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.

September 13, 2013

Contents

Notation and Definitions

Matrices. For positive integers r and p, $\mathbb{R}^{r\times p}$ stands for the class of all matrices of dimension $r\times p$, and $\mathbb{S}^{r\times r}$ denotes the class of all symmetric $r\times r$ matrices. For $\mathbf{A}\in\mathbb{R}^{r\times p}$, $\mathbf{A}^{\dagger}$ indicates the Moore-Penrose inverse of $\mathbf{A}$. Vectors and matrices will typically be written in bold face, while scalars are not bold. The $r\times r$ identity matrix is denoted as $\mathbf{I}_r$.

For $\mathbf{M}\in\mathbb{S}^{r\times r}$ the notation $\mathbf{M}>0$ means that $\mathbf{M}$ is positive definite, $\mathbf{M}\geq 0$ means that $\mathbf{M}$ is positive semi-definite, $|\mathbf{M}|$ denotes the determinant of $\mathbf{M}$ and $|\mathbf{M}|_0$ denotes the product of the non-zero eigenvalues of $\mathbf{M}$. The spectral norm of $\mathbf{M}\in\mathbb{R}^{r\times p}$ is denoted as $\|\mathbf{M}\|$.

The *vector* operator $\mathrm{vec}:\mathbb{R}^{r\times p}\to\mathbb{R}^{rp}$ stacks the columns of the argument matrix. On the symmetric matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ we use the related *vector-half* operator $\mathrm{vech}:\mathbb{S}^{r\times r}\to\mathbb{R}^{r(r+1)/2}$, which stacks only the unique part of each column that lies on or below the diagonal. The operators vec and vech are related through a *contraction* matrix $\mathbf{C}_r\in\mathbb{R}^{r(r+1)/2\times r^2}$ and an *expansion* matrix $\mathbf{E}_r\in\mathbb{R}^{r^2\times r(r+1)/2}$, which are defined so that $\mathrm{vech}(\mathbf{A})=\mathbf{C}_r\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A})=\mathbf{E}_r\mathrm{vech}(\mathbf{A})$ for any $\mathbf{A}\in\mathbb{S}^{r\times r}$. These relations uniquely define $\mathbf{C}_r$ and $\mathbf{E}_r$, and imply $\mathbf{C}_r\mathbf{E}_r=\mathbf{I}_{r(r+1)/2}$. The $pr\times pr$ commutation matrix that maps $\mathrm{vec}(\mathbf{A})$ to $\mathrm{vec}(\mathbf{A}^T)$ is denoted by $\mathbf{K}_{rp}$: $\mathrm{vec}(\mathbf{A}^T)=\mathbf{K}_{rp}\mathrm{vec}(\mathbf{A})$. For further background on these operators, see Henderson and Searle (1979; *Canadian Journal of Statistics*) and Magnus and Neudecker (1979, *Annals of Statistics*; 1983, *Canadian Journal of Statistics*)

Let $\mathbf{A}=(a_{ij})\in\mathbb{R}^{r\times s}$ and $\mathbf{B}\in\mathbb{R}^{t\times u}$. The Kronecker product of two matrices $\otimes:\mathbb{R}^{r\times s}\times\mathbb{R}^{t\times u}\to\mathbb{R}^{rt\times su}$ can be defined block-wise as $\mathbf{A}\otimes\mathbf{B}=(a_{ij}\mathbf{B})$, $i=1\ldots r$, $j=1\ldots s$.

$\mathbf{1}_n$ denotes the $n\times 1$ vector of 1's.

We denote the eigenvalues of $\mathbf{A}\in\mathbb{S}^{p\times p}$ as $\varphi_1(\mathbf{A})\geq\ldots\geq\varphi_p(\mathbf{A})$ with corresponding ordered eigenvector $\boldsymbol{\ell}_1(\mathbf{A}),\ldots,\boldsymbol{\ell}_p(\mathbf{A})$. The arguments to φ_j and $\boldsymbol{\ell}_j$ may be supressed when they are expected to be clear from context.

Subspaces. For $\mathbf{A} \in \mathbb{R}^{p\times r}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^r$, $\mathbf{A}\mathcal{S} \equiv \{\mathbf{Ax} : \mathbf{x} \in \mathcal{S}\}$. For $\mathbf{B} \in \mathbb{R}^{r\times p}$, $\mathrm{span}(\mathbf{B})$ denotes the subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{B}$. We may occasionally use $\mathcal{S}_\mathbf{B}$ or $\mathcal{S}(\mathbf{B})$ as shorthand for $\mathrm{span}(\mathbf{B})$ when $\mathbf{B}$ has been defined. Subscripts $\mathcal{S}_{(\cdot)}$ will also be used to name commonly occurring subspaces. A *basis matrix* for a subspace $\mathcal{S}$ is any matrix whose columns form a basis for $\mathcal{S}$. A *semi-orthogonal matrix* $\mathbf{A} \in \mathbb{R}^{r\times p}$ has orthogonal columns, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$. We will frequently refer to semi-orthogonal basis matrices.

Let $\mathbf{A} \in \mathbb{S}^{r\times r}$ and $\mathbf{B} \in \mathbb{S}^{r\times r}$ with $\mathbf{A} > 0$. Then $\mathcal{S}_d(\mathbf{A}, \mathbf{B})$ equals $\mathbf{A}^{-1/2}$ times the span of the first $d \leq r$ eigenvectors of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. This subspace can also be described as the span of the first d eigenvectors of $\mathbf{B}$ relative to $\mathbf{A}$.

A sum of subspaces of $\mathbb{R}^r$ is indicated with the notation '$\oplus$': $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$. We use $\mathcal{S}_1 \subset \mathcal{S}_2$ to indicate that the subspace $\mathcal{S}_1$ is a proper subset of $\mathcal{S}_2$, while $\mathcal{S}_1 \subseteq \mathcal{S}_2$ allows $\mathcal{S}_1 = \mathcal{S}_2$.

For a positive definite matrix $\mathbf{\Sigma} \in \mathbb{S}^{r\times r}$, the inner product in $\mathbb{R}^r$ defined by $\langle \mathbf{x}_1, \mathbf{x}_2\rangle_\mathbf{\Sigma} = \mathbf{x}_1^\mathbf{T}\mathbf{\Sigma}\mathbf{x}_2$ is referred to as the $\mathbf{\Sigma}$ inner product; when $\mathbf{\Sigma} = \mathbf{I}_r$, the r by r identity matrix, this inner product is be called the usual inner product. A projection relative to the $\mathbf{\Sigma}$ inner product is the projection operator in the inner product space $\{\mathbb{R}^r, \langle\cdot,\cdot\rangle_\mathbf{\Sigma}\}$; that is, if $\mathbf{B} \in \mathbb{R}^{r\times p}$, then the projection onto $\mathrm{span}(\mathbf{B})$ relative to $\mathbf{\Sigma}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})} \equiv \mathbf{B}(\mathbf{B}^\mathbf{T}\mathbf{\Sigma}\mathbf{B})^\dagger\mathbf{B}^\mathbf{T}\mathbf{\Sigma}$. The projection onto the orthogonal complement of $\mathrm{span}(\mathbf{B})$ relative to the $\mathbf{\Sigma}$ inner product, $\mathbf{I}_r - \mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})}$, is denoted by $\mathbf{Q}_{\mathbf{B}(\mathbf{\Sigma})}$. Projection operators employing the usual inner product are written with a single subscript argument $\mathbf{P}_{(\cdot)}$, where the subscript describes the subspace, and $\mathbf{Q}_{(\cdot)} = \mathbf{I}_r - \mathbf{P}_{(\cdot)}$. The orthogonal complement $\mathcal{S}^\perp$ of a subspace $\mathcal{S}$ is constructed with respect to the usual inner product, unless indicated otherwise.

Let $\mathcal{S}_1 \subseteq \mathcal{S}$ be two nested subspaces of $\mathbb{R}^r$. We write $\mathcal{S} \setminus \mathcal{S}_1$ for the part of $\mathcal{S}$ that is orthogonal to $\mathcal{S}_1$. That is, $\mathcal{S} \setminus \mathcal{S}_1 = \mathrm{span}(\mathbf{P}_\mathcal{S} - \mathbf{P}_{\mathcal{S}_1})$.

The set of all u-dimensional subspaces of $\mathbb{R}^r$ is called a Grassmann manifold or Grassmanian and denoted as $\mathcal{G}(u, r)$. A Grassmanian is a compact topological manifold named for Hermann Grassmann (1809–1877).

Envelopes. Let $\mathbf{M} \in \mathbb{S}^{r\times r}$ and let $\mathcal{S} \subseteq \mathrm{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{S}$, to be written as $\mathcal{E}_\mathbf{M}(\mathcal{S})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{S}$ (cf. Definition 5.2). A variety of envelopes will be used in some chapters and to avoid proliferation of notation we may use a matrix as the argument of $\mathcal{E}_\mathbf{M}(\cdot)$: If $\mathbf{B} \in \mathbb{R}^{r\times d}$ and $\mathrm{span}(\mathbf{B}) = \mathcal{S}$

then $\mathcal{E}_{\mathbf{M}}(\mathbf{B}) = \mathcal{E}_{\mathbf{M}}(\text{span}(\mathbf{B})) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$.

Random vectors and their distributions. The notation $\mathbf{Y} \sim \mathbf{X}$ means that the random vectors $\mathbf{Y}$ and $\mathbf{X}$ have the same distribution, and $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}$ means that they are independent. The conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, or equivalently of $\mathbf{Y}|(\mathbf{X} = \mathbf{x})$, varies with the value $\mathbf{x}$ assumed by $\mathbf{X}$. It may occasionally be useful to refer to the distribution of $\mathbf{Y}|(\mathbf{X} = \mathbf{x})$ in terms of its distribution function $F(\mathbf{y}|\mathbf{X} = \mathbf{x})$. Since distribution function will always appear with their arguments, this notation should cause no difficulty. When there is no ambiguity, we will write $\mathbf{Y}|(\mathbf{X} = \mathbf{x})$ and $F(\mathbf{y}|\mathbf{X} = \mathbf{x})$ as simply $\mathbf{Y}|\mathbf{X}$ and $F(\mathbf{y}|\mathbf{X})$, understanding that $\mathbf{X}$ is fixed at some value. The notation $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{V}$ means that $\mathbf{Y}$ and $\mathbf{X}$ are independent given any value for the random vector $\mathbf{V}$.

We often use $\mathbf{\Sigma}_{\mathbf{V}} = \text{var}(\mathbf{V}) \in \mathbb{S}^{r\times r}$ to denote the covariance matrix of the random vector $\mathbf{V} \in \mathbb{R}^r$. Similarly, $\mathbf{\Sigma}_{\mathbf{V}|\mathbf{U}}$ denotes the covariance matrix of $\mathbf{V}$ given $\mathbf{U}$. Sample covariance matrices are denoted with an $\mathbf{S}_{(\cdot)}$ with the argument indicating the random vector. All sample covariances use the sample size as the denominator, unless indicated otherwise.

We write an asymptotic covariance matrix as $\text{avar}(\cdot)$; that is, if $\sqrt{n}(\mathbf{T} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{A})$, then $\text{avar}(\sqrt{n}\,\mathbf{T}) = \mathbf{A}$.

Common abbreviations

AIC Akaike information criterion

BIC Bayes information criterion

CS Central subspace

CMS Central mean subspace

DRS Dimension reduction subspace

LAD Likelihood acquired directions

LRT(α) Refers to the level-α sequential likelihood-ratio estimator of the dimension of an envelope.

MLE Maximum likelihood estimator

OLS Ordinary least squares

PCR Principal components regression

PFC Principal fitted components

PLS Partial least squares

SAVE Sliced average variance estimation

SDR Sufficient dimension reduction

SIR Sliced inverse regression

Chapter 1

Enveloping a multivariate mean

Envelopes, which were introduced by Cook, Li and Chiaromonte (2007) and developed for the multivariate linear model by Cook, Li and Chiaromonte (2010), encompass a class of methods for increasing efficiency in multivariate analyses without altering traditional objectives. Multivariate data can contain information that is material to the goals at hand and information that is immaterial to those goals. Envelopes operate by enveloping the material information and thereby accounting for the immaterial information. Essentially a type of targeted dimension reduction, envelopes can result in substantial gains in efficiency, particularly when the variation of the immaterial information (immaterial variation) is large relative to the variation of the material information (material variation).

Although relatively narrow in scope and applicability, this chapter is intended to provide first intuition in a relatively straightforward statistical context and to set the stage for Chapter 2, which covers envelopes for the coefficient matrix in a multivariate linear model. Some algebraic details in this chapter and the next are presented without justification. The missing development is largely given in Chapter 5, which covers the linear algebra of envelopes.

1.1 Envelope structure

Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots \mathbf{Y}_n$ be independent copies of the normal random vector $\mathbf{Y} \in \mathbb{R}^r$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} > 0$. The usual estimator of $\boldsymbol{\mu}$ is simply the sample mean $\bar{\mathbf{Y}} = n^{-1} \sum_{i=1}^n \mathbf{Y}_i$, which has variance $\mathrm{var}(\bar{\mathbf{Y}}) = n^{-1}\boldsymbol{\Sigma}$. An envelope estimator of $\boldsymbol{\mu}$ can have substantially smaller variance than $\bar{\mathbf{Y}}$ when $\mathbf{Y}$ contains information that is

immaterial to the $\boldsymbol{\mu}$. To develop this idea, let $\mathcal{S} \subseteq \mathbb{R}^r$ denote a subspace with the properties

$$\text{(a) } \boldsymbol{\mu} \in \mathcal{S} \text{ and (b) } \mathbf{P}_{\mathcal{S}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{Y}. \tag{1.1}$$

A subspace with these properties always exists since they hold trivially when $\mathcal{S} = \mathbb{R}^r$. Condition (a) implies that $\mathbf{P}_{\mathcal{S}}\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}})$ and that $\mathbf{Q}_{\mathcal{S}}\mathbf{Y} \sim N(0, \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}})$. Marginal information on $\boldsymbol{\mu}$ is available from $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$, while $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ supplies no marginal information about $\boldsymbol{\mu}$. However, under condition (a) alone $\mathbf{Q}_{\mathcal{S}}\mathbf{Y}$ could still carry information about $\boldsymbol{\mu}$ through an association with $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$. This possibility is ruled out by condition (b). Since $\mathbf{Y}$ is multivariate normal, condition (b) holds if and only if $\mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}} = 0$, which is equivalent to requiring that

$$\boldsymbol{\Sigma} = (\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}})\boldsymbol{\Sigma}(\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}) = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}}.$$

This version of condition (b) implies that $\mathcal{S}$ must be a reducing subspace of $\boldsymbol{\Sigma}$; that is, a subspace spanned by some subset of the eigenvectors of $\boldsymbol{\Sigma}$, without requiring $\boldsymbol{\Sigma}$ to have distinct eigenvalues (cf. Proposition 5.1). We can now state an equivalent structure by requiring the conditions

$$\text{(a) } \boldsymbol{\mu} \in \mathcal{S} \text{ and (b) } \mathcal{S} \text{ is a reducing subspace of } \boldsymbol{\Sigma}.$$

In short, $\mathcal{S}$ must be reducing subspace of $\boldsymbol{\Sigma}$ that contains $\boldsymbol{\mu}$. Conditions (a) and (b) are not useful by themselves but they become serviceable as a variance reduction tool when used together.

The intersection of all reducing subspaces of $\boldsymbol{\Sigma}$ that contain (envelop) $\boldsymbol{\mu}$ is itself a reducing subspace of $\boldsymbol{\Sigma}$ that contains $\boldsymbol{\mu}$ and is the envelope that we seek (see Lemma 5.4 and Definition 5.2). In later chapters we will be enveloping matrices and perhaps other constructs as well. For a general adaptable notation, let $\mathcal{M} = \text{span}(\boldsymbol{\mu})$. Then the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\mathcal{M}$, which equals the intersection of all reducing subspaces of $\boldsymbol{\Sigma}$ that contain $\mathcal{M}$, is called the $\boldsymbol{\Sigma}$-envelope of $\mathcal{M}$ and denoted as $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$. The abbreviated form $\mathcal{E}$ will be used in subscripts when the meaning is clear from context. Then $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ contains all of the material information on $\boldsymbol{\mu}$ with material variation $\mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}}$, and $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ contains the immaterial information with immaterial variation $\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$.

To gain intuition about envelopes, consider a single sample $\mathbf{Y}$ in which $\mathcal{S}$ and $\boldsymbol{\Sigma}$ are known and condition (a) holds, without necessarily requiring condition (b). If $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is to constitute material information then in this special case it seems reasonable to think there must be some connection with Fisher's fundamental concept of sufficiency. By definition

$\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is sufficient for $\boldsymbol{\mu}$ if the conditional distribution of $\mathbf{Y}|\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ does not depend on $\boldsymbol{\mu}$. Since $\mathbf{Y}|\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is normally distributed its distribution is characterized by its first two moments. Its variance does not depend on $\boldsymbol{\mu}$, but its mean can depend on $\boldsymbol{\mu}$:

$$\begin{aligned}
\mathrm{E}(\mathbf{Y}|\mathbf{P}_{\mathcal{E}}\mathbf{Y}) &= \boldsymbol{\mu} + \mathbf{P}^T_{\mathcal{S}(\boldsymbol{\Sigma})}(\mathbf{Y}-\boldsymbol{\mu}) \\
&= \mathbf{P}_{\mathcal{S}}\mathbf{Y} + \mathbf{Q}_{\mathcal{S}}\mathbf{P}^T_{\mathcal{S}(\boldsymbol{\Sigma})}(\mathbf{Y}-\boldsymbol{\mu}) \\
&= \mathbf{P}_{\mathcal{S}}\mathbf{Y} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\mathbf{Y}-\boldsymbol{\mu}) \\
&= \mathbf{P}_{\mathcal{S}}\mathbf{Y} + \boldsymbol{\Gamma}_0\{\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}\}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\mathbf{Y}-\boldsymbol{\mu}),
\end{aligned}$$

where $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal matrix. From this we see that $\mathrm{E}(\mathbf{Y}|\mathbf{P}_{\mathcal{E}}\mathbf{Y})$ is a constant function of $\boldsymbol{\mu}$ if and only if the second term on the right side is 0 for all $\boldsymbol{\mu}$, and this happens if and only if $\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma} = 0$, which leads us back to condition (b). That is, in this simple setting, $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is sufficient for $\boldsymbol{\mu}$ if and only if condition (b) holds. Our notion of material information is distinct from sufficiency because $\mathcal{S}$ is unknown and will eventually be estimated.

To gain intuition about the potential gain from an envelope analysis, suppose that the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ is known. Then the maximum likelihood estimator of $\boldsymbol{\mu}$ is just $\widehat{\boldsymbol{\mu}} = \mathbf{P}_{\mathcal{E}}\bar{\mathbf{Y}}$, which has variance $n^{-1}\mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}}$. Since $\boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$, we have straightforwardly

$$\mathrm{var}(\bar{\mathbf{Y}}) - \mathrm{var}(\widehat{\boldsymbol{\mu}}) = n^{-1}\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}},$$

so the difference between the variance of the standard and envelope estimators of $\boldsymbol{\mu}$ is exactly the immaterial variation. If $\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$ has eigenvalues that are large relative to the eigenvalues of the material variation $\mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}}$ then the envelope estimator will have substantially smaller variation than the standard estimator $\bar{\mathbf{Y}}$. On the other extreme, if $\mathbb{R}^r$ is the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\boldsymbol{\mu}$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M}) = \mathbb{R}^r$, then $\mathbf{Q}_{\mathcal{E}} = 0$ and the envelope estimator reduces to $\bar{\mathbf{Y}}$.

The estimation process with a known envelope is illustrated in Figure 1.1 for two responses. The ellipses on the plot are centered at the origin and represent the contours of $\boldsymbol{\Sigma}$, and the ellipse axes represent the two eigenspaces of $\boldsymbol{\Sigma}$. The population mean lies in the second eigenspace which equals the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ in this illustration. The envelope estimator of $\boldsymbol{\mu}$ is obtained by projecting $\bar{\mathbf{Y}}$ onto the envelope, as represented by the dashed line. The gain in precision comes about because the material variation in the envelope is substantially smaller than the immaterial variation along the orthogonal complement of the envelope, which corresponds to the first eigenspace.

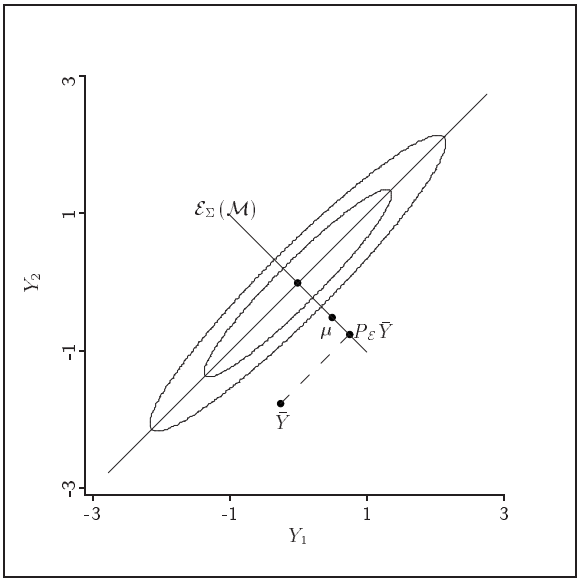

Figure 1.1: Graphical illustration of enveloping a population mean.

If $\boldsymbol{\mu}$ fell in the first eigenspace of $\boldsymbol{\Sigma}$, then the envelope would become the major axis of the ellipses in Figure 1.1. The envelope would still provide estimative gain in this case, although it would not be a great because the eigenvalue of $\mathbf{Q}_\mathcal{E}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{E}$ would now be the smaller eigenvalue of $\boldsymbol{\Sigma}$.

If $\boldsymbol{\mu}$ fell in neither eigenspace of Figure 1.1 then the envelope would be $\mathcal{E}_\boldsymbol{\Sigma}(\mathcal{M}) = \mathbb{R}^2$ and, strictly speaking, there would be no immaterial information. However, gains in mean squared error might still be realized. Still in the context of Figure 1.1, temporarily think of $\mathcal{E}_\boldsymbol{\Sigma}(\mathcal{M})$ as the eigenspace of $\boldsymbol{\Sigma}$ that is closest to $\boldsymbol{\mu}$ that does not necessarily fall into an eigenspace:

$$\begin{aligned} \mathrm{E}(\mathbf{P}_\mathcal{E}\bar{\mathbf{Y}} - \boldsymbol{\mu})(\mathbf{P}_\mathcal{E}\bar{\mathbf{Y}} - \boldsymbol{\mu})^T &= \mathrm{E}\{\mathbf{P}_\mathcal{E}(\bar{\mathbf{Y}} - \boldsymbol{\mu}) - \mathbf{Q}_\mathcal{E}\boldsymbol{\mu}\}\{\mathbf{P}_\mathcal{E}(\bar{\mathbf{Y}} - \boldsymbol{\mu}) - \mathbf{Q}_\mathcal{E}\boldsymbol{\mu}\}^T \\ &= n^{-1}\mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{E} + \mathbf{Q}_\mathcal{E}\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{Q}_\mathcal{E}. \end{aligned}$$

Comparing this mean squared error with $\mathrm{var}(\bar{\mathbf{Y}}) = n^{-1}\mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{E} + n^{-1}\mathbf{Q}_\mathcal{E}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{E}$, we see that there could still be substantial gains provided $\boldsymbol{\mu}$ is not too far from the closest eigenspace.

While the low dimensional representation of Figure 1.1 might leave the impression that the applicability of envelopes is limited because $\boldsymbol{\mu}$ might rarely fall in an eigenspace of $\boldsymbol{\Sigma}$, conditions (1.1) describe a plausible statistical context in which this can happen. Envelopes are generally more serviceable when $r > 2$ because empirically the propensity

for $\mathbf{Y}$ to contain immaterial information tends to increase with r. To illustrate one way in which this might arise, suppose that $\mathbf{Y}$ is a linear combination of $q < r$ latent variables $\mathbf{Z}$ plus an isotropic error, $\mathbf{Y} = \mathbf{A}\mathbf{Z} + \boldsymbol{\delta}$, where $\mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}})$, $\boldsymbol{\delta} \sim N(0, \sigma^2\mathbf{I}_r)$ and $\mathbf{A} \in \mathbb{R}^{r\times q}$ has full column rank. Without loss of generality, we take $\mathbf{A}$ to be a semi-orthogonal matrix. Then $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{Z}} \in \mathrm{span}(\mathbf{A})$ and

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{Z}}\mathbf{A}^T + \sigma^2\mathbf{I}_q \\ &= \mathbf{A}(\boldsymbol{\Sigma}_{\mathbf{Z}} + \sigma^2\mathbf{I}_q)\mathbf{A}^T + \sigma^2\mathbf{A}_0\mathbf{A}_0^T,\end{aligned}$$

where $(\mathbf{A}, \mathbf{A}_0)$ is an orthogonal matrix. Clearly, $\mathrm{span}(\mathbf{A})$ is a reducing subspace of $\boldsymbol{\Sigma}$ that contains $\boldsymbol{\mu}$, although it may not be the smallest. In this illustration then the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ is at most $q < r$.

1.2 Envelope model

The first step in estimating $\boldsymbol{\mu}$ under the envelope model is to formally incorporate a basis for the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$. Let $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M}))$, let $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ and let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be an orthogonal matrix so that $\boldsymbol{\Gamma}_0$ is a basis matrix for the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$. Since $\boldsymbol{\mu} \in \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ by construction, we can write $\boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ for some vector $\boldsymbol{\eta} \in \mathbb{R}^{u\times 1}$ that contains the coordinates of $\boldsymbol{\mu}$ relative to $\boldsymbol{\Gamma}$. To incorporate the basis $\boldsymbol{\Gamma}$ into the model, let $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma} > 0$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0 > 0$. Then

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}} \\ &= \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T \\ &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.\end{aligned}$$

The four parameter matrices $\boldsymbol{\eta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ live on a product space. The positive definite matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ can be viewed as matrices of coordinates for $\mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{E}}$ and $\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}}$ in much the same way as $\boldsymbol{\eta}$ is a vector of coordinates for $\boldsymbol{\mu}$. The envelope model can now be summarized as

$$\mathbf{Y} \sim N(\boldsymbol{\Gamma}\boldsymbol{\eta}, \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T). \tag{1.2}$$

Since $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal matrix, a $\boldsymbol{\Gamma}_0$ can be constructed from $\boldsymbol{\Gamma}$ and thus there are four parameter matrices to estimate, $\boldsymbol{\Gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. The basis $\boldsymbol{\Gamma}$ of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ is not identifiable in this model since mapping $\boldsymbol{\Gamma} \mapsto \boldsymbol{\Gamma}\mathbf{O}$ with an orthogonal matrix $\mathbf{O}$ leads

to an equivalent model. However, the envelope itself is identifiable and, as will be seen later, an estimator of it is all that is required to form the envelope estimator of $\boldsymbol{\mu}$. The parameter space for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$ is the set of all u dimensional subspaces of $\mathbb{R}^r$, which is called a Grassmann manifold or Grassmannian. It takes $u(r-u)$ real numbers to uniquely describe a u dimensional subspace of $\mathbb{R}^r$. With this we can describe the number of real parameters in these four matrices as

$$\begin{aligned} N(r,u) &= u(r-u)+u+u(u+1)/2+(r-u)(r-u+1)/2 \\ &= r(r+1)/2+u. \end{aligned} \tag{1.3}$$

The first count $u(r-u)$ is the number of reals needed to determine $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$, the second u is the dimension of $\boldsymbol{\eta} \in \mathbb{R}^u$ and the final two counts are for the positive definite matrices $\boldsymbol{\Omega} \in \mathbb{S}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{(r-u)\times(r-u)}$.

1.3 Estimation

1.3.1 Maximum likelihood estimation

Using the parameterization given in (1.2) and assuming temporarily that u is known, the log likelihood L_u for the multivariate normal can be represented, apart from the constant $-(nr/2)\log 2\pi$, as

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

where the third inequality comes from the relationships given in Corollary 5.1. Since $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y}$ the likelihood will factor accordingly. This can be done algebraically by

replacing $\mathbf{Y}$ with $\mathbf{P}_{\mathcal{E}}\mathbf{Y} + \mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ and similifying L_u:

$$\begin{aligned}
L_u(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &= -(n/2)\log|\boldsymbol{\Omega}| - (n/2)\log|\boldsymbol{\Omega}_0| \\
&\quad -(1/2)\sum_{i=1}^{n}(\mathbf{P}_{\mathcal{E}}\mathbf{Y}_i + \mathbf{Q}_{\mathcal{E}}\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\eta})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T) \\
&\quad \times(\mathbf{P}_{\mathcal{E}}\mathbf{Y}_i + \mathbf{Q}_{\mathcal{E}}\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\eta}) \\
&= -(n/2)\log|\boldsymbol{\Omega}| - (n/2)\log|\boldsymbol{\Omega}_0| \\
&\quad -(1/2)\sum_{i=1}^{n}\{\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta}) + \mathbf{Q}_{\mathcal{E}}\mathbf{Y}_i\}^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T) \\
&\quad \times\{\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta}) + \mathbf{Q}_{\mathcal{E}}\mathbf{Y}_i\} \\
&= -(n/2)\log|\boldsymbol{\Omega}| - (n/2)\log|\boldsymbol{\Omega}_0| \\
&\quad -(1/2)\sum_{i=1}^{n}\{(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta}) + \mathbf{Y}_i^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{Y}_i\} \\
&= -(n/2)\log|\boldsymbol{\Omega}| - (1/2)\sum_{i=1}^{n}(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\eta}) \\
&\quad -(n/2)\log|\boldsymbol{\Omega}_0| - (1/2)\sum_{i=1}^{n}\mathbf{Y}_i^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{Y}_i
\end{aligned}$$

Let $\mathbf{S}_{\mathbf{Y}} = n^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$ denote the sample covariance matrix of $\mathbf{Y}$ and let $\mathbf{T}_{\mathbf{Y}} = n^{-1}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^T$ denote the matrix of raw second moments of $\mathbf{Y}$. Then applying standard results in multivariate analysis, $L_u(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ is maximized for fixed $\boldsymbol{\Gamma}$ at $\boldsymbol{\eta} = \boldsymbol{\Gamma}^T\bar{\mathbf{Y}}$, $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\mathbf{T}_{\mathbf{Y}}\boldsymbol{\Gamma}_0$. Substituting these relationships into the log likelihood and simplifying leads to the partially maximized log likelihood

$$\begin{aligned}
L_u(\boldsymbol{\Gamma}) &= -(n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}| - (n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{T}_{\mathbf{Y}}\boldsymbol{\Gamma}_0| - nr/2 \\
&= -(n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}| - (n/2)\log|\boldsymbol{\Gamma}^T\mathbf{T}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}| - (n/2)\log|\mathbf{T}_{\mathbf{Y}}| - nr/2,
\end{aligned} \tag{1.4}$$

where the final step comes from Lemma 6.1. For a given dimension u, the maximum likelihood estimators can now be summarized as

$$\begin{aligned}
\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{M}) &= \operatorname{span}\{\arg\max L_u(\mathbf{\Gamma})\} \\
\widehat{\boldsymbol{\eta}} &= \widehat{\mathbf{\Gamma}}^T\bar{\mathbf{Y}} \\
\widehat{\mathbf{\Omega}} &= \widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}\widehat{\mathbf{\Gamma}} \\
\widehat{\mathbf{\Omega}}_0 &= \widehat{\mathbf{\Gamma}}_0^T\mathbf{T}_{\mathbf{Y}}\widehat{\mathbf{\Gamma}}_0 \\
\widehat{\boldsymbol{\mu}} &= \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\bar{\mathbf{Y}} \\
\widehat{\mathbf{\Sigma}} &= \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{\Gamma}}^T + \widehat{\mathbf{\Gamma}}_0\widehat{\mathbf{\Omega}}_0\widehat{\mathbf{\Gamma}}_0^T,
\end{aligned}$$

where the maximum is over the set of all semi-orthogonal matrices $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$. A $\widehat{\mathbf{\Gamma}}$ can then be taken as any semi-orthogonal basis for $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{M})$.

The partially maximized log likelihood $L_u(\mathbf{\Gamma})$ has the property that $L_u(\mathbf{\Gamma}) = L_u(\mathbf{\Gamma}\mathbf{O})$ for all orthogonal matrices $\mathbf{O} \in \mathbb{R}^{u\times u}$. Consequently, a value $\widehat{\mathbf{\Gamma}}$ of the semi-orthogonal matrix $\mathbf{\Gamma}$ that maximizes $L_u(\mathbf{\Gamma})$ is not unique, but $\operatorname{span}(\widehat{\mathbf{\Gamma}})$ is unique. Since $\widehat{\boldsymbol{\mu}} = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\bar{\mathbf{Y}}$ any value of $\mathbf{\Gamma}$ that maximizes $L_u(\mathbf{\Gamma})$ produces the same estimator of $\boldsymbol{\mu}$. Similarly, any value of $\mathbf{\Gamma}$ that maximizes $L_u(\mathbf{\Gamma})$ also gives the same estimate $\widehat{\mathbf{\Sigma}}$. On the other hand, the estimators $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Omega}}_0$ depend on the basis $\widehat{\mathbf{\Gamma}}$ selected. This may be of little consequence, because $\boldsymbol{\eta}$, $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ will typically be nuisance parameters, of minor interest in an analysis.

1.3.2 Asymptotic variance of $\widehat{\boldsymbol{\mu}}$

From standard likelihood theory, $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ is asymptotically normal with mean 0 and variance represented as $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\mu}})$, which can be constructed as the inverse of the Fisher information matrix. Following the derivations of Cook, Li and Chiaromonte (2010),

$$\begin{aligned}
\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\mu}}) &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_0)\mathbf{V}^\dagger(\boldsymbol{\eta} \otimes \mathbf{\Gamma}_0) && (1.5)\\
&\leq \operatorname{avar}(\sqrt{n}\bar{\mathbf{Y}}) && (1.6)\\
&= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T = \mathbf{\Sigma},
\end{aligned}$$

where $\mathbf{V} = \boldsymbol{\eta}\boldsymbol{\eta}^T \otimes \mathbf{\Omega}_0^{-1} + \mathbf{\Omega} \otimes \mathbf{\Omega}_0^{-1} + \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}_0 - 2\mathbf{I}_u \otimes \mathbf{I}_{r-u}$. Equation (1.5) gives the algebraic form of the asymptotic variance, while (1.6) says that it can never be greater than the asymptotic variance of the standard estimator $\bar{\mathbf{Y}}$. To see how (1.6) arises from (1.5), the last three addends of $\mathbf{V}$ have the same eigenvectors and thus a typical eigenvalue of their sum is $(\lambda/\lambda_0) + (\lambda_0/\lambda) - 2 \geq 0$, where λ and λ_0 are typical eigenvectors of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$. Consequently, $\mathbf{V} \geq \boldsymbol{\eta}\boldsymbol{\eta}^T \otimes \mathbf{\Omega}_0^{-1}$ and the conclusion then follows algebraically.

If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_r$ for $\sigma^2 > 0$, then $\boldsymbol{\Omega} = \sigma^2\mathbf{I}_u$, $\boldsymbol{\Omega}_0 = \sigma^2\mathbf{I}_{r-u}$, $\boldsymbol{\Gamma} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and $\boldsymbol{\eta} = \|\boldsymbol{\mu}\| \in \mathbb{R}^1$. Then $\mathbf{V} = \sigma^{-2}\|\boldsymbol{\mu}\|^2\mathbf{I}_{r-u}$, $\mathbf{V}^{-1} = \sigma^2\|\boldsymbol{\mu}\|^{-2}\mathbf{I}_{r-u}$ and $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\mu}}) = \sigma^2\mathbf{I}_r = \mathrm{avar}(\sqrt{n}\bar{\mathbf{Y}})$. Consequently, if $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ then standard estimator is asymptotically equivalent to the envelope estimator, and enveloping offers no gains.

The first term on the right hand side of (1.5) corresponds to the asymptotic variance of the envelope estimator when the envelope is known, $\mathrm{avar}(\sqrt{n}\mathbf{P}_{\mathcal{E}}\bar{\mathbf{Y}}) = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T$, as discussed previously in Section 1.1. Consequently, we can think of the second term as the cost of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$. The previous description of $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\mu}})$ when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_r$ describes an extreme case with maximal cost, so the asymptotic variance of $\widehat{\boldsymbol{\mu}}$ equals the that of $\bar{\mathbf{Y}}$.

1.3.3 Selecting $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M}))$

The development has so far assumed that u, which is essentially a model-selection parameter, is known. There several ways to guide the selection of u in practice, including cross validation, likelihood ratio testing and use of an information criterion, like AIC or BIC.

Using an information criterion, the envelope dimension u is selected as

$$\widehat{u} = \arg\min_u\{-2\widehat{L}_u + h(n)N(r,u)\},$$

where the minimum is over the set $\{0, 1, \ldots, r\}$, $\widehat{L}_u$ is the value of the fully maximized log likelihood function, $N(r,u)$ is the number of parameters in the model as given in (1.3), and $h(n) = 2$ for AIC and $h(n) = \log n$ for BIC.

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 $u = r$. The log likelihood ratio statistic for this hypothesis is $\Lambda(u_0) = 2(\widehat{L}_r - \widehat{L}_{u_0})$, where $\widehat{L}_{u_0} = L_{u_0}(\widehat{\boldsymbol{\Gamma}})$ is the fully maximized envelope log likelihood given by (1.4) evaluated at $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}$ and $\widehat{L}_r = \widehat{L}_r(\mathbf{I}_r)$ is the maximized log likelihood under the standard model, $\widehat{L}_r = -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}}|$, giving

$$\begin{aligned}\Lambda(u_0) &= n\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}| + n\log|\boldsymbol{\Gamma}^T\mathbf{T}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}| + n\log|\mathbf{T}_{\mathbf{Y}}| - n\log|\mathbf{S}_{\mathbf{Y}}| \quad (1.7)\\ &= n\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}_0| + n\log|\boldsymbol{\Gamma}_0^T\mathbf{T}_{\mathbf{Y}}\boldsymbol{\Gamma}_0|. \quad (1.8)\end{aligned}$$

Under the null hypothesis this statistic is distributed asymptotically as a chi-squared random variable with $N(r,r) - N(r,u_0) = r - u_0$ 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.

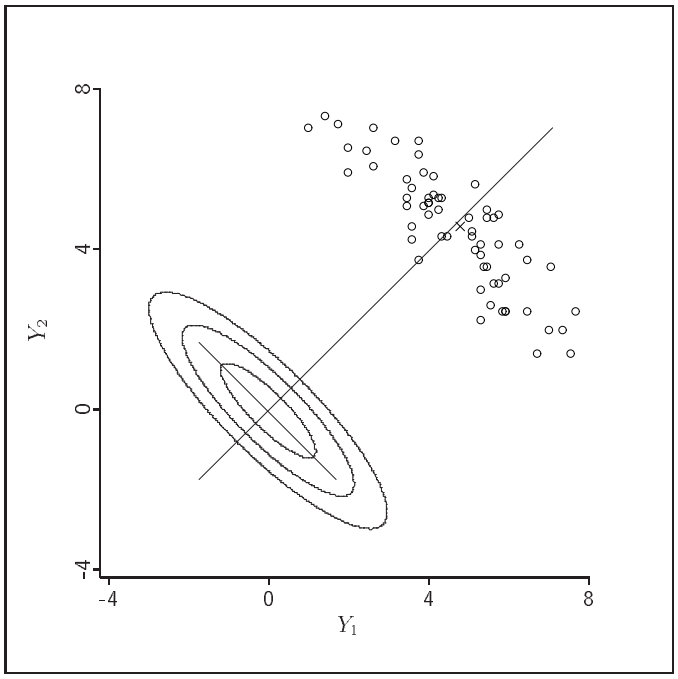

Figure 1.2: Minneapolis School data: Y_1 is the square root of the percentage of sixth graders scoring above average and Y_2 is the square root of the percentage of sixth graders scoring below average. The ellipses are contours of $\mathbf{S_Y}$ with its eigenvectors shown, and ex marks the mean of the data.

1.4 Minneapolis Schools

The Minneapolis school data[1] consists of observations on various characteristics of $n = 63$ elementary schools in Minneapolis, including the percentage of fourth and sixth grade students scoring above and below the national average on standardized reading tests. We use these data to illustrate various aspects of an envelope analysis. The percentages for each grade do not add to 100% because there is also an average performance classification which is not included in the analysis.

[1] A available from Cook, R. D. (1998). *Regression Graphics*. New York: Wiley.

1.4.1 Two transformed responses

We begin by considering $r = 2$ responses, the square root of the percentage of sixth grade students scoring above Y_1 and below Y_2 average, the square root transformation being used to bring the data closer to bivariate normality. Figure 1.2 given an indication of the potential for an envelope to improve efficiency. The ellipses, which are centered at the origin, are contours of $\mathbf{S_Y}$, and the lines extending from the origin mark the two eigenspaces of $\mathbf{S_Y}$. The mean of the data, which are plotted as circles, is marked by an ex in the center of the point cloud. From this representation we can see that the mean $\bar{\mathbf{Y}}$ falls quite close to the second eigenspace of $\mathbf{S_Y}$. Consequently, we can expect to find that $u = 1$, that the estimated variance of $\widehat{\boldsymbol{\mu}}$ from an envelope analysis is noticeably less than the estimated variance of $\bar{\mathbf{Y}}$, and that $\widehat{\boldsymbol{\mu}}$ is close to $\bar{\mathbf{Y}}$.

As expected, AIC and BIC agreed that $u = 1$. The results from an envelope analysis with $u = 1$ are summarized in Table 1.1. The estimates $\bar{\mathbf{Y}}$ and $\widehat{\boldsymbol{\mu}}$ are quite close, as anticipated from Figure 1.2. The standard error for $\widehat{\boldsymbol{\mu}}$ is $\widehat{\text{avar}}^{1/2}(\sqrt{n}\widehat{\boldsymbol{\mu}})/\sqrt{n}$, where the asymptotic variance was computed from (1.5) by plugging in the estimates of the unknown parameters. Table 1.1 shows that the envelope analysis reduced the standard error of $\bar{\mathbf{Y}}$ by about 30 percent. Around 120 observations would be needed in a standard analysis using $\bar{\mathbf{Y}}$ to achieve the standard error of $\widehat{\boldsymbol{\mu}}$ shown in Table 1.1, so the gain from using $\widehat{\boldsymbol{\mu}}$ is roughly equivalent to doubling the sample size for $\bar{\mathbf{Y}}$. The bootstrap standard errors shown in Table 1.1 are in good agreement with the standard errors obtained by using the plugin estimate of the asymptotic variance of $\widehat{\boldsymbol{\mu}}$.

The gains shown in Table 1.1 are indicated qualitatively by the relative magnitudes of the material variation summarized by $\|\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Gamma}}^T\| = \widehat{\boldsymbol{\Omega}} = 0.31$ and the immaterial variation summarized by $\|\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T\| = \widehat{\boldsymbol{\Omega}}_0 = 4.21$. Since the material variation is evidently notably less than the immaterial variation, we can expect gains from an envelope analysis. Of course, these summary statistics agree qualitatively with the representation in Figure 1.2.

The final column of Table 1.1 gives the estimated basis $\widehat{\boldsymbol{\Gamma}}$ for the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$. Since the elements of $\boldsymbol{\Gamma}$ are nearly the same, the material information is essentially contained in the average response, while the immaterial information is captured by the difference between the responses.

1.4.2 Four untransformed responses

The square root transformations in Section 1.4.1 were used to move the data closer to bivariate normality. While normalizing transformations can be useful depending on the

Table 1.1: Summary of an envelope analysis of the square root of the percent of sixth grade students scoring above Y_1 and below Y_2 average in the Minneapolis school data. "se" denotes a standard error, "boot se" indicates a bootstrap standard error based on 100 bootstrap samples, and "se ratio" is the ratio of the standard error for $\bar{\mathbf{Y}}$ to the standard error for $\widehat{\boldsymbol{\mu}}$.

Variable	$\bar{\mathbf{Y}}$	se of $\bar{\mathbf{Y}}$	$\widehat{\boldsymbol{\mu}}$	se of $\widehat{\boldsymbol{\mu}}$	boot se of $\widehat{\boldsymbol{\mu}}$	se ratio	$\widehat{\boldsymbol{\Gamma}}$
Y_1	4.67	0.191	4.61	0.134	0.135	1.42	0.711
Y_2	4.52	0.190	4.57	0.135	0.138	1.40	0.703

application, normality is not strictly required for an envelope analysis to produce gains over a standard analysis based on $\bar{\mathbf{Y}}$. The definition of an envelope requires first and second moments, but does not otherwise impose constraints on the distribution of $\mathbf{Y}$. If $\mathbf{Y}$ is not normal but has finite fourth moments then the normal-theory estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ constructed in Section 1.3.1 are still $\sqrt{n}$ consistent envelope estimators. In particular, $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ is asymptotically normal with mean 0 and finite variance. However, without normality $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\mu}})$ will depend on the fourth moments of $\mathbf{Y}$, and could be quite different from the normal theory variance shown in 1.5, but the bootstrap can still be used for standard errors.

To illustrate these comments we next apply the envelope method to the Minneapolis data using four untransformed responses, the percentages of fourth and sixth grade students scoring above and below average, which we designate as "4Above," "4Below," "6Above," "6Below." A scatterplot matrix of these four variables is shown in Figure 1.3. Deviations from normality seem clear although they do not appear to be dramatic.

Table 1.2: Estimates of u determined by using likelihood ratio tests sequentially for various test levels α on the Minneapolis school data with four untransformed responses.

α	0.4	0.3	0.25	0.1	0.05	0.005
$\widehat{u}$	3	3	2	2	1	1

The first step in an envelope analysis is to determine a value for the dimension u of the envelope. In contrast to the illustration in the previous section, there is now disagreement

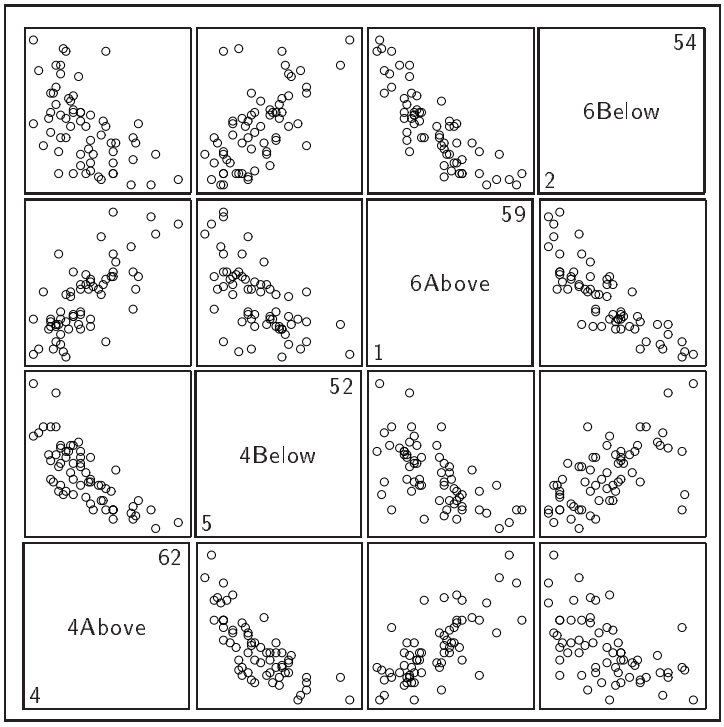

Figure 1.3: Minneapolis School data

between AIC, which yields the estimate $\widehat{u} = 3$, and BIC, which gives $\widehat{u} = 1$. For these data the results of an envelope analysis are sensitive to the choice of u. With $\widehat{u} = 1$ there are clear estimated gains in efficiency, while with $\widehat{u} = 3$ there are negligible gains, the results in this case being essentially identical to those based on $\bar{\mathbf{Y}}$. Theoretical results confirmed by simulations have shown that AIC tends to overestimate u, particularly when the data are non-normal, and that could be the case here. We turn to likelihood ratio tests to help decide the issue. Table 1.2 gives the estimate of u determined by using the sequential test procedure described in Section 1.3.3 for various test levels α. The results show that $\widehat{u} = 2$ or 3 for unusually large $\alpha \geq 0.25$, while $\widehat{u} = 1$ or 2 for the usual levels. This seems to support the notion that AIC overestimated the dimension. In some analyses a compromise value, here $\widehat{u} = 2$, might be reasonable, even if it is larger that the true dimension. Envelope models with $\widehat{u} > u$ are still valid, although they may incorporate some immaterial variation resulting in less efficiency than that possible with the true u.

The results of an envelope analysis with $u = 2$ are shown in Table 1.3. The standard error ratios are similar to those shown in Table 1.1 and, although the data deviate notice-

Table 1.3: Summary of an envelope analysis with $u = 2$ of the percentage of fourth and sixth grade students scoring above and below average in the Minneapolis school data. Column headings are as in Table 1.1.

$\mathbf{Y}$	$\bar{\mathbf{Y}}$	se of $\bar{\mathbf{Y}}$	$\widehat{\boldsymbol{\mu}}$	se† of $\widehat{\boldsymbol{\mu}}$	boot se of $\widehat{\boldsymbol{\mu}}$	se ratio
4Above	24.60	1.587	22.91	1.121	1.160	1.41
4below	22.19	1.273	23.60	0.960	0.929	1.32
6Above	24.05	1.756	21.86	1.096	1.146	1.60
6Below	22.71	1.706	24.85	1.192	1.296	1.42

† computed under normality.

ably from normality, the asymptotic standard errors are still reasonably accurate as judged against the results from 500 bootstrap samples.

Problems

Problem 1.1 *Is it possible in the population to have a setting where $u = 2$ for two responses and yet $u = 1$ after adding a third response?*

Problem 1.2 *Find a new data set where envelopes result in improved estimation of $\boldsymbol{\mu}$ and present the analysis.*

Problem 1.3 *Let $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Gamma}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})$, $\mathbf{Y} \in \mathbb{R}^r$ and $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{M})) < r$. Consider a full rank transformation $\mathbf{Y} \to \mathbf{AY}$. Under what class of transformation $\mathbf{A}$ does $\mathbf{AY}$ still follow an envelope model with an envelope of the same dimension? What does your conclusion say about the types of problems in which envelopes are likely to be effective?*

Problem 1.4 *Contrast Stein estimation of a normal population mean with envelope estimation. (open ended).*

Problem 1.5 *Repeat the analysis of the Minneapolis school data of Section 1.4.2 and attempt to reify the material and immaterial information.*

Problem 1.6 *Let $\bar{\mu} = \mathbf{1}^T\boldsymbol{\mu}/r$. Suppose that we are interested in the deviations $\boldsymbol{\alpha} = \boldsymbol{\mu} - \bar{\mu}\mathbf{1}$. We can represent this in terms of reparameterized normal model $\mathbf{Y} \sim N(\bar{\mu}\mathbf{1} + \boldsymbol{\alpha}, \boldsymbol{\Sigma})$,*

with $\mathbf{1}^T\boldsymbol{\alpha} = 0$. With the goal of estimating $\boldsymbol{\alpha}$, formulate an envelope version of this model based on $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{A})$, where $\mathcal{A} = \text{span}(\boldsymbol{\alpha})$. Derive the maximum likelihood estimator of $\boldsymbol{\alpha}$ under envelope model. How does your estimator differ from the estimator derived in this chapter?