boldsymbol{\beta}_{A,\mathrm{PLS}}^T (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \tag{6}$$

The ordinary PLS estimator with A components is simply a plug-in estimator with population quantities replaced by their sample counterparts. The number of components A is typically selected by cross validation or by use of an independent test sample. The hope is that the sample version of $\boldsymbol{\beta}_{A,\mathrm{PLS}}$ will lead to better predictions than $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. Different ways of understanding the population properties of PLS from this basic algorithm were developed in Helland (1988, 1990). The following proposition, basically from Helland (1988), may assist in forming an appreciation of PLS at the population level. In preparation, let $\mathbf{W}_A = (\mathbf{w}_1, ..., \mathbf{w}_A)$ and let $\mathcal{W}_A = \mathrm{span}(\mathbf{W}_A)$.

Proposition 3.1 *(a) The weight vectors $\mathbf{w}_a$, $a = 1, \ldots, p$, satisfy the recurrence relation*

$$\begin{aligned} \mathbf{w}_{A+1} &= \boldsymbol{\sigma}_{\mathbf{x}y} - \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{W}_A (\mathbf{W}_A^T \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{W}_A)^{-1} \mathbf{W}_A^T \boldsymbol{\sigma}_{\mathbf{x}y} & (7) \\ &= \mathbf{P}_{\mathcal{W}_A^{\perp}(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1})} \boldsymbol{\sigma}_{\mathbf{x}y}. & (8) \end{aligned}$$

(b) The identity (6) for $y_{A,PLS}$ holds with

$$\begin{aligned} \boldsymbol{\beta}_{A,\mathrm{PLS}} &= \mathbf{W}_A (\mathbf{W}_A^T \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{W}_A)^{-1} \mathbf{W}_A^T \boldsymbol{\sigma}_{\mathbf{x}y} & (9) \\ &= \mathbf{P}_{\mathcal{W}_A(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\beta}, & (10) \end{aligned}$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\sigma}_{\mathbf{x}y}$ is the coefficient vector from the population OLS fit.

Useful insights into univariate PLS can be obtained from these representations. First, (7) expresses $\mathbf{w}_A$ as a type of successive residual vector. This idea is represented more explicitly by (8) where $\mathbf{w}_{A+1}$ is depicted as the projection of $\boldsymbol{\sigma}_{\mathbf{x}y}$ onto $\mathcal{W}_A^{\perp}$ in the $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$ inner product. From this we see that the weight vectors are orthogonal, $\mathbf{w}_{A+1}^T\mathbf{W}_A = 0$, and thus the subspaces $\mathcal{W}_A$ form an increasing nested sequence, $\mathcal{W}_A \subseteq \mathcal{W}_{A+1}$. Second, (10) shows that the population PLS coefficients $\boldsymbol{\beta}_{A,\mathrm{PLS}}$ are of the same form as the population envelope coefficients $\boldsymbol{\beta}_{\mathcal{E}}$ shown in (3). In the next section we piece together results from the literature to show that the population PLS stopping point is $A = m$ and then $\mathcal{W}_m = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\boldsymbol{\beta}_{\mathcal{E}} = \boldsymbol{\beta}_{m,\mathrm{PLS}} = \boldsymbol{\beta}$. Third, representations (8) and (10) require that $\boldsymbol{\Sigma}_{\mathbf{x}} > 0$, while (7) and (9) require only that $\mathbf{W}_A^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{W}_A > 0$. Consequently, the PLS algorithm does not necessarily require $\boldsymbol{\Sigma}_{\mathbf{x}} > 0$, depending on A. Nevertheless, Chun and Keleş (2010) recently proved that the PLS estimator of $\boldsymbol{\beta}$ is consistent when $p/n \to k = 0$, but inconsistent when $k > 0$. They also proposed a sparse version of PLS that in simulations seems to do well against competing methods (see also Nadler and Coifman, 2005). Because of these results, we limit discussion of the properties of the sample estimator to the $n > p$ setting. The case $n < p$ certainly is of interest in chemometrics and in genomic applications (Boulesteix and Strimmer, 2006).

3.2 Envelopes, univariate PLS and Krylov sequences

A first connection between envelopes and PLS can be seen by linking the result from Proposition 2.2(c) with results from Helland (1990). Applying Proposition 2.2(c) in the context of reducing the $\mathbf{x}$-dimension in model (1) we have $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \oplus_{i=1}^{q}\mathbf{P}_i\mathcal{B}$, where $\mathbf{P}_i$ is the projection onto the i-th eigenspace of $\boldsymbol{\Sigma}_{\mathbf{x}}$ with the ordering of the eigenspaces being immaterial. It follows immediately that the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is bounded above by the number of eigenspaces of $\boldsymbol{\Sigma}_{\mathbf{x}}$. Further, if $\boldsymbol{\beta}$ has a non-zero projection onto $m \leq q$ eigenspaces then $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \oplus_{i=1}^{m}\mathbf{P}_i\mathcal{B}$. Define the length 1 vectors $\boldsymbol{\ell}_i = \mathbf{P}_j\boldsymbol{\beta}/\|\mathbf{P}_j\boldsymbol{\beta}\|$, $i = 1, \ldots, m$, so each $\boldsymbol{\ell}_i$ is a normalized (unordered) eigenvector of $\boldsymbol{\Sigma}_{\mathbf{x}}$ and together they give an orthogonal basis for the envelope, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathrm{span}(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_m)$. Since $\mathcal{B} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, we have the unique representation

$$\boldsymbol{\beta} = \sum_{i=1}^{m} \gamma_i \boldsymbol{\ell}_i \tag{11}$$

with only non-zero γ_i's. The population PLS model with $A = m$ components is shown in Helland (1990) to be equivalent to the representation (11) with m non-zero terms and consequently the dimension of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is equal to the number of PLS components in unvariate $r = 1$ regressions.

To further elucidate the relationship between PLS and envelopes, and to show the connection to (7)-(10), we introduce the Krylov matrix $\mathbf{K}_A = (\boldsymbol{\sigma}_{\mathbf{x}y}, \mathbf{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, ..., \mathbf{\Sigma}_{\mathbf{x}}^{A-1}\boldsymbol{\sigma}_{\mathbf{x}y})$. Let $\mathcal{K}_A = \mathrm{span}(\mathbf{K}_A)$. This subspace is called a Krylov subspace in numerical analysis and is related to cyclic invariant subspaces in linear algebra. Helland (1988) showed that the sample version $\widehat{\mathcal{W}}_A$ of the subspace $\mathcal{W}_A$ used in Proposition 3.1 is equal to the sample version of the Krylov subspace, $\widehat{\mathcal{K}}_A = \widehat{\mathcal{W}}_A$. Since $\widehat{\mathbf{K}}_A$ and $\widehat{\mathbf{W}}_A$ are consistent estimators of $\mathbf{K}_A$ and $\mathbf{W}_A$, we also have $\mathcal{K}_A = \mathcal{W}_A$, and the population and sample PLS coefficients can be represented as $\boldsymbol{\beta}_{A,\mathrm{PLS}} = \mathbf{P}_{\mathcal{K}_A(\mathbf{\Sigma}_{\mathbf{x}})}\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{A,\mathrm{PLS}} = \widehat{\mathbf{K}}_A(\widehat{\mathbf{K}}_A^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{K}}_A)^{-1}\widehat{\mathbf{K}}_A^T\mathbf{s}_{\mathbf{x}y}$.

To bring envelopes into the picture, let $\mathbf{L} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_m)$, so $\mathrm{span}(\mathbf{L}) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, let φ_j denote the eigenvalue of $\mathbf{\Sigma}_{\mathbf{x}}$ associated with $\boldsymbol{\ell}_j$, $j = 1, \ldots, m$, let $\mathbf{D} = \mathrm{diag}(\varphi_1\gamma_1, \ldots, \varphi_m\gamma_m)$ and let $\mathbf{V}_A \in \mathbb{R}^{m\times A}$ denote the Vandermonde matrix with elements φ_j^{k-1}, $j = 1, \ldots, m$, $k = 1, \ldots, A$. Then we can express $\mathbf{K}_A = \mathbf{L}\mathbf{D}\mathbf{V}_A$. Using well-known properties of the Vandermonde matrix, it follows that $\mathcal{K}_A$ is a strictly increasing sequence of nested subspaces that converges to $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ after m steps and remains constant thereafter,

$$\mathcal{K}_1 \subset \mathcal{K}_2 \subset \cdots \subset \mathcal{K}_{m-1} \subset \mathcal{K}_m = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{K}_{m+1} = \cdots = \mathcal{K}_p. \tag{12}$$

Again, we have the implication that the PLS stopping point m is equal to the smallest integer so that $\mathcal{K}_m = \mathcal{K}_p = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and thus is equal to the dimension of the envelope. This discussion is formally summarized in the following proposition, which includes some additional findings.

Proposition 3.2 *Let $\mathcal{S}_{\mathbf{x}y} = \mathrm{span}(\boldsymbol{\sigma}_{\mathbf{x}y})$ and let $\mathbf{P}_i$ denote the projection onto the i-th eigenspace of $\mathbf{\Sigma}_{\mathbf{x}}$, $i = 1, \ldots, q \leq p$. Then*

(a) $\mathcal{W}_A = \mathcal{K}_A$, $A = 1, \ldots, p$,

(b) $m = \min\{A|\mathbf{\Sigma}_{\mathbf{x}}^A\boldsymbol{\sigma}_{\mathbf{x}y} \in \mathcal{W}_A\} = \min\{A|\boldsymbol{\beta} \in \mathcal{W}_A\}$,

(c) $\mathcal{W}_m = \mathcal{K}_m = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{S}_{\mathbf{x}y}) = \oplus_{i=1}^q\mathbf{P}_i\mathcal{S}_{\mathbf{x}y}$,

(d) When $A = m$, we have $\boldsymbol{\beta}_{A,\mathrm{PLS}} = \boldsymbol{\beta}$.

In (c), exactly m of the spaces $\mathbf{P}_i\mathcal{S}_{\mathbf{x}y}$ are non-trivial. We see from Proposition 3.2 that $\mathcal{W}_m$ (cf.

(10)) can be replaced by $\mathcal{K}_m$ or by $\mathrm{span}\{\mathbf{P}_1\mathcal{S}_{\mathbf{x}y}, \ldots, \mathbf{P}_q\mathcal{S}_{\mathbf{x}y}\}$. Proposition 3.2(d) shows that we must have $m = p$ when the eigenvalues of $\mathbf{\Sigma}_\mathbf{x}$ are distinct and $\boldsymbol{\sigma}_{\mathbf{x}y}$ has a non-zero projection onto each of its p eigenvectors, in which case $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{B}) = \mathbb{R}^p$ and conditions (i) and (ii) given near the end of Section 2 hold only when $\mathcal{S} = \mathbb{R}^p$. If there is a proper envelope $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{B}) \subset \mathbb{R}^p$ then the sample version of the PLS algorithm may yield a more efficient estimator than $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. From (d) the PLS regression vector is equal to the OLS regression vector when $A = m$. When $A < m$, the conclusion of Proposition 3.2(d) does not hold; that is, $\boldsymbol{\beta}_{A,\mathrm{PLS}}$ is different from $\boldsymbol{\beta}$.

Returning to the sample version, since $\mathbf{S}_\mathbf{x}$ and $\mathbf{s}_{\mathbf{x}y}$ are root-n consistent estimators of $\mathbf{\Sigma}_\mathbf{x}$ and $\boldsymbol{\sigma}_{\mathbf{x}y}$, $\widehat{\mathbf{K}}_A$ is also a root-n consistent estimator of $\mathbf{K}_A$, $A = 1, \ldots, p$, which implies that $\mathbf{P}_{\widehat{\mathbf{K}}_m(\mathbf{S}_\mathbf{x})}$ is a root-n consistent estimator of $\mathbf{P}_{\mathcal{E}(\mathbf{\Sigma}_\mathbf{x})}$. In reference to the overview in Section 2.3, we can take $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{K}}_m$, leading to a root-$n$ consistent estimator $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\mathbf{K}}_m(\mathbf{S}_\mathbf{x})}\mathbf{s}_{\mathbf{x}y}$ of $\boldsymbol{\beta}$ when m is known. By the discussion above, this is equal to the sample PLS estimator with m terms.

4 Envelopes and multivariate PLS

There are two main PLS algorithms for the multivariate linear regression of $\mathbf{y} \in \mathbb{R}^r$ on $\mathbf{x} \in \mathbb{R}^p$: NIPLS (Wold, 1966) and SIMPLS (de Jong, 1993). These algorithms are not usually presented as model-based, but instead are regarded as methods for estimating the coefficient matrix $\boldsymbol{\beta} = \mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}$ followed by prediction. Much has been written about these algorithms since their introductions, although there has so far not been proposed any population characterizations analogous to those given in Section 3 for univariate regressions. It is known that these algorithms give distinct sample results when $r > 1$ but they are the same when $\mathbf{y}$ is univariate, $r = 1$.

4.1 Overview

In the following sections we present three different constructs for connecting PLS and envelopes in multivariate regressions when the goal is to reduce $\mathbf{x}$ only. The first is based on a population characterization of the SIMPLS algorithm, the second is based on an extension of the Krylov matrices discussed in Section 3.2, and the third derives from a likelihood-based objective function. At the population level, each approach is designed to produce a basis matrix $\mathbf{\Gamma}$, $\mathrm{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{S}_{\mathbf{xy}})$, where $\mathcal{S}_{\mathbf{xy}} = \mathrm{span}(\mathbf{\Sigma}_{\mathbf{xy}})$ and the equality $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}_\mathbf{x}}(\mathcal{S}_{\mathbf{xy}})$ follows from Proposition 3.1 of CLC. Since $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})$ and $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{S}_{\mathbf{xy}})$ are equal, we may use one or the other in expressions, depending on desired emphasis. Once $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})$ is determined we use (3) to get $\boldsymbol{\beta}_{\mathcal{E}} = \mathbf{P}_{\mathcal{E}(\mathbf{\Sigma_x})}\boldsymbol{\beta}$. Since $\mathcal{B} \subseteq \mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})$ we see that $\boldsymbol{\beta}_{\mathcal{E}} = \boldsymbol{\beta}$ in the population, although that will of course not be so in the sample. For instance, the SIMPLS algorithm produces a sample $\mathbf{\Gamma}$ and then uses (3) to form the envelope estimator $\widehat{\boldsymbol{\beta}}$, replacing $\mathbf{\Gamma}$, $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_{xy}}$ with their sample counterparts. This is then followed by forming the linear predictive equation $\widehat{\mathbf{y}} = \bar{\mathbf{y}} + \widehat{\boldsymbol{\beta}}^T(\mathbf{x} - \bar{\mathbf{x}})$.

Each of the three approaches to be discussed depends on $\mathbf{x}$ and $\mathbf{y}$ only via the dimension m of the envelope $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})$ and smooth functions of $\mathbf{\Sigma_x}$, $\mathbf{\Sigma_{xy}}$ and $\mathbf{\Sigma_y}$. The sample versions thus depend on the data only through $\mathbf{S_x}$, $\mathbf{S_{xy}}$ and $\mathbf{S_y}$, the sample version of $\mathbf{\Sigma_y}$.

4.2 Envelopes for multivariate responses

Before turning to estimators, we discuss the structure of envelopes for multivariate $\mathbf{y}$ as an extension of our discussion for univariate $\mathbf{y}$ in Section 3.2. Applying Proposition 2.2(c) we have $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B}) = \oplus_{i=1}^{q}\mathbf{P}_i\mathcal{B} = \oplus_{i=1}^{q}\mathbf{P}_i\mathcal{S}_{\mathbf{xy}}$, where $\mathbf{P}_i$ is still the projection onto the i-th eigenspace of $\mathbf{\Sigma_x}$. In the univariate case we found that the dimension of $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})$ is bounded above by the number of eigenspaces of $\mathbf{\Sigma_x}$. This is no longer so in the multivariate case. In the extreme, if $\dim(\mathcal{B}) = p$ then regardless of the number of eigenspaces $\dim(\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B})) = p$ and no reduction is possible. This will be avoided when $r < p$ and more generally when $\dim(\mathcal{B}) < p$. Typically $r \ll p$ in chemometrics applications. We assume that $r < p$ in the remainder of this article.

Further, if $\boldsymbol{\beta}$ has a non-zero projection onto $e \leq q$ eigenspaces then $\mathcal{E}_{\mathbf{\Sigma_x}}(\mathcal{B}) = \oplus_{i=1}^{e}\mathbf{P}_i\mathcal{S}_{\mathbf{xy}}$. In the univariate case, e and the dimension m of the envelope are the same. However, this is not necessarily so in the multivariate case. Suppose for instance that $\mathcal{S}_{\mathbf{xy}}$ is contained in one eigenspace, say $\mathrm{span}(\mathbf{P}_1)$. Then $e = 1$, but $m = \dim(\mathbf{P}_1\mathcal{S}_{\mathbf{xy}}) = \dim(\mathcal{S}_{\mathbf{xy}})$, and so $1 \leq m \leq r$.

4.3 SIMPLS

The population SIMPLS algorithm for predictor reduction proceeds by finding a sequence of p-dimensional vectors $\mathbf{w}_0, \ldots, \mathbf{w}_k$ as follows. Set $\mathbf{w}_0 = 0$ and let $\mathbf{W}_k = (\mathbf{w}_0, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$.

Then given $\mathbf{W}_k$, the next vector $\mathbf{w}_{k+1}$ is constructed as

$$\begin{aligned}\mathbf{w}_{k+1} &= \arg\max_{\mathbf{w}} \mathbf{w}^T \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{xy}}^T \mathbf{w}, \text{ subject to} \\ & \qquad \mathbf{w}^T \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{W}_k = 0 \text{ and } \mathbf{w}^T\mathbf{w} = 1.\end{aligned}$$

The following proposition gives a characterization of the population behavior of the SIMPLS algorithm. It shows that the nested structure (12) for univariate PLS holds also for SIMPLS. Recall that $m = \dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}))$.

Proposition 4.1 *The SIMPLS subspaces $\mathcal{W}_k = \mathrm{span}(\mathbf{W}_k)$ are nested and strictly increasing for $k \le m$. They converge to $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ after m steps, $\mathcal{W}_1 \subset \ldots \subset \mathcal{W}_{m-1} \subset \mathcal{W}_m = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, and are constant thereafter, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{W}_{m+1} = \ldots = \mathcal{W}_p$.*

The SIMPLS algorithm is a function of only three population quantities, $\boldsymbol{\Sigma}_{\mathbf{xy}}$, $\boldsymbol{\Sigma}_{\mathbf{x}}$ and m. The sample version of SIMPLS is constructed by replacing $\boldsymbol{\Sigma}_{\mathbf{xy}}$, $\boldsymbol{\Sigma}_{\mathbf{x}}$ by their sample counterparts, terminating after m steps and then setting $\widehat{\boldsymbol{\Gamma}}$ equal to the sample version of $\mathbf{W}_m$ for use in (3). Of course there is no sample counterpart to m. Five or ten-fold cross-validation of predictive performance is often an effective method for choosing an estimate of m. If m is known then the results of Chun and Keleş (2010) can be adapted to show that, with r and p fixed, this algorithm provides a root-n consistent estimator of $\boldsymbol{\beta}$. Generally, $\dim(\mathcal{S}_{\mathbf{xy}}) \le m \le p$, where $\dim(\mathcal{S}_{\mathbf{xy}}) \le r$ since we have assumed that $r < p$. If it turns out that $m = p$ then the SIMPLS estimator of $\boldsymbol{\beta}$ is equal to the OLS estimator.

Let $\boldsymbol{\ell}_{\max}(\mathbf{A})$ be an eigenvector associated with the largest eigenvalue of the symmetric matrix $\mathbf{A}$, $\boldsymbol{\ell}_{\max}(\mathbf{A}) = \arg\max_{\boldsymbol{\ell}^T\boldsymbol{\ell}=1} \boldsymbol{\ell}^T \mathbf{A} \boldsymbol{\ell}$. It can be seen from the proof of Proposition 4.1 given in the appendix that the SIMPLS algorithm can be stated equivalently without explicit constraints as follows. Again set $\mathbf{w}_0 = 0$ and $\mathbf{W}_0 = \mathbf{w}_0$. For $k = 0, \ldots, m-1$, set

$$\begin{aligned}\mathcal{E}_k &= \mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{W}_k) \\ \mathbf{w}_{k+1} &= \boldsymbol{\ell}_{\max}(\mathbf{Q}_{\mathcal{E}_k} \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{xy}}^T \mathbf{Q}_{\mathcal{E}_k}) \\ \mathbf{W}_{k+1} &= (\mathbf{w}_0, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1}).\end{aligned}$$

At termination, $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{W}_m = \mathrm{span}(\mathbf{W}_m)$. Since $\mathbf{W}_k$ has full column rank for $k \leq m$, $\dim(\mathcal{E}_k) = k$ and thus no rank consideration is necessary for $\mathcal{E}_k$.

4.4 Krylov constructions

Define the multivariate Krylov matrix as

$$\mathbf{K}_a^{(r)} = (\mathbf{\Sigma}_{\mathbf{xy}}, \mathbf{\Sigma}_{\mathbf{x}}\mathbf{\Sigma}_{\mathbf{xy}}, \mathbf{\Sigma}_{\mathbf{x}}^2\mathbf{\Sigma}_{\mathbf{xy}}, ..., \mathbf{\Sigma}_{\mathbf{x}}^{a-1}\mathbf{\Sigma}_{\mathbf{xy}}) \in \mathbb{R}^{p\times ar},$$

and let $\mathcal{K}_a^{(r)} = \mathrm{span}(\mathbf{K}_a^{(r)})$ denote the corresponding subspace. Recall that e denotes the number of eigenspaces of $\mathbf{\Sigma}_{\mathbf{x}}$ that are not orthogonal to $\mathcal{B}$. Then Cook, Li and Chiaromonte (2007) showed that there is an interger $b \leq e$ so that $\mathcal{K}_a^{(r)}$ is strictly increasing until $a = b$, and then $\mathcal{K}_b^{(r)} = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\mathcal{K}_a^{(r)}$ is constant for $a \geq b$:

$$\mathcal{K}_1^{(r)} \subset \mathcal{K}_2^{(r)} \subset \cdots \subset \mathcal{K}_{b-1}^{(r)} \subset \mathcal{K}_b^{(r)} = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{K}_{b+1}^{(r)} = \cdots$$

This sequence of Krylov subspaces then has the same general structure as the subspace sequences for univariate (12) and multivariate PLS (Proposition 4.1), but there are consequential differences. In univariate and multivariate PLS, the stopping points m for $\mathbf{K}_m$ and $\mathcal{W}_m$ are the same as the dimension of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, but the stopping point b for the multivariate Krylov matrices $\mathbf{K}_a^{(r)}$ is not necessarily equal to m unless $r = 1$. If b were known then we would also know that $m \leq br$, but we would not know m itself. For instance, if $b = 1$ then $1 \leq m = \dim(\mathcal{S}_{\mathbf{xy}}) \leq r$ and an extra step would be necessary to determine m.

A different aspect of the structure of $\mathcal{K}_a^{(r)}$ can be seen by writing it as $\mathcal{K}_a^{(r)} = \oplus_{j=1}^{r}\mathcal{K}_a^{[j]}$, where $\mathcal{K}_a^{[j]}$ is the univariate Krylov subspace for the j-th response. Since the responses could have very different relationships with $\mathbf{x}$, there is no necessary connection between the subspaces $\mathcal{K}_a^{[j]}$. For example, $r-1$ of the responses could be independent of $\mathbf{x}$, while the remaining response has a simple one-component relationship with $\mathbf{x}$. In that case, $m = 1$.

These shortcomings notwithstanding, the span of the sample version of $\mathbf{K}_b^{(r)}$ is a root-n consistent estimator of $\mathcal{K}_b^{(r)}$ and thus provides for an alternative estimator of $\boldsymbol{\beta}$.

4.5 Likelihood-based estimation

The estimators of $\boldsymbol{\beta}$ that we have discussed so far are all based on sequential estimators of a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$. In this section we describe a non-sequential method of construction that requires searching over the Grassmann manifold $\mathcal{G}_{(m,p)}$, where still $m = \dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}))$. Suppose that we wish to optimize a scalar-valued function $f(\mathbf{A})$ of the matrix argument $\mathbf{A}$, where f has the property that $f(\mathbf{A}) = f(\mathbf{AO})$ for all conforming orthogonal matrices $\mathbf{O}$. Then the solution depends only on $\mathrm{span}(\mathbf{A})$ and the optimization problem is Grassmann.

4.5.1 The estimators

Let $\mathbf{c} = (\mathbf{x}^T, \mathbf{y}^T)^T$ denote the random vector constructed by concatenating $\mathbf{x}$ and $\mathbf{y}$, and let $\mathbf{S}_{\mathbf{c}}$ denote the sample version of $\boldsymbol{\Sigma}_{\mathbf{c}} = \mathrm{var}(\mathbf{c})$. We base estimation on the objective function $F_m(\mathbf{S}_{\mathbf{c}}, \boldsymbol{\Sigma}_{\mathbf{c}}) = \log|\boldsymbol{\Sigma}_{\mathbf{c}}| + \mathrm{trace}(\mathbf{S}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{c}}^{-1})$ that stems from the likelihood of the multivariate normal family, although we do not require $\mathbf{c}$ to have a multivariate normal distribution. Rather we are using F_m as a multi-purpose objective function in the same spirit as it has been used recently for the development of sparse estimates of a covariance matrix. For example, Rothman, et al. (2008) studied a sparse estimator for the inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ of a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ based on its sample version $\widehat{\boldsymbol{\Sigma}}$ by minimizing the penalized normal likelihood $\mathrm{trace}(\boldsymbol{\Omega}\widehat{\boldsymbol{\Sigma}}) - \log|\boldsymbol{\Omega}| + \lambda\sum_{i,j=1}^{p}(\boldsymbol{\Omega} - \mathrm{diag}(\boldsymbol{\Omega}))_{ij}$, where λ is the tuning parameter and $\mathbf{A}_{ij}$ denotes the (i,j)-th element of the matrix $\mathbf{A}$. Although normality was required in their formal development, Rothman et al. (2008, Section 5) also stated that their estimator requires only a tail condition that parallels a condition used by Bickel and Levina (2008) and that it works well as a loss function without normality (See also Levina et al., 2008). We show later in Proposition 4.3 that our use of F_m leads to a root-n consistent envelope estimator of $\boldsymbol{\beta}$ that requires only finite fourth moments for $\mathbf{y}$ and $\mathbf{x}$.

It is traditional in regression to base estimation on the conditional likelihood of $\mathbf{y}|\mathbf{x}$, treating the predictors as fixed even if they were randomly sampled. This practice arose because in many regressions the predictors provide only ancillary information and consequently estimation and inference should be conditioned on their observed values. (See Aldrich, 2005, for a review and an historical perspective.) In contrast, PLS and the likelihood-based method developed in this section both postulate a link – represented here by the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ – between $\boldsymbol{\beta}$, the parameter of interest, and $\boldsymbol{\Sigma}_{\mathbf{x}}$. As a consequence, $\mathbf{x}$ is not ancillary and we used the joint distribution of $\mathbf{y}$ and $\mathbf{x}$.

The structure of the envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{S}_{\mathbf{xy}})$ can be introduced into F_m by using the parameterizations $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and $\mathbf{\Sigma}_{\mathbf{xy}} = \mathbf{\Gamma}\boldsymbol{\eta}$, where $\mathbf{\Gamma} \in \mathbb{R}^{p\times m}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{S}_{\mathbf{xy}})$, $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, and $\mathbf{\Omega} \in \mathbb{R}^{m\times m}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-m)\times(p-m)}$ are symmetric positive definite matrices. Since $\mathcal{S}_{\mathbf{xy}} \subseteq \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{S}_{\mathbf{xy}})$ we can write $\mathbf{\Sigma}_{\mathbf{xy}}$ as linear combinations of the columns of $\mathbf{\Gamma}$. The matrix $\boldsymbol{\eta} \in \mathbb{R}^{m\times r}$ then gives the coordinates of $\mathbf{\Sigma}_{\mathbf{xy}}$ in terms of the basis $\mathbf{\Gamma}$. With this we have

$$\mathbf{\Sigma}_{\mathbf{c}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{x}} & \mathbf{\Sigma}_{\mathbf{xy}} \\ \mathbf{\Sigma}_{\mathbf{xy}}^T & \mathbf{\Sigma}_{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T & \mathbf{\Gamma}\boldsymbol{\eta} \\ \boldsymbol{\eta}^T\mathbf{\Gamma}^T & \mathbf{\Sigma}_{\mathbf{y}} \end{pmatrix}. \tag{13}$$

The objective function $F_m(\mathbf{S}_{\mathbf{c}}, \mathbf{\Sigma}_{\mathbf{c}})$ can now be regarded as a function of the five parameters – $\mathbf{\Gamma}$, $\mathbf{\Omega}$, $\mathbf{\Omega}_0$, $\boldsymbol{\eta}$ and $\mathbf{\Sigma}_{\mathbf{y}}$ – that comprise $\mathbf{\Sigma}_{\mathbf{c}}$. In this paramerization, $\boldsymbol{\alpha} = \mathbf{\Omega}^{-1}\boldsymbol{\eta}$ and $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\alpha} = \mathbf{\Gamma}\mathbf{\Omega}^{-1}\boldsymbol{\eta}$.

Define the jointly standardized response as $\mathbf{z} = \mathbf{S}_{\mathbf{y}}^{-1/2}\mathbf{y}$, let $\mathbf{S}_{\mathbf{xz}}$ be the sample covariance matrix between $\mathbf{x}$ and $\mathbf{z}$ and let $L(\mathbf{G}) = \log|\mathbf{G}^T(\mathbf{S}_{\mathbf{x}} - \mathbf{S}_{\mathbf{xz}}\mathbf{S}_{\mathbf{xz}}^T)\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{G}|$. Minimizing $F_m(\mathbf{S}_{\mathbf{c}}, \mathbf{\Sigma}_{\mathbf{c}})$ over all parameters except $\mathbf{\Gamma}$ we arrive at the estimator

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{G}}\{L(\mathbf{G})\}, \tag{14}$$

where the minimization is over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{p\times m}$. The objective function $L(\mathbf{G})$ is invariant under right orthogonal transformations $\mathbf{G} \to \mathbf{GO}$, where $\mathbf{O} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, so the minimization is over the Grassmann manifold $\mathcal{G}_{(m,p)}$ and the solution is not unique. Following determination of a $\widehat{\mathbf{\Gamma}}$, the remaining parameters that comprise $\mathbf{\Sigma}_{\mathbf{c}}$ are estimated via F_m as $\widehat{\boldsymbol{\eta}} = \widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{xy}}$, $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{\Gamma}}$, $\widehat{\mathbf{\Omega}}_0 = \widehat{\mathbf{\Gamma}}_0^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{\Gamma}}_0$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} = \mathbf{S}_{\mathbf{y}}$, where $(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Gamma}}_0)$ is an orthogonal matrix. Additionally, $\boldsymbol{\beta}$ is estimated as described previously:

$$\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{xy}} = \widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{\Gamma}})^{-1}\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{xy}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Omega}}^{-1}\widehat{\boldsymbol{\eta}}. \tag{15}$$

This estimator of $\boldsymbol{\beta}$ depends only on $\mathrm{span}(\widehat{\mathbf{\Gamma}})$ so the particular solution to (14) does not matter.

There are consequential differences between the estimation method leading to (15) and the previous methods. To see how these differences arise, we first describe some operating characteristics of $L(\mathbf{G})$ and then contrast those characteristics with the behavior of SIMPLS. Let $\mathbf{v} = \mathbf{S}_{\mathbf{x}}^{-1/2}\mathbf{x}$ denote the sample standardized version of $\mathbf{x}$ and let $\mathbf{S}_{\mathbf{vz}} = \mathbf{S}_{\mathbf{x}}^{-1/2}\mathbf{S}_{\mathbf{xz}}$ denote the matrix of sample covariances between $\mathbf{v}$ and $\mathbf{z}$, which can also be interpreted as the sample coefficient matrix from the linear regression of $\mathbf{z}$ on $\mathbf{v}$. Let also $L_1(\mathbf{G}) = \log|\mathbf{G}^T\mathbf{S}_{\mathbf{x}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{G}|$ and $L_2(\mathbf{G}) = \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{vz}}^T\mathbf{P}_{\mathbf{S}_{\mathbf{x}}^{1/2}\mathbf{G}}\mathbf{S}_{\mathbf{vz}}|$. Then the objective function L can be represented as $L(\mathbf{G}) = L_1(\mathbf{G}) + L_2(\mathbf{G})$. The first addend $L_1(\mathbf{G}) \geq 0$ with $L_1(\mathbf{G}) = 0$ when the columns of $\mathbf{G}$ correspond to any subset of m eigenvectors of $\mathbf{S}_{\mathbf{x}}$. Consequently, the role of L_1 is to pull the solution toward subsets of m eigenvectors of $\mathbf{S}_{\mathbf{x}}$. This in effect imposes a sample counterpart of the characterization in Proposition 2.2(c), which states that in the population $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is spanned by a subset of the eigenvectors of $\mathbf{\Sigma}_{\mathbf{x}}$. The second addend $L_2(\mathbf{G})$ of $L(\mathbf{G})$ carries the covariance signal from $\mathbf{S}_{\mathbf{vz}}$ in terms of the standardized variables $\mathbf{v}$ and $\mathbf{z}$. It is minimized alone by choosing the columns of $\mathbf{G}$ to be the first m generalized eigenvectors of $\mathbf{S}_{\mathbf{xz}}\mathbf{S}_{\mathbf{xz}}^T$ relative to $\mathbf{S}_{\mathbf{x}}$, which are the solutions $\boldsymbol{\ell}$ to the generalized eigenvector problem $\mathbf{S}_{\mathbf{xz}}\mathbf{S}_{\mathbf{xz}}^T\boldsymbol{\ell} = \lambda\mathbf{S}_{\mathbf{x}}\boldsymbol{\ell}$. If $m > r$ only the first $m - r$ of these generalized eigenvectors are determined uniquely. An equivalent solution can be obtained by setting $\mathbf{G}$ to be $\mathbf{S}_{\mathbf{x}}^{-1/2}$ times the first m eigenvectors of $\mathbf{S}_{\mathbf{vz}}\mathbf{S}_{\mathbf{vz}}^T$. The full objective function $L(\mathbf{G}) = L_1(\mathbf{G}) + L_2(\mathbf{G})$ can then be viewed as balancing the requirement that the optimal value should stay close to a subset of m eigenvectors of $\mathbf{S}_{\mathbf{x}}$ and to the generalized eigenvectors of $\mathbf{S}_{\mathbf{xz}}\mathbf{S}_{\mathbf{xz}}^T$ relative to $\mathbf{S}_{\mathbf{x}}$.

Turning to comparisons of the likelihood-based method with SIMPLS, we see first that $L(\mathbf{G})$ depends on the response only through its standardized version $\mathbf{z} = \mathbf{S}_{\mathbf{y}}^{-1/2}\mathbf{y}$. On the other hand, SIMPLS depends on the scale of the response: when $m = 1$, the SIMPLS estimator of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is the span of the first eigenvector $\widehat{\mathbf{w}}_1$ of $\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{xy}}^T$. After performing a full rank transformation of the response $\mathbf{y} \rightarrow \mathbf{A}\mathbf{y}$, the SIMPLS estimator of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ is the span of the first eigenvector $\widetilde{\mathbf{w}}_1$ of $\mathbf{S}_{\mathbf{xy}}\mathbf{A}^T\mathbf{A}\mathbf{S}_{\mathbf{xy}}^T$. Generally, $\mathrm{span}(\widehat{\mathbf{w}}_1) \neq \mathrm{span}(\widetilde{\mathbf{w}}_1)$, so the estimates of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ differ, although $\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{xy}}^T$ and $\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}_{\mathbf{xy}}^T$ span the same subspace. It is customary in chemometrics to standardize the individual responses marginally $y_j \rightarrow y_j/\{\widehat{\mathrm{var}}(y_j)\}^{1/2}$, $j = 1, \ldots, r$, prior application of a multivariate PLS algorithm, but it is evidently not customary to standardize the responses jointly $\mathbf{y}_i \rightarrow \mathbf{z}_i = \mathbf{S}_{\mathbf{y}}^{-1/2}\mathbf{y}_i$. Of course, the SIMPLS algorithm could be applied after replacing $\mathbf{y}$ with jointly standardized responses $\mathbf{z}$, leading to a new variation on PLS methodology.

The methods also differ on how they utilize information from $\mathbf{S}_{\mathbf{x}}$. In the likelihood-based objective function, $L_1(\mathbf{G})$ guages how far $\mathrm{span}(\mathbf{G})$ is from subsets of m eigenvectors of $\mathbf{S}_{\mathbf{x}}$, but there is no corresponding operation in the SIMPLS method. The first SIMPLS vector $\widehat{\mathbf{w}}_1$ does not incorporate information about $\mathbf{S}_{\mathbf{x}}$. As indicated by the algorithm at the end of Section 4.3, the second SIMPLS vector incorporates $\mathbf{S}_{\mathbf{x}}$ by essentially removing the subspace $\mathrm{span}(\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{w}}_1)$ from consideration, but the choice of $\mathrm{span}(\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{w}}_1)$ is not guided by the relationship between $\widehat{\mathbf{w}}_1$ and the eigenvectors of $\mathbf{S}_{\mathbf{x}}$. Subsequent SIMPLS vectors operate similarly in successively smaller spaces. We have discovered empirically using cross validation that a single likelihood-based direction is often sufficient for prediction, while SIMPLS requires multiple directions to match its performance. These findings are illustrated in Section 5.

It is also noteworthy that the previous estimators are sequential and their computation is straightforward, but $\widehat{\mathbf{\Gamma}}$ requires full (non-sequential) optimization and its computation is more difficult, although we have not found it to be burdensome. On the other hand, sequential optimization can be notably less efficient than joint optimization and our experience is that the added effort in computing $\widehat{\mathbf{\Gamma}}$ is worthwhile (see Cook and Forzani (2010) for a related discussion of joint versus sequential optimization).

Finally, the likelihood-based estimation produces a full complement of estimators, for example $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} = \widehat{\mathbf{P}}_{\mathbf{\Gamma}}\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{P}}_{\mathbf{\Gamma}} + \widehat{\mathbf{Q}}_{\mathbf{\Gamma}}\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{Q}}_{\mathbf{\Gamma}}$, while the previous methods apparently do not.

4.5.2 Properties of the estimators

When $\mathbf{c}$ is distributed as a multivariate normal random vector, the estimators described previously inherit their properties from standard likelihood theory. Since we are requiring only a sample consisting of independent and identically distributed copies of $\mathbf{c}$ with finite fourth moments, we next present some first results in support of the estimators. We assumed an envelope structure with known m when forming the estimators. This structure always holds for some $1 \leq m \leq p$, and so it does not constitute a modeling constraint in the present context.

The next proposition shows that the envelope estimator is Fisher consistent and gives some alternative population versions of (14). It follows from this result that the estimators of the remaining parameters in $\mathbf{\Sigma}_{\mathbf{c}}$ are also Fisher consistent.

Proposition 4.2 *Assuming that $\mathbf{\Sigma}_{\mathbf{x}} > 0$, the envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ can be construced as*

$$\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \arg\min_{\mathcal{T}\in\mathcal{G}_{(m,p)}} \{\log|\mathbf{P}_{\mathcal{T}}(\mathbf{\Sigma}_{\mathbf{x}} - \mathbf{\Sigma}_{\mathbf{xz}}\mathbf{\Sigma}_{\mathbf{xz}}^T)\mathbf{P}_{\mathcal{T}}|_0 + \log|\mathbf{Q}_{\mathcal{T}}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{Q}_{\mathcal{T}}|_0\},$$

where $|\mathbf{A}|_0$ denotes the product of the non-zero eigenvalues of the symmetric matrix $\mathbf{A}$. A semi-orthogonal basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times m}$ for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ can be obtained as

$$\begin{aligned}\boldsymbol{\Gamma} &= \arg\min_{\mathbf{G}}\{\log|\mathbf{G}^T(\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{xz}}\boldsymbol{\Sigma}_{\mathbf{xz}}^T)\mathbf{G}| + \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{G}_0|\} \\ &= \arg\min_{\mathbf{G}}\{\log|\mathbf{G}^T(\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{xz}}\boldsymbol{\Sigma}_{\mathbf{xz}}^T)\mathbf{G}| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{G}|\},\end{aligned}$$

where $\min_{\mathbf{G}}$ is taken over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{p\times m}$ and $(\mathbf{G}, \mathbf{G}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix.

The next proposition addresses the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ given in (15). If a random vector $\mathbf{v}$ has the property that $\sqrt{n}(\mathbf{v} - \mathbf{b}) \rightarrow N(0, \mathbf{A})$ then we write $\mathrm{avar}(\sqrt{n}\mathbf{v}) = \mathbf{A}$ for its asymptotic covariance matrix.

Proposition 4.3 *Assume that $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are independent and identically distributed copies of $\mathbf{c}$ with finite fourth moments and assume that m is known. Then $\widehat{\boldsymbol{\beta}}$ as defined in (15) is a root-n consistent estimator of $\boldsymbol{\beta}$ and $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta})\}$ converges in distribution to a normal random vector with mean 0 and positive definite covariance matrix represented as $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$.*

The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ depends on fourth moments of $\mathbf{c}$ and seems quite complicated. The bootstrap is a useful option in practice for estimating the covariance matrix of $\widehat{\boldsymbol{\beta}}$. However, informative expressions for $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ are possible when $\mathbf{c}$ is normally distributed. Normality may be a useful context in some chemometrics applications, as we expect could be the case for the data on meat properties considered in Section 5.1. The next proposition gives a form for $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ when $\mathbf{c}$ is normal. In reference to model (2), let $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ be the envelope estimators of a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\boldsymbol{\beta}$ when $\boldsymbol{\alpha}$ is known, let $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Gamma}}$ denote the estimator of $\boldsymbol{\alpha}$ when $\boldsymbol{\Gamma}$ is known and recall that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ denotes the envelope estimator of $\boldsymbol{\beta}$ when $\boldsymbol{\Gamma}$ is known,

Proposition 4.4 *Assume that m is known and that $\mathbf{c}$ is normally distributed with mean $\boldsymbol{\mu}_{\mathbf{c}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{c}} > 0$. Then $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta})\}$ converges in distribution to a normal random vector with mean 0 and covariance matrix*

$$\begin{aligned}\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] &= \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}})] \\ &= \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + (\boldsymbol{\alpha}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\alpha} \otimes \boldsymbol{\Gamma}_0^T),\end{aligned}$$

where $\mathbf{M} = \boldsymbol{\alpha}\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}^{-1}\boldsymbol{\alpha}^T \otimes \boldsymbol{\Omega}_0 + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_m \otimes \mathbf{I}_{p-m}$. *Additionally,* $T_m = n(F_m(\mathbf{S}_{\mathbf{c}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{c}}) - F_m(\mathbf{S}_{\mathbf{c}}, \mathbf{S}_{\mathbf{c}}))$ *converges to a chi-squared random variable with* $(p-m)r$ *degrees of freedom, where* F_m *is as defined at the outset of Section 4.5.1.*

The statistic T_m described in this proposition can be used in a sequential manner to estimate m: Beginning with $m_0 = 0$ test the hypothesis $m = m_0$, terminating the first time it is not rejected. Otherwise, m_0 is incremented by one and then the hypothesis is tested again. The relative advantages of this versus cross validation have not been studied.

The decomposition of $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ shown in Proposition 4.4 has the same algebraic form as the decomposition found by CLC when pursuing reduction in the $\mathbf{y}$-dimension (see their Section 5.1 and Corollary 6.1), although the components of the decomposition of course differ. In particular, it follows that $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] \leq \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})]$, so the envelope estimator never does worse asymptotically than the OLS estimator. The first term in the decomposition of $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ can also be represented as

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] = (\mathbf{I}_r \otimes \boldsymbol{\Gamma})\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Gamma}})](\mathbf{I}_r \otimes \boldsymbol{\Gamma}^T) = \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T,$$

which corresponds to the first term of (4) in univariate regressions. The second term in the decomposition of $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$, which then represents the cost of estimating $\boldsymbol{\Gamma}$, can be rexpressed as $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}})] = (\boldsymbol{\alpha}^T \otimes \mathbf{Q}_{\boldsymbol{\Gamma}})\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}})](\boldsymbol{\alpha} \otimes \mathbf{Q}_{\boldsymbol{\Gamma}})$. We also see from these results that when performing a prediction at $\mathbf{z}_N = \mathbf{x}_N - \boldsymbol{\mu}_{\mathbf{x}}$ the asymptotic covariance $\mathrm{avar}(\sqrt{n}\mathbf{z}_N^T\widehat{\boldsymbol{\beta}})$ depends on the part $\boldsymbol{\Gamma}^T\mathbf{z}_N$ of $\mathbf{z}_N$ that lies in the envelope and on the part $\boldsymbol{\Gamma}_0^T\mathbf{z}_N$ that lies in the orthogonal complement, which is in contrast to the situation when $\boldsymbol{\Gamma}$ is known as discussed previously in conjunction with (4).

4.5.3 Comparisons with OLS

The following corollary to Proposition 4.4 describes $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ when $\boldsymbol{\Sigma}_{\mathbf{x}} = \sigma_{\mathbf{x}}^2\mathbf{I}_p$, and provides a comparison with the OLS estimator.

Corollary 4.1 *Assume the conditions of Proposition 4.4 and additionally that* $\boldsymbol{\Sigma}_{\mathbf{x}} = \sigma_{\mathbf{x}}^2\mathbf{I}_p$ *and that the coefficient matrix* $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ *has rank* r. *Then* $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] = \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})]$.

This corollary says that if there is no collinearity among homoscedastic predictors then the envelope and OLS estimators are asymptotically equivalent. Since this conclusion is based on maximum likelihood estimation, the performance of SIMPLS or other PLS estimators will also be no better asymptotically than OLS, a conclusion that seems at odds with some popular impressions. However, envelope and PLS estimators could still have small sample advantages over OLS, as mentioned previously during the discussion of (4).

To gain insights into the impact of predictor collinearity in a relatively simple context, consider a univariate regression ($r = 1$, $\boldsymbol{\alpha} \in \mathbb{R}^m$) with $\boldsymbol{\Omega} = \omega \mathbf{I}_m$ and $\boldsymbol{\Omega}_0 = \omega_0 \mathbf{I}_{p-m}$. Here the effects of collinearity will be manifested when ω_0 is small relative to ω. Define the signal-to-noise ratio $\tau = \|\boldsymbol{\alpha}\|/(\sigma_{y|\mathbf{x}}/\omega) = \|\boldsymbol{\sigma}_{\mathbf{x}y}\|/\sigma_{y|\mathbf{x}}$. We use the relative excess $R_{\mathrm{OLS}}(\tau, \omega, \omega_0)$ over the asymptotic covariance of the ideal estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ to compare the asymptotic covariances of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\boldsymbol{\beta}}$:

$$R_{\mathrm{OLS}}(\tau, \omega, \omega_0) = \frac{\mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\}}{\mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\}}.$$

The relative excess in the present context is then

Corollary 4.2 *Assume the conditions of Proposition 4.4 with $r = 1$, $\boldsymbol{\Omega} = \omega \mathbf{I}_m$ and $\boldsymbol{\Omega}_0 = \omega_0 \mathbf{I}_{p-m}$. Then*

$$R_{\mathrm{OLS}}(\tau, \omega, \omega_0) = \frac{\tau^2}{\tau^2 + (1 - \omega/\omega_0)^2}. \tag{16}$$

The relative behavior of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ then depends on the signal-to-noise ratio τ and on strength of the collinearity in $\boldsymbol{\Sigma}_{\mathbf{x}}$ as reflected by ω/ω_0. For a fixed τ, R will be small, and thus $\widehat{\boldsymbol{\beta}}$ will dominate $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, when ω/ω_0 is large. Depending on τ, $\widehat{\boldsymbol{\beta}}$ may also have some advantages over $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ when ω/ω_0 is small since then $R(\tau, \omega, \omega_0) \approx \tau^2/(\tau^2 + 1) < 1$. On the other hand, for a fixed level of collinearity, there is a signal τ large enough to make the estimators essentially equivalent.

These cases support a reoccurring thesis: The envelope estimator $\widehat{\boldsymbol{\beta}}$ will be superior to the OLS estimator when there is notable collinearity present in $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\mathrm{span}(\boldsymbol{\beta})$ lies substantially in a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{x}}$ that is associated with its larger eigenvalues. These types of regressions evidently occur frequently in chemometrics.

4.5.4 Comparisons with PLS

In this section we compare the envelope estimator of $\boldsymbol{\beta}$ to the PLS estimator in situations that allow a contrast with the results implied by Corollaries 4.1 and 4.2 where multivariate normality of $\mathbf{c}$ is assumed. Under normality the envelope estimator is the MLE and so will do no worse asymptotically than the PLS estimator. The results of this section may provide some intuition about the magnitude of the difference. We restrict attention to the relatively straightforward setting in which $r = 1$ and $m = 1$ since this is sufficient to allow informative comparisons. While more general results are possible, the level of complexity increases greatly when $r > 1$ and $m > 1$. The next proposition gives the basis for our comparisons.

Proposition 4.5 *Assume the representation of $\boldsymbol{\Sigma}_{\mathbf{c}}$ given in (13) with $r = 1$ and $m = 1$. Since $m = 1$ we use ω to represent $\boldsymbol{\Omega}$ as in Corollary 4.2. Then*

(i) The PLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ of $\boldsymbol{\beta}$ has the expansion

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} - \boldsymbol{\beta}) = n^{-1/2}\omega^{-1}\sum_{i=1}^{n}\{(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})\varepsilon_i + \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})^T\boldsymbol{\beta}\} + O_p(n^{-1/2}),$$

where ε is the error for model (1).

(ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} - \boldsymbol{\beta})$ is asymptotically normal with mean 0 and variance $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}) = \omega^{-2}\{\boldsymbol{\Sigma}_{\mathbf{x}}\sigma^2_{y|\mathbf{x}} + \mathrm{var}(\mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\boldsymbol{\beta})\}$.

(iii) If, in addition, $\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{x}$ is independent of $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{x}$ then $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}) = \omega^{-1}\sigma^2_{y|\mathbf{x}}\mathbf{P}_{\boldsymbol{\Gamma}} + \omega^{-2}\sigma^2_y\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$.

The results of parts (i) and (ii) show as expected that $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is asymptotically normal and that its asymptotic covariance depends on fourth moments of the marginal distribution of $\mathbf{x}$. However, if $\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{x}$ is independent of $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{x}$, as required in part (iii), then only second moments are needed. The condition of part (iii) is of course implied when $\mathbf{x}$ is normally distributed.

The asymptotic covariance in part (iii) of Proposition 4.5 can be expressed equivalently as

$$\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}) = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\{\sigma^2_y\boldsymbol{\Omega}_0^2/(\sigma^2_{y|\mathbf{x}}\omega^2) - \mathbf{I}_{p-1}\}\boldsymbol{\Gamma}_0^T\sigma^2_{y|\mathbf{x}}.$$

From this we see that the performance of PLS relative to OLS depends on the strength of the regression as measured by the ratio $\tau_1 = \sigma^2_{y|\mathbf{x}}/\sigma^2_y \leq 1$ and on the level of collinearity as measured by

$\boldsymbol{\Omega}_0^2/\omega^2$. For every level of collinearity there is a regression so that PLS does worse asymptotically than OLS and for every regression strength there is a level of collinearity so that PLS does better than OLS. For instance, if $\boldsymbol{\Sigma}_{\mathbf{x}} = \sigma^2_{\mathbf{x}}\mathbf{I}_p$ then $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}) = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0(\sigma^2_y - \sigma^2_{y|\mathbf{x}})$ and the asymptotic covariance of the PLS estimator is never less than that of the OLS estimator. In contrast, recall that the envelope estimator never does worse than OLS, $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) \leq \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$, with equality when $\boldsymbol{\Sigma}_{\mathbf{x}} = \sigma^2_{\mathbf{x}}\mathbf{I}_p$.

We compare the envelope and PLS estimators directly in the context of Corollary 4.2 with $m = 1$, so Proposition 4.5(iii) applies as well. Replacing the OLS estimator in (16) with the PLS estimator we find that the resulting relative excess R_{PLS} can be expressed informatively as $R_{\mathrm{PLS}}(\tau_1, \omega, \omega_0) = \tau_1(1-\tau_1)/\{(\tau_1 - \omega_0/\omega)^2 + \tau_1(1-\tau_1)\} \leq 1$. Again, we see that the relationship depends on the signal strength, now measured by τ_1, and the level of collinearity measured by ω_0/ω. The envelope estimator will tend to do much better than PLS in high and low signal regressions, $\tau_1 \to 0$ or $\tau_1 \to 1$ with ω_0/ω fixed. If there is a high level of collinearity and ω_0/ω is small relative to τ_1 then $R_{\mathrm{PLS}} \approx 1 - \tau_1$. If $\omega_0 = \omega$ then $R_{\mathrm{PLS}} = \tau_1$.

5 Numerical Illustrations

We conducted a variety of numerical studies to obtain a qualitative feeling for the relative performance of SIMPLS, OLS and likelihood-based envelopes. In our experience, the envelope prediction error is alway comparable to or smaller than the SIMPLS prediction error, while the performance of these methods relative to OLS depends on the signal strength and the relative magnitudes of the eigenvalues of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, as defined in (13). The eigenvalues in $\boldsymbol{\Omega}$ may be called the relevant eigenvalues, and the eigenvalues in $\boldsymbol{\Omega}_0$ the irrelevant eigenvalues. Let $\varphi_{\max}(\mathbf{A})$ and $\varphi_{\min}(\mathbf{A})$ denote the largest and smallest eigenvalues of the symmetric matrix $\mathbf{A}$. With a modest to strong signal, envelope prediction error was observed to be less than that for OLS when $\varphi_{\max}(\boldsymbol{\Omega}) \gg \varphi_{\max}(\boldsymbol{\Omega}_0)$, and substantially less when $\varphi_{\min}(\boldsymbol{\Omega}) \gg \varphi_{\max}(\boldsymbol{\Omega}_0)$. These empirical findings are in agreement with the indications given by (4) and Corollaries 4.1 and 4.2. Additionally, we found that envelope predictions with $m = 1$ typically outperform SIMPLS predictions regardless of the number of components used, which is in agreement with the discussion in Section 4.5.1.

In this section, we provide numerical examples to illustrate these qualitative conclusions. The

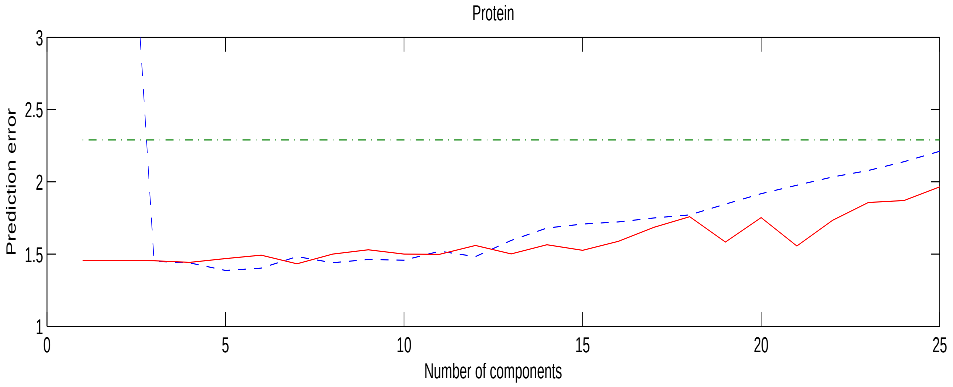

Figure 1: Protein prediction errors for the meat data: The solid line marks the envelope prediction error and the dashed marks the prediction error for SIMPLS. The horizontal dashed dotted line marks the constant prediction errors of OLS.

SIMPLS estimator was obtained with the MATLAB function *plsregress*. The Grassmann minimization needed for the envelope estimator was carried out by using Lippert's MATLAB package *sg_min* 2.4.1 (http://web.mit.edu/$\sim$ripper/www/sgmin.html). We used 5 fold cross validation to calculate the average squared prediction error $(y - \widehat{y})^2$, dividing the data into five parts of equal size. The reported error is then the average from prediction on each part while the remaining four parts were used as training set for estimation. Predictions based on the likelihood method discussed in Section 4.5 are called *envelope predictions* and the dimension of the fitted envelope is called the *number of components* to distinguish it from the true value m and to provide a connection with PLS terminology.

5.1 Meat properties

The meat data describes absorbance spectra from infrared transmittance for fat, protein and water in 103 meat samples. It was analyzed in Sæbø et al. (2007) as an example with collinearity and multiple relevant components for soft-threshold-PLS. We took spectral measurements at every fourth wavelength between 850 nm and 1050 nm as predictors, yielding $p = 50$. Prediction errors with between 1 and 35 components were computed by the 5 fold cross validation method previously described. For these data $\varphi_{\max}(\widehat{\mathbf{\Sigma}}_{\mathbf{x}})/\varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathbf{x}}) = 7.4 \times 10^8$ so there is a potential for PLS and envelope predictions to outperform OLS. We first predicted fat, protein and water in turn as univariate responses.

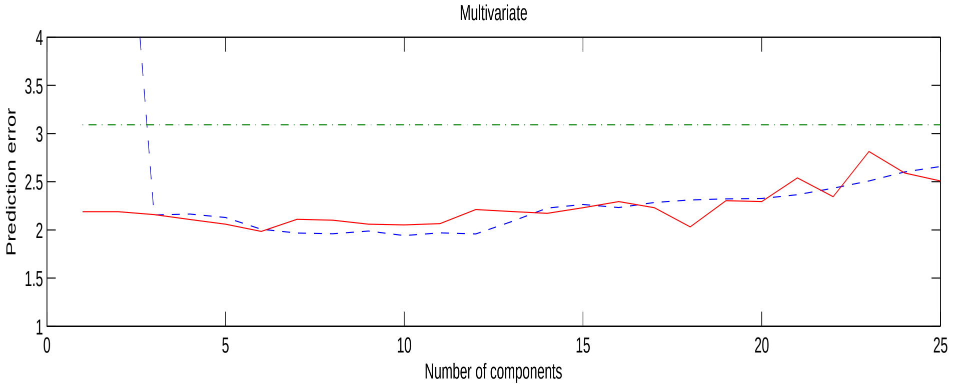

Figure 2: Prediction error $\|\hat{\mathbf{y}} - \mathbf{y}\|^2$ for the meat data with multivariate response. The line markers are the same as Figure 1.

The results for protein are summarized in Figure 1. We cut the y axis at 3 for visual clarity. With a single component, the prediction error of SIMPLS was around 6. With one and two components, the envelope predictor performed much better than SIMPLS and notably better than OLS. SIMPLS and envelope prediction performed about the same with 3 to 15 components, and all three prediction methods were essentially the same with more than 35 components. The results with water as the response were quite similar to those for protein. However, there was little to distinguish between the three prediction methods when using fat as the response. The identity was used to bind the elements of $\mathbf{y}$ when using fat, protein and water as a multivariate response. The prediction results for the multivariate response shown in Figure 2 are similar to those of Figure 1.

5.2 Simulations

In this section we use simulations to illustrate a range of behaviors beyond that shown with the meat data. For the first study we took $n = 100$ observations from a univariate regression with $p = 10$ predictors, an envelope dimension of $m = 8$, generating $\mathbf{c}$ as a multivariate normal vector with mean 0. We generated $\boldsymbol{\Sigma}_{\mathbf{x}}$ as $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} = 200\mathbf{I}_8$, $\boldsymbol{\Omega}_0 = 50\mathbf{I}_2$, $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was constructed by orthogonalizing a matrix of uniform $(0, 1)$ random variables, and $\boldsymbol{\beta}$ was then generated as $\boldsymbol{\Gamma}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{8\times 1}$ was generated a vector of uniform $(0, 1)$ random variables. Finally, we set $\sigma^2_{y|\mathbf{x}} = 0.74$. In this scenario, there is not an appreciable difference between the eigenvalue of $\boldsymbol{\Omega}$ and that of $\boldsymbol{\Omega}_0$ so, judging from Corollaries 4.1 and 4.2, we did not expect substantial differences between the envelope and OLS predictions. We had no conclusions

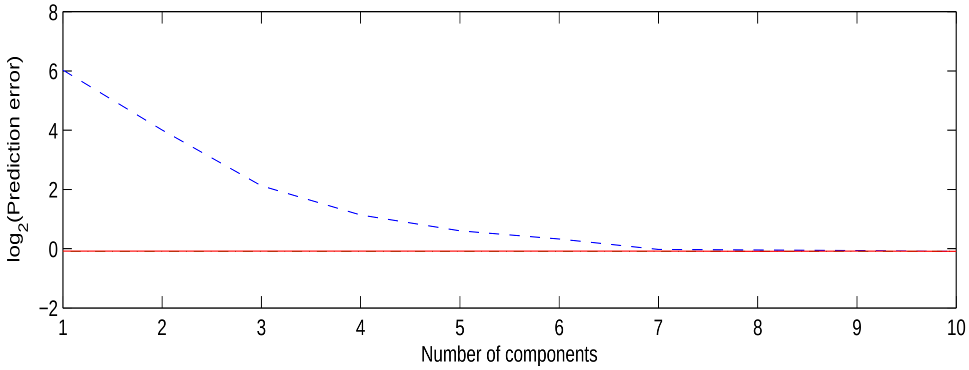

Figure 3: Prediction errors for simulation with $m = 8$. Dashed line gives the SIMPLS prediction error. The envelope and OLS prediction errors overlap at the horizontal line.

on which to base a prior expectation of the SIMPLS predictions. The results are shown in Figure 3. It turned out that the range of SIMPLS prediction errors was quite large. Instead of cutting the prediction error axis as we did previously, the base two logarithms of the prediction error are plotted in Figure 3. The envelope and OLS prediction errors are indistinguishable on the log scale and overlap at the horizontal line on the plot. The SIMPLS prediction error was significantly larger than that for the other two methods until 7 components was reached. Even with 4 or 5 components, the SIMPLS prediction error was about twice that for envelopes and OLS.

In the previous examples, SIMPLS and envelope predictions performed similarly with a sufficiently large number of components. In other regressions envelope prediction may be preferred to SIMPLS prediction regardless of the component number. To illustrate, we generated data following the general scheme we used previously for Figure 3 with a univariate response, 7 predictors, $m = 2$, $n = 60$ and $\sigma^2_{y|\mathbf{x}} = 0.03$. The eigenvalues of $\mathbf{\Omega}$ were 0.068 and 1.58, and $\mathbf{\Omega}_0$ had eigenvalues ranging from 2.9 to 583.9. The results are shown in Figure 4.

In the multivariate case, there are also situations in which envelope prediction is preferred over PLS and OLS regardless of the number of components. Using the previous simulation scheme, we simulated a dataset with 3 responses and 7 predictors. The eigenvalues of $\mathbf{\Omega}$ were 0.0720 and 0.6360, and $\mathbf{\Omega}_0$ had eigenvalues between 4.5 and 457.1. The results are displayed in Figure 5.

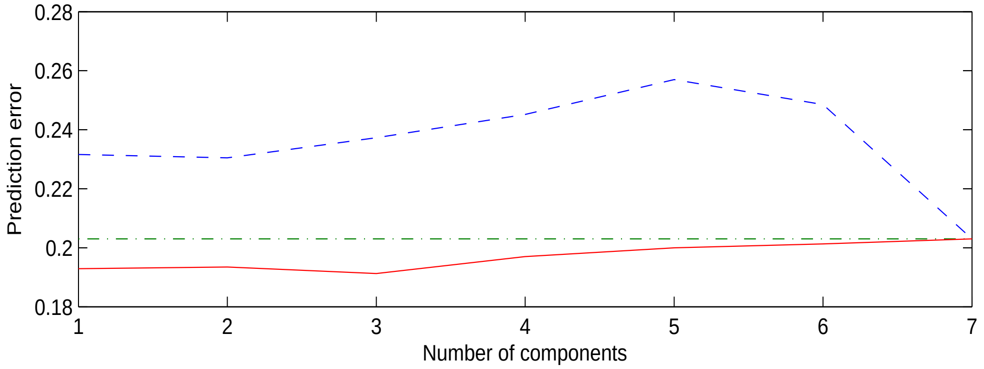

Figure 4: Simulation results on prediction errors of PLS and the envelope estimator: The solid line marks the envelope prediction error and the dashed marks the prediction error of SIMPLS. The horizontal dashed dotted line marks the constant prediction errors of OLS.

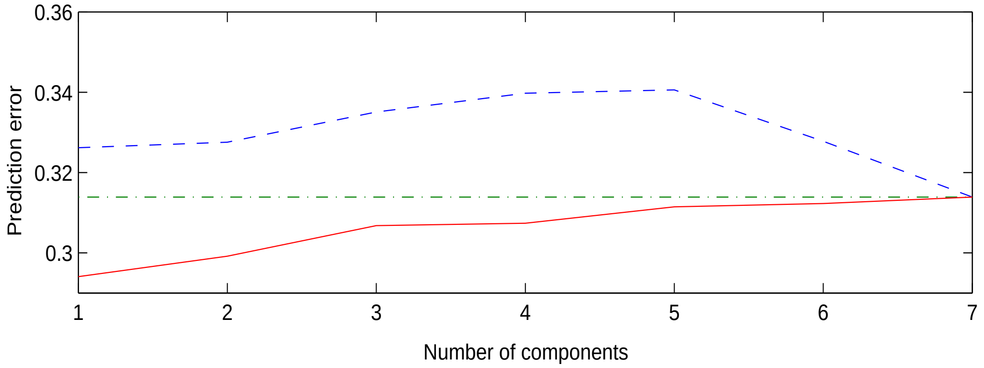

Figure 5: Simulation results on prediction errors of SIMPLS and the envelope estimator with multivariate response. The line markers are the same as Figure 4.

6 Discussion

Partial least squares originated as an algorithm for prediction in chemometrics. Its beginnings can be traced back to Herman Wold's general systems analysis method and much of the development has taken place in the Scandinavian countries. While SIMPLS has existed for decades and has been studied extensively, the conceptual apparatus needed to frame it as a Fisherian parameterization has apparently not existed until now: We have shown that the fundamental goal of SIMPLS is to estimate an envelope. This advance connects two different statistical cultures, allows for deeper understanding of SIMPLS and its properties and opens the door to methodological improvements through the pursuit of better envelope estimators. In addition to the model considered here, we expect that improvements in methodology will be possible in other contexts as well. For instance, Delaigle and Hall (2012) found that PLS does very well in classification problems for functional data and we conjecture that an extension of envelope methodology will offer gains over PLS.

As a general point, exploring the interrelationship between the concepts of different scientific cultures has an independent value. Such explorations may lead to the discovery of new underlying principles. A completely different - but conceptually related - area where this has recently been attempted, is a discussion of the an approach to quantum theory using conceptual variables - a notion generalizing the parameter concept of theoretical statistics; see Helland (2010) and Helland (2012 a,b).

Acknowledgements We are grateful to Trygve Almøy and Solve Sæbøe for providing the meat properties data, to Harald Martens for background on standard scaling methods in PLS, and to the referees for their helpful suggestions. This work was supported in part by grant DMS-1007547 from the U.S. National Science Foundation awarded to RDC, and by the Institute for Mathematical Sciences, National University of Singapore.

Appendix A: Proofs

Proposition 2.1. (a) is equivalent to $\mathrm{cov}(\mathbf{Q}_{\mathcal{S}}\mathbf{x}, \mathbf{P}_{\mathcal{S}}\mathbf{x}) = \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}_{\mathcal{S}} = \mathbf{0}$. The conclusion follows from Proposition 2.2 (a) and the representation $\boldsymbol{\Sigma}_{\mathbf{x}} = (\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}})$.

Model (1) can be written $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\beta}^T(\mathbf{P}_{\mathcal{S}}\mathbf{x} + \mathbf{Q}_{\mathcal{S}}\mathbf{x}) + \boldsymbol{\varepsilon}$, so (b) is equivalent to $\boldsymbol{\beta}^T\mathbf{Q}_{\mathcal{S}} = \mathbf{0}$, or

$\mathbf{Q}_{\mathcal{S}}\boldsymbol{\beta} = \mathbf{0}$, which is equivalent to $\boldsymbol{\beta} \in \mathcal{S}$.

Proposition 2.3. Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ have rows $(\mathbf{x}_i - \bar{\mathbf{x}})^T$. Then $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \mathrm{E}(\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}|\mathbb{X})) + \mathrm{var}(\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}|\mathbb{X})) = \mathrm{E}(\mathbb{X}^T\mathbb{X})^{-1}\boldsymbol{\sigma}^2_{y|\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathrm{E}(\mathbb{K}^{-1})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}$ where $\mathbb{K}$ is a Wishart matrix with covariance $\mathbf{I}_p$ and $n-1$ degrees of freedom. It follows from von Rosen (1988) that $\mathrm{E}(\mathbb{K}^{-1}) = \mathbf{I}_p/(n-p-2)$. Thus $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\sigma^2_{y|\mathbf{x}}/(n-p-2)$. Applying the same reasoning to model (2) we have $\mathrm{var}(\boldsymbol{\Gamma}\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}) = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\sigma^2_{y|\mathbf{x}}/(n-m-2)$. The final expression then follows from the envelope decomposition $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T$.

Proposition 3.1. In Helland (1988) it is proved for the sample PLS algorithm that $\widehat{y}_{A,PLS} = \bar{y} + \widehat{\boldsymbol{\beta}}^T_{A,PLS}(\mathbf{x} - \bar{\mathbf{x}})$. Here $\widehat{\boldsymbol{\beta}}_{A,PLS} = \widehat{\mathbf{W}}_A(\widehat{\mathbf{W}}_A^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{W}}_A)^{-1}\widehat{\mathbf{W}}_A^T\mathbf{s}_{\mathbf{x}y}$, where $\widehat{\mathbf{W}}_A = (\widehat{\mathbf{w}}_1, ..., \widehat{\mathbf{w}}_A)$ is found from the algorithm, or alternatively from the recurrence relation:

$$\widehat{\mathbf{w}}_{A+1} = \mathbf{s}_{\mathbf{x}y} - \mathbf{S}_{\mathbf{x}}\widehat{\mathbf{W}}_A(\widehat{\mathbf{W}}_A^T\mathbf{S}_{\mathbf{x}}\widehat{\mathbf{W}}_A)^{-1}\widehat{\mathbf{W}}_A^T\mathbf{s}_{\mathbf{x}y},$$

Let $n \to \infty$ in these relations.

Proposition 3.2. *Proof of (a):* Simple induction from (7) shows that $\mathbf{w}_1, ..., \mathbf{w}_A$ is a Gram-Schmidt orthogonalization of the Krylov sequence $\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, ..., \boldsymbol{\Sigma}_{\mathbf{x}}^{A-1}\boldsymbol{\sigma}_{\mathbf{x}y}$.

Proof of (b) and (c): $A = m+1$ is the first value for which $\mathbf{w}_1, ..., \mathbf{w}_A$ is linearly dependent. Then by a) it must also be the first value where the Krylov sequence is linearly dependent. This case can always be formulated such that the first member of the sequence is a linear combination of the rest or the last member of the sequence is a linear combination of the rest.

Proof of (d): We prove that the space $\mathcal{W}_m = \mathcal{K}_m$ is also spanned by the relevant eigenvectors $\boldsymbol{\ell}_1, ..., \boldsymbol{\ell}_m$, that is, the minimal set of eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{x}}$ for which $\boldsymbol{\ell}_a^T\boldsymbol{\sigma}_{\mathbf{x}y} \neq 0$. The word 'minimal' here points at the fact that when eigenvectors are multiple, one can always rotate in this subspace so that exactly one eigenvector in this space has a nontrivial component along $\boldsymbol{\sigma}_{\mathbf{x}y}$. Let ν_k be the eigenvalue corresponding to eigenvector $\boldsymbol{\ell}_k$.

We have: $\sum_{j=1}^A c_j\boldsymbol{\Sigma}_{\mathbf{x}}^{j-1}\boldsymbol{\sigma}_{\mathbf{x}y} = \sum_{k=1}^K \boldsymbol{\ell}_k\{\sum_{j=1}^A c_j(\nu_k)^{j-1}\boldsymbol{\ell}_k^T\boldsymbol{\sigma}_{\mathbf{x}y}\}$. This is $\mathbf{0}$ if and only if $\sum_{j=1}^A c_j(\nu_k)^{j-1} = 0$ for all k such that $\boldsymbol{\ell}_k^T\boldsymbol{\sigma}_{\mathbf{x}y} \neq 0$.

Let m' be the number of such ν_k, and look at the system above for $A = m'$. The determinant corresponding to this set of equations will be a Vandermonde determinant, and this determinant is non-zero if and only if $\nu_1, ..., \nu_{m'}$ are really different. It follows that $\mathcal{K}_{m'}$ is spanned by the eigenvectors $\boldsymbol{\ell}_k$ with different eigenvalues such that $\boldsymbol{\ell}_k^T \boldsymbol{\sigma}_{\mathbf{x}y} \neq 0$, and that $m' = m$ by (a) and (b).

Proof of Proposition 4.1. The following lemma will facilitate a demonstration that the SIMPLS sequence has property claimed.

Lemma 6.1 *Let $\mathbf{U} \in \mathbb{R}^{r \times r}$ be a positive semi-definite matrix and let $\mathbf{V} \in \mathbb{R}^{r \times r}$ be a symmetric positive definite matrix. Let $\mathcal{S}$ and $\mathcal{T}$ be orthogonal subspaces of $\mathbb{R}^r$. Then*

$$\begin{aligned} \mathbf{w}_{\max} &= \arg\max_{D_1} \mathbf{w}^T \mathbf{P}_{\mathcal{S}} \mathbf{U} \mathbf{P}_{\mathcal{S}} \mathbf{w} \\ &= \arg\max_{D_2} \mathbf{w}^T \mathbf{P}_{\mathcal{S}} \mathbf{U} \mathbf{P}_{\mathcal{S}} \mathbf{w} \\ &= \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \boldsymbol{\ell}_1 \{ (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \boldsymbol{\Gamma}^T \mathbf{U} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \}, \end{aligned}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$ is a semi-orthogonal basis matrix for $\mathcal{S}$, $\boldsymbol{\ell}_1(\mathbf{A})$ is any eigenvector in the first eigenspace of $\mathbf{A}$,

$$\begin{aligned} D_1 &= \{ \mathbf{w} | \mathbf{w} \in \mathcal{S} + \mathcal{T} \text{ and } \mathbf{w}^T \mathbf{P}_{\mathcal{S}} \mathbf{V} \mathbf{P}_{\mathcal{S}} \mathbf{w} + \mathbf{w}^T \mathbf{P}_{\mathcal{T}} \mathbf{V} \mathbf{P}_{\mathcal{T}} \mathbf{w} = 1 \}, \\ D_2 &= \{ \mathbf{w} | \mathbf{w} \in \mathcal{S} \text{ and } \mathbf{w}^T \mathbf{P}_{\mathcal{S}} \mathbf{V} \mathbf{P}_{\mathcal{S}} \mathbf{w} = 1 \}. \end{aligned}$$

Clearly, $\mathbf{w}_{\max} \in \mathcal{S}$, although it is not necessarily unique.

PROOF: Let $\boldsymbol{\Gamma}_1 \in \mathbb{R}^{r \times u_1}$ be a semi-orthogonal basis matrix for $\mathcal{T}$ so that $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_1) \in \mathbb{R}^{r \times (u+u_1)}$ is also semi-orthogonal; it will be orthogonal if $u + u_1 = r$. Let $\mathbf{s} = \boldsymbol{\Gamma}^T \mathbf{w}$ and $\mathbf{t} = \boldsymbol{\Gamma}_1^T \mathbf{w}$. Then since $\mathbf{w} \in \mathcal{S} + \mathcal{T}$ it can be expressed in the coordinates of $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_1)$ as $\mathbf{w} = \boldsymbol{\Gamma} \mathbf{s} + \boldsymbol{\Gamma}_1 \mathbf{t}$ and $\mathbf{w}_{\max} = \boldsymbol{\Gamma} \mathbf{s}_{\max} + \boldsymbol{\Gamma}_1 \mathbf{t}_{\max}$, where $\mathbf{s}_{\max} = \arg\max \mathbf{s}^T \boldsymbol{\Gamma}^T \mathbf{U} \boldsymbol{\Gamma} \mathbf{s}$ is now over all vectors $\mathbf{s} \in \mathbb{R}^u$ and $\mathbf{t} \in \mathbb{R}^{u_1}$ such that $\mathbf{s}^T \boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma} \mathbf{s} + \mathbf{t}^T \boldsymbol{\Gamma}_1^T \mathbf{V} \boldsymbol{\Gamma}_1 \mathbf{t} = 1$, and $(\mathbf{s}_{\max}, \mathbf{t}_{\max})$ is the pair of values at which the maximum occurs. Since $\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma} > 0$ we can make a change of variable in $\mathbf{s}$ without affecting $\mathbf{t}$. Let $\mathbf{d} = (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{1/2} \mathbf{s}$. Then $\mathbf{s}_{\max} = (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \mathbf{d}_{\max}$, where

$$\mathbf{d}_{\max} = \arg\max \mathbf{d}^T (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \boldsymbol{\Gamma}^T \mathbf{U} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma})^{-1/2} \mathbf{d}$$

and the maximum is over all vectors $\mathbf{d} \in \mathbb{R}^u$ and $\mathbf{t} \in \mathbb{R}^{u_1}$ such that $\mathbf{d}^T\mathbf{d} + \mathbf{t}^T\boldsymbol{\Gamma}_1^T\mathbf{V}\boldsymbol{\Gamma}_1\mathbf{t} = 1$. The conclusion follows since the maximum is achieved when $\mathbf{t} = 0$ and then $\mathbf{d}_{\max}$ is the first eigenvector of $(\boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma})^{-1/2}\boldsymbol{\Gamma}^T\mathbf{U}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma})^{-1/2}$ and $\mathbf{w}_{\max} = \boldsymbol{\Gamma}\mathbf{s}_{\max} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma})^{-1/2}\mathbf{d}_{\max}$.

The first step in proving Proposition 4.1 is to incorporate $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$ into the algorithm. For notational convenience we shorten $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$ to $\mathcal{E}$ when used as a subscript and set $\mathbf{U} = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{xy}}^T$. Since $\mathcal{S}_{\mathbf{xy}} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$, we have $\mathbf{w}^T\mathbf{U}\mathbf{w} = \mathbf{w}^T\mathbf{P}_\mathcal{E}\mathbf{U}\mathbf{P}_\mathcal{E}\mathbf{w}$. We know from Proposition 2.2(a) that $\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{P}_\mathcal{E} + \mathbf{Q}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{Q}_\mathcal{E}$. Consequently we have $\mathbf{w}^T\boldsymbol{\Sigma}_\mathbf{x}\mathbf{W}_k = 0$ if and only if $\mathbf{w}^T\mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{P}_\mathcal{E}\mathbf{W}_k + \mathbf{w}^T\mathbf{Q}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{Q}_\mathcal{E}\mathbf{W}_k = 0$. These considerations lead to the following equivalent statement of the algorithm. For $k = 0, 1, ..., m-1$,

$$\begin{aligned}
\mathbf{w}_{k+1} &= \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{P}_\mathcal{E}\mathbf{U}\mathbf{P}_\mathcal{E}\mathbf{w}, \text{ subject to} && (17)\\
&\quad\ \mathbf{w}^T\mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{P}_\mathcal{E}\mathbf{W}_k + \mathbf{w}^T\mathbf{Q}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{Q}_\mathcal{E}\mathbf{W}_k = 0 && (18)\\
&\quad\ \mathbf{w}^T\mathbf{P}_\mathcal{E}\mathbf{w} + \mathbf{w}^T\mathbf{Q}_\mathcal{E}\mathbf{w} = 1. && (19)
\end{aligned}$$

We next establish Proposition 4.1 by induction, starting with an analysis of (17) - (19) for $k = 0$.

First direction vector $\mathbf{w}_1$. For the first vector $\mathbf{w}_1$, only the length constraint (19) is active since $\mathbf{w}_0 = 0$. It follows from Lemma 6.1 with $\mathbf{V} = \mathbf{I}_p$ and $\mathcal{T} = \mathcal{S}^\perp$ that

$$\mathbf{w}_1 = \boldsymbol{\Gamma}\boldsymbol{\ell}_1(\boldsymbol{\Gamma}^T\mathbf{U}\boldsymbol{\Gamma}) = \boldsymbol{\ell}_1(\mathbf{P}_\mathcal{E}\mathbf{U}\mathbf{P}_\mathcal{E}) = \boldsymbol{\ell}_1(\mathbf{U}) \in \operatorname{span}(\boldsymbol{\Sigma}_\mathbf{x}),$$

where $\boldsymbol{\Gamma}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$. Clearly, $\mathbf{w}_1 \in \mathcal{S} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$, so trivially $\mathcal{W}_0 \subset \mathcal{W}_1 \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$ with equality if and only if $m = 1$.

Second direction vector $\mathbf{w}_2$. Next, assume that $m \geq 2$ and consider the second vector $\mathbf{w}_2$. In that case $\mathcal{W}_1 \subset \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{x}}(\mathcal{B})$ and so the second addend on the left of (18) is 0. Consequently,

$$\begin{aligned}
\mathbf{w}_2 &= \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{P}_\mathcal{E}\mathbf{U}\mathbf{P}_\mathcal{E}\mathbf{w}, \text{ subject to} && (20)\\
&\quad\ \mathbf{w}^T\mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{P}_\mathcal{E}\mathbf{w}_1 = 0 && (21)\\
&\quad\ \mathbf{w}^T\mathbf{P}_\mathcal{E}\mathbf{w} + \mathbf{w}^T\mathbf{Q}_\mathcal{E}\mathbf{w} = 1. && (22)
\end{aligned}$$

Condition (21) holds if and only if $\mathbf{w}$ is orthogonal to $\mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}_{\mathcal{E}}\mathbf{w}_1$. Letting $\mathcal{E}_1 = \mathrm{span}(\mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}_{\mathcal{E}}\mathbf{w}_1)$ for notational convenience, we require $\mathbf{w} \in \mathcal{E}_1^{\perp}$ which satisfies (21) by construction, leaving only the length constraint. The algorithm can be restated as

$$\mathbf{w}_2 \;=\; \arg\max_{\mathbf{w}\in\mathcal{E}_1^{\perp}} \mathbf{w}^T\mathbf{P}_{\mathcal{E}}\mathbf{U}\mathbf{P}_{\mathcal{E}}\mathbf{w}, \text{ subject to } \mathbf{w}^T\mathbf{P}_{\mathcal{E}}\mathbf{w} + \mathbf{w}^T\mathbf{Q}_{\mathcal{E}}\mathbf{w} = 1.$$

Let $\mathcal{D}_1 = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) \setminus \mathcal{E}_1$, which is the part of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ that is orthogonal to $\mathcal{E}_1$. Then $\mathbf{P}_{\mathcal{D}_1} + \mathbf{Q}_{\mathcal{E}} = \mathbf{P}_{\mathcal{E}} - \mathbf{P}_{\mathcal{E}_1} + \mathbf{Q}_{\mathcal{E}} = \mathbf{Q}_{\mathcal{E}_1}$. Consequently, we can rewrite the constraint $\mathbf{w} \in \mathcal{E}_1^{\perp}$ as $\mathbf{w} \in \mathcal{D}_1 + \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}^{\perp}(\mathcal{B})$, where $\mathcal{D}_1$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}^{\perp}(\mathcal{B})$ are orthogonal subspaces. Further, since $\mathbf{Q}_{\mathcal{E}_1}\mathbf{P}_{\mathcal{E}} = \mathbf{P}_{\mathcal{D}_1}$, it follows that $\mathbf{P}_{\mathcal{E}}\mathbf{w} = \mathbf{P}_{\mathcal{D}_1}\mathbf{w}$ for $\mathbf{w} \in \mathcal{E}_1^{\perp}$ and we can restate the algorithm as

$$\mathbf{w}_2 = \arg\max_{\mathbf{w}\in C} \mathbf{w}^T\mathbf{P}_{\mathcal{D}_1}\mathbf{U}\mathbf{P}_{\mathcal{D}_1}\mathbf{w}$$

where $C = \{\mathbf{w} | \mathbf{w} \in \mathcal{D}_1 + \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}^{\perp}(\mathcal{B}) \text{ and } \mathbf{w}^T\mathbf{P}_{\mathcal{D}_1}\mathbf{w} + \mathbf{w}^T\mathbf{Q}_{\mathcal{E}}\mathbf{w} = 1\}$. Let $\boldsymbol{\Gamma}_1$ be a semi-orthogonal basis matrix for $\mathcal{D}_1$. It follows from Lemma 6.1 with $\mathbf{V} = \mathbf{I}_p$, $\mathcal{S} = \mathcal{D}_1$ and $\mathcal{T} = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}^{\perp}(\mathcal{B})$ that

$$\mathbf{w}_2 = \boldsymbol{\Gamma}_1\boldsymbol{\ell}_1(\boldsymbol{\Gamma}_1^T\mathbf{U}\boldsymbol{\Gamma}_1) = \boldsymbol{\ell}_1(\mathbf{P}_{\mathcal{D}_1}\mathbf{U}\mathbf{P}_{\mathcal{D}_1}) = \boldsymbol{\ell}_1(\mathbf{Q}_{\mathcal{E}_1}\mathbf{U}\mathbf{Q}_{\mathcal{E}_1}) \in \mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{x}}).$$

In sum, $\mathbf{w}_1 \in \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, $\mathbf{w}_2 \in \mathcal{D}_1 = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) \setminus \mathcal{E}_1 \subset \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\mathbf{w}_1$ and $\mathbf{w}_2$ are linearly independent. Consequently, we have shown that $\mathcal{W}_0 \subset \mathcal{W}_1 \subset \mathcal{W}_2 \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$, with equality if and only if $m = 2$.

$(q+1)$-st direction vector $\mathbf{w}_{q+1}$, $q < m$. The reasoning here parallels that for $\mathbf{w}_2$ and is omitted. The process will continue until $q = m$, at which point $\mathcal{W}_m = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ and $\mathcal{D}_m$ is the origin; consequently the process must stop with no further change.

Proposition 4.2. The proof of this proposition makes use of the following two lemmas.

Lemma 6.2 *Suppose that $\mathbf{A} \in \mathbb{R}^{t\times t}$ is non-singular and that the column-partitioned matrix $(\mathbf{O}, \mathbf{O}_0) \in \mathbb{R}^{t\times t}$ is orthogonal. Then $|\mathbf{O}_0^T\mathbf{A}\mathbf{O}_0| = |\mathbf{A}| \times |\mathbf{O}^T\mathbf{A}^{-1}\mathbf{O}|$.*

PROOF. Define the $t \times t$ matrix

$$\mathbf{K} = \left(\begin{array}{c} \mathbf{I}_d, \mathbf{O}^T\mathbf{A}\mathbf{O}_0 \\ 0, \mathbf{O}_0^T\mathbf{A}\mathbf{O}_0 \end{array} \right).$$

Since $(\mathbf{O}, \mathbf{O}_0)$ is an orthogonal matrix,

$$\begin{aligned} |\mathbf{O}_0^T\mathbf{A}\mathbf{O}_0| &= |(\mathbf{O}, \mathbf{O}_0)\mathbf{K}(\mathbf{O}, \mathbf{O}_0)^T| = |\mathbf{O}\mathbf{O}^T + \mathbf{O}\mathbf{O}^T\mathbf{A}\mathbf{O}_0\mathbf{O}_0^T + \mathbf{O}_0\mathbf{O}_0^T\mathbf{A}\mathbf{O}_0\mathbf{O}_0^T| \\ &= |\mathbf{A} - (\mathbf{A} - \mathbf{I}_t)\mathbf{O}\mathbf{O}^T| = |\mathbf{A}||\mathbf{I}_d - \mathbf{O}^T(\mathbf{I}_t - \mathbf{A}^{-1})\mathbf{O}| \\ &= |\mathbf{A}||\mathbf{O}^T\mathbf{A}^{-1}\mathbf{O}|. \end{aligned}$$

Lemma 6.3 *Let* $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2) \in \mathbb{R}^{t \times t}$ *be a column partitioned orthogonal matrix and let* $\mathbf{A} \in \mathbb{R}^{t \times t}$ *be symmetric and positive definite. Then* $|\mathbf{A}| \leq |\mathbf{O}_1^T\mathbf{A}\mathbf{O}_1| \times |\mathbf{O}_2^T\mathbf{A}\mathbf{O}_2|$ *with equality if and only if* $\mathrm{span}(\mathbf{O}_1)$ *reduces* $\mathbf{A}$.

PROOF.

$$\begin{aligned} |\mathbf{A}| &= |\mathbf{O}^T\mathbf{A}\mathbf{O}| = \left| \begin{array}{cc} \mathbf{O}_1^T\mathbf{A}\mathbf{O}_1 & \mathbf{O}_1^T\mathbf{A}\mathbf{O}_2 \\ \mathbf{O}_2^T\mathbf{A}\mathbf{O}_1 & \mathbf{O}_2^T\mathbf{A}\mathbf{O}_2 \end{array} \right| \\ &= |\mathbf{O}_1^T\mathbf{A}\mathbf{O}_1| \times |\mathbf{O}_2^T\mathbf{A}\mathbf{O}_2 - \mathbf{O}_2^T\mathbf{A}\mathbf{O}_1(\mathbf{O}_1^T\mathbf{A}\mathbf{O}_1)^{-1}\mathbf{O}_1^T\mathbf{A}\mathbf{O}_2| \\ &\leq |\mathbf{O}_1^T\mathbf{A}\mathbf{O}_1| \times |\mathbf{O}_2^T\mathbf{A}\mathbf{O}_2|. \end{aligned}$$

To prove Proposition 4.2, let $\mathbf{G}$ be a semi-orthogonal basis matrix for $\mathcal{T}$ and let $(\mathbf{G}, \mathbf{G}_0)$ be orthogonal. Additionally, to simplify notation let $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}^T$ and $\mathbf{U} = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}^T$, so that $\mathcal{S}_{\mathbf{xy}} = \mathrm{span}(\mathbf{U})$. Then

$$|\mathbf{P}_{\mathcal{T}}\mathbf{M}\mathbf{P}_{\mathcal{T}}|_0 = \left| (\mathbf{G}, \mathbf{G}_0) \left(\begin{array}{cc} \mathbf{G}^T\mathbf{M}\mathbf{G} & 0 \\ 0 & 0 \end{array} \right) (\mathbf{G}, \mathbf{G}_0)^T \right|_0 = |\mathbf{G}^T\mathbf{M}\mathbf{G}|.$$

Consequently, we can work in terms of bases without loss of generality. Now,

$$\begin{aligned}\log|\mathbf{G}^T\mathbf{M}\mathbf{G}| + \log|\mathbf{G}_0^T(\mathbf{M}+\mathbf{U})\mathbf{G}_0| &= \log|\mathbf{G}^T\mathbf{M}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{M}\mathbf{G}_0 + \mathbf{G}_0^T\mathbf{U}\mathbf{G}_0| \\ &\geq \log|\mathbf{G}^T\mathbf{M}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{M}\mathbf{G}_0| \\ &\geq \log|\mathbf{M}|,\end{aligned}$$

where the second inequality follows from Lemma 6.3. To achieve the lower bound, the second inequality requires that $\mathrm{span}(\mathbf{G})$ reduce $\mathbf{M}$ (cf. Proposition 2.2), while the first inequality requires that $\mathrm{span}(\mathbf{U}) \subseteq \mathrm{span}(\mathbf{G})$. The first representation for $\boldsymbol{\Gamma}$ follows since m is the dimension of the smallest subspace that satisfies these two properties. The second representation for $\boldsymbol{\Gamma}$ follows immediately from Lemma 6.2.

Proposition 4.3. The justification of this proposition involves application of Shapiro's (1986) results on the asymptotic behavior of overparameterized structural models. The shifted objective function $G(\mathbf{S_c}, \boldsymbol{\Sigma}_\mathbf{c}) = F(\mathbf{S_c}, \boldsymbol{\Sigma}_\mathbf{c}) - F(\mathbf{S_c}, \mathbf{S_c})$, is non-zero, twice continuously differentiable in $\mathbf{S_c}$ and $\boldsymbol{\Sigma}_\mathbf{c}$ and is equal to 0 if and only if $\boldsymbol{\Sigma}_\mathbf{c} = \mathbf{S_c}$. Additionally, $\sqrt{n}(\mathrm{vech}(\mathbf{S_c}) - \mathrm{vech}(\boldsymbol{\Sigma}_\mathbf{c}))$ is asymptotically normal, where 'vech' denotes the vector-half operator. These conditions plus Proposition 4.2 and a some minor technical restrictions enable us to apply Shapiro's Propositions 3.1 and 4.1, from which the conclusions can be shown to follow.

Proposition 4.4. The derivation of the results in this proposition is rather long so here we give only a sketch of the main ideas. We assume without loss of generality that $\boldsymbol{\mu}_\mathbf{x} = 0$. Let $\mathrm{vec} : \mathbb{R}^{p\times q} \to \mathbb{R}^{pq}$ denote the operator that maps a matrix to a vector by stacking its columns and let $\mathrm{vech} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p(p+1)/2}$ denote the vector-half operator that maps a symmetric matrix to a vector by stacking the unique elements of each column on and below the diagonal. The operators vec and vech are related through the expansion matrix $\mathbf{E}_p \in \mathbb{R}^{p^2\times p(p+1)/2}$ and the contraction matrix $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2\times p^2}$: For any symmetric matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_p\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_p\mathrm{vech}(\mathbf{A})$.

The multivariate normal density for the concatenate variable $\mathbf{c}$ can be represented uniquely as the product of the conditional normal density of $\mathbf{y}|\mathbf{x}$ and the marginal normal density of $\mathbf{x}$: $\mathbf{y}|\mathbf{x} \sim N_r(\boldsymbol{\mu} + \boldsymbol{\beta}^T\mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}})$ and $\mathbf{x} \sim N(0, \boldsymbol{\Sigma}_\mathbf{x})$. The envelope model is then introduced by setting

$\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0$. The six parameters of the envelope model are then

$$\begin{aligned} \boldsymbol{\phi} &= \{\boldsymbol{\mu}^T, \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}), \mathrm{vec}^T(\boldsymbol{\alpha}), \mathrm{vec}^T(\boldsymbol{\Gamma}), \mathrm{vech}^T(\boldsymbol{\Omega}), \mathrm{vech}^T(\boldsymbol{\Omega}_0)\}^T \\ &\equiv (\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T, \boldsymbol{\phi}_3^T, \boldsymbol{\phi}_4^T, \boldsymbol{\phi}_5^T, \boldsymbol{\phi}_6^T)^T, \end{aligned}$$

and the estimable functions under the envelope model correspond to the parameters in the unconstrained normal model:

$$h(\boldsymbol{\phi}) = \{\boldsymbol{\mu}^T, \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}), \mathrm{vec}^T(\boldsymbol{\beta}), \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{x}})\}^T \equiv (h_1^T(\boldsymbol{\phi}), h_2^T(\boldsymbol{\phi}), h_3^T(\boldsymbol{\phi}), h_4^T(\boldsymbol{\phi}))^T.$$

The asymptotic covariance of $h(\widehat{\boldsymbol{\phi}})$ can then be expressed as $\mathrm{avar}[\sqrt{n}h(\widehat{\boldsymbol{\phi}})] = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T$, where $\mathbf{H} = (\partial h_i/\partial \boldsymbol{\phi}_j)_{i=1,\cdots,4, j=1,\cdots,6}$ is the gradient matrix, $\mathbf{J}$ is the Fisher information matrix for the unconstrained normal model and † denotes the Moore-Penrose generalized inverse. Partitioning $\mathbf{H}$ on its row blocks for corresponding to $(h_1^T(\boldsymbol{\phi}), h_2^T(\boldsymbol{\phi}))^T$ and $(h_3^T(\boldsymbol{\phi}), h_4^T(\boldsymbol{\phi}))^T$, and on its column blocks for $(\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)^T$ and $(\boldsymbol{\phi}_3^T, \ldots, \boldsymbol{\phi}_6^T)^T$, we find that

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{r+r(r+1)/2} & 0 \\ 0 & \mathbf{H}_{22} \end{pmatrix}$$

where

$$\mathbf{H}_{22} = \begin{pmatrix} \mathbf{I}_r \otimes \boldsymbol{\Gamma} & \boldsymbol{\alpha}^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_m & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-m} \end{pmatrix}.$$

The Fisher information $\mathbf{J}$ is a block diagonal matrix with lower right block $\mathbf{J}_{22}$ for $(h_3^T(\boldsymbol{\phi}), h_4^T(\boldsymbol{\phi}))^T$ being

$$\mathbf{J}_{22} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{x}} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{x}}^{-1})\mathbf{E}_p \end{pmatrix}.$$

It follows that $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ it is then the upper left $rp \times rp$ block of $\mathbf{H}_{22}(\mathbf{H}_{22}^T\mathbf{J}_{22}\mathbf{H}_{22})^{\dagger}\mathbf{H}_{22}$. The matrices $\mathbf{H}_{22}$ and $\mathbf{J}_{22}$ are of the same algebraic form as those encountered by CLC and so the rest of the derivation parallels closely the steps in their analysis.

Corollary 4.1. In this corollary we have $\boldsymbol{\Sigma}_{\mathbf{x}} = \sigma_{\mathbf{x}}^2\mathbf{I}_p$ and consequently $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B}) = \mathcal{B}$. We can therefore define $\boldsymbol{\Gamma} = \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1/2}$ and $\boldsymbol{\alpha} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2}$. Then $\mathbf{M} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2}\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}}^{-1}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2} \otimes \mathbf{I}_{p-m}\sigma_{\mathbf{x}}^2$, and $(\boldsymbol{\alpha}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\alpha} \otimes \boldsymbol{\Gamma}_0^T) = \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\sigma_{\mathbf{x}}^{-2}$. The conclusion then follows by adding $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \otimes \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\sigma_{\mathbf{x}}^{-2}$ and $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\sigma_{\mathbf{x}}^{-2}$.

Corollary 4.2. The conclusion follows algebraically with $\boldsymbol{\Omega} = \omega\mathbf{I}_m$, $\boldsymbol{\Omega}_0 = \omega_0\mathbf{I}_{p-m}$ and $r = 1$. Here we give a few intermediate quantities: Let $T = \omega/\omega_0 + \omega_0/\omega - 2 = (\omega_0/\omega)(1 - \omega/\omega_0)^2$. Then $\mathbf{M}^{-1} = (T\mathbf{I}_m + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\omega_0/\sigma_{y|\mathbf{x}})^{-1} \otimes \mathbf{I}_{p-m}$, and $(\boldsymbol{\alpha}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\alpha} \otimes \boldsymbol{\Gamma}_0^T) = \boldsymbol{\alpha}^T(T\mathbf{I}_m + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\omega_0/\sigma_{y|\mathbf{x}}^2)^{-1}\boldsymbol{\alpha} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$. Consequently,

$$\begin{aligned}
\mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}})\} - \mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} &= \mathrm{trace}\{(\boldsymbol{\alpha}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\alpha} \otimes \boldsymbol{\Gamma}_0^T)\} \\
&= \frac{\sigma_{y|\mathbf{x}}^2}{\omega_0} \times \frac{\tau^2(p-m)}{\tau^2 + (1 - \omega/\omega_0)^2}.
\end{aligned}$$

The conclusion follows since

$$\mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\} - \mathrm{trace}\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} = \frac{\sigma_{y|\mathbf{x}}^2(p-m)}{\omega_0}.$$

Proposition 4.5. Full details for this proposition are rather lengthy, so here we only state key steps. Without loss of generality we assume that $\mathbf{x}$ and y have mean zero. We need to find the expansion for $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} - \boldsymbol{\beta})$ where $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} = \widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|^2(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y})^{-1}$. We first expand the factors $\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|^2$ and $(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y})^{-1}$, leading to

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}\|^2 - \boldsymbol{\sigma}_{\mathbf{x}y}\|\boldsymbol{\sigma}_{\mathbf{x}y}\|^2) &= \|\boldsymbol{\sigma}_{\mathbf{x}y}\|\{\mathbf{I}_p + 2\mathbf{P}_{\boldsymbol{\Gamma}}\}n^{-\frac{1}{2}}\sum_{i=1}^{n}(\mathbf{x}_i y_i - \boldsymbol{\sigma}_{\mathbf{x}y}) + O_p(n^{-1/2}) \\
\sqrt{n}\{(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y})^{-1} - (\boldsymbol{\sigma}_{\mathbf{x}y}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y})^{-1}\} &= -\sqrt{n}(\boldsymbol{\sigma}_{\mathbf{x}y}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y})^{-2}\boldsymbol{\sigma}_{\mathbf{x}y}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}})\boldsymbol{\sigma}_{\mathbf{x}y} \\
&\quad -2\sqrt{n}(\boldsymbol{\sigma}_{\mathbf{x}y}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y})^{-2}\boldsymbol{\sigma}_{\mathbf{x}y}^T\boldsymbol{\Sigma}_{\mathbf{x}}(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y} - \boldsymbol{\sigma}_{\mathbf{x}y}) + O_p(n^{-1/2})
\end{aligned}$$

Substituting these into $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ we obtain

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} - \boldsymbol{\beta}) &= \{(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-1}(\mathbf{I}_p + 2\mathbf{P}_{\boldsymbol{\Gamma}}) - 2(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-2}\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}_{\mathbf{x}}\}\sqrt{n}(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y} - \boldsymbol{\sigma}_{\mathbf{x}y}) \\
&\quad -(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma})^{-2}\mathbf{P}_{\boldsymbol{\Gamma}}\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}})\boldsymbol{\sigma}_{\mathbf{x}y} + O_p(n^{-1/2}).
\end{aligned}$$

Substituting the expansions

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\sigma}}_{\mathbf{x}y} - \boldsymbol{\sigma}_{\mathbf{x}y}) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} (\mathbf{x}_i y_i - \boldsymbol{\sigma}_{\mathbf{x}y}) + O_p(n^{-\frac{1}{2}}), \\ \sqrt{n}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}}) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} (\mathbf{x}_i \mathbf{x}_i^T - \boldsymbol{\Sigma}_{\mathbf{x}}) + O_p(n^{-\frac{1}{2}}).\end{aligned}$$

and simplifying leads to the stated results.

Appendix B: Model reduction under symmetry

The conditions of Section 2.1 describe a setting in which we expect PLS and envelope estimation to result in improved prediction. However, a subspace $\mathcal{S}$ that satisfies those conditions is not invariant or equivariant under all linear transformations of $\mathbf{x}$. Similarly, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{x}}}(\mathcal{B})$ does not transform equivariantly for all linear transformations of $\mathbf{x}$, although it does so for symmetric linear transformations that commute with $\boldsymbol{\Sigma}_{\mathbf{x}}$ (CLC, Prop. 2.4). This raises a general question about the kinds of regressions that are logically amenable to reduction of $\mathbf{x}$ via envelopes. We address this by introducing a group of transformations, first giving a little background on this approach generally.

A parametric inference problem related with parameter θ may often have some symmetry property imposed or associated with the corresponding model. Such structure can be formalized by introducing a group G of transformations acting upon the parameter space Θ. When θ is transformed by the group and the observations are transformed accordingly, one should get equivalent results from the statistical analysis. Now fix a point θ_0 in the parameter space Θ. An orbit in this space under G is the set of points of the form $g\theta_0$ as g varies over the group G. The different orbits are disjoint, and θ_0 can be replaced by any parameter on the orbit. Any set in Θ which is an orbit of G or can be written as a union of orbits, is an invariant set under G in Θ, and conversely, all invariant sets can be written in this way. When considering a model reduction that takes the form of a reduction of the parameter space Θ, the parts of Θ that are essential for the inference sought should be retained, but irrelevant parts should be left out. If there is a group G acting upon the parameter space, any model reduction should be to an orbit or to a set of orbits of G. This criterion ensures that G also can be seen as a group acting upon the new parameter space.

To apply these ideas in the context of PLS and envelopes, consider the random $\mathbf{x}$ regression model (1) with centered predictors. Then the regression vector $\boldsymbol{\beta}$ can always be represented as $\boldsymbol{\beta} = \sum_{i=1}^{p} \boldsymbol{\ell}_i \boldsymbol{\gamma}_i^T$ for some vectors $\boldsymbol{\gamma}_i \in \mathbb{R}^r$, where the $\boldsymbol{\ell}_i$'s are eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Now introduce the group G consisting of combinations of the following transformations: (1) rotations in predictor space and hence of $\boldsymbol{\ell}_1, ..., \boldsymbol{\ell}_p$, and (2) separate linear transformations of the $\boldsymbol{\gamma}_i$, i.e., $\boldsymbol{\gamma}_i \rightarrow \mathbf{A}_i \boldsymbol{\gamma}_i$ with $\det(\mathbf{A}_i) \neq 0$. For the group acting on a single $\boldsymbol{\gamma}$-vector by $\boldsymbol{\gamma} \rightarrow \mathbf{A}\boldsymbol{\gamma}$ ($\det(\mathbf{A}) \neq 0$), there are two orbits: $\boldsymbol{\gamma} = \mathbf{0}$ and $\{\boldsymbol{\gamma} : \boldsymbol{\gamma} \neq \mathbf{0}\}$. From this it follows that the orbits of G are indexed by m and are given by $\boldsymbol{\beta} = \sum_{i=1}^{m} \boldsymbol{\ell}_i \boldsymbol{\gamma}_i^T$ with all $\boldsymbol{\gamma}_i \neq \mathbf{0}$. Note again that it is the number m of terms which characterizes this model. This is an alternative way to characterize the projection of $\boldsymbol{\beta}$ into the envelope space of dimension m, a fact which again can be proved from Proposition 2.2 (c). This characterization was used to do Bayesian estimation/ nearly best equivariant estimation under G in Helland et al (2012). Most importantly, the result here suggests that PLS/envelope approach to linear regression may be most effective when the predictors are standardized in some way or are dimensionally homogenous, as is the case in chemometrics applications when the predictors are spectral intensities at selected wave lengths.

Appendix C: Background on Grassmann optimization

In this appendix we give a little background on Grassmann optimization, mainly to aid intuition. The theoretical basis for the basic algorithm discussed here comes primarily from Liu, et al., (2004); see also Edelman, et al., (1998) and the documentation that comes with Lippert's MATLAB package *sg_min* 2.4.1 (http://web.mit.edu/∼ripper/www/sgmin.html).

Perhaps the most common type of optimization algorithm is based on additively updating a starting value, as in Gauss-Newton iteration. However, additive updates are not generally useful for Grassmann optimization of an objective function $L(\mathbf{G})$, $\mathbf{G} \in \mathbb{R}^{p \times u}$, since additively adjusting an orthogonal starting basis will not result in an appropriate update. Let $\mathbf{W}_i = (\mathbf{G}_i, \mathbf{G}_{i,0})$ be an orthogonal basis for $\mathbb{R}^p$ at the i-th iteration, where $\mathbf{G}_i$ is the current approximation of the optimum value and the starting basis is indicated with $i = 1$. A basic Grassmann algorithm proceeds by orthogonally adjusting $\mathbf{W}_i$: $\mathbf{W}_{i+1} = \mathbf{W}_i \mathbf{O}_{i+1}$, $i = 1, 2, \ldots$, continuing until a stopping criterion is met. The orthogonal matrix $\mathbf{O}_{i+1}$ for the $(i+1)$-st iteration depends on the first derivative $\mathbf{B}_i = \{\nabla L(\mathbf{G})\}^T \mathbf{G}_0$ of the objective function evaluated at $\mathbf{W}_i$ and taken on the manifold. It has the specific form $\mathbf{O}_{i+1} = \exp\{\delta_i \mathbf{A}(\mathbf{B}_i)\}$, where $\exp(\cdot)$ denotes the matrix exponential, δ_i is the step size for the i-th iteration, and $\mathbf{A}(\mathbf{B}_i)$ is the skew-symmetric matrix

$$\mathbf{A}(\mathbf{B}) = \begin{pmatrix} 0_{u\times u} & \mathbf{B}_{u\times(p-u)} \\ -\mathbf{B}^T_{(p-u)\times u} & 0_{(p-u)\times(p-u)} \end{pmatrix} \in \mathbb{R}^{p\times p}.$$

There are a variety of algorithms for Grassmann optimization available, most going well beyond the basic algorithm described here. For example, Lippert's *sg_min* algorithm uses first and second derivatives of the objective function.

To illustrate the behavior of Lippert's algorithm, we give an example on the usage of *sg_min* 2.4.1. Suppose we want to perform the minimization in (14). The derivative of $L(\mathbf{G})$ is $dL(\mathbf{G})/d\mathbf{G} = 2(\mathbf{S_x} - \mathbf{S_{xz}}\mathbf{S}_{\mathbf{xz}}^T)\mathbf{G}\{\mathbf{G}^T(\mathbf{S_x} - \mathbf{S_{xz}}\mathbf{S}_{\mathbf{xz}}^T)\mathbf{G}\}^{-1} + 2\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{G})^{-1}$. We define two MATLAB functions `F.m` and `dF.m` containing the objective function and its derivative. Then $\widehat{\mathbf{G}}$ can be obtained by

```
[Lmin Ghat] = sg_min(Ginit)
```

where `Ginit` is a p by u full rank matrix containing the starting value, `Lmin` is the minimized $L(\mathbf{G})$ and `Ghat` returns $\widehat{\mathbf{G}}$, the orthogonal basis of the subspace that minimizes $L(\mathbf{G})$. Random starting values should be avoided because it makes the optimization process very likely to get trapped in local minima. By Small et al. (2000), using any root-n consistent estimator as the starting value will give an estimator that is asymptotically equivalent to the maximum likelihood estimator.

References

Aldrich, J. (2005). Fisher and regression. *Statistical Science* **20**, 401–417.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 199–227.

Boulesteix, A.-L. and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* **7**, 32–44.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society B*, **72**, 3–25.

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

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

Cook, R.D. and Forzani, L. (2010). Letter to the editor. *Journal of the American Statistical Association*, **105**, 881.

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

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society B*, **74**, 267–286.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251–263.

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

Frank, I.E. and Friedman (1993), A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

Helland, I.S. (1988). On the structure of partial least squares regression. *Commun. Statist. Simula.***17**, 581–607.

Helland, I.S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* **17**, 97–114.

Helland, I.S. (1992). Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society B* **54**, 637–647.

Helland, I.S. (2001). Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **58**, 97–107.

Helland, I.S. (2004). Statistical inference under symmetry. *International Statistical Review* **72**, 409-422.

Helland, I.S. (2010). *Steps Towards a Unity of Scientific Models and Methods.* Singapore: Worlds Scientific.

Helland, I.S. (2012a). *A Unified Scientific Basis for Inference*. Submitted. arXiv: 1206.5075.

Helland, I.S. (2012b). A basis for statistical teory and quantum theory. In: Lazinica, A. [Ed.] *Quantum Mechanics.* Rijeka, Croatia: InTech.

Helland, I.S., Sæbø, S. and Tjelmeland, H. (2012). Near optimal prediction from relevant components. *Scandinavian Journal of Statistics* **39**, 695-713.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* New York: Springer.

Levina, E., Rothman, A. J., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics* **2**, 245–263.

Liu, X., Srivistava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 662–666.

Martens, H. and Næs, T. (1989). *Multivariate Calibration.* New York: Wiley.

Nadler, B. and Coifman, R. R. (2005). The prediction error in cls and pls: the importance of feature selection prior to multivariate calibration. *Journal of Chemometrics* **19**, 107–118.

Næs, T. and Helland, I.S. (1993). Relevant components in regression. *Scandinavian Journal of Statistics* **20**, 239–250.

Rothman, A. J., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81**, 142–149.

Small, C.G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statistical Science* **15**, 313–341.

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

Su, Z. and Cook, R. D. (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika* **99**, 687–702.

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