Envelope Models and Methods

Dimension Reduction for Efficient Estimation in Multivariate Statistics

R. Dennis Cook
School of Statistics
University of Minnesota
Minneapolis, MN 55455, U.S.A.

October 9, 2013

Contents

Chapter 4

Envelopes: Reducing the predictors

In Chapter 2 we considered reduction of $\mathbf{Y}$, relying on the notion of material and immaterial information for motivation and using $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ with $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ as the reduction construct. In this section we consider reducing the dimensionality of $\mathbf{X}$, again using envelopes as the essential device.

4.1 Model formulation

Again consider the multivariate linear model (2.1), but now allowing the predictors to be stochastic. Restating it for ease of reference,

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon}, \tag{4.1}$$

where the error vector $\boldsymbol{\varepsilon}$ has mean 0 and covariance matrix $\mathbf{\Sigma}$, the random predictor vector $\mathbf{X}$ has mean $\boldsymbol{\mu}_{\mathbf{X}}$ and variance $\mathbf{\Sigma}_{\mathbf{X}}$, and $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}$. Given n independent copies $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, of $(\mathbf{Y}, \mathbf{X})$, consider regressing $\mathbf{Y}$ on $\mathbf{X}$ in two steps. The first is the reduction step: reduce $\mathbf{X}$ linearly to $\mathbf{\Phi}^T\mathbf{X}$ using some methodology that produces $\mathbf{\Phi} \in \mathbb{R}^{p\times q}$, $q \le p$. The matrix $\mathbf{\Phi}$ will typically depend on the data, although we temporarily assume that $\mathbf{\Phi}$ is non-stochastic to facilitate discussion. The second step consists of using ordinary least squares to estimate the coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ of $\mathbf{X}$. In preparation for describing the form of the resulting estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{\Phi}}$ of $\boldsymbol{\beta}$, recall that $\mathbf{S}_{\mathbf{X}}$ denotes the sample version of $\mathbf{\Sigma}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{XY}}$ denotes the sample version of $\mathbf{\Sigma}_{\mathbf{XY}}$, $\mathbf{S}_{\mathbf{XY}}^T = \mathbf{S}_{\mathbf{YX}}$ and $\mathbf{B} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}$ is the estimator of $\boldsymbol{\beta}$ from the OLS fit of $\mathbf{Y}$ on $\mathbf{X}$ when $\mathbf{S}_{\mathbf{X}} > 0$.

Following the reduction $\mathbf{X} \mapsto \mathbf{\Phi}^T\mathbf{X}$, we use OLS to fit the multivariate regression of $\mathbf{Y}$ on $\mathbf{\Phi}^T\mathbf{X}$, giving coefficient matrix $\mathbf{B}_{\mathbf{\Phi}} = \mathbf{S}_{\mathbf{YX}}\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Phi})^{-1}$, assuming that

$\mathbf{\Phi}^T\mathbf{S_X}\mathbf{\Phi} > 0$. The estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{\Phi}}$ of $\boldsymbol{\beta}$ is then

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathbf{\Phi}} &= \mathbf{B}_{\mathbf{\Phi}}\mathbf{\Phi}^T = \mathbf{S_{YX}}\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{S_X}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T && (4.2)\\
&= \mathbf{S_{YX}}\mathbf{S_X^{-1}}\mathbf{S_X}\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{S_X}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T \\
&= \mathbf{B}\mathbf{P}^T_{\mathcal{W}_q(\mathbf{S_X})} && (4.3)\\
\boldsymbol{\beta}_{\mathbf{\Phi}} &= \boldsymbol{\beta}\mathbf{P}^T_{\mathcal{W}_q(\mathbf{\Sigma_X})} && (4.4)
\end{aligned}$$

where $\mathcal{W}_q = \text{span}(\mathbf{\Phi})$. The form (4.2) does not require computation of $\mathbf{S_X^{-1}}$ and this could be useful when $n < p$, depending on the size of q. Form (4.3) describes $\widehat{\boldsymbol{\beta}}^T_{\mathbf{\Phi}}$ as the projection of $\mathbf{B}^T$ onto $\mathcal{W}_q$ in the $\mathbf{S_X}$ inner product. The population version (4.4) is constructed by replacing the sample quantities with their population values. The relationships shown in (4.2)-(4.4) depend on $\mathbf{\Phi}$ only through its span $\mathcal{W}_q$ and thus we take $\mathbf{\Phi}$ to be a semi-orthogonal matrix without loss of generality. Let $(\mathbf{\Phi}, \mathbf{\Phi}_0) \in \mathbb{R}^{p\times p}$ be an orthogonal matrix. If we choose $q = p$ then $\mathbf{\Phi} = \mathbf{I}_q$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{\Phi}} = \mathbf{B}$, which achieves nothing beyond $\mathbf{B}$. If we choose the columns of $\mathbf{\Phi}$ to be the first q eigenvectors of $\mathbf{S_X}$ then $\mathbf{\Phi}^T\mathbf{X}$ consists of the first q principal components and $\widehat{\boldsymbol{\beta}}_{\mathbf{\Phi}}$ is the standard principal component regression (PCR) estimator.

We use two conditions to guide the choice of $\mathbf{\Phi}$ and define the material and immaterial parts of $\mathbf{X}$. We first require that $\mathbf{Y}$ and $\mathbf{\Phi}_0^T\mathbf{X}$ be uncorrelated given $\mathbf{\Phi}^T\mathbf{X}$, $\text{cov}(\mathbf{Y}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) = 0$. This condition insures that there is no linear relation between $\mathbf{Y}$ and $\mathbf{\Phi}_0^T\mathbf{X}$ when $\mathbf{\Phi}^T\mathbf{X}$ is known. The second condition is that $\mathbf{\Phi}_0^T\mathbf{X}$ and $\mathbf{\Phi}^T\mathbf{X}$ are uncorrelated, which insures that there is no marginal linear relationship between the reduced predictor and its complement. We think of $\mathbf{\Phi}_0^T\mathbf{X}$ as the immaterial part of $\mathbf{X}$. These conditions play the same role in $\mathbf{X}$ reduction as conditions (2.8) did previously for $\mathbf{Y}$ reduction:

Lemma 4.1 *Under model (4.1), (i)* $\text{cov}(\mathbf{Y}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) = 0$ *if and only if* $\text{span}(\boldsymbol{\beta}^T) \in \mathcal{W}_q$*, and (ii)* $\text{cov}(\mathbf{\Phi}^T\mathbf{X}, \mathbf{\Phi}_0^T\mathbf{X}) = 0$ *if and only if* $\mathcal{W}_q$ *reduces* $\mathbf{\Sigma_X}$.

PROOF.

Replacing $\mathbf{Y}$ with its model (4.1), we have

$$\begin{aligned}
\text{cov}(\mathbf{Y}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) &= \text{cov}(\boldsymbol{\beta}\mathbf{X}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) \\
&= \text{cov}(\boldsymbol{\beta}\mathbf{P}_{\mathbf{\Phi}}\mathbf{X} + \boldsymbol{\beta}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{X}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) \\
&= \text{cov}(\boldsymbol{\beta}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{X}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) \\
&= \boldsymbol{\beta}\mathbf{Q}_{\mathbf{\Phi}}\text{var}(\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X})\mathbf{\Phi}_0
\end{aligned}$$

The conclusion follows since $\text{span}\{\text{var}(\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X})\} = \text{span}(\mathbf{\Phi}_0)$ and thus $\text{cov}(\mathbf{Y}, \mathbf{\Phi}_0^T\mathbf{X} \mid \mathbf{\Phi}^T\mathbf{X}) = 0$ if and only if $\text{span}(\boldsymbol{\beta}^T) \in \mathcal{W}_q$.

For statement (ii), $\text{cov}(\mathbf{\Phi}^T\mathbf{X}, \mathbf{\Phi}_0^T\mathbf{X}) = 0$ if and only if $\mathcal{W}_q$ decomposes $\mathbf{\Sigma_X} = \mathbf{P}_{\mathcal{W}_q}\mathbf{\Sigma_X}\mathbf{P}_{\mathcal{W}_q} + \mathbf{Q}_{\mathcal{W}_q}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{W}_q}$. Consequently, $\mathcal{W}_q$ must be a reducing subspace of $\mathbf{\Sigma_X}$ (cf. (2.10) and Proposition 5.1). □

Let $\mathcal{B}' = \text{span}(\boldsymbol{\beta}^T)$ then it follows from Lemma 4.1 that conditions (i) and (ii) hold if and only if $\mathcal{W}_q$ is a reducing subspace of $\mathbf{\Sigma_X}$ that contains $\mathcal{B}'$. The smallest subspace with these properties is the $\mathbf{\Sigma_X}$ envelope of $\mathcal{B}'$, $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ (cf. Definition 5.2). From this point on we work in terms of the envelope parameter $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ with basis $\mathbf{\Phi}$ and dimension q. As before, we use $\mathcal{E}$ to denote this subspace when used as a subscript.

Since $\mathcal{B}' \subseteq \mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$, we can write write $\boldsymbol{\beta}^T = \mathbf{\Phi}\boldsymbol{\eta}$ for some coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{q\times r}$ and then rewrite model (4.1) as

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\alpha} + \boldsymbol{\eta}^T\mathbf{\Phi}^T\mathbf{X} + \boldsymbol{\varepsilon} \\ \mathbf{\Sigma_X} &= \mathbf{\Phi}\mathbf{\Omega}\mathbf{\Phi}^T + \mathbf{\Phi}_0\mathbf{\Omega}_0\mathbf{\Phi}_0^T, \end{aligned} \tag{4.5}$$

where $\mathbf{\Omega} = \mathbf{\Phi}^T\mathbf{\Sigma_X}\mathbf{\Phi}$ and $\mathbf{\Omega}_0 = \mathbf{\Phi}_0^T\mathbf{\Sigma_X}\mathbf{\Phi}_0$. In contrast to the approach when reducing $\mathbf{Y}$, here there are no envelope constraints placed on $\mathbf{\Sigma}$. The number of free parameters in this model is the sum of r for $\boldsymbol{\alpha}$, rq for $\boldsymbol{\eta}$, $q(p-q)$ for $\text{span}(\mathbf{\Phi})$, $q(q+1)/2$ for $\mathbf{\Omega}$, $(p-q)(p-q+1)/2$ for $\mathbf{\Omega}_0$ and $r(r+1)/2$ for $\mathbf{\Sigma}$, giving a total of $N(q) = r + rq + p(p+1)/2 + r(r+1)/2$. This amounts to reduction of $N(p) - N(q) = r(p-q)$ parameters over the standard model.

The following proposition provides intuition about the potential advantages of the envelope estimator by contrasting its variance with that of the OLS estimator when a basis $\mathbf{\Phi}$ of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ is known and the predictor are normally distributed.

Proposition 4.1 *Let $f_p = n - p - 2 > 0$ and $f_q = n - q - 2 > 0$. Assume that $\mathbf{X} \sim N_p(\boldsymbol{\mu}_\mathbf{X}, \mathbf{\Sigma_X})$ and that a semi-orthogonal basis $\mathbf{\Phi}$ of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ is known. Then (i) $\widehat{\boldsymbol{\beta}}_\mathbf{\Phi} = \mathbf{B}\mathbf{P}^T_{\mathcal{E}(\mathbf{S_X})}$, (ii) $\text{var}(\text{vec}(\mathbf{B})) = f_p^{-1}\mathbf{\Sigma_X}^{-1} \otimes \mathbf{\Sigma}$ and (iii) $\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_\mathbf{\Phi})) = f_q^{-1}\mathbf{\Phi}\mathbf{\Omega}^{-1}\mathbf{\Phi}^T \otimes \mathbf{\Sigma}$, where the variances are computed over both $\mathbf{X}$ and $\mathbf{Y}$.*

PROOF.

Part (i) follows immediately from (4.3) by replacing $\mathcal{W}_q$ with the basis $\mathbf{\Phi}$ for the envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$.

To see part (ii) write

$$\text{var}(\text{vec}(\mathbf{B})) \quad = \quad \text{E}(\text{var}(\text{vec}(\mathbf{B})|\mathbf{X}_1, \ldots, \mathbf{X}_n)) + \text{var}(\text{E}(\text{vec}(\mathbf{B})|\mathbf{X}_1, \ldots, \mathbf{X}_n)).$$

Since $\text{E}(\text{vec}(\mathbf{B})|\mathbf{X}_1, \ldots, \mathbf{X}_n) = \text{vec}(\boldsymbol{\beta})$ the second addend on the right side is 0. Then substituting $\text{var}(\text{vec}(\mathbf{B})|\mathbf{X}_1, \ldots, \mathbf{X}_n) = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$ from (2.6) we have

$$\begin{aligned} \text{var}(\text{vec}(\mathbf{B})) &= \text{E}(n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}) \\ &= \text{E}(n^{-1}\mathbf{S}_{\mathbf{X}}^{-1}) \otimes \boldsymbol{\Sigma} \\ &= f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}, \end{aligned}$$

where the last step follows because $n^{-1}\mathbf{S}_{\mathbf{X}}^{-1}$ follows an inverse Wishart distribution[1] with mean $f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$.

Part (iii) follows from part (ii) after a little preparation. Write $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Phi}} = \widehat{\boldsymbol{\eta}}_{\boldsymbol{\Phi}}^T\boldsymbol{\Phi}^T$, where $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Phi}}^T$ is the coefficient matrix from the OLS fit of $\mathbf{Y}$ on $\boldsymbol{\Phi}^T\mathbf{X}$. Replacing $\mathbf{X} \in \mathbb{R}^p$ with $\boldsymbol{\Phi}^T\mathbf{X} \in \mathbb{R}^q$, it follows from part (ii) that

$$\begin{aligned} \text{var}(\text{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Phi}}^T)) &= f_q^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\Phi}^T\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} \\ &= f_q^{-1}(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Phi})^{-1} \otimes \boldsymbol{\Sigma} \\ &= f_q^{-1}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Sigma}. \end{aligned}$$

Then

$$\begin{aligned} \text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Phi}})) &= (\boldsymbol{\Phi} \otimes \mathbf{I}_r)\text{var}(\text{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Phi}}^T))(\boldsymbol{\Phi}^T \otimes \mathbf{I}_r) \\ &= f_q^{-1}\boldsymbol{\Phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\Phi}^T \otimes \boldsymbol{\Sigma}. \end{aligned}$$

□

It follows from Proposition 4.1 that

$$\text{var}(\text{vec}(\mathbf{B})) = f_q f_p^{-1}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Phi}})) + f_p^{-1}\boldsymbol{\Phi}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Phi}_0^T \otimes \boldsymbol{\Sigma}.$$

Since $f_q/f_p \geq 1$, the first addend on the right hand side reflects gain though reducing the dimension. This term could be large if u is small relative to p. But experience has shown that the largest gains are associated with the second addend: If $\boldsymbol{\Sigma}_{\mathbf{X}}$ has some small eigenvalues and if some of the corresponding eigenvectors are associated with the immaterial

[1] See von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics* **15**, 97–109. for background on inverse Wishart moments.

variation, then $\mathbf{\Omega}_0^{-1}$ may be quite large. In short, if collinearity among the predictors is associated with immaterial variation then predictor envelopes could result in substantial gains over OLS. In practice it will be necessary to estimate a basis $\mathbf{\Phi}$ for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ and this will mitigate the gains suggested here. Nevertheless, these calculations should give a qualitative indication for the kinds of regressions in which envelope reduction of the predictors will be useful.

We next turn to methods for estimating the predictor envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$.

4.2 SIMPLS

Historically, partial least squares (PLS) has been defined in terms of the iterative algorithms, the two most common being the NIPALS[2] and SIMPLS[3] algorithms. It is now widely used across the applied sciences as a method that improves prediction performance over ordinary least square regression, particularly in chemometrics[4] where it originated.

While PLS algorithms have existed for decades and much has been written about them, their theoretical statistical properties have been largely unknown. Cook, Helland and Su (2013)[5] showed recently that SIMPLS algorithm produces a $\sqrt{n}$-consistent estimator of a basis for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$, and thus it provides a first method for completing the estimation method of Section 4.1. We focus mostly on SIMPLS in this section, but point out variations needed for NIPALS.

The algorithm discussed in this section is an instance of a more general algorithm studied in Section 6.5.1. Attributes claimed in this section are justified in that section.

The population SIMPLS algorithm for predictor reduction[6] produces a sequence of p-dimensional vectors $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_u$, with initial $\mathbf{w}_0 = 0$, whose cumulative spans are strictly increasing and converge after q steps to the predictor envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$. Let $\mathbf{W}_k = (\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p\times(k+1)}$ and let $\mathcal{W}_k = \text{span}(\mathbf{W}_k)$. Then the $\mathbf{w}_k$'s are

[2] non-linear iterative partial least squares: Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, et al,. editor, *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, 307–357. Academic Press.

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

[4] Martens and Næs (1989), *Multivariate Calibration.* New York: Wiley, is a classical reference for PLS within the chemometrics community.

[5] Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, B* **75**, to appear

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

constructed so that $\mathcal{W}_0 \subset \mathcal{W}_1 \subset \ldots \subset \mathcal{W}_{q-1} \subset \mathcal{W}_q = \mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$. Given $\mathbf{W}_k$, $k < q$, the $k+1$-st direction is constructed as

$$\begin{aligned} \mathbf{w}_{k+1} &= \arg\max_{\mathbf{w}} \mathbf{w}^T \mathbf{\Sigma_{XY}} \mathbf{\Sigma}_{\mathbf{XY}}^T \mathbf{w}, \text{ subject to} \\ & \quad \mathbf{w}^T \mathbf{\Sigma_X} \mathbf{W}_k = 0 \text{ and } \mathbf{w}^T \mathbf{w} = 1. \end{aligned}$$

At termination, $\mathbf{\Phi}$ can be any semi-orthogonal basis for $\mathcal{W}_q$. If the length constraint $\mathbf{w}^T\mathbf{w} = 1$ is modified to $\mathbf{w}^T \mathbf{Q}_{\mathcal{W}_k} \mathbf{w} = 1$ then we obtain the population version of the NIPALS algorithm, which also produces the desired envelope after q steps.

Since $\mathbf{w}^T \mathbf{\Sigma_{XY}} \mathbf{\Sigma}_{\mathbf{XY}}^T \mathbf{w} = \mathbf{\Sigma}_{(\mathbf{w}^T\mathbf{X})\mathbf{Y}} \mathbf{\Sigma}_{(\mathbf{w}^T\mathbf{X})\mathbf{Y}}^T$, the SIMPLS algorithm is sometimes described as a sequential method for producing linear combinations of the predictors that have maximal covariances with the response. However, the constraints are more than just a convenience, but in fact are key in producing an envelope solution. If the length constraint is changed to $\mathbf{w}^T \mathbf{\Sigma_X} \mathbf{w} = 1$, which might be considered more natural in view of the orthogonality constraint, then $\mathcal{W}_q$ is equal to $\mathbf{\Sigma}_{\mathbf{X}}^{-1/2}$ times the span of the first q eigenvectors of $\mathbf{A} = \mathbf{\Sigma}_{\mathbf{X}}^{-1/2} \mathbf{\Sigma_{XY}} \mathbf{\Sigma}_{\mathbf{XY}}^T \mathbf{\Sigma}_{\mathbf{X}}^{-1/2}$. But $q \geq \text{rank}(\mathbf{\Sigma_{XY}}) = \text{rank}(\mathbf{A})$ and thus the last $q - \text{rank}(\mathbf{\Sigma_{XY}})$ eigenvectors of $\mathbf{A}$ are not uniquely defined so the algorithm cannot return the envelope.

Let $\boldsymbol{\ell}_{\max}(\mathbf{A})$ be an eigenvector associated with the largest eigenvalue of the symmetric matrix $\mathbf{A}$, $\boldsymbol{\ell}_{\max}(\mathbf{A}) = \arg\max_{\boldsymbol{\ell}^T\boldsymbol{\ell}=1} \boldsymbol{\ell}^T \mathbf{A} \boldsymbol{\ell}$. 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 $k = 0, \ldots, q-1$, set

$$\begin{aligned} \mathcal{S}_k &= \text{span}(\mathbf{\Sigma_X} \mathbf{W}_k) \\ \mathbf{w}_{k+1} &= \boldsymbol{\ell}_{\max}(\mathbf{Q}_{\mathcal{S}_k} \mathbf{\Sigma_{XY}} \mathbf{\Sigma}_{\mathbf{XY}}^T \mathbf{Q}_{\mathcal{S}_k}) \\ \mathbf{W}_{k+1} &= (\mathbf{w}_0, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1}). \end{aligned}$$

At termination, $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}') = \mathcal{W}_q = \text{span}(\mathbf{W}_q)$. Since $\mathbf{W}_k$ has full column rank for $k \leq q$, $\dim(\mathcal{S}_k) = k$ and thus no rank consideration is necessary for $\mathcal{S}_k$.

(NOTE: connect with Proposition 5.5 after the prop.)

The SIMPLS algorithm is a function of only three population quantities, $\mathbf{\Sigma_{XY}}$, $\mathbf{\Sigma_X}$ and $q = \dim(\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}'))$. The sample version of SIMPLS is constructed straightforwardly by replacing $\mathbf{\Sigma_{XY}}$, $\mathbf{\Sigma_X}$ by their sample counterparts, terminating after q steps with $\widehat{\mathbf{\Phi}} = \mathbf{W}_q$ (cf. equation 4.3) or generally any basis matrix for $\mathcal{W}_q$. Of course there is no sample counterpart to q, which must be inferred using one of the many available rules. Five or ten-fold cross-validation of predictive performance seems to be a commonly used method for

choosing a suitable q. Given $q = \dim(\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}'))$ then, with r and p fixed, this algorithm provides a $\sqrt{n}$ consistent estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ because $\mathbf{W}_q$ is a smooth function of $\mathbf{S_{XY}}$ and $\mathbf{S_X}$ which are themselves $\sqrt{n}$ consistent.

Generally, $\mathrm{rank}(\mathbf{\Sigma_{XY}}) \leq q \leq p$, where $\mathrm{rank}(\mathbf{\Sigma_{XY}}) \leq \min(r, p)$. Consequently, if $\mathrm{rank}(\mathbf{\Sigma_{XY}}) = p$ then $q = p$ and the SIMPLS estimator of $\boldsymbol{\beta}$ is the same as the OLS estimator. In particular, if $r \geq p$ and $\mathbf{\Sigma_{XY}}$ has full row rank then again SIMPLS and OLS are the same. The SIMPLS algorithm is most useful for reducing the dimension of $\mathbf{X}$ when $r < p$.

Chun and Keleş (2010) showed recently that the partial least squares estimator of the coefficient vector in the univariate linear regression of Y on $\mathbf{X}$ is inconsistent unless $p/n \to 0$ and this raises questions about the usefulness of the SIMPLS algorithm when $n < p$. They also developed a penalized version that has better asymptotic behavior and may mitigate this concern. On the other hand, SIMPLS is typically used for prediction rather than parameter estimation and inconsistency of this estimator does not necessarily imply inconsistency of its predictions. It is possible that, while SIMPLS may not be sufficiently accurate for parameter estimation when $n < p$, it could still be quite useful for prediction. The predictive behavior of SIMPLS as n and p grow is apparently unknown.

4.3 Likelihood-based envelopes.

It is traditional in regression to base estimation on the conditional likelihood from $\mathbf{Y}|\mathbf{X}$, 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.[7] In contrast, PLS and the likelihood-based method developed in this section both postulate a link – represented here by the envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ – between $\boldsymbol{\beta}$, the parameter of interest, and $\mathbf{\Sigma_X}$. As a consequence, $\mathbf{X}$ is not ancillary and we pursue estimation through the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$.

4.3.1 Estimation

Let $\mathbf{C} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ denote the random vector constructed by concatenating $\mathbf{X}$ and $\mathbf{Y}$, and let $\mathbf{S_C}$ denote the sample version of $\mathbf{\Sigma_C} = \mathrm{var}(\mathbf{C})$. We base estimation on the objec-

[7] For a review and an historical perspective, See Aldrich, J. (2005), Fisher and regression. *Statistical Science* **20**, 401–417.

tive function $F(\mathbf{S_C}, \mathbf{\Sigma_C}) = \log|\mathbf{\Sigma_C}| + \mathrm{tr}(\mathbf{S_C}\mathbf{\Sigma_C}^{-1})$ that stems from the log likelihood of the multivariate normal family after replacing the population mean vector with the vector of sample means, although we do not require $\mathbf{C}$ to have a multivariate normal distribution. Rather we are using F as a multi-purpose objective function in the same spirit as least squares objective functions are often used. The structure of the envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ can be introduced into F by using the parameterizations $\mathbf{\Sigma_X} = \mathbf{\Phi\Omega\Phi}^T + \mathbf{\Phi}_0\mathbf{\Omega}_0\mathbf{\Phi}_0^T$ and $\mathbf{\Sigma_{XY}} = \mathbf{\Phi\gamma}$, where $\mathbf{\Phi} \in \mathbb{R}^{p\times q}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$, $(\mathbf{\Phi}, \mathbf{\Phi}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, and $\mathbf{\Omega} \in \mathbb{R}^{q\times q}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-q)\times(p-q)}$ are symmetric positive definite matrices. Since $\mathrm{span}(\mathbf{\Sigma_{XY}}) \subseteq \mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ (cf. Corollary 5.2) we can write $\mathbf{\Sigma_{XY}}$ as linear combinations of the columns of $\mathbf{\Phi}$. The matrix $\mathbf{\gamma} \in \mathbb{R}^{q\times r}$ then gives the coordinates of $\mathbf{\Sigma_{XY}}$ in terms of the basis $\mathbf{\Phi}$. With this we have

$$\begin{aligned}
\mathbf{\Sigma_C} &= \begin{pmatrix} \mathbf{\Sigma_X} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{XY}}^T & \mathbf{\Sigma_Y} \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi\Omega\Phi}^T + \mathbf{\Phi}_0\mathbf{\Omega}_0\mathbf{\Phi}_0^T & \mathbf{\Phi\gamma} \\ \mathbf{\gamma}^T\mathbf{\Phi}^T & \mathbf{\Sigma_Y} \end{pmatrix} \qquad (4.6)\\
&= \begin{pmatrix} \mathbf{\Phi} & \mathbf{\Phi}_0 & 0 \\ 0 & 0 & \mathbf{I}_r \end{pmatrix} \begin{pmatrix} \mathbf{\Omega} & 0 & \mathbf{\gamma} \\ 0 & \mathbf{\Omega}_0 & 0 \\ \mathbf{\gamma}^T & 0 & \mathbf{\Sigma_Y} \end{pmatrix} \begin{pmatrix} \mathbf{\Phi}^T & 0 \\ \mathbf{\Phi}_0^T & 0 \\ 0 & \mathbf{I}_r \end{pmatrix} \\
&= \mathbf{O}\mathbf{\Sigma}_{\mathbf{O}^T\mathbf{C}}\mathbf{O}^T, \qquad (4.7)
\end{aligned}$$

where

$$\mathbf{O} = \begin{pmatrix} \mathbf{\Phi} & \mathbf{\Phi}_0 & 0 \\ 0 & 0 & \mathbf{I}_r \end{pmatrix} \in \mathbb{R}^{(p+r)\times(p+r)}$$

is an orthogonal matrix and

$$\mathbf{\Sigma}_{\mathbf{O}^T\mathbf{C}} = \begin{pmatrix} \mathbf{\Omega} & 0 & \mathbf{\gamma} \\ 0 & \mathbf{\Omega}_0 & 0 \\ \mathbf{\gamma}^T & 0 & \mathbf{\Sigma_Y} \end{pmatrix} \in \mathbb{R}^{(p+r)\times(p+r)}$$

is the covariance matrix of the transformed vector $\mathbf{O}^T\mathbf{C}$. The objective function $F(\mathbf{S_C}, \mathbf{\Sigma_C})$ can now regarded as a function of the five constituent parameters – $\mathbf{\Phi}$, $\mathbf{\Omega}$, $\mathbf{\Omega}_0$, $\mathbf{\gamma}$ and $\mathbf{\Sigma_Y}$ – that comprise $\mathbf{\Sigma_C}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ of model (4.5) can be written as $\boldsymbol{\eta} = \mathbf{\Omega}^{-1}\mathbf{\gamma}$ and $\boldsymbol{\beta}^T = \mathbf{\Phi}\boldsymbol{\eta} = \mathbf{\Phi\Omega}^{-1}\mathbf{\gamma}$.

To minimize $F(\mathbf{S_C}, \mathbf{\Sigma_C})$ we first hold $\mathbf{\Phi}$ fixed and substitute (4.7),

$$\begin{aligned}
F(\mathbf{S_C}, \mathbf{\Sigma_C}) &= \log|\mathbf{O}\mathbf{\Sigma}_{\mathbf{O}^T\mathbf{C}}\mathbf{O}^T| + \mathrm{tr}(\mathbf{O}^T\mathbf{S_C}\mathbf{O}\mathbf{\Sigma}^{-1}_{\mathbf{O}^T\mathbf{C}}) \\
&= \log|\mathbf{\Sigma}_{\mathbf{O}^T\mathbf{C}}| + \mathrm{tr}(\mathbf{S}_{\mathbf{O}^T\mathbf{C}}\mathbf{\Sigma}^{-1}_{\mathbf{O}^T\mathbf{C}}).
\end{aligned}$$

The form of $\mathbf{\Sigma}_{\mathbf{O}^T\mathbf{C}}$ allows us to divide this into the negative log likelihood from the marginal of $\mathbf{\Phi}_0^T\mathbf{X}$, which depends only on its covariance matrix $\mathbf{\Omega}_0$, and the negative of the log likelihood from $(\mathbf{\Phi}^T\mathbf{X}, \mathbf{Y})$, which depends on $\mathbf{\Omega}$, $\boldsymbol{\gamma}$ and $\mathbf{\Sigma}_{\mathbf{Y}}$. The values of these parameters that minimize F for fixed $\mathbf{\Phi}$ are then straightforwardly, $\mathbf{\Sigma}_{\mathbf{Y}} = \mathbf{S}_{\mathbf{Y}}$, $\mathbf{\Omega} = \mathbf{\Phi}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Phi}$, $\mathbf{\Omega}_0 = \mathbf{\Phi}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{\Phi}_0$ and $\boldsymbol{\gamma} = \mathbf{\Phi}^T\mathbf{S}_{\mathbf{XY}}$. Substituting these forms into F then leads to the following estimator of the envelope when q is assumed known:

$$\begin{aligned} \widehat{\mathcal{E}}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{B}') &= \text{span}\{\arg\min_{\mathbf{F}} L_q(\mathbf{F})\}, \text{ where} && (4.8) \\ L_q(\mathbf{F}) &= \log|\mathbf{F}^T(\mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{XZ}}\mathbf{S}_{\mathbf{XZ}}^T)\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{F}|, && (4.9) \end{aligned}$$

$\mathbf{Z} = \mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{Y}$ is the standardized response vector and the minimization in (4.8) is taken over all semi-orthogonal matrices $\mathbf{F} \in \mathbb{R}^{p\times q}$. This optimization is the same as that encountered for $\mathbf{Y}$ reduction in Chapter 2 (cf. equation (2.15)), except that the roles of $\mathbf{Y}$ and $\mathbf{X}$ have been interchanged. Let $\widehat{\mathbf{\Phi}}$ be any semi-orthogonal basis of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$. The estimators of the remaining constituent parameters are then

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{\mathbf{Y}} &= \mathbf{S}_{\mathbf{Y}}, \\ \widehat{\mathbf{\Omega}} &= \widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Phi}}, \\ \widehat{\mathbf{\Omega}}_0 &= \widehat{\mathbf{\Phi}}_0^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Phi}}_0, \\ \widehat{\boldsymbol{\gamma}} &= \widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{XY}} \\ \widehat{\boldsymbol{\eta}} &= \widehat{\mathbf{\Omega}}^{-1}\widehat{\boldsymbol{\gamma}}. \end{aligned}$$

From these we construct the estimators of the parameters of interest:

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{\mathbf{XY}} &= \mathbf{P}_{\widehat{\mathbf{\Phi}}}\mathbf{S}_{\mathbf{XY}}, \\ \widehat{\mathbf{\Sigma}}_{\mathbf{X}} &= \widehat{\mathbf{\Phi}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{\Phi}}^T + \widehat{\mathbf{\Phi}}_0\widehat{\mathbf{\Omega}}_0\widehat{\mathbf{\Phi}}_0^T = \mathbf{P}_{\widehat{\mathbf{\Phi}}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\widehat{\mathbf{\Phi}}} + \mathbf{Q}_{\widehat{\mathbf{\Phi}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\mathbf{\Phi}}}, \\ \widehat{\boldsymbol{\beta}}^T &= \widehat{\mathbf{\Phi}}\widehat{\mathbf{\Omega}}^{-1}\widehat{\boldsymbol{\gamma}} = \widehat{\mathbf{\Phi}}(\widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Phi}})^{-1}\widehat{\mathbf{\Phi}}^T\mathbf{S}_{\mathbf{XY}} = \mathbf{P}_{\widehat{\mathbf{\Phi}}(\mathbf{S}_{\mathbf{X}})}\mathbf{B}^T. \end{aligned}$$

The estimators $\widehat{\mathbf{\Omega}}$, $\widehat{\mathbf{\Omega}}_0$ and $\widehat{\boldsymbol{\gamma}}$ depend on the selected basis $\widehat{\mathbf{\Phi}}$. The parameters of interest – $\widehat{\mathbf{\Sigma}}_{\mathbf{XY}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ and $\widehat{\boldsymbol{\beta}}$ – depend on $\widehat{\mathcal{E}}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$ but do not depend on the particular basis selected.

4.3.2 Comparisions with SIMPLS and PCR

There are consequential differences between the likelihood-based estimation method and SIMPLS. To see how these differences arise, we first describe some operating characteristics of $L_q(\mathbf{F})$ and then contrast those characteristics with the behavior of SIMPLS. Let

$L_q^{(1)}(\mathbf{F}) = \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}|$ and $L_q^{(2)}(\mathbf{F}) = \log|\mathbf{S}_{\mathbf{Z}|\mathbf{F}^T\mathbf{X}}|$, where $\mathbf{S}_{\mathbf{Z}|\mathbf{F}^T\mathbf{X}}$ is the sample covariance matrix of the residual vectors from the OLS fit of $\mathbf{Z}$ on $\mathbf{F}^T\mathbf{X}$. Then the objective function L_q can be represented as $L_q(\mathbf{F}) = L_q^{(1)}(\mathbf{F}) + L_q^{(2)}(\mathbf{F})$:

$$\begin{aligned} L_q(\mathbf{F}) &= \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}| + \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{XZ}}^T\mathbf{F}(\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{S}_{\mathbf{XZ}}| \\ &= \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}| + \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{Z},\mathbf{F}^T\mathbf{X}}\mathbf{S}_{\mathbf{F}^T\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{Z},\mathbf{F}^T\mathbf{X}}^T| \\ &= \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}| + \log|\mathbf{S}_{\mathbf{Z}|\mathbf{F}^T\mathbf{X}}|. \end{aligned}$$

The first addend $L_q^{(1)}(\mathbf{F}) \geq 0$ with $L_q^{(1)}(\mathbf{F}) = 0$ when the columns of $\mathbf{F}$ correspond to any subset of q eigenvectors of $\mathbf{S}_\mathbf{X}$ (cf. Lemma 6.3). Consequently, the role of $L_q^{(1)}$ is to pull the solution toward subsets of q eigenvectors of $\mathbf{S}_\mathbf{X}$. This in effect imposes a sample counterpart of the characterization in Proposition 5.3, 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 $L_q^{(2)}(\mathbf{F}) = \log|\mathbf{S}_{\mathbf{Z}|\mathbf{F}^T\mathbf{X}}|$ of $L_q(\mathbf{F})$ measures the goodness of fit of regression of the standardized response $\mathbf{Z}$ on $\mathbf{F}^T\mathbf{X}$. As a consequence, $L_q(\mathbf{F})$ can be seen as balancing the closeness of $\mathrm{span}(\mathbf{F})$ to a reducing subspace of $\mathbf{S}_\mathbf{X}$ and the fit of $\mathbf{Z}$ on $\mathbf{F}^T\mathbf{X}$.

Let $\mathbf{V} = \mathbf{S}_\mathbf{X}^{-1/2}\mathbf{X}$ denote the sample standardized version of $\mathbf{X}$, 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 $\mathbf{Z}$, and let $L_q^{(3)} = \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{VZ}}^T\mathbf{P}_{\mathbf{S}_\mathbf{X}^{1/2}\mathbf{F}}\mathbf{S}_{\mathbf{VZ}}|$, then L_q can be expressed also as $L_q(\mathbf{F}) = L_q^{(1)}(\mathbf{F}) + L_q^{(3)}(\mathbf{F})$.

$$\begin{aligned} L_q(\mathbf{F}) &= \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}| + \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{XZ}}^T\mathbf{S}_\mathbf{X}^{-1/2}\mathbf{P}_{\mathbf{S}_\mathbf{X}^{1/2}\mathbf{F}}\mathbf{S}_\mathbf{X}^{-1/2}\mathbf{S}_{\mathbf{XZ}}| \\ &= \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}\mathbf{F}| + \log|\mathbf{F}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{F}| + \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{VZ}}^T\mathbf{P}_{\mathbf{S}_\mathbf{X}^{1/2}\mathbf{F}}\mathbf{S}_{\mathbf{VZ}}|. \end{aligned}$$

The addend $L_q^{(3)}(\mathbf{F})$ of $L_q(\mathbf{F})$ 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 setting $\mathbf{F}$ to be $\mathbf{S}_\mathbf{X}^{-1/2}$ times the first q eigenvectors of $\mathbf{S}_{\mathbf{VZ}}\mathbf{S}_{\mathbf{VZ}}^T$. If $q > r$ only the first $q - r$ of these generalized eigenvectors are determined uniquely. The full objective function $L_q(\mathbf{F}) = L_q^{(1)}(\mathbf{F}) + L_q^{(3)}(\mathbf{F})$ can also be viewed as balancing the requirement that the optimal value should stay close to a subset of q eigenvectors of $\mathbf{S}_\mathbf{X}$ and to the generalized eigenvectors of $\mathbf{S}_{\mathbf{XZ}}\mathbf{S}_{\mathbf{XZ}}^T$ relative to $\mathbf{S}_\mathbf{X}$.

The PCR estimator of $\boldsymbol{\beta}$ is obtained by setting the columns of $\widehat{\mathbf{\Phi}}$ to be the first few eigenvectors of $\mathbf{S}_\mathbf{X}$. While $L_q^{(1)}(\mathbf{F})$ pulls the solution toward the eigenspaces of $\mathbf{S}_\mathbf{X}$ there is no particular preference for the principal eigenspaces. In fact, $L_q^{(1)}(\mathbf{F})$ alone places no special preference on any ordering of the eigenspaces. The function $L_q^{(2)}(\mathbf{F})$ can guide the

solution toward any q-dimensional eigenspace. In short, while the eigenspaces of $\mathbf{S}_\mathbf{X}$ play a role in both envelopes and PCR, those roles are quite different.

Turning to comparisons of the likelihood-based method with SIMPLS, we see first that $L_q(\mathbf{F})$ 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 $q = 1$, the SIMPLS estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ 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 $\mathbf{Y} \to \mathbf{AY}$, the SIMPLS estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ 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 $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ 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 to standardize the individual responses marginally $y_j \to y_j/\{\widehat{\mathrm{var}}(y_j)\}^{1/2}$, $j = 1, \ldots, r$, prior application of SIMPLS, but it is evidently not so customary to standardize the responses jointly $\mathbf{Y}_i \to \mathbf{Z}_i = \mathbf{S}_\mathbf{Y}^{-1/2}\mathbf{Y}_i$. Of course, the SIMPLS algorithm could be applied after replacing $\mathbf{Y}$ with jointly standardized responses $\mathbf{Z}$.

The methods also differ on how they utilize information from $\mathbf{S}_\mathbf{X}$. In the likelihood-based objective function, $L_q^{(1)}(\mathbf{F})$ guages how far $\mathrm{span}(\mathbf{F})$ is from subsets of q 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 $\mathbf{S}_\mathbf{X}$. 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. SIMPLS often requires more directions to match the performance of the likelihood-based method.

4.3.3 Asymptotic properties

In this section we describe asymptotic properties of the envelope estimator, starting with the case in which $\mathbf{C}$ is normally distributed.

Proposition 4.2 *Assume that q is known and that $\mathbf{C}$ is normally distributed with mean $\boldsymbol{\mu}_\mathbf{C}$ and covariance matrix $\mathbf{\Sigma}_\mathbf{C} > 0$. 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*

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

where $\mathbf{M} = \boldsymbol{\eta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0 + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_q \otimes \mathbf{I}_{p-q}$. *Additionally,* $T_q = n(F(\mathbf{S_C}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}}) - F(\mathbf{S_C}, \mathbf{S_C}))$ *converges to a chi-squared random variable with* $(p-q)r$ *degrees of freedom.*

The decomposition of $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ shown in Proposition 4.2 has the same algebraic form as the decomposition found by Cook, Li and Chiaromonte (2010; Section 5.1 and Corollary 6.1) when pursuing reduction in the $\mathbf{Y}$-dimension (see Section 2.5), although the components of the decomposition of course differ. In particular, it follows that

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] \leq \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{B})],$$

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

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Phi}})] = (\mathbf{I}_r \otimes \boldsymbol{\Phi})\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Phi}})](\mathbf{I}_r \otimes \boldsymbol{\Phi}^T) = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\Phi}^T.$$

The second term in the decomposition of $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ then represents the cost of estimating the envelope. We also see from these results that when performing a prediction at $\mathbf{X}_N - \boldsymbol{\mu}_{\mathbf{X}}$ the asymptotic covariance $\mathrm{avar}(\sqrt{n}(\mathbf{X}_N - \boldsymbol{\mu}_{\mathbf{X}})^T\widehat{\boldsymbol{\beta}})$ depends on the part $\boldsymbol{\Phi}^T(\mathbf{X}_N - \boldsymbol{\mu}_{\mathbf{X}})$ of $\mathbf{X}_N - \boldsymbol{\mu}_{\mathbf{X}}$ that lies in the envelope and on the part $\boldsymbol{\Phi}_0^T(\mathbf{X}_N - \boldsymbol{\mu}_{\mathbf{X}})$ that lies in the orthogonal complement, which is in contrast to the situation when $\boldsymbol{\Phi}$ is known as discussed previously.

The following corollary to Proposition 4.2 describes $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})]$ when $\boldsymbol{\Sigma}_{\mathbf{X}} = \sigma^2_{\mathbf{X}}\mathbf{I}_p$, and provides a comparison with the OLS estimator.

Corollary 4.1 *Assume the conditions of Proposition 4.2 and additionally that* $\boldsymbol{\Sigma}_{\mathbf{X}} = \sigma^2_{\mathbf{X}}\mathbf{I}_p$ *and that the coefficient matrix* $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ *has rank* r. *Then* $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] = \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{B})]$.

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.

The next proposition describes the asymptotic properties of the envelope estimator when $\mathbf{C}$ is not necessarily normal.

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* $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta})\}$ *converges in distribution to a normal random vector with mean 0 and positive definite covariance matrix.*

PROOF. The justification of this proposition involves application of Shapiro's (1986) results on the asymptotic behavior of overparameterized structural models. The shifted objective function $F^*(\mathbf{S_C}, \boldsymbol{\Sigma}_\mathbf{C}) = F(\mathbf{S_C}, \boldsymbol{\Sigma}_\mathbf{C}) - F(\mathbf{S_C}, \mathbf{S_C})$, 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 some 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. □

This proposition says that the envelope estimator $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal when the original data are non-normal. The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ depends on fourth moments of $\mathbf{C}$ and seems complicated. The bootstrap is a useful option in practice for estimating the covariance matrix of $\widehat{\boldsymbol{\beta}}$, as demonstrated in previous chapters.

4.3.4 Choice of dimension

The statistic T_q described in Proposition 4.2 can be used in a sequential manner to estimate q: beginning with $q_0 = 0$ test the hypothesis $q = q_0$, terminating the first time it is not rejected. Otherwise, q_0 is incremented by one and then the hypothesis is tested again. The relative advantages of this versus cross validation have not been studied. Cross-validation, a hold-out sample and an information criterion are also options that may be useful in practice.

4.4 Illustrations

In this section we provide examples to illustrate the operating characteristics of $\mathbf{X}$ envelopes in relatively simple settings. More complicated examples involving predictive performance were given by Cook, et al. (2013)[8] One example from Cook et al. (2013) is

[8]Cook, R.D., Helland, I. and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society B*, to appear.

repeated below for ease of reference.

4.4.1 Australian Institute of Sport, again

In this first example we consider the regression of red cell count on the $p = 2$ hematological measurements hematocrit and hemoglobin. Since there are only two predictors there are only three options for a non-trivial envelope. Either $q = 1$ and $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ aligns with the first or second eigenvector of $\mathbf{\Sigma_X}$ or $q = 2$ and $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}') = \mathbb{R}^2$ aligns with neither eigenvector. In higher dimensional examples, the envelope need not correspond to an eigenspace, but will be contained in a reducing subspace of $\mathbf{\Sigma_X}$.

As in previous examples, a first step is to select a dimension for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$. In this example, AIC, BIC and LRT(0.05) all indicated that $q = 1$ and thus that $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ aligns with either the first or second eigenvector of $\mathbf{\Sigma_X}$. Shown in Figure 4.1 is a plot of the $n = 202$ observations along with the contours of $\mathbf{S_X}$, the estimated envelope $\widehat{\mathcal{E}}_{\mathbf{\Sigma_X}}(\mathcal{B}')$ and its orthogonal complement $\widehat{\mathcal{E}}^{\perp}_{\mathbf{\Sigma_X}}(\mathcal{B}')$, and the subspace spanned by the OLS coefficient vector $\mathbf{B}$. The plot indicates that the span of the first eigenvector of $\mathbf{S_X}$ and the envelope are quite close, and are not far from the OLS subspace $\mathrm{span}(\mathbf{B}^T)$. Evidently, the response changes as we move along the enveloped from left to right, which reflects the material variation in the data, but is relatively constant as we move along the orthogonal complement, which reflects the immaterial variation. The eigenvalues of $\mathbf{S_X}$ are 3.04 and 0.032, so according to the intuition provided by Propositions 4.1 and 4.2 we anticipate notable reduction in estimative variation.

Table 4.1 shows the estimated coefficients with their standard errors from the standard and envelope analyses. While the coefficients are quite close, as suggested by Figure 4.1, the standard errors from the standard analysis are about two and a half times the standard errors from the envelope analysis. We would need a sample size about six times as large for the standard errors of $\mathbf{B}$ to match those of $\widehat{\boldsymbol{\beta}}$ based on the current sample size. Additionally, the sample $\mathbf{S_X}$ version of $\mathbf{\Sigma_X}$ is nearly identical to the envelope estimate $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ in this example:

$$\mathbf{S_X} = \begin{pmatrix} 13.3513 & 4.7211 \\ 4.7211 & 1.8468 \end{pmatrix}, \quad \widehat{\mathbf{\Sigma}}_{\mathbf{X}} = \begin{pmatrix} 13.3511 & 4.7214 \\ 4.7214 & 1.8471 \end{pmatrix}.$$

Since $q = 1$, the subspace estimated by SIMPLS is simply

$$\mathrm{span}(\mathbf{S_{XY}}) = \mathrm{span}((1.544, 0.552)^T).$$

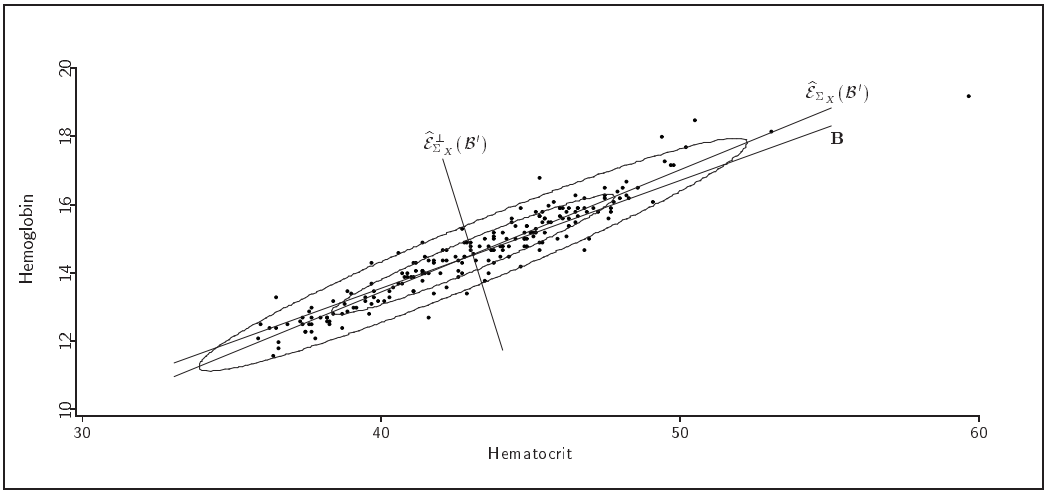

Figure 4.1: Australian Institute of Sport Data: Scatterplot of the predictors hematocrit and hemoglobin, along with the estimated envelope and $\mathrm{span}(\mathbf{B}^T)$ from the regression with red cell count as the response.

Table 4.1: Australian Institute of Sport: Coefficient estimates and their standard errors from the envelope model with $q = 1$ and the standard model.

$\mathbf{X}$	$\widehat{\boldsymbol{\beta}}$	$\mathrm{se}(\widehat{\boldsymbol{\beta}})$	$\mathbf{B}$	$\mathrm{se}(\mathbf{B})$	$\mathrm{se}(\mathbf{B})/\mathrm{se}(\widehat{\boldsymbol{\beta}})$
hematocrit	0.103	0.005	0.104	0.011	2.25
hemoglobin	0.037	0.010	0.033	0.029	2.77

The ratio of the first to the second element of $\mathbf{S}_{\mathbf{X}Y}$ is 2.798, while the same ratio from a basis $\widehat{\boldsymbol{\Phi}}$ for the estimated envelope is 2.7946. Consequently, the estimated envelope and SIMPLS subspaces are nearly identical in this illustration, although that will not always be so, particularly when $q > 1$.

4.4.2 Wheat protein, again

In Section 2.10.1 we used the wheat protein data by setting the response vector to be the logarithms of near infrared reflectance at six wavelengths and the univariate predictor to be the protein content. Here we interchange the roles of the variables, setting the univariate response to be protein content and the bivariate predictor to be measurements at the third and fourth wavelengths. Since there are only $p = 2$ predictor we again have only three

options for the envelope, and the three dimension selection methods again indicated that $q = 1$.

Figure 4.2 shows a scatterplot of the data along with $\mathrm{span}(\mathbf{B}^T)$ from the regression with protein content as the response, and the estimated envelope and its orthogonal complement. In contrast to the previous illustration, the envelope is now estimated to coincide with the span of the second (rather than the first) eigenvector of $\mathbf{\Sigma_X}$, while $\mathrm{span}(\mathbf{B}^T)$ is still quite close to the envelope. The eigenvalues of $\mathbf{S_X}$ are 2.051 and 0.018 so there might be little if any gain over OLS. The SIMPLS subspace $\mathrm{span}(\mathbf{S_{XY}})$ is distinctly different than the estimated envelope subspace.

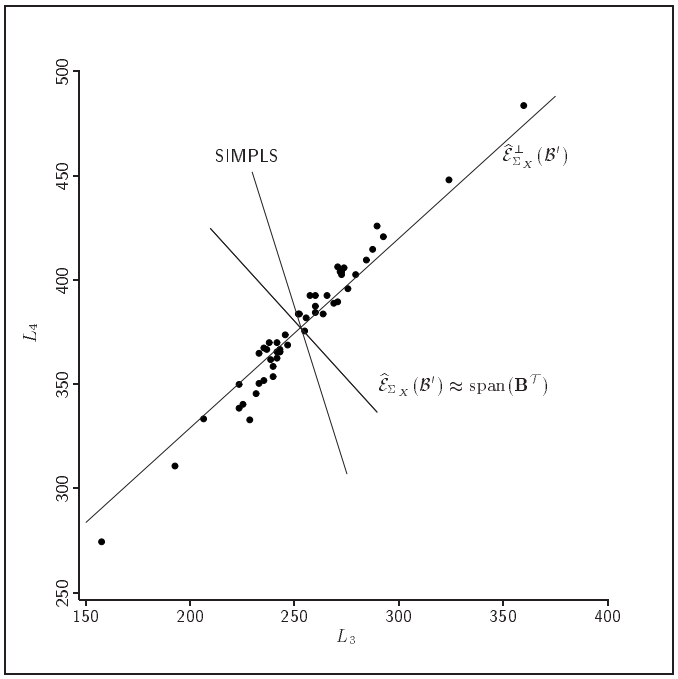

Figure 4.2: Wheat protein data: Scatterplot of the predictors at the third and fourth wavelength, along with the estimated envelope and $\mathrm{span}(\mathbf{B}^T)$ from the regression with protein content as the response.

Table 4.2 gives the coefficient estimates and their standard errors. In this case, although there is a proper envelope, it essentially reproduced the OLS analysis because the immaterial variation in $\mathbf{X}$ is large relative to the material variation in $\mathbf{X}$.

Turning to the regression of protein content on all six wavelengths, we find a consistent dimension of $q = 4$ and the fitted envelope with $q = 4$ again reproduces the OLS analysis

with unimportant differences, so the relationships illustrated in Figure 4.2 may be present in the larger data set and in generally in spectral data of this type.

Table 4.2: Wheat protein data: Coefficient estimates and their standard errors from the envelope model with $q = 1$, the standard model and SIMPLS.

$\mathbf{X}$	$\widehat{\boldsymbol{\beta}}$	$\mathrm{se}(\widehat{\boldsymbol{\beta}})$	$\mathbf{B}$	$\mathrm{se}(\mathbf{B})$	SIMPLS
L_3	0.2470	0.0072	0.24763	0.0074	0.0165
L_4	-0.2249	0.0066	-0.22486	0.0068	- 0.0052

4.4.3 Meat properties

Sæbø et al. (2007) analyzed the absorbance spectra from infrared transmittance for fat, protein and water in $n = 103$ pork or beef samples as an example with collinearity and multiple relevant components for soft-threshold-PLS. Cook, et al. (2013) took the measurements at every fourth wavelength between 850 nm and 1050 nm as predictors, yielding $p = 50$. Predictions of protein were constructed as $\widehat{Y} = \bar{Y} + \widehat{\boldsymbol{\beta}}(\mathbf{X} - \bar{\mathbf{X}})$ with $\widehat{\boldsymbol{\beta}}$ obtained by using envelopes and SIMPLS.[9] Five-fold cross validation was used to estimate the average prediction error, dividing the data into five equal parts.

The prediction errors shown in Figure 4.3 for $1 \leq q \leq 25$ are then the average from the predictions on each split of the data while the remaining four parts were used for fitting. The range of the y-axis was truncated to improve visualization. The SIMPLS prediction error was about 6 with $q = 1$. The results in Figure 4.3 indicate that envelopes perform much better than SIMPLS for a small q while the two methods have similar performance for q sufficiently large.

Problems

Problem 4.1 *One of often cited practical advantages PLS is that is does not run into computational problems when $n < p$. In particular, the population SIMPLS algorithm described in Section 4.2 is serviceable even when $\mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{X}}) < p$. The same holds of the sample algorithm when $\mathrm{rank}(\mathbf{S}_{\mathbf{X}}) < p$. (a) Describe how the algorithm manages to*

[9] The SIMPLS estimator was obtained by using the MATLAB function *plsregress*.

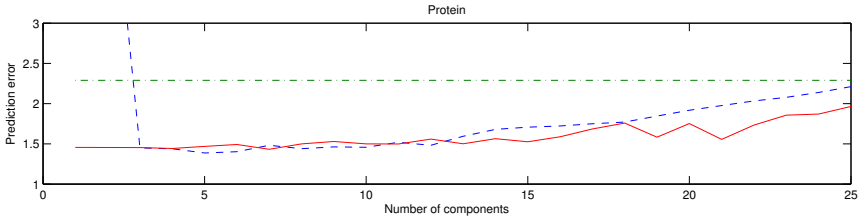

Figure 4.3: Prediction error for the meat data with protein as the response. The solid line marks the envelope prediction error and the dashed marks the prediction error for PLS. The horizontal dashed dotted line marks the constant prediction errors of OLS. (From Cook, et al. (2013)

avoid computational issues in the presence of singular covariance matrices (Hint: principal components). Provide justification (proof) for your conclusions and a critique of the method itself. (b) Can the same technique be used for the construction of envelope described in Section 4.3.1.

Problem 4.2 *The point with the largest value of Hematocrit in the Australian Institute of Sport Data may be overly influencing the analysis, as suggested by Figure 4.1. Remove that point and recompute the analysis of the data. Did the point prove to be influential?*