Chapter 6

Connections Between Envelopes and Other Multivariate Methods

In Chapter 4 we studied the connections between envelopes and partial least squares regression. In this chapter we discuss relationships between envelopes and other dimension reduction methods, starting with canonical correlations. In each case we give an introduction to a method, followed by its connection to envelopes. These introductions are not in-depth treatments of the methods themselves.

6.1 Canonical correlations

Suppose that we have \( n \) independent and identically distributed observations \( Z_i \) on a random vector \( Z \in \mathbb{R}^{p+r} \) partitioned as

\[
Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \quad i = 1, \ldots, n.
\]

Canonical correlations and the corresponding canonical variates are used to characterize the strength of the linear relationships between the two vector-valued random variables \( X \in \mathbb{R}^p \) and \( Y \in \mathbb{R}^r \). Although the \( XY \)-notation for these variates is the same as that used for the multivariate linear model (1.1), the context here is not regression, so \( Y \) is not designated as a response variable.
6.1.1 Construction of canonical variates and correlations

Canonical variates are constructed as pairs \((a_j^T X, b_j^T Y)\), where the canonical vectors \(a_j \in \mathbb{R}^p\) and \(b_j \in \mathbb{R}^r\), \(j = 1, \ldots, m\) with \(m = \min(p, r)\), are linear transformations of \(X\) and \(Y\) so that \((a_1^T X, b_1^T Y)\) is the most highly correlated pair, \((a_2^T X, b_2^T Y)\) is the second most highly correlated pair, and so on, subject to certain constraints on the construction process. In this way the pairs of canonical variates give a characterization of the linear dependencies between \(X\) and \(Y\). We begin by describing how the canonical variates are constructed and then we sketch some of their properties. As in previous chapters, we let \(S_X\), \(S_Y\) and \(S_{XY}\) denote the sample versions of \(\Sigma_X = \text{var}(X)\), \(\Sigma_Y = \text{var}(Y)\) and \(\Sigma_{XY} = \text{cov}(X, Y)\).

The correlation structure between \(X\) and \(Y\) is reflected indirectly by the matrix of sample cross correlations \(C_{X,Y} = S_X^{-1/2}S_{XY}S_Y^{-1/2}\) between the elements of the standardized vectors \(S_X^{-1/2}X\) and \(S_Y^{-1/2}Y\). Canonical correlation analysis can be viewed as first constructing the desired canonical vectors in the standardized scale of \(C_{X,Y}\) and then transforming back to the original scale to obtain the canonical vectors \(\{a_j\}\) and \(\{b_j\}\), followed by forming the canonical variates themselves. Let \(C_{X,Y} = UDV^T\) denote the compact singular value decomposition of \(C_{X,Y}\):

\[
U = (u_1, \ldots, u_m), \quad U^T U = I_m
\]
\[
V = (v_1, \ldots, v_m), \quad V^T V = I_m
\]
\[
D = \text{diag}(d_1, \ldots, d_m),
\]

where we assume that \(d_1 > d_2 > \ldots > d_m > 0\). Then the pairs of canonical variates \((a_j^T X, b_j^T Y)\) are constructed from the canonical vectors

\[
a_j = S_X^{-1/2}u_j, \quad b_j = S_Y^{-1/2}v_j, \quad j = 1, \ldots, m. \tag{6.1}
\]

The sample variance of the canonical variate \(a_j^T X\) equals \(u_j^T S_X^{-1/2}S_X S_X^{-1/2}u_j = 1\), and in the same way the sample variance of \(b_j^T Y\) is equal to 1. It can be seen similarly that the pairs of canonical variates \((a_j^T X, a_k^T X)\), \((b_j^T Y, b_k^T Y)\) and \((a_j^T X, b_k^T Y)\) are uncorrelated for \(j \neq k\), while the sample correlation between \(a_j^T X\) and \(b_j Y\) is the \(j\)-th canonical correlation:

\[
\hat{\rho}(a_j^T X, b_j^T Y) = u_j^T S_X^{-1/2}S_{XY}S_Y^{-1/2}v_j = u_j^T C_{X,Y} v_j = d_j,
\]

where \(\hat{\rho}(:, :)\) denotes the sample correlation between its arguments. Summarizing these
6.1. CANONICAL CORRELATIONS

Accordingly, the canonical correlation for the $j$-th pair of canonical variates is equal to the $j$-th singular value of $C_{X,Y}$, the first canonical pair having the largest correlation possible between a linear transformation of $X$ and a linear transformation of $Y$. The canonical variates $\{a^T X_j\}$ and $\{b^T Y_j\}$ are uncorrelated and standardized to have variance 1. In this way the canonical pairs $(a^T X_j, b^T Y_j)$ serve to represent the linear relationship between $X$ and $Y$ in terms of one-dimensional projections ordered by their correlations.

The population version of a canonical correlation analysis is obtained by replacing the sample matrices $S_X, S_Y$ and $S_{XY}$ with their population counterparts $\Sigma_X, \Sigma_Y$ and $\Sigma_{XY}$.

### 6.1.2 Derivation of canonical variates

In this section we sketch how the canonical variates and vectors are derived in the population, starting with the standardized variates $\Sigma^{-1/2} X$ and $\Sigma^{-1/2} Y$. Let

$$\rho_{XY} = \Sigma^{-1/2} X \Sigma^{-1/2} Y = \Upsilon \Delta \Psi^T$$

denote the singular value decomposition of population version $\rho_{XY}$ of the sample cross correlation matrix $C_{X,Y}$:

$$\Upsilon = (v_1, \ldots, v_d), \quad \Upsilon^T \Upsilon = I_p$$

$$\Psi = (\psi_1, \ldots, \psi_d), \quad \Psi^T \Psi = I_r$$

$$\Delta = \text{diag}(\delta_1, \ldots, \delta_d),$$

where $d = \text{rank}(\rho_{XY}) \leq m$ and we assume that the singular values $\delta_1 > \delta_2 > \ldots > \delta_d > 0$ are unique. Consider finding the first pair of canonical vectors in the standardized scales of $\rho_{XY}$.
CHAPTER 6. CONNECTIONS

For any pair of unit-length vectors \( \alpha_1 \in \mathbb{R}^p \) and \( \beta_1 \in \mathbb{R}^r \), the correlation between 
\[
\rho(\alpha_1^T \Sigma_X^{-1/2} X, \beta_1^T \Sigma_Y^{-1/2} Y) = \frac{\sum_{k=1}^d (\alpha_1^T \upsilon_k)(\beta_1^T \psi_k \delta_k)}{\sum_{k=1}^d (\beta_1^T \psi_k)^2 \delta_k^2}.
\]

(6.5)

The restriction to unit length vectors \( \alpha_1 \) and \( \beta_1 \) serves to standardize the canonical variates:
\[
\text{var}(\alpha_1^T \Sigma_X^{-1/2} X) = \text{var}(\beta_1^T \Sigma_Y^{-1/2} Y) = 1.
\]

Using the Cauchy-Schwarz inequality on the right-hand side of (6.5) we have
\[
\rho^2(\alpha_1^T \Sigma_X^{-1/2} X, \beta_1^T \Sigma_Y^{-1/2} Y) \leq \left\{ \sum_{k=1}^d (\alpha_1^T \upsilon_k)^2 \right\} \left\{ \sum_{k=1}^d (\beta_1^T \psi_k)^2 \delta_k^2 \right\}
\leq \left\{ \sum_{k=1}^d (\beta_1^T \psi_k)^2 \delta_k^2 \right\}
\leq \delta_1^2 \left\{ \sum_{k=1}^d (\beta_1^T \psi_k)^2 \right\}
\leq \delta_1^2.
\]

The second equality follows because
\[
\sum_{k=1}^d (\alpha_1^T \upsilon_k)^2 = \alpha_1^T \Upsilon \Upsilon^T \alpha_1 \leq \alpha_1^T \alpha_1 = 1.
\]

The fourth inequality follows similarly. The third inequality follows because \( \delta_1^2 > \delta_j^2 \), \( j = 2, \ldots, d \). The upper bound is attained when we choose \( \alpha_1 = \upsilon_1 \) and \( \beta_1 = \psi_1 \). Accordingly the first pair of canonical variates is \( (\upsilon_1^T \Sigma_X^{-1/2} X, \psi_1^T \Sigma_Y^{-1/2} Y) \), which implies that the first pair of canonical vectors in the original scale of \( X \) and \( Y \) consists of \( \Sigma_X^{-1/2} \upsilon_1 \) and \( \Sigma_Y^{-1/2} \psi_1 \), and that \( \delta_1 \) is the first canonical correlation.

To extend this treatment to all \( d \) pairs of population canonical variates, we seek semi-orthogonal matrices \( \alpha = (\alpha_1, \ldots, \alpha_d) \) and \( \beta = (\beta_1, \ldots, \beta_d) \) so that \( \alpha^T \rho_{XY} \beta \) is a diagonal matrix with maximal and ordered diagonal elements. The solution \( \alpha = \Upsilon \) and \( \beta = \Psi \) follows from the uniqueness of the singular value decompositions of \( \rho_{XY} \), recalling that we have assumed the singular values to be distinct. The canonical vectors in the scales of \( X \) and \( Y \) are then the columns of \( \Sigma_X^{-1/2} \Upsilon \) and \( \Sigma_Y^{-1/2} \Psi \).
6.2. REDUCED-RANK REGRESSION

6.1.3 Connection to envelopes

Canonical correlations are related to the response envelope $\mathcal{E}_\Sigma(B)$ and predictor envelope $\mathcal{E}_{\Sigma X}(B')$ that occurred in Chapter 1 for response envelopes, and in Chapter 4 for predictor and simultaneous predictor-response envelopes. To develop this connection, recall from Corollary A.2, Proposition A.5 and the subsequent discussion that, in the context of model (1.1),

$$\mathcal{E}_\Sigma(B) = \mathcal{E}_{\Sigma Y}(B) = \mathcal{E}_{\Sigma Y}(\Sigma_Y^{-1/2} B \Sigma_X^{1/2}) = \mathcal{E}_{\Sigma Y} \{\text{span}(\rho_{YX})\}.$$  

Similarly, $\mathcal{E}_{\Sigma X}(B') = \mathcal{E}_{\Sigma X}(\rho_{XY})$. It follows from these relationships that $\text{span}(\Psi) \subseteq \mathcal{E}_\Sigma(B)$ and $\text{span}(Y) \subseteq \mathcal{E}_{\Sigma X}(B')$, and consequently

$$\Sigma_Y^{-1/2} \text{span}(\Psi) \subseteq \Sigma_Y^{-1/2} \mathcal{E}_\Sigma(B) = \mathcal{E}_\Sigma(B)$$
$$\Sigma_X^{-1/2} \text{span}(Y) \subseteq \Sigma_X^{-1/2} \mathcal{E}_{\Sigma X}(B') = \mathcal{E}_{\Sigma X}(B'),$$

where the final equality in each relation follows from Proposition A.4. While the population canonical vectors $\Sigma_Y^{-1/2} \text{span}(\Psi)$ and $\Sigma_X^{-1/2} \text{span}(Y)$ are in $\mathcal{E}_\Sigma(B)$ and $\mathcal{E}_{\Sigma X}(B')$ respectively, they will not span these envelopes unless $\text{span}(Y)$ and $\text{span}(\Psi)$ are themselves reducing subspaces of $\Sigma_X$ and $\Sigma_Y$. Even when $\Sigma_Y^{-1/2} \text{span}(\Psi) = \mathcal{E}_\Sigma(B)$ and $\Sigma_X^{-1/2} \text{span}(Y) = \mathcal{E}_{\Sigma X}(B')$, the sample versions of $\Sigma_Y^{-1/2} \text{span}(\Psi)$ and $\Sigma_X^{-1/2} \text{span}(Y)$ are often unreliable estimators of the corresponding envelopes (Cook and Zhang 2015).

6.2 Reduced-rank regression

6.2.1 Reduced-rank model and estimation

Reduced-rank regression begins with the multivariate linear model (1.1) and then introduces dimension reduction by constraining $\beta$ to be less than full rank, say $\text{rank}(\beta) = d < m$, where $m = \min(p, r)$ as defined in Section 6.1.1. Consequently, we can write $\beta = Hh$ as the product of two rank $d$ matrices, $H \in \mathbb{R}^{r \times d}$ and $h \in \mathbb{R}^{d \times p}$, leading to the reduced-rank model

$$Y = \alpha + HhX + \varepsilon, \var(\varepsilon) = \Sigma$$  \hspace{1cm} (6.6)

This is different from an envelope model since there is no required connection between \( \beta \) and the error variance \( \Sigma \). This distinction can lead to very different performance in practice, as described later. There are no essential constraints placed on \( H \) and \( h \) as part of the formal model, although constraints might be introduced as part of the estimation process to promote identifiability.

Assuming normal errors, maximum likelihood estimators were derived by Anderson\(^2\) (1999), Reinsel and Velu\(^3\) (1998) and Stoica and Viberg\(^4\) (1996) under various constraints on \( H \) and \( h \) for identifiability, such as \( hh^T = I_d \) or \( H^T H = I_d \). The decomposition \( \beta = H h \) is not unique even with such identifiable constraints: for any orthogonal matrix \( O \in \mathbb{R}^{d \times d} \), \( H_1 = HO \) and \( h_1 = O^T h \) offer another valid decomposition that satisfies the rank constraint. The parameters of interest, \( \beta \) and \( \Sigma \), are nevertheless identifiable, as well as \( \text{span}(H) = \text{span}(\beta) \) and \( \text{span}(h^T) = \text{span}(\beta^T) \). Cook, Forzani and Zhang\(^5\) (2015) gave a unified framework so that every statement involving \( H \) or \( h \) holds universally for any decomposition of \( \beta \) satisfying the rank constraint.

Recall from Section 6.1.1 that \( C_{X,Y} = S_X^{-1/2} S_{XY} S_Y^{-1/2} \) is the matrix of sample correlations between the elements of the standardized vectors \( S_X^{-1/2} X \) and \( S_Y^{-1/2} Y \), with singular value decomposition \( C_{X,Y} = UV^T \). Extending this notation, let \( C^{(d)}_{X,Y} = U_d D_d V_d^T \), where \( U_d \in \mathbb{R}^{r \times d} \) and \( V_d \in \mathbb{R}^{p \times d} \) consist of the first \( d \) columns of \( U \) and \( V \), and \( D_d \) is the diagonal matrix consisting of the first \( d \) singular values of \( C_{X,Y} \). We also use \( C_{Y,X} = C^{T}_{X,Y} \). Then, assuming normal errors and that \( d \) is given, the maximum likelihood estimators of the parameters in the reduced-rank regression model (6.6) are

\[
\hat{\alpha}_{RR}(Y \mid X, d) = \bar{Y} \quad (6.7)
\]
\[
\hat{\beta}_{RR}(Y \mid X, d) = S_Y^{1/2} C^{(d)}_{Y,X} S_X^{-1/2} \quad (6.8)
\]
\[
\hat{\Sigma}_{RR}(Y \mid X, d) = S_Y - \hat{\beta}_{RR} S_{XY} \quad (6.9)
\]

\[
= S_Y^{1/2} (I_r - C^{(d)}_{Y,X} C^{(d)}_{X,Y}) S_Y^{1/2}.
\]

The non-traditional part of this notation ‘\( (Y \mid X, d) \)’ is meant to indicate that the estimators arise from the rank \( d \) linear regression of \( Y \) on \( X \). It will be suppressed until needed in


6.2. REDUCED-RANK REGRESSION

Section 6.2.3 when combing reduced-rank regression and envelopes. The usual estimators are obtained by setting \( d = m \) so there is no rank reduction. The asymptotic covariance matrix of the reduced-rank estimator of \( \hat{\beta}_{RR} \) is (Cook, Forzani and Zhang 2015, equation (A17))

\[
\text{avar} \{ \sqrt{n} \text{vec}(\hat{\beta}_{RR}) \} = \Sigma_X^{-1/2} \otimes \Sigma^{1/2} \{ \mathbf{I}_{pr} - Q^{1/2} \mathbf{h}^T \otimes Q \Sigma^{-1/2} \mathbf{H} \} \Sigma_X^{-1/2} \otimes \Sigma^{1/2},
\]

where \( \Sigma_X^{-1} \otimes \Sigma = \text{avar} \{ \sqrt{n} \text{vec}(\mathbf{B}) \} \) is the asymptotic covariance matrix of the maximum likelihood estimator (cf. (1.17)).

Let \( \mathbf{R} \) denote the asymptotic variation of the reduced-rank estimator relative to that of the maximum likelihood estimator:

\[
\mathbf{R} = \text{avar}^{-1/2} \{ \sqrt{n} \text{vec}(\mathbf{B}) \} [\text{avar} \{ \sqrt{n} \text{vec}(\hat{\beta}_{RR}) \}] \text{avar}^{-1/2} \{ \sqrt{n} \text{vec}(\mathbf{B}) \}
\]

\[
= \mathbf{I}_{pr} - Q^{1/2} \mathbf{h}^T \otimes Q \Sigma^{-1/2} \mathbf{H}.
\]

This relationship shows that the gain over the maximum likelihood estimator is due primarily to the reduction in the number of parameters being estimated, and not to any internal structure associated with the parameters themselves. For instance, since \( \mathbf{R} \) is a projection its eigenvalues are 0 or 1, and \( \text{tr}(\mathbf{R}) = \text{pr} - (p - d)(r - d) = d(p + r - d) \), where \( \text{pr} \) is the number of real parameters need to determine \( \beta \) in multivariate linear model (1.1) and \( d(p + r - d) \) is number of real parameters need to determine \( \beta = \mathbf{H}_h \) in reduced-rank model (6.6). Consequently, the number of parameters needed for \( \beta \) in the reduced-rank model is a indicator of its relative gain in efficiency.

The methodology implied by (6.7)-(6.9) requires that \( d = \text{rank}(\beta) \) be selected first. While there are several ways in which this might be done, including adaptation of information criterial like AIC and BIC, we have found that the relatively straightforward method described in Section 1.8.2 works well in this context.

6.2.2 Contrasts with envelopes

As mentioned previously, the reduced-rank model (6.6) is distinctly different from the envelope models for the responses (1.20) or predictors (4.5) since there is no required connection between \( \beta \) and the error covariance matrix \( \Sigma \) or the predictor covariance matrix \( \Sigma_X \). This distinction has notable consequences. If \( r > 1 \) and \( p = 1 \) then \( \beta \in \mathbb{R}^r \) and its only possible ranks are 0 and 1. In this case reduced-rank regression is not useful, while a response envelope can still lead to substantial gains as illustrated by some of the examples
in Chapter 2. A similar conclusion holds when \( r = 1 \) and \( p > 1 \). More generally, reduced-rank regression offers no gains when \( \beta \) is full rank, while envelopes can still produce substantial gains. On the other hand, it is also possible to have situations where envelopes offer no gain, while reduced-rank regression provides notable gains. For instance, if \( r > 1 \), \( p > 1 \) and \( \text{rank}(\beta) = 1 \), so reduced-rank regression gives maximal gain, it is possible that \( \mathcal{E}_\Sigma(B) = \mathbb{R}^r \) so that response envelopes produce no gains.

There are also notable differences between the potential gains produced by envelope and reduced-rank regressions. We saw near the end of Section 6.2.1 that the gain from reduced-rank regression is controlled largely by the reduction the number of real parameters need to specify \( \beta \). On the other hand, the gain from a response envelope is due to the reduction in parameters and to the structure of \( \Sigma = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \), with massive gains possible when \( \| \Omega \| \ll \| \Omega_0 \| \).

These contrasts lead to the conclusion that envelopes and reduced-rank regression are distinctly different methods of dimension reduction with different operating characteristics. Cook, Forzani and Zhang (2015) combined reduced-rank regression and response envelopes, leading to a new dimension reduction paradigm – reduced-rank response envelopes – that can automatically choose the better of the two methods if appropriate and can also give an estimator that does better than both of them.

### 6.2.3 Reduced-rank response envelopes

The reduced-rank envelope model follows by applying an envelope to the reduced rank representation (6.6). As in Chapter 1, let \( \Gamma \in \mathbb{R}^{r \times u} \) be a semi-orthogonal basis matrix for \( \mathcal{E}_\Sigma(B) = \mathcal{E}_\Sigma(H) \), where \( u = \text{dim}(\mathcal{E}_\Sigma(B)) \geq d = \text{rank}(\beta) \). Then we can represent \( H \) in terms of \( \Gamma \), \( H = \Gamma \eta \), where \( \eta \in \mathbb{R}^{u \times d} \), and write the reduced-rank envelope model as

\[
Y = \alpha + \Gamma \eta h X + \epsilon, \quad \Sigma = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T.
\]  

(6.10)

This model represents \( \beta \) as the product of three factors arising from the reduced dimension \( d = \text{rank}(\beta) \) and the envelope structure. It contains two unknown dimensions \( d \) and \( u \). The number of real parameters needed to specify \( \beta \) is \( d(p + u - d) \), which is a reduction of \( (p - d)(u - d) \) parameters from the envelope model and \( d(r - u) \) parameters from the reduced-rank model. If \( u = r \) then \( \Gamma = I_r \), \( H = \eta \) and the model reduces to reduced-rank regression. If \( d = u \) then (6.10) reduces to the response envelope model (1.20).

Assuming normal errors in model (6.10) and that \( \Gamma \) and \( d \) are known, the maximum
likelihood estimators of $\alpha$, $\beta$ and $\Sigma$ are

\[
\begin{align*}
\hat{\alpha}_\Gamma &= \bar{Y} \\
\hat{\beta}_\Gamma &= \Gamma \hat{\beta}_{RR}(\Gamma^T Y | X, d) \\
\hat{\Sigma}_\Gamma &= \Gamma \left\{ \hat{\Sigma}_{RR}(\Gamma^T Y | X, d) \right\}^{\Gamma^T} + Q_Y S_Y Q_\Gamma,
\end{align*}
\]

where we make use of the notation established in (6.8) and (6.9). To construct these estimators, we first reduce to a reduced-rank envelope model (6.10) by transforming both sides by $\Gamma^T$ to obtain the reduced-rank model with coefficient matrix $\eta_h$:

\[
\Gamma^T Y = \Gamma^T \alpha + \eta_h X + \Gamma^T \epsilon, \quad \text{var}(\Gamma^T \epsilon) = \Sigma(\Gamma^T Y | X, d) = \Omega.
\]

This model is then fitted by using the reduced-rank estimators (6.7)–(6.9) and resulting estimators transformed back to the scale of $Y$ to obtain the estimators. Once the estimator of $\Gamma$ is determined it can be substituted into (6.11) and (6.12) to obtain the full reduced-rank envelope estimators.

A little additional notation is needed for a compact description of the maximum likelihood estimator of $\Gamma$. Within the class of semi-orthogonal matrices $G \in \mathbb{R}^{r \times u}$, let $Z_G = (G^T S_Y G)^{-1/2} G^T Y$ denote the standardized version of $G^T Y$ with sample covariance $I_u$, let $S_G^{(d)} | X$ denote the sample covariance of the residuals from the fit of the reduced-rank model (6.6), and let $\tilde{\phi}_j(G)$ denote the $j$-th eigenvalue of $S_G^{-1} | X$. Define also the objective function

\[
F_{\alpha}(G | d, u) = \log |G^T S_Y G| + \log |G^T S_Y^{-1} G| + \log |S_G^{(d)} | X| \\
= \log |G^T S_Y | X| + \log |G^T S_Y^{-1} G| + \sum_{j=d+1}^{u} \log(\tilde{\phi}_j(G)),
\]

which corresponds to the negative of the log likelihood for model (6.10) maximized over all parameters except $\Gamma$. Then a maximum likelihood estimator of $\Gamma$ is obtained as $\hat{\Gamma} = \arg\min_G F(G | d, u)$, where the minimization is over the Grassmannian of dimension $u$ in $\mathbb{R}^r$. These two forms give different insights into the nature of $\hat{\Gamma}$. The term $\log |G^T S_Y G| + \log |G^T S_Y^{-1} G|$ that occurs in (6.13) is minimized when $\text{span}(G)$ is any $u$-dimensional reducing subspace of $S_Y$ (Lemma A.15). The remaining term $\log |S_G^{(d)} | X|$ in (6.13) measures the goodness of fit of the reduced-rank regression of the reduced standardized response $Z_G$ on $X$. In this way the terms in (6.13) balance the requirements of the fit, the first two terms pulling the solution toward the reducing subspaces of $S_Y$ and the last term ensuring a reasonable fit of the reduced-rank regression.
Cook, Forzani and Zhang (2015) use the second form for $F_n(G \mid d, u)$ as a basis for computation. The also show that the term $\sum_{j=d+1}^{u} \log(\hat{\phi}_j(G))$ converges to 0 uniformly in $G$ and thus that $F_n(G \mid d, u)$ converges uniformly to $F(G \mid u) := \log |G^T \Sigma_{Y,X} G| + \log |G^T \Sigma_{Y}^{-1} G|$ which does not depend on $d$. The sample version

$$F_n(G \mid u) = \log |G^T \Sigma_{Y,X} G| + \log |G^T \Sigma_{Y}^{-1} G|$$

of $F(G \mid u)$ is the same as objective function (1.24) derived for response envelopes in Chapter 1. The implication of this result is that in large samples the reduced-rank envelope estimator moves toward a two stage estimator: first estimate the envelope from $F_n(G \mid u)$ ignoring the rank $d$ and then obtain the reduced-rank estimator from within the estimated envelope.

The methodology described in this section requires the selection of two dimensions $d = \text{rank}(\beta)$ and $u = \text{dim}(\mathcal{E}_{\Sigma}(B))$ subject to the constraint $0 \leq d \leq u \leq p$. As mentioned at the end of Section 6.2.1, the method outlined in Section 1.8.2 seems to work well for choosing $d$. Following that, Cook, Forzani and Zhang (2015) recommended that $u$ be chosen using BIC over the range $d \leq u \leq p$.

More discussion of this method is available from Cook, Forzani and Zhang (2015), who show that the reduced-rank envelope estimators are $\sqrt{n}$-consistent without normality, and give asymptotic comparisons, methods for selecting $u$ and $d$, simulation results and examples.

### 6.2.4 Reduced-rank predictor envelopes

The results sketched in the previous section combine reduced-rank regression and response envelopes. In principal, reduced-rank regression can also be combined with the predictor and simultaneous response-predictor envelopes of Chapter 4.

To develop reduced-rank predictor envelopes, we assume as in Chapter 4 that the predictor are random with mean $\mu_X$ and variance $\Sigma_X$. Then we extend the parametrization of the reduced-rank regression model (6.6) to include $\mathcal{E}_{\Sigma_X}(h^T)$, the $\Sigma_X$-envelope of $\text{span}(\beta^T) = \text{span}(h^T)$. Let $\Phi \in \mathbb{R}^{p \times q}$ denote a semi-orthogonal basis for this subspace, so that we can write $h^T = \Phi \eta$, where $\eta \in \mathbb{R}^{q \times d}$. Then, recalling that $H \in \mathbb{R}^{r \times d}$ and
6.3. SUFFICIENT DIMENSION REDUCTION, SDR

$h \in \mathbb{R}^{d \times p}$, the model becomes

\[
Y = \alpha + \beta (X - \mu_X) + \epsilon = \alpha + Hh(X - \mu_X) + \epsilon = \alpha + H\eta^T \Phi^T (X - \mu_X) + \epsilon
\]

\[
\text{var}(\epsilon) = \Sigma = \Phi \Delta \Phi^T + \Phi_0 \Delta_0 \Phi_0^T.
\]

This model can now be studied following the general steps of previous chapters.

6.3 Sufficient dimension reduction, SDR

The origins of sufficient dimension reduction go back to the early 1990’s with the revolution in desktop computing and computer graphics. The area originated from the pursuit of low dimensional projective views that contains all or nearly all of the relevant information on the regression of a univariate response $Y \in \mathbb{R}^1$ and predictor vector $X \in \mathbb{R}^p$ (Cook 1998). Stated more formally, the issue was to find a low-dimensional subspace $S$ of the predictor space with the property that $Y$ is independent of $X$ given the projection $P_S X$ of $X$ onto $S$ without pre-specifying a parametric model. Subspaces with this property are called dimension reduction subspaces. If $S$ is a dimension reduction subspace then $P_S X$ contains all of the information that $X$ has about $Y$ and, letting $\rho \in \mathbb{R}^{p \times d}$ with $d < p$ denote a basis for $S$, a plot of $Y$ versus $\rho^T X$ could be used as a diagnostic guide to the regression. The following definition (Cook 2007) formalizes this idea and gives a basis for many standard linear SDR methods.

**Definition 6.1** A projection $P_S : \mathbb{R}^p \to S \subseteq \mathbb{R}^p$ onto a $q$-dimensional subspace $S$ is a sufficient linear reduction if it satisfies at least one of the following three statements:

(i) inverse reduction, $X \mid (Y, P_S X) \sim X \mid P_S X$,

(ii) forward reduction, $Y \mid X \sim Y \mid P_S X$,

(iii) joint reduction, $Y, X \mid P_S X$.

The subspace $S$ is then called a dimension reduction subspace.

Each of the three conditions in this definition conveys the idea that the reduction $P_S X$ carries all the information that $X$ has about $Y$, and consequently all the information available
to estimate the conditional mean $E(Y \mid X)$ and variance $\text{var}(Y \mid X)$ functions. They are equivalent when $(Y, X)$ has a joint distribution. We are then free to determine a reduction inversely or jointly and pass it to the forward regression without additional structure; for instance, $E(Y \mid X) = E(Y \mid P_S X)$. A sufficient summary plot of $Y$ versus coordinates of the projection $P_S X$ is often an effective tool for guiding the regression.

If $S$ is a dimension reduction subspace and $S \subseteq S_1$ then $S_1$ is also a dimension reduction subspace. Obviously, within the class of linear reductions, we would like to find the smallest dimension reduction subspace (Cook 1994, Cook1998):

**Definition 6.2** The intersection of all dimension reduction subspaces, when it is itself a dimension reduction subspace, is called the central subspace, $S_{Y \mid X}$.

The central subspace doesn’t always exist, but it does so under mild regularity conditions that we assume throughout this section (Cook 1998, Yin, Li and Cook^6 2008). This area is now widely know as sufficient dimension reduction because of the similarity between the driving condition $Y \perp \!\!\!\!\perp X \mid P_{S_{Y \mid X}} X$ and Fisher’s fundamental notion of sufficiency. The name also serves to distinguish it from other approaches to dimension reduction.

The central subspace turned out be an effective construct, and over the past 20 years much work has been devoted to methods for estimating it, the first two methods being sliced inverse regression (Li 1991) and sliced average variance estimation (Cook and Weisberg 1991). These methods, like nearly all of the subsequent methods, require the so called linearity and constant covariance conditions on the marginal distribution of the predictors:

**Linearity condition:** $E(X \mid \eta^T X)$ is a linear function of $\eta^T X$,

**Constant covariance condition:** $\text{var}(X \mid \eta^T X)$ is a non-random matrix,

where $\eta \in \mathbb{R}^{p \times d}$ is a semi-orthogonal basis matrix for $S_{Y \mid X}$. These conditions are required only at a basis for $S_{Y \mid X}$ and not over all $\eta$. Diaconis and Freedman^7 (1984; see also Hall and Li^8 (1993)) showed that almost all projections of high-dimensional data are

---


6.3. SUFFICIENT DIMENSION REDUCTION, SDR

approximately normal, which is often used as partial justification for these conditions. See Cook (1998) and Li and Wang\(^9\) (2007) for additional discussion.

Although the linearity and constant covariance conditions are largely seen as mild, they are essentially uncheckable and thus can be a nag in application. Ma and Zhu\(^10\) (2012) developed a semi-parametric approach that allows modifications of previous methods so they no longer depend on these conditions. The fundamental restriction to linear reduction \(P_{S_{Y|X}}X\) has also been long recognized as a limitation. Lee et al.\(^11\) (2013) extended the foundations of sufficient dimension reduction to allow for non-linear reduction. This breakthrough, like that from Ma and Zhu, opened new frontiers in dimension reduction that promises further significant advances. Although SDR methods were originally developed as comprehensive graphical diagnostics, they are now serviceable outside of that context.

The envelope methods described so far in this book require a parametric model or a method to estimate a parametric characteristic of the regression, while SDR methods do not require a pre-specified model. This distinction has both advantages and disadvantages depending on the context. For instance, if we are dealing with a univariate linear regression, \(Y = \alpha + \beta^T X + \epsilon\), then clearly \(Y \perp X \mid \beta^T X\), so \(S_{Y|X} = \text{span}(\beta)\) and SDR offers little beyond the model itself, while envelopes can offer substantial gain as described in Chapter 4. On the other hand, if we wish to study the regression of \(Y\) on \(X\) without specifying a model or estimator then envelopes at present have little to offer, while SDR methods are applicable. Nevertheless, it is possible to adapt envelopes for use in SDR contexts to provide more efficient estimators. The motivating condition underlying SDR is (i) \(Y \perp X \mid P_{S}X\). If we require in addition that (ii) \(P_{S}X, \perp Q_{S}X\) then we arrive at an envelope-type structure that is analogous to that described in Chapter 4 for predictor reduction.

We review SIR in the next section and then return to envelopes.

---


6.4 Sliced inverse regression, SIR

6.4.1 SIR methodology

Let $\Sigma_X = \text{var}(X)$, let $\nu(y) = \{E(X \mid Y = y) - E(X)\}$ and let $\eta$ be a basis matrix for $S_{Y \mid X}$. SIR, like all of the early SDR methods, is based on the following result (Li, 1991).

**Proposition 6.1** Assume that linearity condition holds. Then for all $y$ in the sample space of $Y$

$$\nu(y) = P^T_{\eta (\Sigma_X)} \nu(y).$$

Consequently $\xi(y) := \Sigma_X^{-1} \nu(y) \in S_{Y \mid X}$ for all $y$.

Using this result we should in principle be able to estimate $S_{Y \mid X}$ if we can find a way to estimate $\xi(y)$. If $Y$ is discrete or categorical, a sample version of $\xi(y)$ can be constructed by substituting sample versions of $\Sigma_X, E(X \mid Y = y)$ and $E(X)$. When $Y$ is continuous, SIR replaces $Y$ with a sliced version constructed by slicing the range of $Y$ into $h$ slices.

Let $\nu_s = \{E(X \mid J_s(Y) = 1) - E(X)\}$, where $J_s(Y)$ is the indicator function for slice $s = 1, \ldots, h$. Then $\xi_s := \Sigma_X^{-1} \nu_s \in S_{Y \mid X}$, for $s = 1, \ldots, h$, which can again be estimated by substituting sample moments. For a given dimension $d = \text{dim}(S_{Y \mid X})$ and assuming that $Y$ is categorical possibly induced by slicing, the typical estimation algorithm is as follows.

1. Standardize $X$ to $\hat{Z} = S_X^{-1/2}(X - \bar{X})$.

2. Form $\hat{\text{var}}\{E(\hat{Z} \mid Y)\}$.

3. Set $\hat{S}_{Y \mid X} = S_X^{-1/2} \times \text{span of the first } d \text{ eigenvectors of } \hat{\text{var}}\{E(\hat{Z} \mid Y)\}$.

4. Select a basis $\hat{\eta}$ for $\hat{S}_{Y \mid X}$ and construct a display of $Y$ vs $\hat{\eta}^T X$.

The dimension $d$ of $S_{Y \mid X}$ is often select based on a sequence of hypothesis tests: starting with $d_0 = 0$, test the hypothesis $d = d_0$. If the hypothesis is rejected, increment $d_0$ by 1 and test again. If not, stop and use $d_0$ as the value of $d$. The test statistic for these hypotheses is $\hat{\Lambda}_n = n \sum_{j=d_0+1}^p \hat{\lambda}_j$, where the $\hat{\lambda}_j$s are the eigenvalue of $\hat{\text{var}}\{E(\hat{Z} \mid Y)\}$ ordered from largest to smallest. Under the hypothesis $d = d_0$, $\hat{\Lambda}_n$ converges in distribution to $\sum_{k=1}^{(p-d_0)(h-d_0)} \omega_k C_k$, where the $C_k$s are iid $\chi^2_1$ and the $\omega_k$s are eigenvector of a certain covariance matrix (Cook 1998, Section 11.3). If the constant covariance assumption holds then $\hat{\Lambda}_n$ converges to a chi-square random variable with $(p-d_0)(h-d_0)$ degrees of freedom.
6.4. SLICED INVERSE REGRESSION, SIR

6.4.2 Mussels’ muscles II

For a simple illustration of SIR and SDR methods generally, we consider data on horse mussels introduced in Section 4.4.3, but now we take the response to be muscle mass $M$ to allow non-linearity in the mean function. The predictors $X$ are still the logarithms of the length, width, height and mass of the mussel’s shell. Fitting with $h = 8$ slices, SIR’s dimension tests indicated clearly that $d = 1$ so only a single linear combination of the four predictors is needed to describe the regression of $M$ on $X$, which is consistent with the finding in Section 4.4.3. Shown in Figure 6.1 is a plot of $M$ versus $\hat{\eta}^T X$, the estimated reduction of $X$ with $\|\hat{\eta}\| = 1$. The curve on the plot is the estimated mean function from fitting a cubic polynomial to the points on the plot; that is, from fitting

$$Y = \mu + \beta_1 \eta^T X + \beta_2 (\eta^T X)^2 + \beta_3 (\eta^T X)^3 + \epsilon.$$  

Standard inferences based on this model will likely be optimistic, since they do not take the variation in $\hat{\eta}$ into account. To avoid this issue, we can fit the model

$$Y = \mu + \beta_1 \eta^T X + \beta_2 (\eta^T X)^2 + \beta_3 (\eta^T X)^3 + \epsilon$$

with the constraint $\|\eta\| = 1$. This serves to re-estimate $\eta$ given the model form inferred from the summary plot. To avoid the constraint $\|\eta\| = 1$, re-parameterize the previous

Figure 6.1: Summary plot for the mussels data
model in terms of \( \alpha = \eta \beta_1 \), \( \gamma_2 = \beta_2 / \beta_1^2 \) and \( \gamma_3 = \beta_3 / \beta_1^3 \):

\[
Y = \mu + \alpha^T X + \gamma_2(\alpha^T X)^2 + \gamma_3(\alpha^T X)^3 + \epsilon.
\]

This model can now be fitted and studied using traditional methods for non-linear models.

Generally, inspection of a low dimensional summary plot following an SDR reduction can lead to a model specification in the form

\[
Y = f(\eta_1^T X, \ldots, \eta_d^T X, \beta) + \epsilon,
\]

where \( f \) is a function of the SDR variates \( \eta_1^T X, \ldots, \eta_d^T X \) and model parameters \( \beta \).

### 6.4.3 Envelopes and SIR

Let \( \mathcal{V} = \text{span}(\nu_1, \ldots, \nu_h) \). The linearity condition guarantees only that \( \Sigma_X^{-1} \mathcal{V} \subseteq S_Y|X \). We assume for simplicity that \( S_Y|X = \Sigma_X^{-1} \mathcal{V} \). SIR’s estimator of \( S_Y|X \) might be improved by using \( E_{\Sigma_X}(\mathcal{V}) \), the \( \Sigma_X \)-envelope of \( \mathcal{V} \). To see this possibility, let \( \Gamma \in \mathbb{R}^{p \times u} \) be a basis for \( E_{\Sigma_X}(\mathcal{V}) \) and let \( (\Gamma, \Gamma_0) \in \mathbb{R}^{p \times p} \) be an orthogonal matrix. Then, as in other envelope contexts,

\[
\Sigma_X = P_\mathcal{V} \Sigma_X P_\mathcal{V} + Q_\mathcal{V} \Sigma Q_\mathcal{V} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T
\]

\[
\Sigma_X^{-1} = \Gamma \Omega^{-1} \Gamma^T + \Gamma_0 \Omega_0^{-1} \Gamma_0^T
\]

\[
S_Y|X = \Sigma_X^{-1} \mathcal{V} = \Gamma \Omega^{-1} \Gamma^T \mathcal{V} = \Gamma (\Gamma \Sigma_X \Gamma^T)^{-1} \Gamma^T \mathcal{V}.
\]

Consequently, we may be able to improve the SIR estimator by using the envelope variant

\[
\hat{S}_Y|X = \hat{\Gamma} (\hat{\Gamma} \Sigma_X \hat{\Gamma}^T)^{-1} \hat{\Gamma}^T \hat{\mathcal{V}},
\]

where \( \hat{\Gamma} \) is an estimator of a semi-orthogonal basis matrix for \( E_{\Sigma_X}(\mathcal{V}) \) and \( \hat{\mathcal{V}} \) is an estimator of \( \mathcal{V} \). The rationale here is similar to that for predictor envelopes discussed in Chapter 4. Any of the envelope algorithms discussed in Chapter 8 could be used to form \( \hat{\Gamma} \).

Restricting consideration to \( d = 1 \), Li et al.\(^{12}\) (2007) estimated \( E_{\Sigma_X}(\mathcal{V}) \) by using a Krylov matrix (cf. Section 8.4.3) because the resulting method does not require inverting \( S_X \) and is thus serviceable when \( n < p \). This treatment was extended by Cook et al.\(^{13}\) (2007) to allow for \( d > 1 \). Cook, et al. (2013) compared the use of Krylov matrices with other methods of envelope construction.

---


6.5 Principal Components

6.5.1 Introduction

Principal component analysis (PCA) has a long tradition as a data-analytic method of unsupervised reduction of multivariate observations, including predictors in regression. Its first uses have been traced back to Adcock\(^\text{14}\) (1878), Pearson\(^\text{15}\) (1901) and Hotelling\(^\text{16}\) (1933), and it may now be the most widely used dimension reduction method in the applied sciences. PCA has been justified on various grounds (Jolliffe\(^\text{17}\) 2002), but mainly users sought a few uncorrelated linear combinations of the original variables that capture maximal variation, with the hope that those linear combinations would, in some sense, preserve relevant information.

Suppose we have \(n\) observations on the random vector \(Y \in \mathbb{R}^r\) with mean \(\mu\) and variance \(\Sigma_Y > 0\). Let \(q = \text{rank}(S_Y) \leq r\). Additionally, let \(\hat{\phi}_1 > \hat{\phi}_2 > \cdots > \hat{\phi}_q > 0\) and \(\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_q\) be the non-zero eigenvalues and corresponding eigenvectors of \(S_Y\). The linear combinations \(\hat{\ell}_j^T Y, j = 1, \ldots, q\) are the sample principal components of \(Y\); we refer to the eigenvectors \(\hat{\ell}_j, j = 1, \ldots, q\), as principal component directions. The population principal components are computed in the same way with \(S_Y\) replaced by \(\Sigma_Y\).

The sample variance of the \(j\)-th principal component \(\hat{\ell}_j^T Y\) is \(\hat{\phi}_j\). If the fraction of explained variation \(f_d = \sum_{j=1}^{d} \hat{\phi}_j / \sum_{j=1}^{q} \hat{\phi}_j\) is sufficiently close to 1, then the first \(d \leq q\) principal components \(\hat{\ell}_j^T Y, j = 1, \ldots, d\), are uncorrelated linear combinations of the original variables that capture most of the variation in the data. There are many methods for selecting the number of principal components \(d\), depending on application specific requirements. Jolliffe (2002) gave a comprehensive account of principal components, including methods for selecting \(d\). Despite its popularity, PCA was not based on a probability model until Tipping and Bishop\(^\text{18}\) (1999) introduced a model in which the principal component directions arise through maximum likelihood estimation.

The developments in this section are based on the latent variable model

where $\mu \in \mathbb{R}^r$, $\theta \in \mathbb{R}^{r \times d}$ with $d \leq r$, and the errors $\epsilon_i$ are independent copies of $\epsilon \sim N(0, \Delta)$. The $\nu_i \in \mathbb{R}^d$ are latent random vectors that are assumed to be independent copies of a random vector $\nu \sim N(0, I_d)$ that is independent of $\epsilon$, but we will also consider in Section 6.5.3 the case where the $\nu_i$’s are non-stochastic. The identity covariance matrix in the former assumption can always be achieved by normalization: If $\text{var}(\nu) = A$ then write $\theta A^{1/2} A^{-1/2} \nu$ and redefine $\theta$ and $\nu$ accordingly. The parameter $\theta$ is not identified since $\theta \nu = (\theta O)(O^T \nu)$ for any orthogonal matrix $O$, resulting in an equivalent model. Model (6.15) is like the multivariate linear model (1.1) except for the crucial difference that the “predictors” $\nu$ are unobserved random vectors.

We think of $\nu$ as representing variation that is caused by latent extrinsic factors, while $\epsilon$ represents intrinsic variation that would be present if all extrinsic factors were held fixed. The goal of an analysis is to extract the part of $Y$ that is affected by the extrinsic factors. More specifically, under model (6.15) it can be shown that

$$Y \mid \theta^T \Delta^{-1} Y$$

and thus, given $\theta^T \Delta^{-1} Y$, $Y$ is unaffected by the extrinsic vector $\nu$. Consequently, $R(Y) := \theta^T \Delta^{-1} Y$ is the part of $Y$ affected by $\nu$ and is the minimal reduction we would like to estimate (Cook and Forzani 2008, Thm 2.1), reasoning that the intrinsic variation is generally of little relevance to the dimension reduction goals. Full rank linear transformations of $R$ do not matter in this context: for any full rank matrix $A \in \mathbb{R}^{d \times d}$, $Y \perp \nu \mid \theta^T \Delta^{-1} Y$ if and only if $Y \perp \nu \mid A^T \theta^T \Delta^{-1} Y$ and so the fundamental goal is to estimate $T := \text{span}(\Delta^{-1} \theta)$. Because $\nu$ is not observable, only the marginal distribution of $Y$ is available to estimate $T$. Under model (6.15), $Y$ is normal with mean $\mu$ and variance $\Sigma_Y = \Delta + \theta \theta^T$. The maximum likelihood estimator of $\mu$ is simply the sample mean of $Y$. However $\Delta$ and $\theta$ are confounded, and thus $T$ can not be estimated without additional structure.

For example, principal components have been used in face recognition to construct eigenface representations of individual images. Beginning with a set of $np \times c$ greyscale images normalized to have eyes and mouths aligned, the images are then vectorized so each face is represented as a vector $Y$ of length $r = pc$. The mean $\mu$ in model (6.15)
represents the average face and $\theta \nu_i$ models the deviations of individual faces from the average. The error $\varepsilon$ reflects the intrinsic variation in grayscale measurements across many images of the same face. An additional assumption of isotropic variation $\Delta = \sigma^2 I_r$ might be reasonable in this case. It could also be reasonable in applications involving measurement error, microarray data and calibration.

### 6.5.2 Random latent variables

Tipping and Bishop (1999) studied the version of model (6.15) with $\nu \sim N(0, I_d)$ and isotropic errors, $\Delta = \sigma^2 I_r$. In this case $T = \text{span}(\theta)$ is identified and Tipping and Bishop showed that its maximum likelihood estimator is the span of the first $d$ sample principal component directions of $Y$, leading to their notion of probabilistic principal components.

The maximum likelihood estimator of $T$ is perhaps easiest to find by first reformulating the isotropic version of model (6.15) as

$$Y = \mu + \Theta \delta \nu + \varepsilon,$$  \hspace{1cm} (6.17)

where $\Theta \in \mathbb{R}^{r \times d}$ is a semi-orthogonal basis matrix for $T$, $\delta \in \mathbb{R}^{d \times d}$ is a full rank coordinate matrix, $\nu \sim N(0, I_d)$ and $\varepsilon \sim N(0, \sigma^2 I_r)$. The variance $\Delta = \sigma^2 I_r$ says that, conditional on the extrinsic variables $\nu$, the measurements are independent and have the same variance. This requirement may be reasonable in a variety of applications, including calibration, some models for gene expression data and face recognition. The covariance matrix of $Y$ can be re-expressed as

$$\Sigma_Y = \Theta \delta \delta^T \Theta^T + \sigma^2 I_r = \Theta \Theta^T + \sigma^2 \Theta_0 \Theta_0^T,$$  \hspace{1cm} (6.18)

where $V = \delta \delta^T + \sigma^2 I_d$, and $(\Theta, \Theta_0)$ is an orthogonal matrix. From this we see that (6.17) is a trivial version of an envelope model since $T = \mathcal{E}_\Delta(T)$. Since $\delta$ is full rank, we parameterize the model in terms of $V$, $T$ and $\sigma^2$.

The maximum likelihood estimator of $\mu$ is $\hat{\mu} = \bar{Y}$. The log-likelihood $L(T, V, \sigma^2)$ maximized over $\mu$ is then

$$(2/n)L(T, V, \sigma^2) = c - \log |V| - \text{tr}(\Theta^T S_Y \Theta V^{-1}) - (r-d) \log(\sigma^2) - \sigma^{-2} \text{tr}(\Theta_0^T S_Y \Theta_0),$$

where $c = -r \log(2\pi)$. This log-likelihood, like others encountered in this book, depends only on $T$ and not on a particular basis $\Theta$.  

---

6.5. PRINCIPAL COMPONENTS
Maximizing over $V$ and $\sigma^2$ separately leads to the same partially maximized likelihood function as found by Tipping and Bishop (1999). However, the parameters $V$ and $\sigma^2$ are not in a proper product space because the eigenvalues of $V$ are bounded below by $\sigma^2$. Thus it seems inappropriate to maximize over $V$ and $\sigma^2$ separately and instead we re-express the log likelihood as

$$(2/n) L(T, V, \sigma^2) = c - \log |V| - (r - d) \log(\sigma^2) + \text{tr}\{\Theta^T S_Y \Theta (\sigma^{-2} I_r - V^{-1})\},$$

where $(\sigma^{-2} I_r - V^{-1}) > 0$. It now follows immediately from Lemma A.19 that this likelihood is maximized over $T$ by the span of the first $d$ eigenvectors of $S_Y$ regardless of the particular value of $(\sigma^{-2} I_r - V^{-1}) > 0$. We then have

**Proposition 6.2** The maximum likelihood estimator of $T$ under model (6.17) is the subspace spanned by the first $d$ principal component directions of $S_Y$.

The result of Proposition 6.2 is the same as that given by Tipping and Bishop (1999), but the justification here is different. The proposition gives no notion of the relative importance of the principal component directions, but they can still be ordered using traditional reasoning (Jolliffe 2002). In the next section we use envelopes to expand the probabilistic principal component model (6.17)

**Envelopes**

Instead of assuming an isotropic error structure, consider the unrestricted version of model (6.15) with $\Delta > 0$ and $\nu \sim N(0, I_d)$. We still think of the latent variable $\nu$ as representing extrinsic variation in $Y$ and the error $\varepsilon$ representing intrinsic variation, but now the intrinsic variation is anisotropic and might have correlated elements. This leads us back to the general confounding between $\Delta$ and $\theta$ mentioned previously. Nevertheless, useful results might still be possible if we can estimate an upper bound on $T$. By doing so, we would not lose any information on its extrinsic variation, but we might retain some of its intrinsic variation. Following Chen (2010), we construct an upper bound on $T$ by using an envelope.

To apply these ideas in the context of model (6.15), consider the $\Delta$-envelope of $T$, $E_\Delta(T)$. It follows from Proposition A.5 that $E_\Delta(T) = E_\Delta(\theta)$, so we can equivalently consider the $\Delta$-envelope of $\text{span}(\theta)$. Let the columns of the semi-orthogonal matrix

---

\( \Gamma \in \mathbb{R}^{p \times u} \) be a basis for \( \mathcal{E}_\Delta(\theta) \) and let \((\Gamma, \Gamma_0)\) be an orthogonal matrix where, \( u = \dim(\mathcal{E}_\Delta(\theta)) \geq d \). We have \( \text{span}(\theta) \subseteq \mathcal{E}_\Delta(\theta) \). Consequently there is a coordinate matrix \( \eta \in \mathbb{R}^{u \times d} \) with rank \( d \) so that \( \theta = \Gamma \eta \). From this we obtain an envelope component model

\[
\begin{align*}
Y &= \mu + \Gamma \eta \nu + \varepsilon, \\
\Delta &= \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T,
\end{align*}
\] (6.19)

Under this model we see that \( \Gamma^T Y \perp \Gamma_0^T Y | \nu \) and that \( \Gamma_0^T Y \perp (\Gamma^T Y, \nu) \) and \( Y \perp \nu | \Gamma^T Y \). Consequently, we think of \( \Gamma_0^T Y \) with \( \text{var}(\Gamma_0^T Y) = \text{var}(\Gamma^T Y) = \Omega_0 \) as capturing just intrinsic variation, while \( \Gamma^T Y \) with \( \text{var}(\Gamma^T Y | \nu) = \Omega \) and \( \Psi := \text{var}(\Gamma^T Y) = \Omega + \eta \eta^T \) holds all of the extrinsic variation. The intrinsic and extrinsic variation are both represented in the marginal covariance matrix of \( Y \): \( \Sigma_Y = \Gamma \Psi \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \). Under model (6.19), \( \Delta^{-1} \theta = \Gamma \Omega^{-1} \eta \) and so the desired reduction is \( R = \eta^T \Omega^{-1} \Gamma^T Y \in \mathbb{R}^d \). However, \( \eta \) is not estimable from the marginal of \( Y \) as it is confounded with \( \Omega \) and thus we pursue the sufficient reduction \( R = \Gamma^T Y \in \mathbb{R}^u \). Of course, if \( u = d \) then \( \mathcal{T} \) and the envelope coincide, \( \mathcal{T} = \mathcal{E}_\Delta(\mathcal{T}) \).

An estimate of the envelope \( \mathcal{E}_\Delta(\mathcal{T}) = \text{span}(\Gamma) \) then provides an estimated upper bound on \( \mathcal{T} \), which is the target of our inquiry. In short, we need to estimate \( \mathcal{E}_\Delta(\mathcal{T}) \) based on a sample \( Y_1, \ldots, Y_n \) from a \( N_r(\mu, \Sigma_Y) \) distribution. The estimated reduction is then \( \hat{R} = \hat{\Gamma}^T Y \), where the columns of \( \hat{\Gamma} \) form a basis for the estimated envelope. We turn next to maximum likelihood estimation.

After some algebra, the log-likelihood function \( L_1 \) with \( \hat{\mu} = \hat{Y} \) can be represented in the form

\[
(2/n)L_1(\mathcal{E}, \Psi, \Omega_0) = -\log |\Psi| - \text{tr}(\Gamma^T S_Y \Gamma \Psi^{-1}) - \log |\Omega_0| - \text{tr}(\Gamma_0^T S_Y \Gamma_0 \Omega_0^{-1}),
\]

where \( \mathcal{E}_\Delta(\mathcal{T}) \) has been shortened to \( \mathcal{E} \) for use as an argument. Maximizing over \( \Psi \) and \( \Omega_0 \), which are defined on a product space, we have the partially maximized log-likelihood function

\[
(2/n)L_{11}(\mathcal{E}) = -\log |\Gamma^T S_Y \Gamma| - \log |\Gamma_0^T S_Y \Gamma_0| - r
= -\log |\Gamma^T S_Y \Gamma| - \log |\Gamma^T S_Y^{-1} \Gamma| - r - \log |S_Y|.
\]

The function \( L_{11}(\mathcal{E}) \) requires \( n > r \) as \( S_Y \) must not be singular. Also, as expected it is invariant under right orthogonal transformations of \( \Gamma \) and thus it determines only a subspace and not a particular coordinate system. It follows immediately from Proposition A.15 that
the span of any \( u \) sample principal component directions is a maximum likelihood estimator of \( \mathcal{E}_\Delta(T) \). Model (6.19) represents a demarcation point for the effectiveness of envelopes. Because any subset of principal component directions is equally supported by the likelihood, there seems little hope of obtaining useful estimates of \( T \) or its envelope \( \mathcal{E}_\Delta(T) \) with less structure. However, useful results might be obtained when additional structure is imposed. For instance, suppose that \( u = d \), \( \text{var}(\Gamma^T Y|\nu) = \Omega = \sigma^2 I_d \) and \( \text{var}(\Gamma^T_0 Y) = \Omega_0 = \sigma_0^2 I_{r-d} \), so that the conditional extrinsic and intrinsic variation have the same structure. Then \( \Delta = \sigma^2 I_r \) and model (6.19) reduces to probabilistic principal component model (6.17). In effect, models (6.17) and (6.19) represent extremes of a modeling environment for multivariate reduction. Perhaps there is sufficient flexibility between these extremes for application-specific requirements. Two possibilities are described in the next two sections.

**Envelopes with Isotropic intrinsic and extrinsic variation**

We begin with a relatively small generalization of model (6.15) that nevertheless directs us away from the leading principal components as the desired reduction of \( Y \). Assume that the marginal variation \( \Sigma_Y = \sigma^2 \Gamma^T \Gamma + \sigma_0^2 \Gamma_0^T \Gamma_0^T \), where the intrinsic variation \( \Omega_0 = \sigma_0^2 I_{r-u} \). Then the log-likelihood \( L_2 \) can be written as

\[
(2/n)L_2(\mathcal{E}, \sigma^2, \sigma_0^2) = -u \log(\sigma^2) - \sigma^{-2} \text{tr}(\Gamma^T S_Y \Gamma) - (r-u) \log(\sigma_0^2) - \sigma_0^{-2} \text{tr}(\Gamma^T_0 S_Y \Gamma_0).
\]

Maximizing over \( \sigma^2 \) and \( \sigma_0^2 \), we have the partially maximized log-likelihood function

\[
(2/n)L_{21}(\mathcal{E}) = -u \log\{\text{tr}(\Gamma^T S_Y \Gamma)\} - (r-u) \log\{\text{tr}(S_Y) - \text{tr}(\Gamma^T S_Y \Gamma)\}
- r + u \log(u) + (r-u) \log(r-u).
\]

The function \( L_{21}(\mathcal{E}) \) requires \( n > r - u + 1 \) to ensure that \( \text{tr}(\Gamma^T S_Y \Gamma) > 0 \) for all \( \Gamma \).

**Proposition 6.3** Assume envelope model (6.19) with \( \Sigma_Y = \sigma^2 \Gamma^T \Gamma + \sigma_0^2 \Gamma_0^T \Gamma_0^T \), where \( \Omega_0 = \sigma_0^2 I_{r-u} \). Then \( L_{21}(\mathcal{E}) \) is maximized by the span of either the first \( u \) principal component directions or the last \( u \) principal component directions.

**Proof.** Maximizing \( L_{21}(\mathcal{E}) \) is equivalent to minimizing

\[
\log\{\text{tr}(\Gamma^T S_Y \Gamma)\} + \frac{r - u}{u} \log\{\text{tr}(S_Y) - \text{tr}(\Gamma^T S_Y \Gamma)\}.
\]

Let \( x = \text{tr}(\Gamma^T S_Y \Gamma), C = (r-u)/u \) and \( K = \text{tr}(S_Y) \). With probability one, we have

\[
\min\{\text{tr}(\Gamma^T S_Y \Gamma)\} < \frac{K}{1 + C} = \frac{u}{r} \text{tr}(S_Y) < \max\{\text{tr}(\Gamma^T S_Y \Gamma)\}
\]

CHAPTER 6. CONNECTIONS
subject to $\Gamma^T\Gamma = I_u$. Following Lemma A.20, we know that the maximum value of $L_{21}(\mathcal{E})$ is reached at either $\max\{\text{tr}(\Gamma^T\mathbf{S}Y\Gamma)\}$ or $\min\{\text{tr}(\Gamma^T\mathbf{S}Y\Gamma)\}$. That is, the maximum likelihood estimator of $\mathcal{E}_\Delta(T)$ is the span of either the first $u$ principal component directions or the last $u$ principal component directions.

Using this proposition in combination with the likelihood function $L_{21}$ it can be seen that the maximum likelihood estimator is the span of the first $u$ principal components if either (a) $f_u > 1/2$ and $r - 2u > 0$ or (b) $f_u < 1/2$ and $r - 2u < 0$, where $f_u$ is the fraction of explained variation as defined in Section 6.5.1. Otherwise, the maximum likelihood estimator is the span of the last $u$ principal components.

**Envelopes with isotropic intrinsic variation**

Additional flexibility can be incorporated by requiring only that $\Omega_0 = \sigma_0^2 I_{r-u}$, leaving the extrinsic component $\Psi > 0$ arbitrary. In this case model (6.19) reduces to

$$
\begin{align*}
\mathbf{Y} &= \mu + \Gamma\eta\nu + \varepsilon \quad (6.20) \\
\Delta &= \Gamma\Omega^{\Gamma^T} + \sigma_0^2 \Gamma_0\Gamma_0^T
\end{align*}
$$

The log-likelihood function $L_3$ for model (6.20) can be written as

$$(2/n)L_3(\mathcal{E}, \Delta, \sigma_0^2) = -\log|\Psi| - \text{tr}(\Gamma^T\mathbf{S}Y\Gamma\Psi^{-1}) - (r-u)\log(\sigma_0^2) - \sigma_0^{-2}\text{tr}(\Gamma_0^T\mathbf{S}Y\Gamma_0).$$

Maximizing over $\Psi$ and $\sigma_0^2$, we have the partially log-likelihood function

$$(2/n)L_{31}(\mathcal{E}) = -\log |\Gamma^T\mathbf{S}Y\Gamma| - (r-u)\log(\text{tr}(\mathbf{S}Y) - \text{tr}(\Gamma^T\mathbf{S}Y\Gamma))$$

$$-r + (r-u)\log(r-u).$$

The function $L_{31}(\mathcal{E})$ requires $n > r$ as $\Gamma^T\mathbf{S}Y\Gamma$ must not be singular. The next proposition (Chen 2010) describes how to maximize $L_{31}$.

**Proposition 6.4** Under model (6.20), the maximum likelihood estimator of $\mathcal{E}_\Delta(T)$ is the span of the first $k$ and last $u-k$ principal component directions for some $k \in \{0, 1, \ldots, u\}$.

**Proof.** We sketch the main ideas in the proof given by Chen (2010).

Maximizing $L_{31}(\mathcal{E})$ is equivalent to minimizing

$$L_{32}(\mathcal{E}) = \log |\Gamma^T\mathbf{S}Y\Gamma| + (r-u)\log\{\text{tr}(\mathbf{S}Y) - \text{tr}(\Gamma^T\mathbf{S}Y\Gamma)\} \quad (6.21)$$
subject to $\mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_u$. Using Lagrange multipliers, a solution can be obtained by finding the stationary points of the unconstrained function
\[
\log|\mathbf{\Gamma}^T \mathbf{S}_Y \mathbf{\Gamma}| + (r - u) \log\{\text{tr}(\mathbf{S}_Y) - \text{tr}(\mathbf{\Gamma}^T \mathbf{S}_Y \mathbf{\Gamma})\} + \text{tr}\{\mathbf{U}(\mathbf{\Gamma}^T \mathbf{\Gamma} - \mathbf{I}_u)\},
\]
where $\mathbf{U}$ is a $u \times u$ matrix of Lagrange multipliers. Let
\[
\mathbf{A}(\mathbf{\Gamma}) = (\mathbf{\Gamma}^T \mathbf{S}_Y \mathbf{\Gamma})^{-1} - \frac{(r - u)}{\text{tr}(\mathbf{S}_Y) - \text{tr}(\mathbf{\Gamma}^T \mathbf{S}_Y \mathbf{\Gamma})} \mathbf{I}_u.
\]
Then the Lagrange analysis leads to the conclusion that the minimizers of (6.21) must satisfy $\mathbf{Q}_r \mathbf{S}_Y \mathbf{\Gamma} \mathbf{A}(\mathbf{\Gamma}) = 0$. Chen (2010) showed that $\mathbf{A}(\mathbf{\Gamma})$ must be of full rank, which implies that $\text{span}(\mathbf{S}_Y \mathbf{\Gamma}) = \text{span}(\mathbf{\Gamma})$ and thus that $\mathbf{\Gamma}$ spans a reducing subspace of $\mathbf{S}_Y$. Consequently we can take the columns of $\mathbf{\Gamma}$ to be a subset of eigenvectors of $\mathbf{S}_Y$ with eigenvalues $\hat{\sigma}_1, \ldots, \hat{\sigma}_u$, which now are still not necessarily ordered. Let $\hat{\sigma}_{(u+1)}, \ldots, \hat{\sigma}_{(r)}$ denote the complement of $\hat{\sigma}_1, \ldots, \hat{\sigma}_u$. Then objective function (6.21) becomes
\[
L_{32}(\mathcal{E}) = \sum_{i=1}^{u} \log(\hat{\sigma}_i) + (r - u) \log\{\hat{\sigma}_{(u+1)} + \ldots + \hat{\sigma}_{(r)}\}. \tag{6.22}
\]
Suppose there exists a configuration so that $\hat{\sigma}_{(i)} < \hat{\sigma}_{(l)} < \hat{\sigma}_{(j)}$. It follows from Lemma A.20 that (6.22) can be reduced by replacing $\hat{\sigma}_{(l)}$ with either $\hat{\sigma}_{(i)}$ or $\hat{\sigma}_{(j)}$. This tells us that the complementary set of eigenvalues $\hat{\sigma}_{(u+1)}, \ldots, \hat{\sigma}_{(r)}$ must form a contiguous block. In other words, there exist an integer $1 \leq k \leq r$ so that the maximum likelihood estimator of $\mathcal{E}$ is the span of the first $k$ and last $u - k$ principal component directions. \hfill \Box

Let $\sigma_i, i = 1, 2, \ldots, u$ be the population eigenvalues of $\mathbf{\Psi}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_u$. The setting of model (6.20) basically says that the extrinsic signals can have different scales and be correlated, both conditionally $\text{var}(\mathbf{\Gamma}^T \mathbf{Y}|\mathbf{\nu}) = \mathbf{\Omega}$ and unconditionally $\text{var}(\mathbf{\Gamma}^T \mathbf{Y}) = \mathbf{\Psi}$, but the intrinsic noise has a circular normal distribution $\text{var}(\mathbf{\Gamma}_0) = \sigma_0^2 \mathbf{I}$. If $\sigma_i > \sigma_0$ for all $i = 1, 2, \ldots, u$, then in the population we have the same solution as the usual principal component analysis. In effect, if the extrinsic signal is strong enough, the usual principal component analysis is doing a sensible thing. If $\sigma_0 > \sigma_i$ for all $i = 1, 2, \ldots, u$, then we have the last $u$ principal component directions as the solution. If $\sigma_0$ lies among $\sigma_i$ for $i = 1, 2, \ldots, u$, then the solution is the span of the first $k$ principal component directions and $u - k$ last principal component directions for some $k$, where $k$ ranges from 0 to $u$. This provides a fast algorithm to search for the maximizer of $L_{31}$. If $u = d$, so that $\mathbf{T}$ is itself
a reducing subspace of $\Delta$ and $\mathcal{T} = \mathcal{E}_\Delta(\mathcal{T})$ then this reduces to the extreme components studied by Welling et. al\textsuperscript{21} (2003).

**Selection of the dimension $u$**

The sequence likelihood ratio test can be used to help determine the dimensionality of the envelope denoted by $u$. For example, consider model (6.20). The hypothesis $u = u_0$ can be tested by using the likelihood ratio statistic $\Lambda(u_0) = 2(\hat{L}_{\text{fm}} - \hat{L}^{(u_0)})$, where $\hat{L}_{\text{fm}}$ denotes the maximum value of the log likelihood for the full model ($u = r$), and $\hat{L}^{(u_0)}$ the maximum value of the log likelihood when $u = u_0$. In fact, $\hat{L}_{\text{fm}} = -(nr)/2 - (n/2) \log |S_Y|$. The total number of parameters needed to estimate model (6.20) is $r + u(u + 1)/2 + u(r - u) + 1$. The first term on the right hand side corresponds to the estimation of the grand mean $\mu$. The second term corresponds to the estimation of the unconstrained symmetric matrix $\Psi$. The third term corresponds the number of parameters needed to describe the subspace $\mathcal{E}_\Delta(\mathcal{T})$. The last term corresponds to $\sigma^2_0$. Following standard likelihood theory, under the null hypothesis, $\Lambda(u_0)$ is distributed asymptotically as a chi-squared random variable with $(r - u_0 + 2)(r - u_0 - 1)/2$ degrees of freedom. This test can be used sequentially to choose $u$: Starting with $u_0 = 0$, test the hypothesis $u = u_0$ against the alternative $u > u_0$. If the test is rejected, increment $u_0$ by 1 and test again, stopping at the first hypothesis that is not rejected.

**6.5.3 Fixed latent variables and isotropic errors**

Model (6.17) is based on the assumption that $\nu \sim N(0, I_r)$. However, we obtain the same reduction if we condition on the realized (but unobserved) values $\nu_1, \ldots, \nu_n$ of $\nu$ and treat them as fixed centered vectors so that $\sum_{i=1}^n \nu_i = 0$. Since the $\nu_i$’s are now fixed, we absorb $\delta$ into them and write the isotropic model as

$$\begin{align*}
Y_i = \mu + \Theta \nu_i + \varepsilon_i, \quad i = 1, \ldots, n. 
\end{align*}$$

The log likelihood for this setting is

$$L(\mu, \mathcal{T}, \sigma^2, \nu_1, \ldots, \nu_n) = -(nr/2) \log(\sigma^2) - (1/2\sigma^2) \sum_{i=1}^n \| Y_i - \mu - \Theta \nu_i \|^2. $$

Because $\sum_{i=1}^{n} \nu_i = 0$, it follows that $\hat{\mu} = \bar{Y}$. To estimate $\nu_i$ for a selected value of $i$ with the remaining parameters held fixed we need to minimize $\|Y_i - \bar{Y} - \Theta \nu_i\|^2$. This just ordinary least squares and consequently is minimized by setting $\nu_i = \Theta^T (Y_i - \bar{Y})$.

These automatically satisfy the constraint $\sum_{i=1}^{n} \nu_i = 0$. Substituting back, we get the partially maximized log likelihood

$$L_1(T, \sigma^2) = -(nr/2) \log(\sigma^2) - (n/2\sigma^2)\text{tr}(S_Y) + (n/2\sigma^2)\text{tr}(P_T S_Y).$$

It follows from Lemma A.17 that $\hat{T} = \text{span}(\hat{\ell}_1, \ldots, \hat{\ell}_d)$. Again there is no notion of relative importance of vectors in $\hat{T}$. Substituting this estimator into $L_1$ we get

$$L_2(\sigma^2) = -(nr/2) \log(\sigma^2) - (n/2\sigma^2) \sum_{i=d+1}^{r} \hat{\phi}_i,$$

which is maximized at $\hat{\sigma}^2 = \sum_{i=d+1}^{r} \hat{\phi}_i/r$.

It follows from the above results that the estimated reduction evaluated at the data points $\hat{R}(Y_i) = (\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_d)^T Y_i$ consists of the first $d$ sample principal components of $Y$. Consequently, we now have two related models that produce principal components as maximum likelihood estimators.

### 6.5.4 Numerical illustrations

Figure 6.2 is a schematic representation of why principal components work in the isotropic model (6.17) based on identity (6.18). $T$ is a reducing subspace of $\Sigma_Y$ and consequently the eigenvectors of $\Sigma_Y$ are in either $T$ or $T^\perp$. And the eigenvalues of $\Sigma_Y$ corresponding to $T$ are uniformly large than the remaining eigenvalues, assuming that $\var(\nu) > 0$. Intuitively, we begin with the circular contours of $\var(Y|\nu) = \Delta = \sigma^2 I_r$. Adding the signal $\Theta \delta \delta^T \Theta^T$ then distorts the circular contours into ellipses with the span of its first axis being equal to $T$.

The following simulation example may help to further fix ideas. Observations on $Y$ were generated from model (6.15) as $Y_i = \theta \nu_i + \epsilon_i$, where $\nu_i$ was sampled uniformly from the boundary of the square $[-1, 1]^2$ and the elements of the $r \times 2$ matrix $\theta$ were sampled independently from a standard normal distribution. Then $\theta$ was normalized to give $\Theta = \theta (\theta^T \theta)^{-1/2}$, leading to the representation

$$Y_i = \Theta (\theta^T \theta)^{1/2} \nu_i + \epsilon_i = \Theta \delta \nu_i + \epsilon_i,$$

where $\delta = (\theta^T \theta)^{1/2}$ and the error vector $\epsilon$ was sampled from a normal distribution with mean 0 and variance matrix $I_r$. The sampling used to construct $\Theta$ is for convenience
6.5. PRINCIPAL COMPONENTS

Figure 6.2: Schematic representation showing why principal components work in isotropic models.

only; the model is still conditional on $\Theta$ regardless of how it was obtained. This sampling process was repeated $n = 80$ times for various values of $r$. If PCA works as predicted by the theory then a plot of the sample principal components should recover the square that holds the $\nu$’s.

Figure 6.3 shows plots of the first two principal components for four values of $r$. We see that for small $r$ the square is not recognizable, but for larger values of $r$ the square is quite clear. In Figure 6.3d, there are $r = 500$ predictors, while the number of observations is still $n = 80$. The sides of the estimated square in Figure 6.3d do not align with the coordinate axes because the method is designed to estimate only the subspace $\mathcal{T}$ with isotropic errors.

Intuition about why principal components are apparently working in this example can be found in the work of Johnson and Lu\textsuperscript{22} (2009) who studied the asymptotic behavior of the first principal component direction under model (6.15) with $d = 1$ and $\nu \sim N(0, 1)$. In this case, $\delta = \|\theta\|$ is simply the length of $\theta$. Let $r(n)$ denote the dimension of $Y$, which is allowed to grow with the sample size. Let $R(\hat{\ell}_1, \theta)$ denote the cosine of the angle between the first principal component direction $\hat{\ell}_1$ and $\theta$, and let $\omega = \lim_{n \to \infty} \|\theta\|^2/\sigma^2$ and let $c = \lim_{n \to \infty} r(n)/n$. This structure imagines that the number of predictors $r(n)$ grows with the sample size and so the previous limits involve only $n$. Then (Johnson and

Figure 6.3: Plots of the estimated sufficient reduction from the PC model with $n = 80$ observations and varying number of predictors $r$. Each plot was constructed from one simulated dataset.

Lu, 2009, Thm. 1)

$$\lim_{n \to \infty} R^2(\hat{\ell}_1, \theta) = R^2_{\infty}(\omega, c) = \frac{(\omega^2 - c)_{+}}{\omega^2 + c\omega}. $$

This implies that if $\omega$ is finite then $R^2_{\infty} < 1$ if and only if $c > 0$ and consequently $\hat{\ell}_1$ is a consistent estimator of $\text{span}(\theta)$ if and only if $r(n)/n \to 0$. In an extreme case, if $\omega^2 < c \left(\lim_{n \to \infty} r\sigma^4/(n\|\theta\|^4) \geq 1\right)$ then $\hat{\ell}_1$ and $\theta$ are asymptotically orthogonal and the first principal component direction contains no useful information on $\theta$. These results can be taken to imply that the sample size should be large relative to the number of variables for principal components to work well in practice when $\omega$ is finite.

However, if $\omega$ is large and $\omega \gg c$ then $R^2 \approx 1$ and principal components can be expected to work well. Suppose for example that $r(n)/n$ is bounded above and that the
elements of $\theta$ are sampled from a distribution with mean 0 and finite fourth moment. Then $\omega^2$ will diverge and principal components should work well. This is effectively what seems to have happened in the illustration of Figure 6.3, although $d = 2$ that illustration. Because the elements of $\theta \in \mathbb{R}^{r \times 2}$ were sampled from a standard normal distribution with mean 0 and $\|\theta\|^2$ will diverge, leading to an ever increasing signal. This suggests that principal components should also work well if information accumulates as $n$ increases.

However, the previous results also suggests that principal components will not work as well when only finitely many variables are relevant – that is, $\theta$ has only finitely many non-zero rows – or the signal grows very slowly, unless $n \gg r$. For example, Figure 6.4 shows the results of a simulation conducted like that for Figure 6.3 except the $i$-th row of $\theta$ was sampled from a normal distribution with mean 0 and variance $(10/i)I_2$. Consequently, $\omega^2$ involves the fourth moments of $N(0,10/i)$ random variables and thus will be finite. This suggests that $\hat{\ell}_1$ will be an inconsistent estimator of $\text{span}(\theta)$, unless $c = 0$. The patterns shown in Figure 6.4 seem to support this conclusion. Although $r$ has been increased from 5 to 1500 the signal does not seem to be increasing. Indeed, the best of the results shown are for $r = 250$, suggesting that there could be value in screening to remove the less informative variables. A screening method for $d = 1$ was developed by Johnson and Lu (2009).

6.6 Principal fitted components, PFC

Assuming the linearity condition, Proposition 6.1 tells us that the centered conditional means are contained in $\Sigma_X S_{Y|X}$. Principal fitted components (PFC) pursues dimension reduction by modeling the inverse regression of the predictors $X$ on the response to obtain an estimator of $\Sigma_X S_{Y|X}$ and subsequently an estimator of $S_{Y|X}$. Let $H \in \mathbb{R}^{p \times d}$ be a semi-orthogonal basis matrix for $\Sigma_X S_{Y|X}$, so $d = \text{dim}(S_{Y|X})$. Then we model the inverse mean as $E(X \mid Y = y) = \alpha + Hh_{f}(y)$, leading to the inverse model

$$
X = \alpha + Hh_{f}(y) + \varepsilon, \quad (6.24)
$$

$$
\text{var}(\varepsilon) = \Sigma_{X|Y},
$$

$$
\Sigma_{X} = HH\Sigma_{f}h_{f}^{T}H^{T} + \Sigma_{X|Y},
$$

where $f(y) \in \mathbb{R}^{r}$ is a user-selected function of the response with linearly independent elements, $h \in \mathbb{R}^{d \times r}$ is an unrestricted rank $d$ matrix, $\Sigma_{f} = \text{var}(f(Y))$ and $\beta := Hf$ is
Figure 6.4: Plots of the estimated sufficient reduction from the PC model with \( n = 80 \) observations, varying number of predictors \( r \) and decreasing signal. Each plot was constructed from one simulated dataset.

the coefficient matrix. In the notation of this model, the quantity that we desire to estimate is

\[
\text{span}( \Sigma^{-1}_X H ) = \text{span}( \Sigma^{-1}_{X|Y} H ) = \text{span}( \Sigma^{-1}_{X|Y} H h ) = \text{span}( \Sigma^{-1}_{X|Y} \beta ) \subseteq S_{Y|X}. \quad (6.25)
\]

Accordingly, we can estimate the central subspace from a fit of model (6.24) by assuming equality in (6.25) and substituting the estimates of \( \Sigma_{X|Y}, H \) and \( h \) into the left-hand side of (6.25). We next consider estimation of \( S_{Y|X} \) under two versions of model (6.24).

### 6.6.1 Isotropic errors, \( \Sigma_{X|Y} = \sigma^2 I_p \)

With isotropic errors, model (6.24) is like model (6.23), but with the latent effects modeled as \( \nu_i = hf(y_i) \) and \( H \) denoting the basis matrix instead of \( \Theta \). From (6.25) the central subspace is simply \( S_{Y|X} = \text{span}(H) \) so we need estimate only \( \text{span}(H) \). Assuming normal
errors and known dimension $d$, the maximum likelihood estimator of $S_{Y|X}$ is the span of the first $d$ eigenvectors of the sample covariance matrix $S_{X_{of}}$ of the fitted vectors from the linear regression of $X$ on $f$ (Cook\textsuperscript{23} 2007; Cook and Forzani 2009). More specifically, for a sample $(X_1, Y_1) \ldots (X_n, Y_n)$, let $X$ denote the $n \times p$ matrix with rows $(X_i - \bar{X})^T$ and let $F$ denote the $n \times r$ matrix with rows $(f(y_i) - \bar{f})^T$. Then the maximum likelihood estimator of $S_{Y|X}$ is the subspace spanned of the first $d$ eigenvectors of $S_{X_{of}} = X^T P F X / n$. The corresponding estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \left( \sum_{j=1}^{p} \hat{\phi}_i - \sum_{j=1}^{d} \hat{\phi}_{i}^{fit} \right) / p,$$

where the $\hat{\phi}_{j}$'s are the eigenvalues of $S_{X_{of}}$ and the $\hat{\phi}_{i}^{fit}$'s are the eigenvalues of $S_X$.

The estimated reduction evaluated at the data points $P_{\hat{S}_{Y|X}} X_i$ are called principal fitted components since they consists of the first $d$ sample principal components of the fitted vectors $\tilde{X}_i = X + \hat{f} f_i$. The rank of the sample covariance matrix $S_{X_{of}}$ will typically equal $r$ and so it will have only $r$ positive eigenvalues. This means that the choice of $f(y)$ automatically bounds the dimension of $S_{Y|X}$, $d \leq r$.

We can arrive at an envelope model by using a slightly different line of reasoning. Beginning with a version of the probabilistic principal component model (6.15) for $X = \mu + H \nu + \epsilon$ with random latent effects $\nu$ having $\text{var}(\nu) = \Sigma_{\nu}$, we write $\nu = \mathbb{E}(\nu | Y) + \epsilon$ where $\epsilon \mathcal{L} (\epsilon, Y)$. Substituting this into the model, we get

$$X = \mu + H \nu + \epsilon$$

$$= \mu + \mathbb{H} \mathbb{E}(\nu | Y) + \mathbb{H} \epsilon + \epsilon$$

$$= \mu + \mathbb{H} h f(y) + \mathbb{H} \epsilon + \epsilon,$$

where in the last step we modeled $\mathbb{E}(\nu | Y = y) = hf(y)$. Since $\text{span}(\mathbb{H})$ is a reducing subspace of $\text{var}(\mathbb{H} \epsilon + \epsilon) = \mathbb{H} (\Sigma_{\epsilon} + \sigma^2 I) \mathbb{H} + \sigma^2 \mathbb{H}_0 \mathbb{H}_0^T$, we have an envelope model. Estimation under this model can be developed following the general steps outlined elsewhere in this book.

### 6.6.2 Anisotropic errors, $\Sigma_{X|Y} > 0$

With general errors $\Sigma_{X|Y} > 0$, model (6.24) becomes an instance of the reduced-rank regression (6.6), with response $X$ and predictor $f(y)$. Adapting the estimators of $\beta$ and $\Sigma$

CHAPTER 6. CONNECTIONS

given at (6.8) and (6.9), we have the maximum likelihood estimators for model (6.24),

\[
\hat{\beta} = \hat{H} = S_{X}^{1/2}C_{X,f}S_{f}^{-1/2}
\]

(6.26)
\[
\hat{\Sigma}_{X|Y} = S_{X}^{1/2}(I_{p} - C_{X,f}^{(d)}C_{f,X}^{(d)})S_{X}^{1/2}.
\]

(6.27)

The estimator of the central subspace with known dimension \(d\) is then

\[
\hat{S}_{Y|X} = \text{span}(\hat{\Sigma}_{X|Y}^{-1}\hat{\beta}) = \text{span}(S_{X}^{-1/2}C_{X,f}^{(d)}),
\]

(6.28)

where the second equality follows from re-expressing \((I_{p} - C_{X,f}^{(d)}C_{f,X}^{(d)})^{-1}\) by using the Woodbury matrix identity.

As in Section 6.2.3, the envelope version of model (6.24) is obtained by parameterizing it in terms of the \(\Sigma_{X|Y}\) envelope of \(\text{span}(H)\), leading to the adaptation of model (6.10),

\[
X = \alpha + \Gamma \eta h f(y) + \varepsilon, \quad \Sigma_{X|Y} = \Gamma \Omega^{T} + \Gamma_{0} \Omega_{0}^{T}.\]

(6.29)

Accordingly, the central subspace is given by

\[
S_{Y|X} = \text{span}(\Sigma_{X|Y}^{-1}\hat{\beta}) = \text{span}(\Gamma^{T} \eta h) \subseteq \mathcal{E}_{\Sigma_{X|Y}}(\beta).
\]

The central subspace can then be estimated by using the following steps.

1. Select \(d\) and \(u\) using one of the methods referenced at the end of Section 6.2.3.

2. Estimate \(\Gamma\) by finding an argument \(\hat{\Gamma}\) that minimizes objective function (6.13) with \(Y\) replaced by \(X\), and \(X\) replaced by \(f(y)\).

3. Regress the reduced predictor vector \(\hat{\Gamma}^{T}X\) on \(f\), and then use (6.26) and 6.27 to estimate \(\Omega\) and \(\eta h\):

\[
\hat{\eta}h = S_{\Gamma^{T}X, f}^{-1/2}C_{\Gamma^{T}X, f}^{(d)}S_{f}^{-1/2}
\]
\[
\hat{\Omega} = S_{\Gamma^{T}X, f}^{-1/2}(I_{u} - C_{\Gamma^{T}X, f}^{(d)}C_{f, \Gamma^{T}X}^{(d)})S_{\Gamma^{T}X}^{-1/2}
\]

4. The central subspace is then estimated as

\[
\hat{S}_{Y|X} = \text{span}(\hat{\Gamma} \hat{\Omega}^{-1} \hat{\eta}) = \text{span}(\hat{\Gamma}S_{\Gamma^{T}X, f}^{-1/2}C_{\Gamma^{T}X, f}^{(d)}).
\]
6.6. PRINCIPAL FITTED COMPONENTS, PFC

6.6.3 Non-normal errors and the choice of f

The estimators in the preceding sections were developed under the assumptions that \( f(y) \) correctly models \( \mathbb{E}(X \mid Y) \) and that the errors in model (6.24) are normal. Neither of these assumptions is crucial for the large-sample success of those methods.

We relax normality by assuming that the errors \( \varepsilon \) for model (6.24) have finite fourth moments and are independent of \( Y \). We allow for misspecification of \( f \) by specifying that \( \mathbb{E}(X \mid Y) = \alpha + \mathbf{H}\phi(y) \) for \( \phi \in \mathbb{R}^d \). Let \( \rho_{\phi,f} \in \mathbb{R}^{d \times r} \) denote the matrix of cross correlations between the elements of \( \phi \) and \( f \). Then the estimators in Section 6.6.2 are \( \sqrt{n} \)-consistent if and only if \( \rho_{\phi,f} \) has rank \( d \) (Cook and Forzani, 2009, Thm. 3.5). As a consequence, there is considerable flexibility in the choice of \( f \) and some misspecification is not necessarily worrisome. As long as \( f \) is sufficiently correlated with the true function \( \phi \) we may expect useful results. We describe some possibilities for the choice of \( f \).

**Graphical choices.** Under model (6.24), the individual elements elements \( X_j \) of \( X \) each follow a univariate linear model with predictor \( f(y) \):

\[
X_j = \alpha_j + \beta_j^T f(y) + \varepsilon_j.
\]

Consequently, we are able to use inverse response plots (Cook, 1998, Ch. 10) of \( X_j \) versus \( Y, j = 1, \ldots, p \), to gain graphical information about suitable choices for \( f(y) \), which is an ability that is not generally available in the forward regression of \( Y \) on \( X \).

**Basis functions.** There are several generic possibilities for choice of \( f \), perhaps guided by graphics. Polynomials deriving from a Taylor approximation,

\[
f(y) = \{ y, y^2, y^3, \ldots, y^r \}^T,
\]

are one possibility. Periodic behavior could be modeled using a Fourier series form

\[
f(y) = \{ \cos(2\pi y), \sin(2\pi y), \ldots, \cos(2\pi ky), \sin(2\pi ky) \}^T
\]
as perhaps in signal processing applications. Here, \( k \) is a user-selected intergre and \( r = 2k \). Splines and other types of non-parametric constructions could also be used to form a suitable \( f \).

**Categorical response.** In some regressions there may be a natural choice for \( f(y) \). Suppose for instance that \( Y \) is categorical, taking values in one of \( h \) categories \( C_k, k = 1, \ldots, c \). We can then set \( r = c - 1 \) and specify the \( k \)-th element of \( f(y) \) to be \( J(y \in C_k) \), where \( J \) is the indicator function. Here there is no approximation in \( f(y) \).
It may help fix ideas to see how we arrive at an inverse model with a categorical response starting from the beginning. Let \( E(X|Y ∈ C_k) = \mu_k = \mu J(y) \), where the columns of the \( p \times h \) matrix \( \mu \) are \( \mu_k, k = 1, \ldots, c \), and \( J(y) \) is the \( c \times 1 \) indicator vector with elements \( J(y ∈ C_k), k = 1, \ldots, c \).

Centering the columns of \( \mu \) we have

\[
E(X|Y ∈ C_k) = (\mu - \bar{\mu} 1_c^T + \mu 1_c^T)J(y) = \bar{\mu} 1_c^T J(y) + (\mu - \bar{\mu} 1_c^T)J(y) = \bar{\mu} + (\mu - \bar{\mu} 1_c^T)J(y) = \bar{\mu} + H_h J(y),
\]

where \( 1_c^T J(y) = 1 \) and \( H \) is a semi-orthogonal basis matrix for \( \text{span}(\mu - \bar{\mu} 1_c^T) \). This model is not estimable because the elements of \( J \) always sum to 1. To remove this linear dependency we can just remove the last element of \( J \) and define

\[
f(y) = \{J(y ∈ C_1), \ldots, J(y ∈ C_{c-1})\}^T.
\]

(6.30)

**Sliced inverse regression.** Another option with continuous responses consists of “slicing” the observed values of \( Y \) into \( h \) bins (categories) \( C_k, k = 1, \ldots, c \), and then specifying the \( k \)-th coordinate of \( f(y) \) as for the case of a categorical \( Y \). This has the effect of approximating each conditional mean \( E(X_j | Y) \) as a step function of \( y \) with \( c \) steps,

\[
E(X_j | Y = y) ≈ \mu_j + \sum_{k=1}^{c-1} H_j^T h_k J(y ∈ C_k),
\]

where \( H_j^T \) is the \( j \)-th row of \( H \) and \( h_k \) is the \( k \)-th column of \( h \).

The estimator of the central subspace given in (6.28) is the same as that given by sliced inverse regression (Li 1991) when the response is categorical or slicing is used to categorize continuous response and (6.30) in conjunction with model (6.24) (Cook 2007). Additionally if the errors are normal and the response is categorical then sliced inverse regression gives the maximum likelihood estimator of the central subspace (Cook and Forzani 2009).