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 26, 2013

Contents

Chapter 2

Envelopes: Reducing the response

This chapter is on enveloping the coefficient matrix in a multivariate linear model. Section 2.1 contains a very brief review of the multivariate linear model, with emphasis on aspects that will play a role in later developments. An introductory illustration to provide some intuition on the operating characteristics of envelopes is given in Section 2.2. In Section 2.3 we turn envelopes as applied to response reduction in multivariate linear regression. The basic ideas of this chapter are the same as those introduced in Chapter 1, but the context is more general. Most of the technical material in this chapter comes from Cook, Li and Chiaromonte (2010).

2.1 The multivariate linear model

Consider the multivariate regression of a response vector $\mathbf{Y} \in \mathbb{R}^r$ on a vector of non-stochastic predictors $\mathbf{X} \in \mathbb{R}^p$. The standard multivariate linear model for describing a sample $(\mathbf{Y}_i, \mathbf{X}_i)$ can be represented in vector form as

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where the predictors are centered in the sample $\sum_{i=1}^n \mathbf{X}_i = 0$, the error vectors $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ are independently and identically distributed normal vectors with mean 0 and covariance matrix $\boldsymbol{\Sigma} > 0$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ is an unknown vector of intercepts and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is an unknown matrix of regression coefficients. Centering the predictors facilitates discussion and presentation of some results, but is unnecessary. If $\mathbf{X}$ is stochastic, $\mathbf{X}$ and $\mathbf{Y}$ have a joint distribution, but we still condition on the observed values of $\mathbf{X}$ since the predictors are

ancillary under model (2.1). The normality requirement for $\boldsymbol{\varepsilon}$ is not essential, as discussed in Section 2.8, but it does facilitate exposition of the ideas underlying envelopes.

Let $\mathbb{Y}$ denote the $n \times r$ matrix with rows $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$ and let $\mathbb{X}$ denote the $n \times p$ matrix with rows $\mathbf{X}_i^T$. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ in model (2.1), which is also the ordinary least squares estimator, is

$$\mathbf{B} = \mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}. \tag{2.2}$$

From this we notice that $\mathbf{B}$ can be constructed by doing r separate multiple linear regressions, one for each element of $\mathbf{Y}$ on $\mathbf{X}$. The coefficients from the j-th regression then form the j-th row of $\mathbf{B}$, $j = 1, \ldots, r$. The stochastic relationships among the elements of $\mathbf{Y}$ are not used in forming these estimators. However, as will be seen later, relationships among the elements of $\mathbf{Y}$ play a central role in envelope estimation. Standard inference on β_{jk}, the (j, k)-th element of $\boldsymbol{\beta}$, under model (2.1) is the same as inference obtained under the univariate linear regression of Y_j, the j-th element of $\mathbf{Y}$, on $\mathbf{X}$. Model (2.1) becomes operational as a multivariate construction only when inferring simultaneously about elements in different rows of $\boldsymbol{\beta}$ or when predicting elements of $\mathbf{Y}$ jointly.

Let $\widehat{\mathbf{Y}}_i = \bar{\mathbf{Y}} + \mathbf{B}\mathbf{X}_i$ and $\mathbf{r}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$ denote the i-th vectors of fitted values and residuals, $i = 1, \ldots, n$. Then the sample covariance matrices of $\widehat{\mathbf{Y}}$, $\mathbf{Y}$ and $\mathbf{r}$, can be expressed as

$$\begin{aligned}
\mathbf{S}_{\widehat{\mathbf{Y}}} &= n^{-1}\mathbb{Y}^T\mathbf{P}_{\mathbb{X}}\mathbb{Y} = \mathbf{B}\mathbf{S}_{\mathbf{X}}\mathbf{B}^T, && (2.3)\\
\mathbf{S}_{\mathbf{Y}} &= n^{-1}\mathbb{Y}^T\mathbb{Y} = \mathbf{S}_{\widehat{\mathbf{Y}}} + \mathbf{S}_{\mathbf{Y}|\mathbf{X}}, && (2.4)\\
\mathbf{S}_{\mathbf{Y}|\mathbf{X}} &= n^{-1}\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^T = n^{-1}\mathbb{Y}^T\mathbf{Q}_{\mathbb{X}}\mathbb{Y}, && (2.5)\\
&= \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{Y}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{Y}\mathbf{X}}^T
\end{aligned}$$

where $\mathbf{S}_{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T$ is non-stochastic, $\mathbf{P}_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ denotes the projection onto the column space of $\mathbb{X}$ and $\mathbf{Q}_{\mathbb{X}} = \mathbf{I}_n - \mathbf{P}_{\mathbb{X}}$.

The covariance matrix of $\mathbf{B}$ can be represented using the vec operator to stack its columns,

$$\mathrm{var}\{\mathrm{vec}(\mathbf{B})\} = (\mathbb{X}^T\mathbb{X})^{-1} \otimes \boldsymbol{\Sigma} = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}, \tag{2.6}$$

where $\otimes$ represents the Kronecker product. This variance is typically estimated by substituting the residual covariance matrix for $\boldsymbol{\Sigma}$,

$$\widehat{\mathrm{var}}\{\mathrm{vec}(\mathbf{B})\} = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \mathbf{S}_{\mathbf{Y}|\mathbf{X}}. \tag{2.7}$$

The covariance matrix can be represented also in terms of $\mathbf{B}^T$ by using the $rp \times rp$ commutation matrix $\mathbf{K}_{rp}$ to convert $\mathrm{vec}(\mathbf{B})$ to $\mathrm{vec}(\mathbf{B}^T)$: $\mathrm{vec}(\mathbf{B}^T) = \mathbf{K}_{rp}\mathrm{vec}(\mathbf{B})$ and

$$\mathrm{var}\{\mathrm{vec}(\mathbf{B}^T)\} = n^{-1}\mathbf{K}_{rp}(\mathbf{S}_{\mathbf{X}}^{-1} \otimes \mathbf{\Sigma})\mathbf{K}_{rp}^T = n^{-1}\mathbf{\Sigma} \otimes \mathbf{S}_{\mathbf{X}}^{-1}.$$

Let $\mathbf{e}_j \in \mathbb{R}^r$ denote the indicator vector with a 1 in the j-th position and 0's elsewhere. Then the covariance matrix for the j-th row of $\mathbf{B}$ is

$$\mathrm{var}\{\mathrm{vec}(\mathbf{e}_j^T\mathbf{B})\} = (\mathbf{I}_p \otimes \mathbf{e}_j^T)\mathrm{var}\{\mathrm{vec}(\mathbf{B})\}(\mathbf{I}_p \otimes \mathbf{e}_j) = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes (\mathbf{\Sigma})_{jj},$$

where $(\mathbf{A})_{jj}$ represents the jj-th element of the matrix $\mathbf{A}$. We see from this that the covariance matrix for the j-th row of $\mathbf{B}$ is the same as that from the marginal linear regression of Y_j on $\mathbf{X}$.

Asymptotically $\sqrt{n}\{\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta})\}$ is distributed as a $rp \times 1$ normal random vector with mean 0 and covariance matrix $\mathbf{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{\Sigma}$, where $\mathbf{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \mathbf{S}_{\mathbf{X}}$.

2.2 Introductory illustration

To gain intuition about the working mechanism of envelopes in the context of model (2.1), consider comparing the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of two bivariate normal populations. This problem can be cast into the framework of model (2.1) by letting $\mathbf{Y} = (Y_1, Y_2)^T$ denote the bivariate response vectors and letting X be an indicator variable taking value $X = 0$ in population 1 and $X = 1$ in population 2. Then $\boldsymbol{\alpha} = \boldsymbol{\mu}_1$ is the mean of the first population and $\boldsymbol{\beta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ is the mean difference. The standard estimator of $\boldsymbol{\beta}$ is just the difference in the sample means for the two populations $\mathbf{B} = \bar{\mathbf{Y}}_2 - \bar{\mathbf{Y}}_1$ and the corresponding estimator of $\mathbf{\Sigma}$ is the pooled intra-sample covariance matrix. As in the general multivariate normal model (2.1), this estimator of $\boldsymbol{\beta}$ does not make use of the dependence between the responses and is equivalent to performing two univariate regressions of Y_j on X, $j = 1, 2$. Bringing envelopes into play can lead to a very different estimator of $\boldsymbol{\beta}$, one with substantially smaller variation than the maximum likelihood estimator $\mathbf{B}$ under model (2.1). In the remainder of this section, we illustrate one way in which this can happen.

Figure 2.1 provides a graphical illustration of the working mechanism of envelopes in this setting. In both panels, the two ellipses represent the two normal populations indicated by the predictor $X = 0$ and $X = 1$, and the axes are the responses Y_1 and Y_2. The left panel represents a standard analysis. For inference on β_2, the second element of $\boldsymbol{\beta}$, a standard analysis directly projects the data "y" onto the Y_2 axis following the dashed line

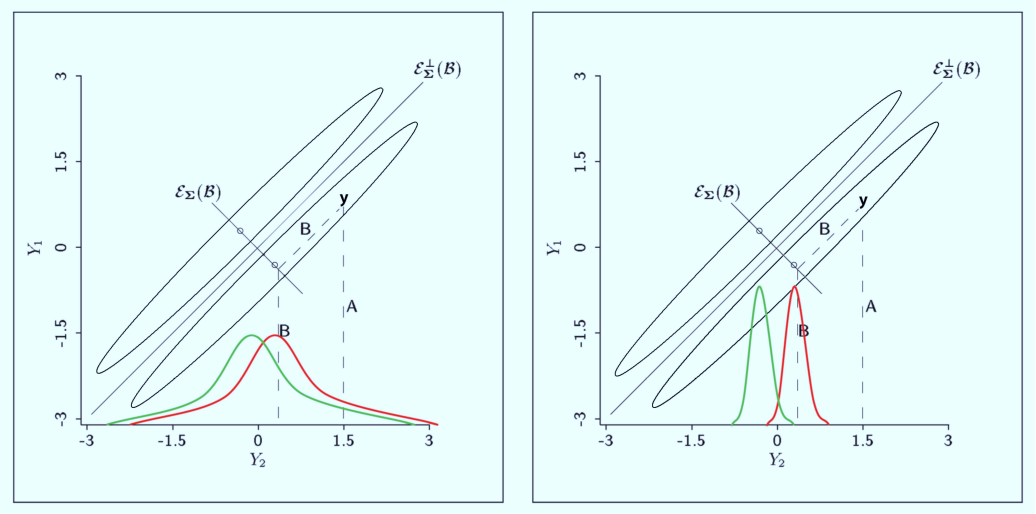


Figure 2.1: Graphical illustration of the working mechanism of the envelope model.

marked A and then proceeds with inference based on the resulting univariate samples. The curves in the left panel stand for the two projected distributions from the two populations. A standard analysis might involve constructing a two-sample t-test on samples drawn from these populations. There is considerable overlap between the two projected distributions, so it may take a large sample size to infer that $\beta_2 \neq 0$ in a standard analysis, while it is clear that the bivariate populations have different means. This illustration is based on $\boldsymbol{\beta}_2$ to facilitate visualization; the same conclusions could be reached using a different linear combination of the elements of $\boldsymbol{\beta}$.

The two populations depicted in Figure 2.1 have the same eigenvectors, as they must because they have equal covariance matrices. The eigenvector corresponding to the smaller eigenvalue is marked by the notation for the envelope $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, where $\mathcal{B} = \text{span}(\boldsymbol{\beta})$. Similarly, the first eigenvector is marked $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$. The two populations have been arranged so that they have equal distributions when projected onto $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$; that is, $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (X = 0) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (X = 1)$, where $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is shortened to $\mathcal{E}$ for use in subscripts. In other words, (i) $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid X \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y}$. Since the populations are normal and $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$ are spanned by eigevectors we also have that (ii) $\mathbf{P}_{\mathcal{E}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid X$. Conditions (i) and (ii) are equivalent to the single condition (iii) $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (\mathbf{P}_{\mathcal{E}}\mathbf{Y}, X) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ which implies that $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ is immaterial to the estimation of $\boldsymbol{\beta}$ and that only $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is material to the same purpose.

Since only $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is material to the estimation of $\boldsymbol{\beta}$, the envelope estimator of β_2 is formed by first projecting the data onto $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ to remove the immaterial information $\mathbf{Q}_{\mathcal{E}}Y$

and extract the material information $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$, and then projecting onto the horizontal axis, as illustrated by the paths marked "B" in Figure 2.1b. Figure 2.1b also shows the resulting projected distributions corresponding to the projected distributions in Figure 2.1a. Now the projected distributions are relatively well-separated and the envelope estimator of β_2 should be much more efficient that the estimator represented in Figure 2.1a. In practice, $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ will be estimated and so the projected distributions illustrated in Figure 2.1b will have a degree of wobble that is taken into account when estimative variation is discussed later in this chapter.

The classic wheat protein data (Fearn, 1983) contains measurements on protein content and the logarithms of near infrared reflectance at six wavelengths across the range 1680-2310 nm, measured on samples ground wheat. To illustrate these ideas associated with Figure 2.1 in data analysis, we use $r = 2$ wavelengths as responses $\mathbf{Y} = (Y_1, Y_2)^T$ and convert the continuous measure of protein content into a categorical predictor X indicating low and high protein (24 and 26 samples, respectively). The complete wheat protein data will be use for illustration later in this chapter.

The mean difference $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ corresponds to the parameter vector $\boldsymbol{\beta}$ in model (2.1), with X representing a binary indicator: $X = 0$ for high protein, and $X = 1$ for low protein wheat. For these data, which are shown in Figure 2.2, the standard estimate of β_2 is -2.1 with a standard error of 9.4 (Figure 2.1a), while the envelope estimate is -4.7 with a standard error of 0.46 (Figure 2.1b). To more fully appreciate the magnitude of this drop in standard errors, we would need for a standard analysis a sample size of $n \sim 20,000$ to reduce the standard error from 9.4 to 0.46. Figure 2.3 shows the projected distributions of the data like those from Figure 2.1. This example will be revisited later.

2.3 The envelope model

To extend the previous discussion to the multivariate linear model (2.1) and characterize situations like that displayed in Figure 2.1, consider subspaces $\mathcal{S} \subseteq \mathbb{R}^r$ with the properties

$$\text{(i) } \mathbf{Q}_{\mathcal{S}}\mathbf{Y}|\mathbf{X} \sim \mathbf{Q}_{\mathcal{S}}\mathbf{Y} \text{ and (ii) } \mathbf{P}_{\mathcal{S}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{Y}|\mathbf{X}. \tag{2.8}$$

These conditions are the extensions of conditions (1.1) to model (2.1). The intersection of all subspace with these properties, which also has these properties, is called the $\mathbf{\Sigma}$-envelope of $\mathcal{B} := \text{span}(\boldsymbol{\beta})$ and represented by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, with $u = \dim(\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}))$. Let $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ and $\mathbf{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ be semi-orthogonal basis matrices for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathcal{E}^{\perp}_{\mathbf{\Sigma}}(\mathcal{B})$. Then

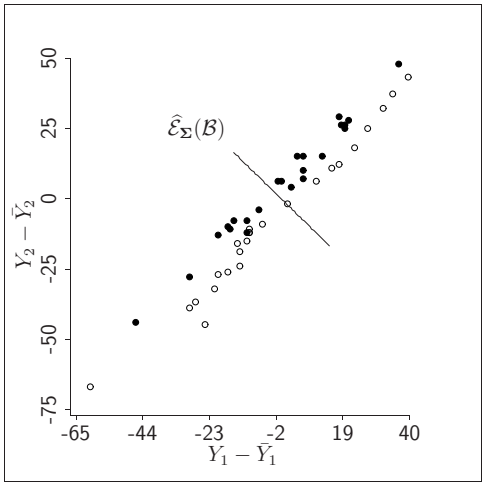
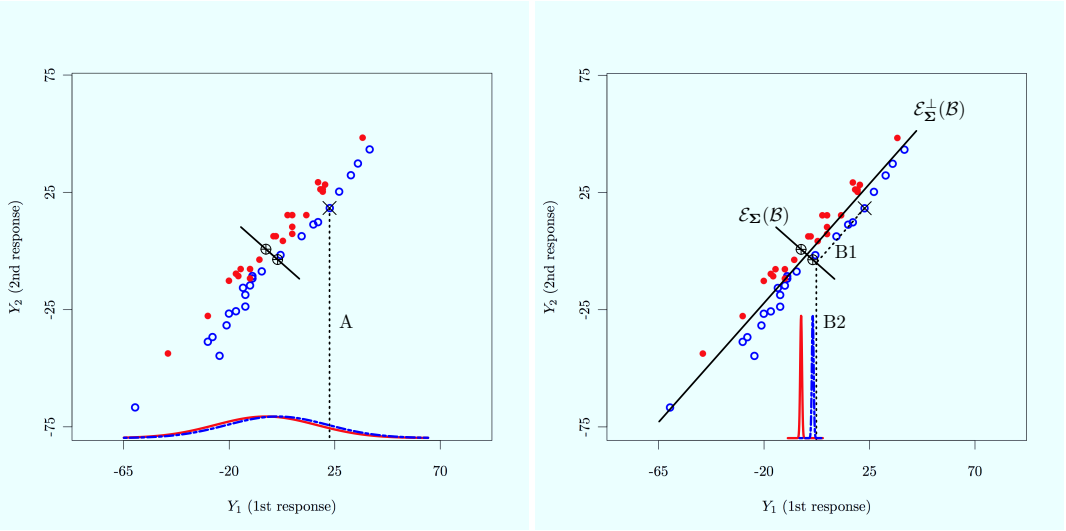


Figure 2.2: Wheat protein data with the estimated envelope superimposed.



a. Standard analysis b. Envelope analysis

Figure 2.3: Standard and envelope analyses for the wheat protein data.

these bases must satisfy

$$\text{(i) } \mathbf{\Gamma}_0^T\mathbf{Y}|\mathbf{X} \sim \mathbf{\Gamma}_0^T\mathbf{Y} \text{ and (ii) } \mathbf{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{\Gamma}_0^T\mathbf{Y}|\mathbf{X}. \tag{2.9}$$

The first condition of (2.9) requires that the distribution of $\mathbf{\Gamma}_0^T\mathbf{Y} = \mathbf{\Gamma}_0^T\boldsymbol{\alpha} + \mathbf{\Gamma}_0^T\boldsymbol{\beta}\mathbf{X} + \mathbf{\Gamma}_0^T\boldsymbol{\varepsilon}$ not depend on $\mathbf{X}$, which holds if and only if $\mathbf{\Gamma}_0^T\boldsymbol{\beta} = 0$. Consequently, the first condition is equivalent to requiring that $\mathcal{B} \subseteq \text{span}(\mathbf{\Gamma})$. The subspace $\text{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$

is called an envelope because it must envelop (contain) $\mathcal{B}$. In Figure 2.1, $\mathcal{B} = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, but in higher dimensions only containment is required.

The second condition of (2.9) holds if and only if $\mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma} = 0$, and consequently we must also have

$$\begin{aligned}\mathbf{\Sigma} &= (\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}})\mathbf{\Sigma}(\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}}) \\ &= \mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{Q}_{\mathbf{\Gamma}} \\ &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T,\end{aligned} \tag{2.10}$$

where $\mathbf{\Omega} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}$ and $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0$ can be interpreted as coordinate matrices relative to $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$. This condition was encountered previously in the envelope model (1.2) for a multivariate mean and it was induced in Figure 2.1 by choosing $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$ to correspond to eigenvectors of $\mathbf{\Sigma}$. The independence in the second condition of (2.9) is stronger than needed for this decomposition, since it requires only that $\mathrm{cov}(\mathbf{\Gamma}^T\mathbf{Y}, \mathbf{\Gamma}_0^T\mathbf{Y}|\mathbf{X}) = 0$. This will be relevant in Section 2.8 when discussing envelopes with non-normal errors.

Separate application of either of the two conditions in (2.9) does not necessarily lead to progress. A subspace $\mathcal{S}$ that decomposes $\mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$ is called a *reducing subspace* of $\mathbf{\Sigma}$ (cf. Proposition 5.1). Such a subspace may have no useful connection with $\mathcal{B}$. If $\mathcal{B}$ is a proper subspace of $\mathbb{R}^r$ then there can be many subspaces $\mathcal{S}$ that contain $\mathcal{B}$, while any particular subspace may not decompose $\mathbf{\Sigma}$. It is only when the two conditions of (2.9) are used in concert that we may obtain useful reductions. Together the two condition of (2.9) are equivalent to the single condition that $\mathbf{\Gamma}_0^T\mathbf{Y}|(\mathbf{\Gamma}^T\mathbf{Y}, \mathbf{X}) \sim \mathbf{\Gamma}_0^T\mathbf{Y}$, which is the general version of the setting for Figure 2.1. Since the distribution of $\mathbf{\Gamma}_0^T\mathbf{Y}$ depends on neither $\mathbf{\Gamma}^T\mathbf{Y}$ nor $\mathbf{X}$, we think of $\mathbf{\Gamma}_0^T\mathbf{Y}$ as being the part of $\mathbf{Y}$ that is immaterial to the regression. Correspondingly, $\mathbf{\Gamma}^T\mathbf{Y}$ is the part of $\mathbf{Y}$ that is material to the regression. The immaterial component $\mathbf{\Gamma}_0^T\mathbf{Y}$ contains no useful information on $\boldsymbol{\beta}$, but it induces extraneous variation in the maximum likelihood estimator of $\boldsymbol{\beta}$ under model (2.1). Envelopes can be use to distinguish the material and immaterial parts of $\mathbf{Y}$ in the estimation process. As discussed later in this chapter, the envelope estimator of $\boldsymbol{\beta}$ may then be more efficient than $\mathbf{B}$, as the variation from the immaterial part is effectively removed.

The actual construction of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ can be characterized in terms of $\mathcal{B}$ and the spectral structure of $\mathbf{\Sigma}$. Suppose that $\mathbf{\Sigma}$ has $q \leq r$ distinct eigenvalues and thus q eigenspaces whose projections are represented by $\mathbf{P}_k$, $k = 1, \ldots, q$. Then, as shown in Proposition 5.2, $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \oplus_{k=1}^q \mathbf{P}_i\mathcal{B}$. This result shows that $\mathcal{B}$ is in fact enveloped by using the eigenspaces of $\mathbf{\Sigma}$. There are only two eigenspaces in Figure 2.1 and $\mathcal{B}$ is the same as the second

eigenspace. Consequently, $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is also equal to the second eigenspace, as shown in the figure.

The coordinate form of an envelope model can now be written as

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \;\; \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \tag{2.11}$$

where the coefficients $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, u denotes the dimension of the envelope $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ is as defined previously. The positive definite matrices $\mathbf{\Omega} \in \mathbb{S}^{u\times u}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ carry the coordinates of $\mathbf{\Sigma}$ with respect to $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$, and $\boldsymbol{\eta} \in \mathbb{S}^{u\times p}$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\mathbf{\Gamma}$.

The total number of free parameters required for the envelope model is

$$\begin{aligned} N_u &= r + pu + u(r-u) + u(u+1)/2 + (r-u)(r-u+1)/2 \\ &= r + pu + r(r+1)/2. \end{aligned} \tag{2.12}$$

This count arises as follows. The first term on the right hand side of (2.12) corresponds to the intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$ and the second term corresponds to the unconstrained coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$. The last two terms correspond to $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$. Their parameter counts arise because, for any integer $k > 0$, it takes $k(k+1)/2$ numbers to specify a nonsingular $k \times k$ symmetric matrix. The third term, $u(r-u)$, which corresponds to $\mathbf{\Gamma}$, arises as follows. The matrix $\mathbf{\Gamma}$ is not identifiable, since, for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$, replacing $\mathbf{\Gamma}$ with $\mathbf{\Gamma O}$ results in an equivalent model. However, $\mathrm{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is identifiable. The parameter space for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is a Grassmann manifold $\mathcal{G}(u,r)$ of dimension u in $\mathbb{R}^r$; that is, the collection of all u-dimensional subspaces of $\mathbb{R}^r$. From basic properties of Grassmann manifolds it is known that $u(r-u)$ parameters are needed to specify an element of $\mathcal{G}(u,r)$ (Edelman, Tomás, and Smith, 1998). Once $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is determined, so is its orthogonal complement $\mathrm{span}(\mathbf{\Gamma}_0)$, and no additional free parameters are required. The difference between the total parameter count for the full model (2.1) with $r = u$ and the envelope model (2.11) with $u < r$ is therefore $p(r-u)$.

The number of parameters $u(r-u)$ needed to identify a subspace can be seen intuitively as follows. It requries ru parameters to specify uniquely an unconstrained matrix $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$. But when dealing with subspaces $\mathrm{span}(\mathbf{\Gamma}) = \mathrm{span}(\mathbf{\Gamma A})$ for any non-singular $\mathbf{A} \in \mathbb{R}^{u\times u}$. It requires u^2 parameters to specify an $\mathbf{A}$. When dealing with subspaces we need to specify $\mathbf{\Gamma}$ only up to transformations $\mathbf{A}$. Consequently, specifying a subspace takes $ru - u^2 = u(r-u)$ parameters.

A specific envelope model is identified by the value of u, with the full model (2.1) occurring when $u = r$. All envelope models are nested within the standard model, but

two envelope models with different values of u are not necessarily nested. To see this, it is enough to realize that the number of free parameters needed to specify an element of $\mathcal{G}(u,r)$ is the same for $u=1$ and $u=r-1$. In full generality, u is a model selection parameter that can be chosen using traditional reasoning, as discussed in Section 2.6.

If $u=r$, then $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})=\mathbb{R}^r$, the envelope model degenerates to the standard model and enveloping offers no gain. If $r\leq p$ and $\dim(\mathcal{B})=r$ then again the envelope model reduces to the standard model. However, if (i) $r>p$ or (ii) if $\dim(\mathcal{B})<r\leq p$ then efficiency gains are possible. These gains arise from two sources. The first is the parameter count. Since the number of parameters in the envelope model is less than that in the standard model, we can expect some efficiency gains from parsimony. But the second source is where we have the to potential to realize massive gains. Recall that one role of the envelope model is to in effect remove the immaterial component from the estimation of $\boldsymbol{\beta}=\mathbf{\Gamma}\boldsymbol{\eta}$. This immaterial element is manifested in the model through the term $\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ that appears in the covariance structure. If it is substantially larger than the corresponding element $\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T$ that is associated with the material information and retained in the estimation process then large gains are possible by enveloping.

To explore these notions a little further and in anticipation of maximum likelihood estimation discussed in the next section, we conclude this section with a brief discussion of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{\Gamma}}$ of $\boldsymbol{\beta}$ when $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is known and represented by a semi-orthogonal basis matrix $\mathbf{\Gamma}$. In this case it can be demonstrated straightforwardly that the MLE of $\boldsymbol{\beta}$ under the envelope model (2.11) is simply the projection of the standard estimator onto $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$: $\widehat{\boldsymbol{\beta}}_{\mathbf{\Gamma}}=\mathbf{P}_{\mathbf{\Gamma}}\mathbf{B}$. Using (2.6), the variance of this estimator is

$$\begin{aligned}\mathrm{var}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{\Gamma}})\} &= \mathrm{var}\{\mathrm{vec}(\mathbf{P}_{\mathbf{\Gamma}}\mathbf{B})\}\\ &= (\mathbf{I}_p\otimes\mathbf{P}_{\mathbf{\Gamma}})\mathrm{var}\{\mathrm{vec}(\mathbf{B})\}(\mathbf{I}_p\otimes\mathbf{P}_{\mathbf{\Gamma}})\\ &= n^{-1}(\mathbf{I}_p\otimes\mathbf{P}_{\mathbf{\Gamma}})(\mathbf{S}_{\mathbf{X}}^{-1}\otimes\mathbf{\Sigma})(\mathbf{I}_p\otimes\mathbf{P}_{\mathbf{\Gamma}})\\ &= n^{-1}\mathbf{S}_{\mathbf{X}}^{-1}\otimes\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T. \qquad (2.13)\end{aligned}$$

Comparing this to the variance of the standard estimator (2.6),

$$\mathrm{var}\{\mathrm{vec}(\mathbf{B})\}-\mathrm{var}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{\Gamma}})\}=n^{-1}\mathbf{S}_{\mathbf{X}}^{-1}\otimes\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T\geq 0.$$

From this we conclude that if the variance of the immaterial part of $\mathbf{Y}$, $\mathrm{var}(\mathbf{\Gamma}_0^T\mathbf{Y})=\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ is large relative to the variance of the material part of $\mathbf{Y}$, $\mathrm{var}(\mathbf{\Gamma}^T\mathbf{Y})=\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T$ then the gain from the envelope model can be substantial. This corresponds to the population gain represented schematically in Figure 2.1. Using the spectral norm $\|\cdot\|$ as

a measure of overall size, the envelope model will be advantageous when $\|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\| = \|\boldsymbol{\Omega}\| \ll \|\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\| = \|\boldsymbol{\Omega}_0\|$. It may be clear from Figure 2.3 that $\|\boldsymbol{\Omega}\| \ll \|\boldsymbol{\Omega}_0\|$ for the wheat protein data. The envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ will normally be estimated in practice and these results will then be mitigated by the variability in its estimator. Nevertheless, experience has shown that they are a useful indicator of the kinds of regressions in which envelopes offer substantial gains.

2.4 Maximum likelihood estimation

2.4.1 Derivation

In this section we discuss the derivation of the parameters in the envelope model (2.11) assuming that the dimension $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$ of the envelope is known. The dimension u is essentially a model-selection parameter; methods for selecting it are discussed in Section 2.6.

The log likelihood $L_u(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ under model (2.11) with known u can be expressed as

$$\begin{aligned} L_u &= -(nr/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| \\ &\quad -(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_i)^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_i) \\ &= -(nr/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Omega}| - (n/2)\log|\boldsymbol{\Omega}_0| \\ &\quad -(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_i)^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_i), \end{aligned}$$

where the second equality arises by applying the third and fourth conclusions of Corollary 5.1. Also, while the likelihood function depends on $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, we have written it in terms of the semi-orthognal basis matrix $\boldsymbol{\Gamma}$ since Grassmann optimization will be performed in terms of bases.

Since the predictors are centered, $\sum_{i=1}^{n}\mathbf{X}_i = 0$, it follows that the MLE of $\boldsymbol{\alpha}$ is $\widehat{\boldsymbol{\alpha}} = \bar{\mathbf{Y}}$. Substituting this into the likelihood function and then decomposing $\mathbf{Y}_i - \bar{\mathbf{Y}} = \mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \bar{\mathbf{Y}}) + \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \bar{\mathbf{Y}})$ and simplifying, we arrive at the first partially maximized log likelihood,

$$L_1(\boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) = -(nr/2)\log(2\pi) + L_{11}(\boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}) + L_{12}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}_0),$$

where

$$\begin{aligned} L_{11} &= -(n/2)\log|\boldsymbol{\Omega}| \\ &\quad -(1/2)\sum_{i=1}^{n}\{\boldsymbol{\Gamma}^T(\mathbf{Y}_i-\bar{\mathbf{Y}})-\boldsymbol{\eta}\mathbf{X}_i\}^T\boldsymbol{\Omega}^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{Y}_i-\bar{\mathbf{Y}})-\boldsymbol{\eta}\mathbf{X}_i\} \\ L_{12} &= -(n/2)\log|\boldsymbol{\Omega}_0| - (1/2)\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbf{Y}_i-\bar{\mathbf{Y}}). \end{aligned}$$

Holding $\boldsymbol{\Gamma}$ fixed, L_{11} can be seen as the log likelihood for the multivariate regression of $\boldsymbol{\Gamma}^T(\mathbf{Y}_i-\bar{\mathbf{Y}})$ on $\mathbf{X}_i$, and thus L_{11} is maximized over $\boldsymbol{\eta}$ at the value $\boldsymbol{\eta}=\boldsymbol{\Gamma}^T\mathbf{B}$. Substituting this into L_{11} and simplifying we obtain a partially maximized version of L_{11}

$$L_{21}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}),\boldsymbol{\Omega}) = -(n/2)\log|\boldsymbol{\Omega}| - (1/2)\sum_{i=1}^{n}(\boldsymbol{\Gamma}^T\mathbf{r}_i)^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{r}_i,$$

where, as defined previously, $\mathbf{r}_i$ is the i-th residual vector from the fit of the standard model. From this it follows immediately that, still with $\boldsymbol{\Gamma}$ fixed, L_{21} is maximized over $\boldsymbol{\Omega}$ at $\boldsymbol{\Omega}=\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}$. Consequently, we arrive at the third partially maximized log likelihood $L_{31}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = -(n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| - nu/2$. By similar reasoning, that value of $\boldsymbol{\Omega}_0$ that maximizes $L_{12}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}),\boldsymbol{\Omega}_0)$ is $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0$. This lead to the maximization of $L_{22}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = -(n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0| - n(r-u)/2$.

Combining the above steps, we arrive at the partially maximized form

$$L_2(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| - (n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0|.$$

Lemma 6.1 enables us to conclude that $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ given in (2.15) is equal to $\arg\max L_2(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$ Using this result, the fully maximized log likelihood at a selected value of u can be written as

$$\begin{aligned} \widehat{L}_u &= -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}}| \\ &\quad -(n/2)\log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\boldsymbol{\Gamma}}| - (n/2)\log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\widehat{\boldsymbol{\Gamma}}|. \end{aligned} \tag{2.14}$$

Summarizing, the MLEs $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B})$ of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and of the remaining parameters are deter-

mined as

$$\begin{aligned}
\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) &= \text{span}\{\arg\min_{\mathbf{G}} (\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0|)\} \qquad (2.15)\\
\widehat{\boldsymbol{\eta}} &= \widehat{\mathbf{\Gamma}}^T\mathbf{B},\\
\widehat{\boldsymbol{\beta}} &= \widehat{\mathbf{\Gamma}}\widehat{\boldsymbol{\eta}} = \widehat{\mathbf{P}}_{\mathcal{E}}\mathbf{B},\\
\widehat{\mathbf{\Omega}} &= \widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\mathbf{\Gamma}},\\
\widehat{\mathbf{\Omega}}_0 &= \widehat{\mathbf{\Gamma}}_0^T\mathbf{S}_{\mathbf{Y}}\widehat{\mathbf{\Gamma}}_0,\\
\widehat{\mathbf{\Sigma}} &= \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{\Gamma}}^T + \widehat{\mathbf{\Gamma}}_0\widehat{\mathbf{\Omega}}_0\widehat{\mathbf{\Gamma}}_0^T,
\end{aligned}$$

where $\min_{\mathbf{G}}$ is over all partitioned orthogonal matrices $(\mathbf{G}, \mathbf{G}_0)$ with $\mathbf{G} \in \mathbb{R}^{r\times u}$, $\widehat{\mathbf{\Gamma}}$ is any semi-orthogonal basis matrix of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\widehat{\mathbf{\Gamma}}_0$ is any semi-orthogonal basis matrix for the orthogonal complement of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$. The envelope estimator of $\boldsymbol{\beta}$ is a form of shrinkage estimator since it is simply the projection of the standard model estimator onto the estimated envelope. As in Chapter 1, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{\Sigma}}$ are invariant to the selection of basis $\widehat{\mathbf{\Gamma}}$ and thus are unique estimators, but the remaining estimators $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Omega}}_0$ are basis dependent and thus not unique.

2.4.2 Insights into $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$

The estimator $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ is a novel construction, so in this section we provide different forms for it that might aid intuition. First, $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ can be re-expressed as

$$\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \arg\min_{\mathcal{S}\in\mathcal{G}(u,r)} \{\log|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathcal{S}}|_0 + \log|\mathbf{Q}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\mathcal{S}}|_0\}, \qquad (2.16)$$

where $\|\cdot\|_0$ denotes the product of the non-zero eigenvalues of the matrix argument, and minimization $\min_{\mathcal{S}\in\mathcal{G}(u,r)}$ is over the Grassmann manifold $\mathcal{G}(u,r)$ of dimension u in $\mathbb{R}^r$.

Second, using Lemma 6.1 on (2.15) the estimator can be written as

$$\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \text{span}\{\arg\min_{\mathbf{G}} \log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|\}, \qquad (2.17)$$

where the minimization is over semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{r\times u}$. The term $\log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}|$ will pull the solutions $\mathbf{G}$ for which the residuals from the regression of $\mathbf{G}^T\mathbf{Y}$ on $\mathbf{X}$ are relatively small, while term $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$ will exert a preference for the subspaces associated with the eigenvectors associated with the larger eigenvalues of $\mathbf{S}_{\mathbf{Y}}$. In this way the solution provides a balance between the fit of the reduced responses $\mathbf{G}^T\mathbf{Y}$ and the principal components of $\mathbf{Y}$.

To express the next form, let $\mathbf{V} = \mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{X}$ and $\mathbf{Z}_{\mathbf{G}} = \mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{-1/2}\mathbf{G}^T\mathbf{Y}$ denote the standardized versions of $\mathbf{X}$ and $\mathbf{G}^T\mathbf{Y}$. Then

$$\begin{aligned}
\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G} &= |\mathbf{G}^T(\mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{YX}}^T)\mathbf{G}| \\
&= |\mathbf{S}_{\mathbf{G}^T\mathbf{Y}} - \mathbf{S}_{\mathbf{G}^T\mathbf{Y},\mathbf{V}}\mathbf{S}_{\mathbf{G}^T\mathbf{Y},\mathbf{V}}^T| \\
|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| &= |\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}| \times |\mathbf{I}_p - \mathbf{S}_{\mathbf{G}^T\mathbf{Y},\mathbf{V}}^T\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{G}^T\mathbf{Y},\mathbf{V}}| \\
&= |\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}| \times |\mathbf{I}_p - \mathbf{S}_{\mathbf{V}\mathbf{Z}_{\mathbf{G}}}\mathbf{S}_{\mathbf{V}\mathbf{Z}_{\mathbf{G}}}^T| \\
&= |\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}| \times |\mathbf{S}_{\mathbf{V}|\mathbf{Z}_{\mathbf{G}}}|
\end{aligned}$$

Combining this form with (2.17) we have

$$\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}\{\arg\min_{\mathbf{G}} \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}| + \log|\mathbf{S}_{\mathbf{V}|\mathbf{Z}_{\mathbf{G}}}|\}, \qquad (2.18)$$

The sum of the first two terms $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$ of this objective function is always non-negative and it equals zero when $\mathrm{span}(\mathbf{G})$ is any u-dimensional reducing subspace of $\mathbf{S}_{\mathbf{Y}}$. The role of these terms then is to pull the solutions toward the reducing subspaces of $\mathbf{S}_{\mathbf{Y}}$. The third term $\log|\mathbf{S}_{\mathbf{V}|\mathbf{Z}_{\mathbf{G}}}|$ measures the goodness of fit of the inverse regression of the standardized predictors $\mathbf{V}$ on the standardized reduced response $\mathbf{Z}_{\mathbf{G}}$. The inverse regression of predictors on responses plays a prominent role in other dimension reduction contexts, like sufficient dimension reduction (see Cook, 1998). The solution to (2.18) will then be a balance between its closeness to a reducing subspace of $\mathbf{S}_{\mathbf{Y}}$ and the fit of the corresponding inverse regression.

Objective functions defined on subspaces, as is the case here, typically have multiple local optima, and (2.15)–(2.18) are no exception. For this reason starting values for iteration tend to be quite important, particularly when the signal is weak. Effective starting can often be found by using the sequential algorithm of Section 6.5 with $\mathbf{U} = \mathbf{S}_{\widehat{\mathbf{Y}}}$ and $\mathbf{M} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. Since this sample sequential algorithm depends continuously only on $\mathbf{S}_{\widehat{\mathbf{Y}}}$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, which are root-$n$ consistent estimators of their population counterparts, it provides a root-n consistent estimator of the envelope assuming that u is known. One Newton-Raphson iteration from the algorithm's solution should then provide an estimator of the envelope that is asymptotically equivalent to the normal-theory MLE which depends on the same quantities.

2.5 Asymptotic variance of $\widehat{\beta}$

In this section we discuss the asymptotic distribution of $\mathrm{vec}(\widehat{\boldsymbol{\beta}})$. The asymptotic distribution was derived and discussed by Cook, Chiaromonte and Li (2010). Here we discuss only the final result, since that will be likely be of interest in most applications.

In preparation, recall that the Fisher information $\mathbf{J}$ for $(\mathrm{vec}^T(\boldsymbol{\beta}), \mathrm{vech}^T(\boldsymbol{\Sigma}))^T$ in the standard model (2.1) is

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix},$$

where $\mathbf{E}_r$ is the expansion matrix that satisfies $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_r \mathrm{vech}(\mathbf{A})$ for $\mathbf{A} \in \mathbb{S}^{r\times r}$, and $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n > 0$. Let $\mathbf{V}_{\mathrm{sm}} = \mathbf{J}^{-1}$ be the asymptotic variance of the MLE under the standard model. Define

$$\mathbf{U} = \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_u \otimes \mathbf{I}_{r-u}.$$

Then $\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta}))$ is asymptotically normal with mean 0 and variance

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{U}^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T), \tag{2.19}$$

If $u = r$, then $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$, and the second term on the right hand side of (2.19) does not appear. The first term on the right hand side of (2.19) is the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ when $\boldsymbol{\Gamma}$ is known, and the second term can be interpreted as the "cost" of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The total on the right does not exceed $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$, which is the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ from the standard model; that is,

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{B})] - \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] \geq 0.$$

The asymptotic variance (2.19) can be re-expressed informatively as

$$\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})] = \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})]. \tag{2.20}$$

The first term $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})]$ on the right hand side is the asymptotic variance of the MLE $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ of $\boldsymbol{\beta}$ when $\boldsymbol{\Gamma}$ is known (cf. (2.13)), and in the second term $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})]$ is the asymptotic variance of the MLE $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}$ of $\boldsymbol{\beta}$ when $\boldsymbol{\eta}$ is known, both terms corresponding to asymptotic variances in multivariate linear models. The role played by $\mathbf{Q}_{\boldsymbol{\Gamma}}$ is to orthogonalize these random matrices so that their contributions to the net asymptotic variance are additive.

To gain some intuition about settings in which the envelope model offers gains over the standard model, write the asymptotic variance of the estimate $\mathbf{B}$ of $\boldsymbol{\beta}$ under the standard model in terms of the envelope parameters:

$$\text{avar}[\sqrt{n}\text{vec}(\mathbf{B})] = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.$$

Let $\boldsymbol{\Delta} = \text{avar}[\sqrt{n}\text{vec}(\mathbf{B})] - \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})]$. Then we have

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T - (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{U}^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T) \geq 0.$$

Next consider the special case in which $\boldsymbol{\beta}$ has full column rank p, $\boldsymbol{\Omega} = \omega\mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = \omega_0\mathbf{I}_{r-u}$. As a consequence of this structure, we see that $\boldsymbol{\Sigma}$ has two eigenspaces, one corresponding to ω and the other to ω_0. Since $\mathcal{B}$ is contained in the eigenspace corresponding to ω, we must have $\mathcal{B} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, $u = p$, $\boldsymbol{\Gamma} = \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1/2}$ and $\boldsymbol{\eta} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2}$. Then after a little algebra we can write

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 - \boldsymbol{\eta}^T\{\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\omega_0^{-1} + (\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2)\mathbf{I}_u\}^{-1}\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T.$$

Simplifying $\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2 = (\omega - \omega_0)^2/\omega\omega_0$ and using the fact that $\boldsymbol{\eta} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2} \in \mathbb{R}^{p\times p}$ is non-singular,

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 - \{\boldsymbol{\Sigma}_{\mathbf{X}} + \omega^{-1}(\omega - \omega_0)^2(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \geq 0$$

Recall that if $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is known, the envelope model will offer substantial gains when $\omega \ll \omega_0$. The second term on the right hand side of the last expression shows the cost of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. From the expression we see that $\boldsymbol{\Delta}$ will be relatively large when again $\omega \ll \omega_0$, so the cost is relatively small. The gain $\boldsymbol{\Delta}$ can also be small if $\omega \neq \omega_0$ and the signal represented by $\boldsymbol{\beta}^T\boldsymbol{\beta}$ is sufficiently small. The gain $\boldsymbol{\Delta}$ will be small for a sufficiently large signal $\boldsymbol{\beta}^T\boldsymbol{\beta}$. In short, we can expect notable gains from envelopes when the immaterial variation $\boldsymbol{\omega}_0$ is large relative to the material variation ω, particularly in the presence of a weak signal. The gains can be small when the immaterial variation is small or the signal is strong.

If $\boldsymbol{\beta}$ has full column rank and $\omega = \omega_0$, so $\boldsymbol{\Sigma} = \omega\mathbf{I}_r$, then the asymptotic variance of the envelope estimator is the same as the asymptotic variance of the usual MLE, $\text{avar}[\sqrt{n}\text{vec}(\mathbf{B})] = \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})]$.

However, gains are still possible when $\omega = \omega_0$ if the rank of $\boldsymbol{\beta}$ is less than p. In that case, $\mathcal{B} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, $\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2 = 0$ and

$$\begin{aligned}\boldsymbol{\Delta} &= \{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \boldsymbol{\eta}^T(\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T)^{-1}\boldsymbol{\eta}\} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \\ &= \mathbf{Q}_{\boldsymbol{\eta}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \geq 0.\end{aligned}$$

2.6 Selecting u

The dimension u of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ was assumed to be known in the discussion of $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}))$, but this will likely not be so in applications. u is essentially a model-selection parameter, and it can be selected by using sequential likelihood ratio testing, an information criterion like AIC or BIC, cross validation or a holdout sample.

As mentioned in Section 2.3, two envelope models with different value for u are not necessarily nested, but an envelope model is always nested within the standard model, which arises when $u = r$. The likelihood ratio for testing an envelope model against the standard model can be cast as a test of the hypothesis $u = u_0$ versus the alternative $u = r$, $u_0 < r$. The likelihood ratio statistic for this hypothesis is $\Lambda(u_0) = 2(\widehat{L}_r - \widehat{L}_{u_0})$, where $\widehat{L}_{u_0}$ is the fully maximized envelope log likelihood given in (2.14) and $\widehat{L}_r$ is the maximized log likelihood under the standard model, $\widehat{L}_r = -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|$, giving

$$\Lambda(u_0) = n\log|\mathbf{S}_{\mathbf{Y}}| + n\log|\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\mathbf{\Gamma}}| + n\log|\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\widehat{\mathbf{\Gamma}}| - n\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|. \quad (2.21)$$

Under the null hypothesis this statistic is distributed asymptotically as a chi-squared random variable with $p(r - u_0)$ degrees of freedom. These likelihood ratio tests can be used sequentially to estimate u: Starting with $u_0 = 0$, test the hypothesis $u = u_0$ against $u = r$ at a selected level. If the hypothesis is rejected, increment u_0 by 1 and test again. The estimate of u is the first hypothesized value that is not rejected. We indicate this estimator using the notation LRT(α) .

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

$$\widehat{u} = \arg\min_u\{-2\widehat{L}_u + h(n)N_u\}, \quad (2.22)$$

where N_u is the number of envelope parameters given in (2.12) and $h(n) = \log n$ for BIC and $h(n) = 2$ for AIC.

2.7 Fitted values and predictions

The previous asymptotic results can be used to derive the asymptotic distribution of the fitted values, as well as the asymptotic prediction variance. The fitted values at a particular $\mathbf{X}$ can be written as $\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\beta}}\mathbf{X} = (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{vec}(\widehat{\boldsymbol{\beta}})$. Hence the fitted value $\widehat{\mathbf{Y}}$ has the following asymptotic distribution

$$\sqrt{n}(\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})) \xrightarrow{\mathcal{L}} N(0, \mathrm{avar}[\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})]. \quad (2.23)$$

Using (2.20) the asymptotic variance in this distribution can be expressed informatively as

$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})] &= (\mathbf{X}^T \otimes \mathbf{I}_r)\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})](\mathbf{X} \otimes \mathbf{I}_r)) \\ &= (\mathbf{X}^T \otimes \mathbf{I}_r)\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})](\mathbf{X} \otimes \mathbf{I}_r) \\ &\quad + (\mathbf{X}^T \otimes \mathbf{I}_r)\text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})](\mathbf{X} \otimes \mathbf{I}_r) \\ &= \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}\mathbf{X})] + \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}\mathbf{X})].\end{aligned}$$

Consequently, the variance of a fitted value has the same decomposition as the variance of $\widehat{\boldsymbol{\beta}}$ discussed previously.

Turning to prediction, suppose that, at some value of $\mathbf{X}$ we observe a new $\mathbf{Y}$, say $\mathbf{Y}_{\text{new}}$, independently of the past observations. Then

$$\begin{aligned}&\text{E}[(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{new}})(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{new}})^T] \\ &\quad = \text{E}[(\widehat{\mathbf{Y}} - \text{E}(\widehat{\mathbf{Y}}))(\widehat{\mathbf{Y}} - \text{E}(\widehat{\mathbf{Y}}))^T] + \text{E}[(\text{E}(\widehat{\mathbf{Y}}) - \mathbf{Y}_{\text{new}})(\text{E}(\widehat{\mathbf{Y}}) - \mathbf{Y}_{\text{new}})^T],\end{aligned}$$

where the cross-product terms vanish because $\mathbf{Y}_{\text{new}}$ and $\widehat{\mathbf{Y}}$ are independent. Combining this with expression (2.23), we see that the mean squared error of the prediction is approximated by

$$\begin{aligned}\text{E}[(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{new}})(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{new}})^T] &= n^{-1}\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})] + \boldsymbol{\Sigma} + o(n^{-1}) \qquad (2.24) \\ &= n^{-1}\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}\mathbf{X})] + n^{-1}\text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}\mathbf{X})] \\ &\quad + \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T + o(n^{-1}).\end{aligned}$$

Envelope models can be quite effective at reducing $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})]$, but they have no impact on the underlying variance $\boldsymbol{\Sigma}$. Envelopes give greatest gain when the immaterial variation $\text{var}(\boldsymbol{\Gamma}^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Omega}_0$ is large relative to the material variation $\text{var}(\boldsymbol{\Gamma}^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Omega}_0$, the immaterial variation being effectively ruled out by envelopes during estimation. Nevertheless, the immaterial variation is still be present in $\boldsymbol{\Sigma}$ and consequently the advantages that envelopes bring in the estimation of $\boldsymbol{\beta}$ may not be present to the same degree in prediction. This can be seen in the schematic illustration of Figure 2.1, where the distributions represented in the left-hand display contribute to prediction, but not to estimation as shown in the right-hand display. Greater predictive gain might be realized by using partial envelopes for prediction, as discussed later in Section 3.4, and by using envelopes for predictor reduction, as discussed in a later chapter.

2.8 Non-normal errors

Again consider model (2.1), but now relax the condition that the errors are normally distributed. The structure of an envelope described in Definition 5.2 requires only $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$; it does not require normality. This implies that the coordinate form of the envelope model (2.11) is still applicable with non-normal errors, although we no longer necessarily have that $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y}|\mathbf{X}$. Nevertheless, the goal under model (2.11) remains the estimation of $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$.

Lacking knowledge of the distribution of the errors, we need to decide how to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. One natural route is to base estimation on the least squares estimators $\mathbf{B}$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ by selecting an objective function to fit the mean and variance structures of model (2.11). There are likely many ways to proceed, but one good way is to use the partially maximized log likelihood $\widehat{L}_u$ (2.14) to fill this role for the purpose of estimating the envelope. It can be used straightforwardly since it is a function of only $\mathbf{B}$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. The remaining parameters are then estimated as described in Section 2.4. Since we are not assuming normality, these estimators no longer inherit optimality properties from general likelihood theory, so a different approach is needed to study their properties.

Lemma 2.1 *The sample matrices* $\mathbf{B}$, $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ *and* $\mathbf{S}_{\mathbf{Y}}$ *are* $\sqrt{n}$ *consistent estimators of their population counterparts* $\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ *and* $\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Sigma} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T$.

Recall from (2.15) that $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \operatorname{span}\{\arg\min_{\mathbf{G}} (\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0|)\}$. It follows from this lemma that $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}|+\log|\mathbf{G}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0|$ converges in probability to $\log|\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T(\boldsymbol{\Sigma} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T)\mathbf{G}_0|$. This population objective function is covered by Proposition 6.2, and consequently it follows that

$$\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \operatorname{span}\{\arg\min_{\mathbf{G}} (\log|\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{G}_0|)\},$$

and thus that the normal-theory objective function recovers $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ in the population without actually assuming normality.

Going further, assume that the errors have finite fourth moments and that as $n \to \infty$ the maximum diagonal element of $\mathbf{P}_{\mathbb{X}}$ converges to 0 and the minimum eigenvalue of $\mathbf{S}_{\mathbf{X}}$ is bounded away from 0. Under these conditions

$$\sqrt{n}\begin{pmatrix} \operatorname{vec}(\mathbf{B}) - \operatorname{vec}(\boldsymbol{\beta}) \\ \operatorname{vech}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}) - \operatorname{vech}(\boldsymbol{\Sigma}) \end{pmatrix}$$

converges to a normal random vector with mean 0 and non-singular covariance matrix (Su and Cook, Biometrika inner envelope paper). It then follows from Shapiro (1986) that

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\widehat{\boldsymbol{\Sigma}}) - \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}$$

also converges to a normal random vector with mean 0 and non-singular covariance matrix. Consequently, using the normal likelihood for estimation under non-normality still produces asymptotically normal $\sqrt{n}$-consistent estimators.

Efficiency gains, as illustrated in Figure 2.1, still accrue without normality, but now they are judged relative to the least squares estimators $\mathbf{B}$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ rather than maximum likelihood estimators. In effect, the contours in Figure 2.1 are no longer the contours of the distribution of $\mathbf{Y}|X$, but are now contours of the quadratic function $\mathbf{x}^T\boldsymbol{\Sigma}\mathbf{x}$. Additionally, the normal-theory asymptotic variances given in Section 2.5 are no longer applicable. While expressions for the asymptotic variances can be derived, it will likely be difficult use them as the basis for estimated variances in practice. The bootstrap offers a practically useful alternative.

2.9 Bootstrap

The residual bootstrap can be implemented in the context of model (2.11) as follows.

1 Fit the envelope model, and get the MLE's $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$. The fitted value for the i-th data is $\widehat{\mathbf{Y}}_i = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\beta}}\mathbf{X}_i$, $i = 1, \cdots, n$. The residuals are $\mathbf{r}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$, $i = 1, \cdots, n$.

2 For each bootstrap replication, sample the residual $\mathbf{r}_i$'s with replacement to get the bootstrap residuals $\mathbf{r}_i^*, \cdots, \mathbf{r}_n^*$.

3 Create a new dataset as $\mathbf{Y}_i^* = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\beta}}\mathbf{X}_i + \mathbf{r}_i^*$, $i = 1, \cdots, n$.

4 Fit the envelope model to the new data $(\mathbf{X}_i, \mathbf{Y}_i^*)$, $i = 1, \cdots, n$, and get $\widehat{\boldsymbol{\alpha}}^*$ and $\widehat{\boldsymbol{\beta}}^*$.

5 Repeat 3-5 B times, getting B $\widehat{\boldsymbol{\alpha}}^*$'s and $\widehat{\boldsymbol{\beta}}^*$'s. Calculate the sample covariance matrix of the $\widehat{\boldsymbol{\beta}}^*$'s. The square root of its diagonal elements gives the estimated standard errors of the elements of $\widehat{\boldsymbol{\beta}}$.

This procedure works the same way for the standard model.

To illustrate application of the bootstrap, we consider an expanded version of the wheat protein data discussed at the end of Section 2.2, now with responses measured at six wavelengths instead of two. Two hundred (B) bootstrap samples were used throughout. The first part of Table 2.1 shows the estimated coefficients under the standard model along with good agreement between their bootstrap and asymptotic standard errors. The dimension of the envelope model was estimated to be $u = 1$ by BIC. Conditioning on this value, the second part of Table 2.1 shows the estimated envelope coefficients and the corresponding bootstrap and asymptotic standard errors. There is again good agreement between the standard errors, which are much smaller than those for the standard model. The advantages of the envelope model in this fit are indicated roughly by the sizes of $\widehat{\mathbf{\Omega}} = 7.88$ and $\|\widehat{\mathbf{\Omega}}_0\| = 4855$. Thus the envelope model has an apparent advantage because the variation $\widehat{\mathbf{\Omega}}_0$ in the estimated immaterial part of $\mathbf{Y}$ is considerably larger than the variation $\widehat{\mathbf{\Omega}}$ in the estimated material part of $\mathbf{Y}$.

Table 2.1: Bootstrap and estimated asymptotic standard errors of the six elements in $\widehat{\boldsymbol{\beta}}$ under standard model and envelope model for the wheat protein data with six responses.

1. Full model

B	3.27	8.03	7.52	-2.06	3.22	0.65
Bootstrap se	9.87	8.12	8.70	9.65	13.90	5.48
Asymptotic se	9.78	8.12	8.70	9.49	13.65	5.39

2. Envelope model with $u = 1$

$\widehat{\boldsymbol{\beta}}$	-1.06	4.47	3.68	-5.97	0.69	-1.60
Bootstrap se	0.35	0.48	0.39	0.64	0.20	0.69
Asymptotic se	0.35	0.43	0.35	0.59	0.21	0.86

2.10 Illustrations of envelopes for response reduction

In this section we present several illustrations on the use of envelopes for reduction of the response. All illustrations show some advantages for envelopes, but this is not meant to imply that there are advantages in all multivariate regressions. The goal here is to reinforce the theory and illustrate broadly how envelopes work. Several of the examples involve a bivariate response $\mathbf{Y} \in \mathbb{R}^2$ and a single binary predictor because comprehensive

graphical representations are possible in this setting. We begin by revisiting the motivating illustration of Section 2.2.

We know from Lemma 5.2 and Proposition 5.2 that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is spanned by some subset of the eigenvectors of $\mathbf{\Sigma}$. In the schematic illustration of Section 2.2 the eigenvalues of $\mathbf{\Sigma}$ are distinct and $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ equals the span of the eigenvector corresponding to the smaller eigenvalue, since the distribution of $\mathbf{Y}$ does not depend on $\mathbf{X}$ in the direction of the other eigenvector of $\mathbf{\Sigma}$. So $u = 1$, and $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is marked on Figure 2.1. Recall that the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{P}}_{\mathcal{E}}\mathbf{B} = \widehat{\mathbf{P}}_{\mathcal{E}}(\bar{\mathbf{Y}}_2 - \bar{\mathbf{Y}}_1)$. Accordingly, for inference on β_2, the data "Y" are first projected onto the estimated envelope to obtain $\widehat{\mathbf{P}}_{\mathcal{E}}\mathsf{Y}$ and then $\widehat{\mathbf{P}}_{\mathcal{E}}\mathsf{Y}$ is projected onto the Y_2 axis following the dashed lines marked B. The distributions at the bottom stand for the estimated projected distributions for the two populations, which are now well separated. This indicates that by getting rid of the immaterial information, the envelope model can achieve substantial efficiency gains compared to the standard model, even when the envelope is estimated. These ideas are relevant in the examples of Sections 2.10.4 and 2.10.2.

2.10.1 Wheat protein, again

The wheat protein data used for illustration in Section 2.2 contains two responses and a binary predictor, while the wheat protein data of Section 2.9 contains six responses and the same binary predictor. These data sets are subsets of a larger data set with $r = 6$ wavelengths as responses and actual wheat protein content as the $p = 1$ continuous predictor X (Fearn 1983). For these complete data, AIC, BIC and LRT(0.05), all produce $\widehat{u} = 4$ and consequently that changes in wheat protein affect the distribution of only 4 linear combinations of $\mathbf{Y}$. Estimates of the 6×1 coefficient vector $\boldsymbol{\beta}$ are given in Table 2.2, $\widehat{\boldsymbol{\beta}}$ under the envelope model with $u = 4$ and $\mathbf{B}$ under the standard model. The magnitudes of the elements of $\widehat{\boldsymbol{\beta}}$ all increased relative to their standard errors. Three of the coefficient estimates that were non-significant under the standard model are significant under the envelope model. The standard error ratios for the coefficients are shown in the last column of the table. Those ratios mostly have a magnitude similar to that of the ratios given previously in Section 2.2.

One implication of this illustration is that u does not have to be small relative to r for envelope models to offer solid gains. Rather, envelope gains are often controlled more by the relative magnitudes of the material and immaterial variation. In this example, $\|\widehat{\mathbf{\Omega}}\| = 196.6$, while $\|\widehat{\mathbf{\Omega}}_0\| = 6,516.6$.

Table 2.2: Coefficient estimates and their asymptotic standard errors from the envelope model with $u = 4$ and the standard model fitted to the complete wheat protein data.

$\mathbf{Y}$	$\widehat{\boldsymbol{\beta}}$	$\mathrm{se}(\widehat{\boldsymbol{\beta}})$	$\mathbf{B}$	$\mathrm{se}(\mathbf{B})$	$\mathrm{se}(\mathbf{B})/\mathrm{se}(\widehat{\boldsymbol{\beta}})$
Y_1	0.46	0.50	1.41	3.35	6.72
Y_2	2.43	0.36	3.21	2.76	7.56
Y_3	2.18	0.35	3.02	2.96	8.45
Y_4	-1.86	0.36	-0.94	3.26	9.19
Y_5	-1.69	1.11	-0.40	4.70	4.23
Y_6	-0.47	0.46	0.04	1.85	4.04

2.10.2 Berkeley Guidance Study

The Berkeley Guidance Study was designed to monitor the growth of children born in Berkeley, California between 1928 and 1929 (Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of California boys and girls from birth to age 18, *University of California Publications in Child Development*, **1**, 183-364. The data are available at http://rss.acs.unt.edu/Rdoc/library/fda/html/growth.html.) Data on $n = 93$ children, 39 boys and 54 girls, are available.

We consider first the bivariate regression of the heights in centimeters at ages 13 and 14 on the single binary predictor indicating gender. Using BIC (cf. Section 2.6) in conjunction with a fit of the envelope model (2.11) led to the inference that $u = 1$. A plot of the data along with the estimated envelope and its orthogonal complement are shown in Figure 2.4. It can be seen from this plot that the variation in the immaterial information, which lies in the direction of $\mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\mathcal{B})$, is large relative to the variation in the material information, which lies in the direction of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Consequently, we expect a substantial reduction in the standard errors relative to those from the usual bivariate linear model, as discussed broadly in Section 2.2. This is reflected also by the relative sizes of estimated variation in the material information $\|\widehat{\boldsymbol{\Omega}}\| = 1.56$ and the immaterial information $\|\widehat{\boldsymbol{\Omega}}_0\| = 118.7$. The standard errors of the estimated elements of $\boldsymbol{\beta} \in \mathbb{R}^2$ under the standard and envelope models shown in Table 2.3 indicate that the envelope standard errors are about 11% of those from a fit of the standard model.

We next consider the bivariate regression of the heights at ages 17 and 18 on gender. Again using BIC (cf. Section 2.6) in conjunction with a fit of the envelope model (2.11)

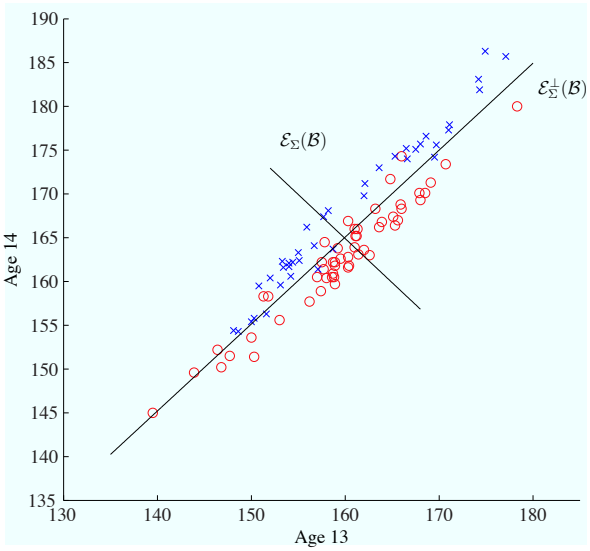


Figure 2.4: Envelope construction for the regression of height at ages 13 and 14, plotted on the horizontal and vertical axes, on gender $X \in \mathbb{R}^2$. Blue exes represent males and red circles represent females. The lines marked by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$ denote the estimated envelope and its orthogonal complement.

led to the inference that $u = 1$. A plot of the data along with the estimated envelope and its orthogonal complement are shown in Figure 2.5. The plot indicates that an envelope model may again be appropriate. However now the variation in the immaterial information, which lies in the direction of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$, is small relative to the variation in the material information, which lies in the direction of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$. This is because the projections of the data directly onto the coordinate axes, as happens in an analysis based on the standard model, will be quite close to the projections of the data first onto $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ and then onto the coordinate axes, as happens in an envelope analysis (cf. Section 2.2). Consequently, we expect little reduction in the standard errors relative to those from the usual bivariate linear model. This expectation is supported by the relative sizes of estimated variation in the material information $\|\widehat{\mathbf{\Omega}}\| = 79.5$ and the immaterial information $\|\widehat{\mathbf{\Omega}}_0\| = 0.156$. The standard errors of the estimated elements of $\boldsymbol{\beta} \in \mathbb{R}^2$ under the standard and envelope models shown in Table 2.3 indicate that there is no advantage to an envelope model in this regression.

Table 2.3: Bootstrap and estimated asymptotic standard errors of the two elements in $\widehat{\boldsymbol{\beta}}$ under the standard model (SM) and envelope model (EM) for two bivariate regressions from the Berkeley data. Rows 2 and 3 give the results from the regression of the heights at ages 13 and 14 on gender; rows 4 and 5 are for the regression of heights at ages 17 and 18 on gender. BSM and BEM designate the bootstrap standard errors for the standard and envelope models based on 200 bootstrap samples.

Response	SM	BSM	EM	BEM	SM/EM	BSM/BEM
Age 13	1.60	1.80	0.188	0.191	8.49	9.44
Age 14	1.61	1.81	0.187	0.190	8.61	9.64
Age 17	1.32	1.36	1.31	1.30	1.01	1.04
Age 18	1.33	1.37	1.34	1.37	0.99	1.01

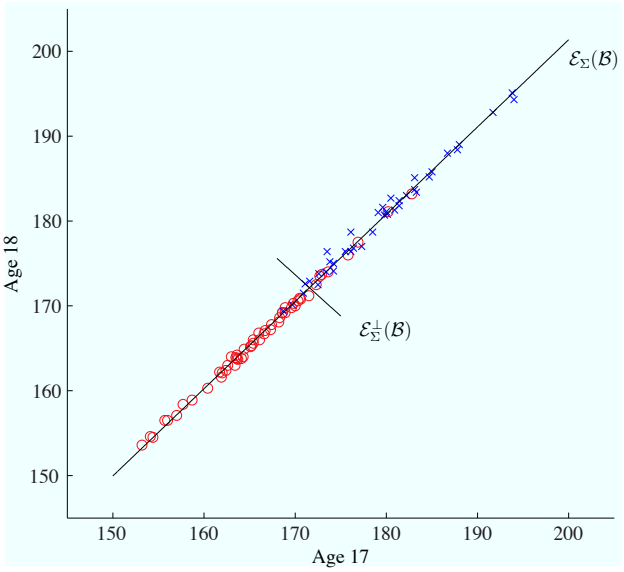


Figure 2.5: Envelope construction for the regression of height at ages 17 and 18, plotted on the horizontal and vertical axes, on gender $X \in \mathbb{R}^2$. Blue exes represent males and red circles represent females. The lines marked by $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\mathcal{B})$ denote the estimated envelope and its orthogonal complement.

2.10.3 Egyptian skulls

The data for this illustration are a subset of measurements taken on Egyptian crania from five different epochs that cover a period of at least 4,000 years. The crania were collected during excavations carried out between 1898 and 1901 in the region of Upper Egypt known as the Thebiad. (Thomson, A. and Randall-Maciver, R., 1905, Ancient Races of the Thebaid, Oxford: Oxford University Press.) The data set is unique because, according to Thomson Randall-Maciver (1905), it is the first instance of data capable of tracing the physical history of a people over a comparable period of 4,000 years.

The data[1] used for this illustration is known in the statistical literature and can be found in D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994, A Handbook of Small Data Sets, New York: Chapman & Hall, pp. 299-301) and Manly, B. F. J. (1986, Multivariate Statistical Methods, New York: Chapman & Hall). The response vector $\mathbf{Y}$ consists of four measurements each in mm on male crania:

- BH: Basibregmatic Height of Skull
- BL: Basialveolar Length of Skull
- MB: Maximal Breadth of Skull
- NH: Nasal Height of Skull

The responses were measured on 30 skulls in each of the five epochs indicated by the nominal dates 4000, 3300, 1850 and 200 B.C., and 150 A.D., giving a total sample size of $n = 150$. We regard epoch as categorical, leading to a regression of $\mathbf{Y}$ on four indicator variables X_j, $j = 1, \ldots, 4$, plus a constant for the intercept:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{3300} X_1 + \boldsymbol{\beta}_{1850} X_2 + \boldsymbol{\beta}_{200} X_3 + \boldsymbol{\beta}_{150} X_4 + \boldsymbol{\varepsilon}, \tag{2.25}$$

where $X_j = 1$ if $\mathbf{Y}$ was measured in the epoch indicated by the subscript on the corresponding coefficient vector and $X_j = 0$ otherwise. Let $\boldsymbol{\mu}_{(\cdot)} \in \mathbb{R}^4$ denote the epoch means. Model (2.25) is parameterized so that the coefficient vectors are the differences between the indicated epoch means and the mean for epoch 4000B.C, $\boldsymbol{\beta}_{3300} = \boldsymbol{\mu}_{3300} - \boldsymbol{\mu}_{4000}$, $\boldsymbol{\beta}_{150} = \boldsymbol{\mu}_{150} - \boldsymbol{\mu}_{4000}$, and so on. In this way, an envelope structure is posited for the differences in epoch means and not the means themselves, which is similar to the parameterization used for the wheat protein data in Section 2.2.

[1]The data used here were obtained from http://www.dm.unibo.it/ simoncin/EgyptianSkulls.html. See also http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html.

Table 2.4: $\widehat{\boldsymbol{\Gamma}}$ and standard error ratios – se of the OLS estimator divided by the se of the envelope estimator with $u = 1$ – for the Egyptian skull data.

$\mathbf{Y}$	$\widehat{\boldsymbol{\Gamma}}$	$\widehat{\boldsymbol{\beta}}_{3300}$	$\widehat{\boldsymbol{\beta}}_{1850}$	$\widehat{\boldsymbol{\beta}}_{200}$	$\widehat{\boldsymbol{\beta}}_{150}$
BH	0.28	3.73	2.09	1.51	1.25
BL	0.72	1.36	1.29	1.19	1.12
MB	-0.62	1.48	1.33	1.17	1.06
NH	-0.11	5.23	2.67	1.84	1.50

As in other examples, AIC, BIC and LRT(0.05) all indicated that $u = 1$, leading to the envelope model

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\eta_{3300}X_1 + \boldsymbol{\Gamma}\eta_{1850}X_2 + \boldsymbol{\Gamma}\eta_{200}X_3 + \boldsymbol{\Gamma}\eta_{150}X_4 + \boldsymbol{\varepsilon}, \tag{2.26}$$

with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{4\times 1}$ is a basis for the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Fitting with $u = 1$, $\widehat{\boldsymbol{\Gamma}}$ is shown in Table 2.4 along with the ratios of the OLS standard errors to the standard errors from the envelope model. The standard error ratios are all greater than one and some are large, indicating a strong overall improvement in efficiency. Table 2.5 gives the coefficients and t-values for the OLS fit of the full model and the envelope fit with $u = 1$. One general impression is that the t-values for the envelope analysis are generally greater than those for the standard analysis, roughly an indication of the a stronger analysis. Additionally, the t-values for $\widehat{\boldsymbol{\beta}}_{3300}$ are all relatively small, indicating that there may be no detectable difference between $\boldsymbol{\mu}_{4000}$ and $\boldsymbol{\mu}_{3300}$ based on model (2.25). This difference could be tested more formally by using the likelihood ratio statistic to test the hypothesis that $\boldsymbol{\eta}_{3300} = 0$ in model (2.26).

Table 2.6 shows the estimates of the residual covariance matrix $\boldsymbol{\Sigma}$ from the full model and from the envelope model. The variance estimates for the two models are close, but some of the estimated covariances have different signs. Those covariances are not significantly different from 0 and so the change in sign is not necessarily noteworthy.

One particularly useful part of this analysis is the inference that $u = 1$ and thus that the material information is embodied in the univariate response $\boldsymbol{\Gamma}^T\mathbf{Y}$. Boxplots of $\boldsymbol{\Gamma}^T\mathbf{Y}$ for each epoch are given in Figure 2.6. Although the envelope fit was based on epoch indicators, the box plots still shows a strong linear relation between location and epoch. This analysis then may lead to useful findings, particularly if it was possible to reify the

Table 2.5: Coefficients and nominal absolute t-value in parentheses for fits to the Egyptian skull data. Top half: envelope analysis; bottom half: OLS analysis

$\mathbf{Y}$	$\widehat{\boldsymbol{\beta}}_{3300}$	$\widehat{\boldsymbol{\beta}}_{1850}$	$\widehat{\boldsymbol{\beta}}_{200}$	$\widehat{\boldsymbol{\beta}}_{150}$
Envelope fit of (2.26)				
BH	-0.25 (0.68)	-1.15 (1.96)	-1.79 (2.20)	-2.24 (2.27)
BL	-0.66 (0.72)	-3.00 (3.09)	-4.67 (4.67)	-5.86 (5.26)
MB	0.56 (0.71)	2.57 (2.93)	4.00 (4.02)	5.01 (4.56)
NH	0.10 (0.65)	0.47 (1.55)	0.73 (1.66)	0.92 (1.83)
OLS fit of (2.25)				
BH	-0.90 (0.72)	0.20 (0.16)	-1.30 (1.04)	-3.27 (2.61)
BL	-0.10 (0.08)	-3.13 (2.47)	-4.63 (3.65)	-5.67 (4.46)
MB	1.00 (0.84)	3.10 (2.61)	4.13 (3.48)	4.80 (4.05)
NH	-0.30 (0.36)	0.03 (0.04)	1.43 (1.74)	0.83 (1.01)

Table 2.6: Estimated covariance matrices $\widehat{\boldsymbol{\Sigma}}$ from fits of the multivariate linear model and envelope model with $u = 1$ to the ozone data. † indicates correlations significantly different from 0 at level 0.05 using the t-test.

	Full model (2.25)				Envelope model, $u = 1$ (2.26)			
$\mathbf{Y}$	BH	BL	MB	NH	BH	BL	MB	NH
BH	22.70	5.03†	0.04	2.75†	21.75	2.47	1.48	2.57
BL		23.37	0.08	1.10		22.17	-0.82	-0.24
MB			20.41	1.94			22.95	3.32
NH				9.81				10.42

estimated material response $\boldsymbol{\Gamma}^T\mathbf{Y}$.

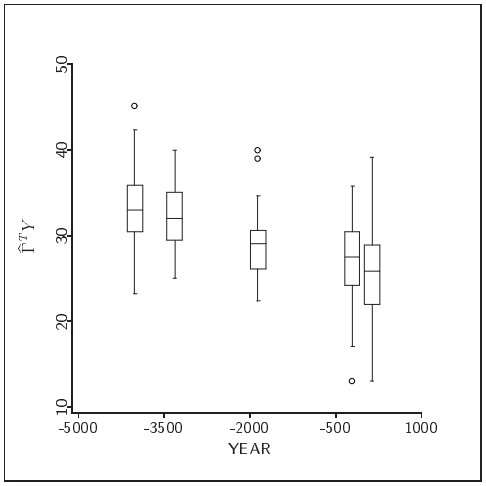


Figure 2.6: Skull data: Boxplots of $\widehat{\mathbf{\Gamma}}^T\mathbf{Y}$ versus year.

2.10.4 Australian Institute of Sport

Various allometric and hematological measurements were taken on 102 male and 100 female athletes at the Australian Institute of Sport[2] We use a binary predictor – $X = 1$ for female and $X = 0$ for male – to indicate gender. The response $\mathbf{Y} \in \mathbb{R}^2$ is bivariate consisting of the logarithms of white cell count (WCC) and Hematocrit (Hc), leading to the multivariate linear model $\mathbf{Y} = \alpha + \boldsymbol{\beta} X + \boldsymbol{\varepsilon}$.

A plot of $\log(Hc)$ versus $\log(WCC)$ with points marked by gender is shown in Figure 2.7. The mean difference, females minus males, for $\log(Hc)$ is -0.12 with a standard error of 0.0084, giving an absolute ratio of 14.4. The mean difference for $\log(WCC)$ is -0.027 with a standard error of 0.035, giving an absolute ratio of 0.77. Consequently, we see from the figure and the analysis that there is a clear gender difference in $\log(Hc)$, while the observed gender difference for $\log(WCC)$ is well within random variation.

The first step in an envelope analysis is to infer the dimension $u = \dim(\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}))$ of envelope. In this case, the LRT(0.05), AIC and BIC all selected $u = 1$. Since $u = 1$ was inferred, we next fit the envelope model $\mathbf{Y} = \alpha + \mathbf{\Gamma}\eta X + \boldsymbol{\varepsilon}$, where $\mathbf{\Gamma} \in \mathbb{R}^{2\times 1}$ and η is a scalar. The estimated envelope, its orthogonal complement and $\mathcal{B}$ are shown in Figure 2.7. The estimated basis $\widehat{\mathbf{\Gamma}}$ and the coefficient estimates under the full model and the envelope model are shown in Table 2.7. We see from the table that the estimates and standard errors

[2]The data are available from the web site for Cook, R. D. and Weisberg, S. *Applied Regression, Including Computing and Graphics*, New York: Wiley

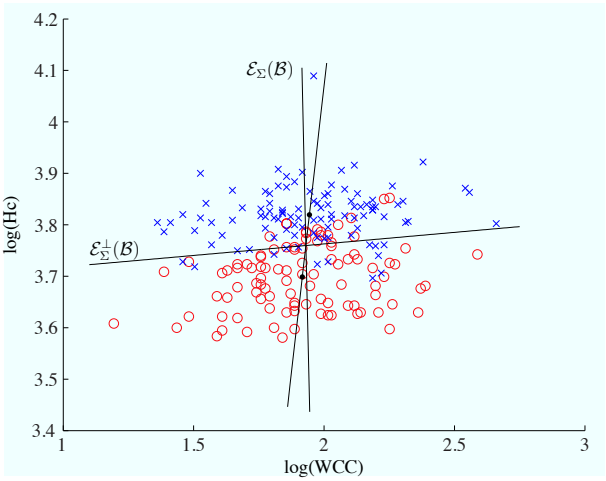


Figure 2.7: Envelope construction for the regression of $(\log(Hc), \log(WCC))^T$ on gender X. Blue exes represent females, red circles represent males and the two black dots represent the gender means. The lines marked by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$ denote the estimated envelope and its orthogonal complement.

for $\log(\mathrm{Hc})$ are essentially the same, while the estimates for $\log(\mathrm{WCC})$ have different signs and the standard error from the fit of the full model is about 17 times the standard error from the envelope model.

Table 2.7: $\widehat{\mathbf{\Gamma}}$ and coefficient estimates for the sports data with standard errors in parentheses.

Y	$\widehat{\mathbf{\Gamma}}$	**B**	$\widehat{\boldsymbol{\beta}}$
$\log(\mathrm{Hc})$	-0.9990	-0.121 (0.008)	-0.119 (0.008)
$\log(\mathrm{WCC})$	0.0448	-0.027 (0.035)	0.005 (0.002)

2.10.5 Air pollution

For this illustration we use data on air pollution obtained from Johnson and Wichern (2007, *Applied multivariate analysis*, Pearson Prentice Hall, New Jersey, p. 39.) The five responses are atmospheric concentrations of CO, NO, NO_2, O_3 and HC recorded at noon

in the Los Angeles area on $n = 42$ different days. The two predictors are measurements on wind speed W and solar radiation S. We explore these data in more detail than other illustration to give a feeling for various aspects of an envelope analysis.

Table 2.8: Estimated coefficients from fits of the multivariate linear model and envelope model with $u = 1$ to the ozone data. S – solar ratiation; W – wind speed. The columns of the ordinary least squares estimator of $\boldsymbol{\beta}$ has colums $\mathbf{B} = (\mathbf{B}_W, \mathbf{B}_S)$. The top half of the table is from the fit of the usual multivariate linear model and the bottom half is from the fit of the envelope model with $u = 1$.

Full model (2.27)				
$\mathbf{Y}$	$\mathbf{B}_W$	$\mathbf{B}_S$	$\mathbf{B}_W/\mathrm{se}(\mathbf{B}_W)$	$\mathbf{B}_S/\mathrm{se}(\mathbf{B}_S)$
CO	-0.1382	0.0117	-1.1804	1.0987
NO	-0.1925	-0.0064	-1.8843	-0.6865
NO_2	-0.2113	0.0205	-0.6465	0.6894
O_3	-0.7868	0.0952	-1.5643	2.0747
HC	0.0713	0.0027	1.0668	0.4485
Envelope model, $u = 1$ (2.28)				
$\mathbf{Y}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{W}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{W}})$	$\widehat{\boldsymbol{\beta}}_{\mathrm{S}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{S}})$
CO	0.0712	0.0008	2.7565	0.4340
NO	-0.0751	-0.0009	-2.8860	-0.4344
NO_2	-0.0166	-0.0002	-2.6224	-0.4334
O_3	-0.0106	-0.0001	-2.7048	-0.4338
HC	0.1172	0.0013	3.1240	0.4352

We consider first the usual multivariate linear regression of the vector $\mathbf{Y}$ of five responses on the two predictors:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\beta}_{\mathrm{W}} W + \boldsymbol{\beta}_{\mathrm{S}} S + \boldsymbol{\varepsilon}, \tag{2.27}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathrm{W}}, \boldsymbol{\beta}_{\mathrm{S}})$ and $\mathbf{X} = (W, S)^T$. The likelihood ratio test statistic for $\boldsymbol{\beta} = 0$ has the value $\Lambda = 19.14$ with 10 degrees of freedom, giving a nominal chi-square p-value of 0.038. This suggests that the mean of $\mathbf{Y}|\mathbf{X}$ depends on $\mathbf{X}$ nontrivially, although that dependence is not likely to be strong. The estimated coefficients and the coefficients relative to their standard errors are shown in the top half of Table 2.8. There seems to be

some indication of an effect for solar radiation on O_3 and perhaps for wind speed on NO, but generally the individual significant characteristics of the regression seem unclear. The eigenvalues of $\widehat{\mathbf{\Sigma}}$ range between 26.2 and 0.21 so in view of (2.13) there is a possibility that an envelope will help clarify the analysis, particularly if the larger eigenvalues are associated with immaterial information.

Turning to envelope regression, BIC, AIC and LRT(0.05) all indicate that $u = 1$ (cf. Section 2.6). Consequently, we base the envelope analysis on the model

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{\Gamma}\eta_{\mathrm{W}} W + \mathbf{\Gamma}\eta_{\mathrm{S}} S + \boldsymbol{\varepsilon}, \tag{2.28}$$

where the coefficient vectors in (2.27) are now modeled as $\boldsymbol{\beta}_{\mathrm{W}} = \mathbf{\Gamma}\eta_{\mathrm{W}}$ and $\boldsymbol{\beta}_{\mathrm{S}} = \mathbf{\Gamma}\eta_{\mathrm{S}}$, and $\mathbf{\Gamma} \in \mathbb{R}^5$. Fitting with $u = 1$ then, the relative magnitudes of the estimated material $\widehat{\mathbf{\Omega}} = 0.21$ and immaterial $\|\widehat{\mathbf{\Omega}}_0\| = 31.1$ variation indicate that the envelope model results in a substantial reduction. Indeed, the ratios of the standard errors for the elements of $\widehat{\boldsymbol{\beta}}$ from the full model to those of the envelope model range between 1.78 and 163.17, shown in Table 2.9.

Table 2.9: Standard error ratios – standard error from the full model fit divided by the standard error from the envelope model fit with $u = 1$ – for the estimated coefficients from the ozone regression.

$\mathbf{Y}$	W	S
CO	4.53	5.67
NO	3.92	4.69
NO_2	51.76	67.94
O_3	128.13	163.17
HC	1.78	1.97

The estimated coefficients and the coefficients relative to their standard errors are shown in the bottom half of Table 2.8. Compared to the coefficients from the full model, we see all but one have been shrunk toward 0, and now wind speed seems relevant for all responses while solar radiation given wind speed seems irrelevant for all responses. The contributions of wind speed and solar radiation can be tested by comparing the envelope model (2.28) to the envelope models with only one of the predictors. The likelihood ratio statistic for $\eta_{\mathrm{W}} = 0$ has the value 4.77 , and the likelihood ratio statistic for $\eta_{\mathrm{S}} = 0$ has the

value 0.17, each on 1 degree of freedom. These statistics then support the indication in the lower half of Table 2.8 that solar radiation is not contributing significantly the regression given that wind speed is in the model. Since the estimated material part of the response is $\widehat{\mathbf{\Gamma}}^T\mathbf{Y}$, we can explore the implications of the envelope further by using the univariate linear regression fit of $\widehat{\mathbf{\Gamma}}^T\mathbf{Y}$ on (W, S). The added variable plots (Cook and Weisberg, 1982) shown in Figure 2.8 seem to be in qualitative agreement with the envelope fit of Table 2.8.

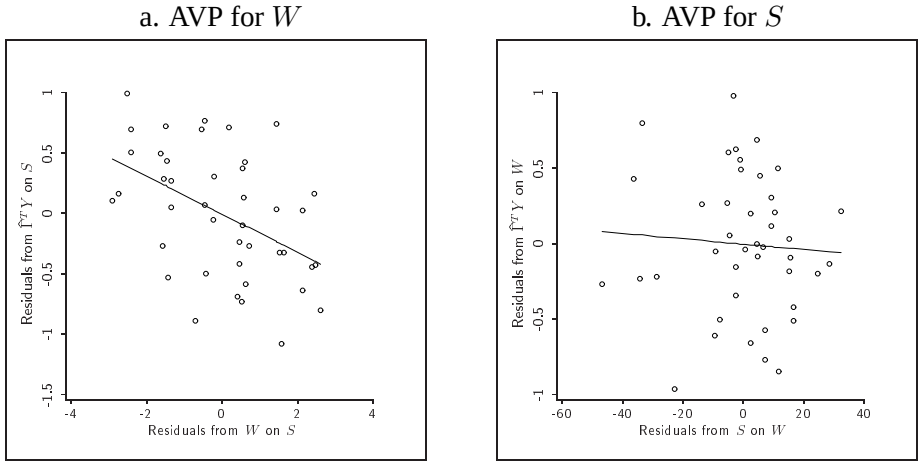


Figure 2.8: Added variable plots (AVP) for W and S from the fit of $\widehat{\mathbf{\Gamma}}^T\mathbf{Y}$ on (W, S).

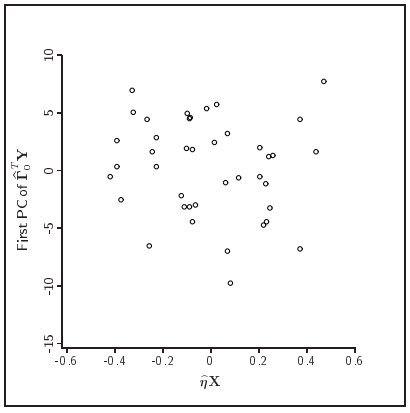


Figure 2.9: Scatterplot illustrating the immaterial variation in the ozone data.

For contrast, Table 2.10 gives the estimated coefficients and their standard error ratios for separate fits of $\mathbf{Y}$ on W and $\mathbf{Y}$ on S. The three dimension selection methods again all

indicated that $u = 1$ for these two regressions. We see from Table 2.10 that the coefficients and standard errors for W are quite similar to those for the envelope fit with both variables in the lower half of Table 2.8. The corresponding quantities for S are very different, however.

Table 2.10: Estimated coefficients from separate envelope fits with $u = 1$ of $\mathbf{Y}$ on W and $\mathbf{Y}$ on S. S – solar ratiation; W – wind speed; $\widehat{\boldsymbol{\beta}}_{\mathrm{W}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{W}})$ – estimated coefficients for wind speed divided by their standard errors; $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{S}})$ – estimated coefficients for solar radiation divided by their standard errors.

$\mathbf{Y}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{W}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{W}})$	$\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{S}}/\mathrm{se}(\widehat{\boldsymbol{\beta}}_{\mathrm{S}})$
CO	0.07	2.74	0.01	1.76
NO	0.07	-2.87	-0.002	-0.50
NO_2	-0.016	-2.62	-0.65	1.07
O_3	-0.011	-2.63	-1.56	2.22
HC	0.12	3.13	0.003	1.05

The estimated immaterial information in the ozone data can be visualized by plotting linear combinations $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$ versus $\widehat{\boldsymbol{\eta}}\mathbf{X}$ or other linear combinations of $\mathbf{X}$, where $\mathbf{b} \in \mathbb{R}^4$ and $\|\mathbf{b}\| = 1$. To be consistent with the envelope model, such plots should leave the impression that $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y} \perp\!\!\!\perp \widehat{\boldsymbol{\eta}}\mathbf{X}$, and the variation of $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$ should be substantially greater than that of $\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y}$. Figure 2.9 shows a plot of $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}_i$ versus $\widehat{\boldsymbol{\eta}}\mathbf{X}$ with $\mathbf{b}$ chosen to be the first eigenvector of the sample variance matrix of $\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$, so $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$ is simply the first principal component of $\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$. There is no clear dependence between $\mathbf{b}^T\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}$ and $\widehat{\boldsymbol{\eta}}\mathbf{X}$ that is discernible from the plot, and the range of the vertical axis is much larger than that for the added variable plots in Figure 2.8.

Table 2.11 shows the estimated covariance matrices $\widehat{\boldsymbol{\Sigma}}$ from the fits of the full model and the envelope model with $u = 1$. The two matrices seem to be in good agreement.

2.10.6 Multivariate bioassay

In this illustration we consider the results of a bioassay of insulin by the rabbit blood sugar method (Vølund, 1980; only the results for day 2 are used in this illustration). The test and standard preparations were each represented at two levels, coded -1 and $+1$. Each of the four treatment combinations was administered to nine rabbits whose blood

Table 2.11: Estimated covariance matrices $\widehat{\mathbf{\Sigma}}$ from fits of the multivariate linear model and envelope model with $u = 1$ to the ozone data.

	Full model (2.27)					Envelope model, $u = 1$ (2.28)				
$\mathbf{Y}$	CO	NO	NO_2	O_3	HC	CO	NO	NO_2	O_3	HC
CO	1.39	0.61	2.10	2.10	0.15	1.55	0.65	2.28	2.88	0.19
NO		1.05	0.99	-1.02	0.21		1.09	1.02	-0.93	0.22
NO_2			10.84	1.97	1.04			11.08	3.02	1.07
O_3				25.67	0.65				30.20	0.79
HC					0.45					0.46

sugar concentration (mg/100 ml) was measured at 0, 1, 2, 3, 4 and 5 hours. Let $\mathbf{Y} = (Y_0, Y_1, \ldots, Y_5)^T \in \mathbb{R}^6$ denote the random vector of blood sugar concentrations, let $X_1 = \pm 1$ indicate the treatment and standard preparations and let $X_2 = \pm 1$ indicate the dose level. The model relating the treatments to the concentrations has three predictors, the two treatment indicators plus their interaction,

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \boldsymbol{\beta}_{12} X_1 X_2 + \boldsymbol{\varepsilon} \tag{2.29}$$

A scatterplot matrix of the six responses is given in Figure 2.10 as background.

The presence of an interaction $\boldsymbol{\beta}_{12} \neq 0$ is often a general concern in models like (2.29). Twice the difference of the log likelihoods Λ under the alternative $\boldsymbol{\beta}_{12} \neq 0$ and null $\boldsymbol{\beta}_{12} = 0$ models has the value $\Lambda = 5.8$ on 6 degrees of freedom and thus there is little to suggest that $\boldsymbol{\beta}_{12} \neq 0$. Turning to the envelope models, AIC, BIC and LRT(0.05) all indicated that $u = 1$, so the envelope version of (2.29) becomes

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{\Gamma}\eta_1 X_1 + \mathbf{\Gamma}\eta_2 X_2 + \mathbf{\Gamma}\eta_{12} X_1 X_2 + \boldsymbol{\varepsilon}, \tag{2.30}$$

where $\mathbf{\Gamma} \in \mathbb{R}^6$, the η's are scalars and $\mathbf{\Sigma}$ has the corresponding envelope structure. The presence of an interaction in this model can be checked by using the log likelihood ratio to test $\boldsymbol{\eta}_{12} = 0$. The test statistic has the value $\Lambda = 0.008$ on one degree of freedom. This test suggests strongly that the effects in model (2.30) are additive. Shown in Figure 2.11 is a plot of $\widehat{\mathbf{\Gamma}}^T \bar{\mathbf{Y}}$ versus (X_1, X_2). The lines on the plot appear to be parallel, although they are not exactly so, which give visual support for outcome of the test of $\boldsymbol{\eta}_{12} = 0$. Depending on application-specific goals, the analysis could now be continued based on

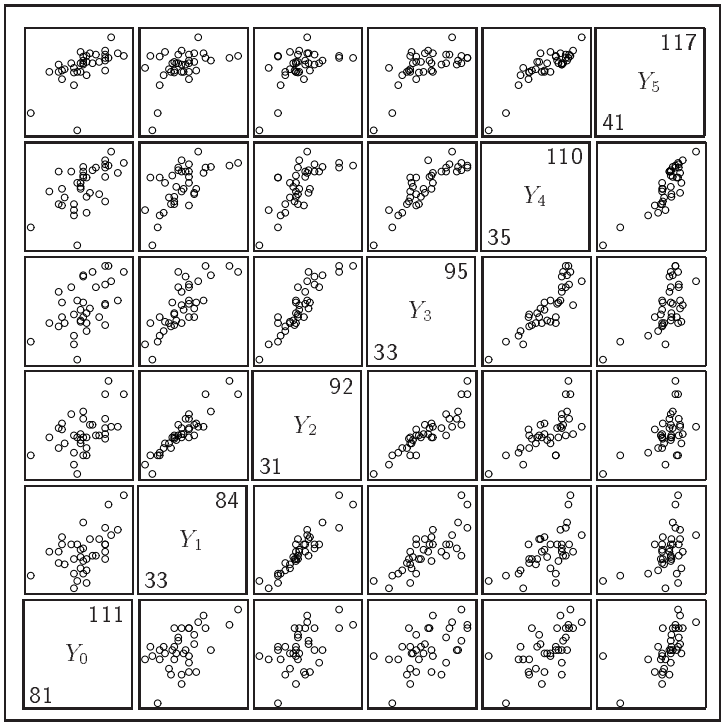


Figure 2.10: Scatterplot matrix of the six responses in the rabbit assay. The subscripts on Y_j indicate the hour at which the response was measured.

the additive model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\eta_1 X_1 + \boldsymbol{\Gamma}\eta_2 X_2 + \boldsymbol{\varepsilon}$, for which AIC, BIC and LRT(0.05) still estimate $u = 1$.

The scatterplot matrix of Figure 2.10 has a few observations that seem to be set apart from the trends exhibited by the remaining data. This leads to the possibility of outliers and influential observations having an impact on the conclusions. Although influence measures have not been developed specifically for envelope models, influence can still be studied by adapting the case deletion and local influence measures proposed by Cook (1979, 1986; see also Cook and Weisberg 1982).

Problems

Problem 2.1 *Show that (2.9) is equivalent to the conditions* $\mathcal{B} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ *and* $\boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$.

Problem 2.2 *Let* $\mathbf{V} = \mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{X}$ *denote the standardized predictor and let* $\mathbf{Z} = \mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{Y}$

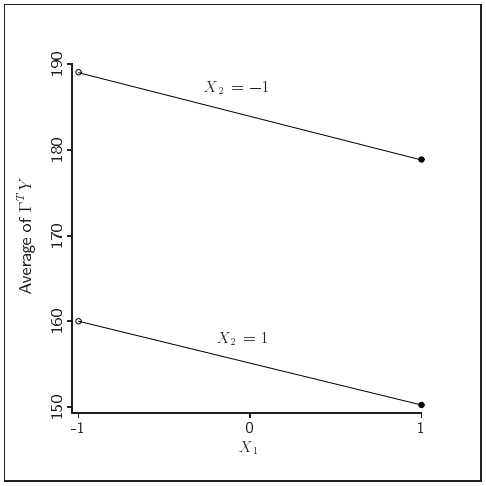


Figure 2.11: Scatterplot of $\widehat{\mathbf{\Gamma}}^T\bar{\mathbf{Y}}$ versus X_1 and X_2 from an envelope analysis of the rabbit assay using model (2.29).

denote the standardized response vector. Then the objective function in (2.15) can be represented conveniently as

$$L(\mathbf{G}) = \log|\mathbf{G}^T(\mathbf{S_Y} - \mathbf{S_{YV}}\mathbf{S}_{\mathbf{YV}}^T)\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|.$$

Let $L_1(\mathbf{G}) = \log|\mathbf{G}^T\mathbf{S_Y}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$ *and* $L_2(\mathbf{G}) = \log|\mathbf{I}_p - \mathbf{S}_{\mathbf{ZV}}^T\mathbf{P}_{\mathbf{S}_{\mathbf{Y}}^{1/2}\mathbf{G}}\mathbf{S_{ZV}}|$.

1. *Show that* $L(\mathbf{G}) = L_1(\mathbf{G}) + L_2(\mathbf{G})$.
2. *Show that* $L_1(\mathbf{G}) \geq 0$ *and describe* $\{\mathbf{G}|L_1(\mathbf{G}) = 0\}$.
3. *Show that* $L_2(\mathbf{G})$ *is minimized when* $\mathbf{G}$ *is equal to* $\mathbf{S}_{\mathbf{Y}}^{-1/2}$ *times the matrix whose columns are the first* d *eigenvectors of* $\mathbf{S_{ZV}}\mathbf{S}_{\mathbf{ZV}}^T$.
4. *Discuss how the terms* L_1 *and* L_2 *guide the minimization of* L.

Problem 2.3 *The data for this problem are a subset of measurements taken on Egyptian crania from five different epochs that cover a period of at least 4,000 years. The crania were collected during excavations carried out between 1898 and 1901 in the region of Upper Egypt known as the Thebiad. (Thomson, A. and Randall-Maciver, R., 1905, Ancient Races of the Thebaid, Oxford: Oxford University Press.) The dataset is unique because, according to Thomson Randall-Maciver (1905), it is the first instance of data capable of tracing the physical history of a people over a comparable period of 4,000 years.*

These data[3] *are known in the statistical literature and can be found in D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994, A Handbook of Small Data Sets, New York: Chapman & Hall, pp. 299-301) and Manly, B. F. J. (1986, Multivariate Statistical Methods, New York: Chapman & Hall). The response vector* $\mathbf{Y}$ *consists of four measurements each in mm on male crania:*

- *BH: Basibregmatic Height of Skull*
- *BL: Basialveolar Length of Skull*
- *MB: Maximal Breadth of Skull*
- *NH: Nasal Height of Skull*

The responses were measured on 30 *skulls in each of the five epochs indicated by the nominal dates 4000, 3300, 1850 and 200 B.C., and 150 A.D., giving a total sample size of* $n = 150$*. We regard epoch as categorical, leading to a regression of* $\mathbf{Y}$ *on four indicator variables* X_j*,* $j = 1, \ldots, 4$ *, plus a constant for the intercept:*

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{3300} X_1 + \boldsymbol{\beta}_{1850} X_2 + \boldsymbol{\beta}_{200} X_3 + \boldsymbol{\beta}_{150} X_4 + \boldsymbol{\varepsilon}, \tag{2.31}$$

where $X_j = 1$ *if* $\mathbf{Y}$ *ws measured in the epoch indicated by the subscript on the corresponding coefficient vector and* $X_j = 0$ *otherwise. Let* $\boldsymbol{\mu}_{(\cdot)} \in \mathbb{R}^4$ *denote the epoch means. Model (2.31) is parameterized so that the coefficient vectors are the differences between the indicated epoch means and the mean for epoch 4000B.C,* $\boldsymbol{\beta}_{3300} = \boldsymbol{\mu}_{3300} - \boldsymbol{\mu}_{4000}$*,* $\boldsymbol{\beta}_{150} = \boldsymbol{\mu}_{150} - \boldsymbol{\mu}_{4000}$*, and so on. In this way, an envelope structure is posited for the differences in epoch means and not the means themselves, which is similar the parameterization used for the wheat protein data in Section 2.2.*

Conduct an envelope analysis of these data with emphasis on developing a parsimonious characterization of the cranial changes over time, if any. Include in the context of the envelope model you develop, tests the hypotheses that $\boldsymbol{\beta}_{3300} = 0$ *and* $\boldsymbol{\beta}_{3300} - \boldsymbol{\beta}_{\mathbf{1}850} = 0$*.*

Problem 2.4 *Derive the variance of* $\mathbf{B}$ *given in (2.6). Recalling that this variance is conditional on the predictors, derive also the variance of* $\mathbf{B}$ *with respect to the joint distribution of* $\mathbf{X}$ *and* $\mathbf{Y}$ *when* $\mathbf{X}$ *is normally distributed with mean* $\boldsymbol{\mu}_{\mathbf{X}}$ *and variance* $\boldsymbol{\Sigma}_{\mathbf{X}}$*. Here you will find helpful a paper by von Rosen (*Scandinavian Journal of Statistics *Vol. 15, No. 2 (1988), pp. 97-109).*

[3]The data used here were obtained from http://www.dm.unibo.it/ simoncin/EgyptianSkulls.html. See also http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html.