Two sections eliminated by the author from "Constrained Principal Component Analysis (CPCA) and Related Techniques"

Yoshio Takane
April 12, 2014

We now show how (4.278), (4.279), and (4.280) can be derived (Yanai and Mukherjee 1987). Let $\mathbf{z}_j$ and $\mathbf{Z}_{(j)}$ denote the jth column vector of $\mathbf{Z}$ and the matrix obtained by eliminating $\mathbf{z}_j$ from $\mathbf{Z}$, respectively. The matrix $\mathbf{P}_{(j)}$ denotes the orthogonal projector defined by $\mathbf{Z}_{(j)}$, and $\mathbf{Q}_{(j)}$ denotes its orthogonal complement. Then, since $\mathrm{Sp}(\mathbf{Z}_{(j)}) \subset \mathrm{Sp}(\mathbf{Z})$,

$$\mathbf{P}_Z \mathbf{P}_{(j)} = \mathbf{P}_{(j)}, \tag{4.286}$$

and

$$\mathbf{P}_Z \mathbf{Q}_{(j)} \mathbf{z}_j = (\mathbf{P}_Z - \mathbf{P}_{(j)}) \mathbf{z}_j = \mathbf{Q}_{(j)} \mathbf{z}_j \tag{4.287}$$

for $j = 1, \cdots, m$. Note that $\mathbf{P}_Z \mathbf{z}_j = \mathbf{z}_j$. These results imply that

$$\mathbf{Z}_I = \mathbf{P}_Z \mathbf{Z}_I, \tag{4.288}$$

and

$$\mathbf{Z}_A = \mathbf{P}_Z \mathbf{Z}_A. \tag{4.289}$$

We now show

$$\mathbf{Z}' \mathbf{Z}_A = \mathbf{D} = \mathrm{diag}(d_1, d_2, \cdots, d_m), \tag{4.290}$$

where $\mathbf{D}$ is diagonal with $d_j = 1 - r^2_{j(j)}$ as the jth diagonal element, and $r^2_{j(j)}$ is the squared multiple correlation coefficient in predicting $\mathbf{z}_j$ from $\mathbf{Z}_{(j)}$. Note that

$$\mathbf{z}'_j \mathbf{z}_{Aj} = \mathbf{z}'_j \mathbf{Q}_{(j)} \mathbf{z}_j = 1 - \mathbf{z}'_j \mathbf{P}_{(j)} \mathbf{z}_j = 1 - r^2_{j(j)} = d_j,$$

and that for $i \neq j$,

$$\mathbf{z}'_j \mathbf{z}_{Ai} = \mathbf{z}_j \mathbf{Q}_{(i)} \mathbf{z}_i = \mathbf{z}'_j \mathbf{z}_i - \mathbf{z}'_j \mathbf{P}_{(i)} \mathbf{z}_i = 0.$$

Then, from (4.288), (4.289), and (4.290), we obtain

$$\mathbf{Z}_A = \mathbf{P}_Z \mathbf{Z}_A = \mathbf{Z} \mathbf{R}^{-1} \mathbf{D}, \tag{4.291}$$

and

$$\mathbf{Z}_I = \mathbf{P}_Z \mathbf{Z}_I = \mathbf{P}_Z (\mathbf{Z} - \mathbf{Z}_A) = \mathbf{Z} - \mathbf{Z} \mathbf{R}^{-1} \mathbf{D}, \tag{4.292}$$

which are identical to (4.279) and (4.278). To show (4.280), we note that $\mathbf{z}'_j \mathbf{z}_{Aj} = \mathbf{z}'_{Aj} \mathbf{z}_{Aj}$, which implies

$$\mathbf{D} = \mathrm{diag}(\mathbf{Z}' \mathbf{Z}_A) = \mathrm{diag}(\mathbf{Z}'_A \mathbf{Z}_A) = \mathrm{diag}(\mathbf{D} \mathbf{R}^{-1} \mathbf{D}) = \mathbf{D} \mathrm{diag}(\mathbf{R}^{-1}) \mathbf{D}, \tag{4.293}$$

or $\mathbf{D} = (\mathrm{diag}(\mathbf{R}^{-1}))^{-1}$. Yanai and Mukherjee (1987) extend the above results to a singular $\mathbf{R}$. See also Tucker et al. (1972) and Kaiser (1976).

4.21 Restricted Maximum Likelihood (REML)

Let

$$\mathbf{z} = \mathbf{G}\mathbf{b} + \mathbf{e} \tag{4.294}$$

denote a linear regression model, where $\mathbf{z}$ is a vector of observations on the criterion variable, $\mathbf{G}$ is the fixed-effect model matrix with the vector of regression coefficients

b, and **e** the vector of disturbance terms such that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. The model of this form often arises in the guise of a linear mixed-effect model (LMM; e.g., McCulloch and Searle 2001; Pinheiro and Bates 2000), in which it is further assumed that $\mathbf{e} = \mathbf{C}\gamma + \mathbf{e}^*$, where **C** is the model matrix for the random coefficient vector γ, and $\mathbf{e}^*$ the vector of independently distributed disturbance terms such that $\gamma \sim \mathcal{N}(\mathbf{0}, \Lambda)$, $\mathbf{e}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and $\mathrm{Cov}(\gamma, \mathbf{e}) = \mathbf{O}$. Then,

$$\mathbf{z} \sim \mathcal{N}(\mathbf{Gb}, \Sigma), \tag{4.295}$$

where

$$\Sigma = \mathbf{C}\Lambda\mathbf{C}' + \sigma^2\mathbf{I}. \tag{4.296}$$

A main goal of REML is the estimation of Σ under a special structural assumption on Σ like (4.296). However, our interest here is primarily theoretical; we would like to explain some interesting relationship between REML and what we know about linear estimation. So we do not assume any special structure for Σ. Rather we assume that Σ is known. Let

$$\hat{\mathbf{b}}_{BLUE} = (\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}\mathbf{G}'\Sigma^{-1}\mathbf{z} \tag{4.297}$$

be the GLSE of **b**, which is also the BLUE under the present setup.

The following decomposition of $\mathrm{SS}(\mathbf{z} - \mathbf{Gb})_{\Sigma^{-1}}$ holds:

$$\begin{aligned}
&\mathrm{SS}(\mathbf{z} - \mathbf{Gb})_{\Sigma^{-1}} \\
&\quad = \mathrm{SS}(\mathbf{G}\hat{\mathbf{b}}_{BLUE} - \mathbf{Gb})_{\Sigma^{-1}} + \mathrm{SS}(\mathbf{z} - \mathbf{G}\hat{\mathbf{b}}_{BLUE})_{\Sigma^{-1}} \\
&\quad = \mathrm{SS}(\hat{\mathbf{b}}_{BLUE} - \mathbf{b})_{G'\Sigma^{-1}G} + \mathrm{SS}(\mathbf{z} - \mathbf{P}_{G/\Sigma^{-1}}\mathbf{z})_{\Sigma^{-1}} \\
&\quad = \mathrm{SS}(\hat{\mathbf{b}}_{BLUE} - \mathbf{b})_{G'\Sigma^{-1}G} + \mathrm{SS}(\mathbf{z})_{\Sigma^{-1}Q_{G/\Sigma^{-1}}}.
\end{aligned} \tag{4.298}$$

We show that the second term in (4.298) is equal to the SS term in the REML likelihood (restricted or residual likelihood; e.g., LaMotte (2007)). Note first that

$$\Sigma^{-1}\mathbf{Q}_{G/\Sigma^{-1}} = \Sigma^{-1} - \Sigma^{-1}\mathbf{G}(\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}\mathbf{G}'\Sigma^{-1}. \tag{4.299}$$

Let **Y** denote a matrix such that $\mathrm{Sp}(\mathbf{Y}) = \mathrm{Ker}(\mathbf{G}')$. Then, by Khatri's lemma (Section 2.2.8), we have

$$\Sigma^{-1} - \Sigma^{-1}\mathbf{G}(\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}\mathbf{G}'\Sigma^{-1} = \mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}', \tag{4.300}$$

so the second term in (4.298) can be rewritten as

$$\mathrm{SS}(\mathbf{z})_{\Sigma^{-1}Q_{G/\Sigma^{-1}}} = \mathrm{SS}(\mathbf{z})_{Y(Y'\Sigma Y)^{-1}Y'} = \mathrm{SS}(\mathbf{Y}'\mathbf{z})_{(Y'\Sigma Y)^{-1}}. \tag{4.301}$$

This represents the SS of **z** in the space orthogonal to **G** (the residual space), and is equal to the SS term in the REML likelihood. It represents the SS of $\mathbf{Y}'\mathbf{z}$ in the metric of the inverse of its variance-covariance matrix ($\mathrm{Var}[\mathbf{Y}'\mathbf{z}] = \mathbf{Y}'\Sigma\mathbf{Y}$).

Now consider another transformation of **z**, namely $\mathbf{X}'\mathbf{z}$, where **X** is any matrix such that $\mathbf{X}'\mathbf{G} = \mathbf{I}$. An example of such **X** is $\mathbf{X} = \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}$, which we choose to use in what follows. This means that $\hat{\mathbf{b}}_{OLSE} = \mathbf{X}'\mathbf{z}$ is the OLSE of **b**. We show that

the first term in (4.298) is equal to the SS in the conditional likelihood of $\mathbf{X}'\mathbf{z}$ given $\mathbf{Y}'\mathbf{z}$ (Verbyla 1990). Note that the joint distribution of $\mathbf{X}'\mathbf{z}$ and $\mathbf{Y}'\mathbf{z}$ is given by

$$\begin{pmatrix} \mathbf{X}'\mathbf{z} \\ \mathbf{Y}'\mathbf{z} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \mathbf{X}'\Sigma\mathbf{X} & \mathbf{X}'\Sigma\mathbf{Y} \\ \mathbf{Y}'\Sigma\mathbf{X} & \mathbf{Y}'\Sigma\mathbf{Y} \end{bmatrix} \right). \tag{4.302}$$

Then the conditional distribution of $\mathbf{X}'\mathbf{z}$ given $\mathbf{Y}'\mathbf{z}$ is a multivariate normal distribution with the expected value of

$$\mathrm{Ex}[\mathbf{X}'\mathbf{z}|\mathbf{Y}'\mathbf{z}] = \mathbf{b} + \mathbf{X}'\Sigma\mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{z}, \tag{4.303}$$

and the variance-covariance matrix

$$\mathrm{Var}[\mathbf{X}'\mathbf{z}|\mathbf{Y}'\mathbf{z}] = \mathbf{X}'(\Sigma - \Sigma\mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}'\Sigma)\mathbf{X}) = (\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}. \tag{4.304}$$

See Note 4.7. The second equality in (4.304) holds due to Khatri's lemma. Note also that $\mathbf{X}'\mathbf{G} = \mathbf{I}$.

The SS term in the conditional distribution of $\mathbf{X}'\mathbf{z}$ given $\mathbf{Y}'\mathbf{z}$ is given by

$$\begin{aligned} &\mathrm{SS}(\mathbf{X}'\mathbf{z} - \mathbf{b} - \mathbf{X}'\Sigma\mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{z})_{G'\Sigma^{-1}G} \\ &\quad = \mathrm{SS}(\mathbf{X}'(\mathbf{I} - \Sigma\mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{z} - \mathbf{b})_{G'\Sigma^{-1}G} \\ &\quad = \mathrm{SS}(\mathbf{X}'\mathbf{G}(\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}\mathbf{G}'\Sigma^{-1}\mathbf{z} - \mathbf{b})_{G'\Sigma^{-1}G} \\ &\quad = \mathrm{SS}(\hat{\mathbf{b}}_{BLUE} - \mathbf{b})_{G'\Sigma^{-1}G}, \end{aligned} \tag{4.305}$$

which is equal to the first term in (4.298). The second equality in (4.305) holds due to Khatri's lemma. Again, note that $\mathbf{X}'\mathbf{G} = \mathbf{I}$.

Note 4.11 The use of the conditional distribution to modify $\hat{\mathbf{b}}_{OLSE} = \mathbf{X}'\mathbf{z}$ into $\hat{\mathbf{b}}_{BLUE}$ given by (4.305), is called a covariance adjustment (Rao 1967). It is well known (Rao 1967, Lemma 2c) that

$$\mathrm{Var}[\hat{\mathbf{b}}_{OLSE}] = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\Sigma\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} \geq (\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1} = \mathrm{Var}[\hat{\mathbf{b}}_{BLUE}]. \tag{4.306}$$

This relation indicates that in general $\hat{\mathbf{b}}_{BLUE}$ is a better estimator of $\mathbf{b}$ than $\hat{\mathbf{b}}_{OLSE}$ because it has smaller variances. The equality in the middle holds if and only if $\Sigma = \mathbf{G}\Lambda_1\mathbf{G}' + \mathbf{Y}\Lambda_2\mathbf{Y}' + \sigma^2\mathbf{I}$ for arbitrary *nnd* matrices Λ_1 and Λ_2 (Rao 1967). (Note that under this condition $\mathbf{X}'\Sigma\mathbf{Y} = (\mathbf{Y}'\Sigma\mathbf{X})' = \mathbf{O}$ in (4.302), the second term in (4.303) is $\mathbf{0}$, $\mathbf{X}'\Sigma\mathbf{X} = (\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}$ in (4.304), the conditional distribution of $\mathbf{X}'\mathbf{z}$ given $\mathbf{Y}'\mathbf{z}$ is equal to the marginal distribution of $\mathbf{X}'\mathbf{z}$, and that the first SS in (4.305) is simply $\mathrm{SS}(\hat{\mathbf{b}}_{OLSE} - \mathbf{b})_{G'\Sigma^{-1}G}$, which is equal to $\mathrm{SS}(\hat{\mathbf{b}}_{BLUE} - \mathbf{b})_{G'\Sigma^{-1}G}$.) Under this condition, the OLSE is identical to the BLUE. Note, however, that (4.306) holds only if Σ is known, as assumed here. That is, (4.306) may not hold if Σ has to be estimated.

Note 4.12 One final point to make is if the determinant of Σ also factors out in a manner conformable to the partition of SS, namely,

$$|\Sigma| = c|(\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}||\mathbf{Y}'\Sigma\mathbf{Y}| \tag{4.307}$$

for some c. The value of c, although it is not essential, can be determined as follows. Since $[\mathbf{G}, \mathbf{Y}]$ is square and nonsingular (LaMotte 2007), we have

$$\begin{aligned}\det\left(\begin{bmatrix}\mathbf{G}'\\ \mathbf{Y}'\end{bmatrix}\Sigma[\mathbf{G},\mathbf{Y}]\right) &= \det\left(\begin{bmatrix}\mathbf{G}'\\ \mathbf{Y}'\end{bmatrix}\right)|\Sigma|\det([\mathbf{G},\mathbf{Y}])\\ &= |\Sigma|\det\left(\begin{bmatrix}\mathbf{G}'\mathbf{G} & \mathbf{O}\\ \mathbf{O} & \mathbf{Y}'\mathbf{Y}\end{bmatrix}\right)\\ &= |\Sigma||\mathbf{G}'\mathbf{G}||\mathbf{Y}'\mathbf{Y}|. \end{aligned} \tag{4.308}$$

(Note that $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| = |\mathbf{B}||\mathbf{A}|$.) We also have

$$\begin{aligned}\det\left(\begin{bmatrix}\mathbf{G}'\\ \mathbf{Y}'\end{bmatrix}\Sigma[\mathbf{G},\mathbf{Y}]\right) &= \det\left(\begin{bmatrix}\mathbf{G}'\Sigma\mathbf{G} & \mathbf{G}'\Sigma\mathbf{Y}\\ \mathbf{Y}'\Sigma\mathbf{G} & \mathbf{Y}'\Sigma\mathbf{Y}\end{bmatrix}\right)\\ &= |\mathbf{Y}'\Sigma\mathbf{Y}||\mathbf{G}'(\Sigma-\Sigma\mathbf{Y}(\mathbf{Y}'\Sigma\mathbf{Y})^{-1}\mathbf{Y}'\Sigma)\mathbf{G}|\\ &= |\mathbf{Y}'\Sigma\mathbf{Y}||\mathbf{G}'\mathbf{G}(\mathbf{G}'\Sigma^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{G}|\\ &= |\mathbf{Y}'\Sigma\mathbf{Y}||\mathbf{G}'\mathbf{G}|^2|(\mathbf{G}'\mathbf{G})^{-1}|, \end{aligned} \tag{4.309}$$

so that $c = |\mathbf{G}'\mathbf{G}||(\mathbf{Y}'\mathbf{Y})^{-1}|$. (Note that $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$. Note also that the third equality in (4.309) holds due to Khatri's lemma.)

4.22 Partial Least Squares (PLS)

Partial least squares (PLS) was first introduced by Wold (1966) as an algorithm to solve linear LS problems. Since then, it has been elaborated in many directions. In this section we discuss some interesting properties of PLS regression (both univariate and multivariate). We also briefly mention recent developments in PLS path analysis, a topic closely related to Sections 6.7 and 6.8. See Rosipal and Krämer (2006) for more comprehensive reviews of the methodology.

Consider a linear regression model

$$\mathbf{z} = \mathbf{G}\mathbf{b} + \mathbf{e}, \tag{4.310}$$

where, as before, $\mathbf{z}$ is the n-component vector of observations on the criterion variable, $\mathbf{G}$ the $n \times p$ matrix of predictor variables, $\mathbf{b}$ the p-component vector of regression coefficients, and $\mathbf{e}$ the n-component vector of disturbance terms. The ordinary LS (OLS) is often employed to estimate $\mathbf{b}$ under the *iid* normal assumption on $\mathbf{e}$. This is fine if n is reasonably large compared to p, and columns of $\mathbf{G}$ are not highly collinear. However, if this condition is not satisfied, the use of OLSE is not recommended. The OLSE tends to have large variances in such cases. Principal component regression (PCR) is often employed in such situations. In PCR, PCA is first applied to $\mathbf{G}$ to find its low rank (say, rank s) approximation, which is subsequently used as the new predictor variables in the regression. One potential problem with this method is that the low rank approximation of $\mathbf{G}$ aims to account for $\mathbf{G}$, and not necessarily to predict $\mathbf{z}$. In contrast, PLS extracts components of $\mathbf{G}$ that are good predictors of $\mathbf{z}$.

For the case of univariate regression, the PLS algorithm (called PLS1) proceeds

as follows:

PLS1 Algorithm

Step 1. Columnwise center $\mathbf{G}$ and $\mathbf{z}$, and set $\mathbf{G}_0 = \mathbf{G}$.
Step 2. Repeat the following substeps for $i = 1, \cdots, s$ ($s \leq \text{rank}(\mathbf{G})$):

Step 2.1. Set $\mathbf{w}_i = \mathbf{G}'_{i-1}\mathbf{z}/\|\mathbf{G}'_{i-1}\mathbf{z}\|$.
Step 2.2. Set $\mathbf{t}_i = \mathbf{G}_{i-1}\mathbf{w}_i/\|\mathbf{G}_{i-1}\mathbf{w}_i\|$.
Step 2.3. Set $\mathbf{v}_i = \mathbf{G}'_{i-1}\mathbf{t}_i$.
Step 2.4. Set $\mathbf{G}_i = \mathbf{G}_{i-1} - \mathbf{t}_i\mathbf{v}'_i$ (deflation).

Vectors $\mathbf{w}_i$, $\mathbf{t}_i$, and $\mathbf{v}_i$ are called weights, scores, and loadings, respectively, and are collected in matrices $\mathbf{W}_s$, $\mathbf{T}_s$, and $\mathbf{V}_s$. For a given s, the PLS estimator (PLSE) of $\mathbf{b}$ is given by

$$\hat{\mathbf{b}}^{(s)}_{PLSE} = \mathbf{W}_s(\mathbf{V}'_s\mathbf{W}_s)^{-1}\mathbf{T}'_s\mathbf{z}. \tag{4.311}$$

The algorithm above assumes that s is known. The choice of its value is very important for the performance of PLSE. A cross validation method is often used to choose the best value of s (see Section 5.3). It has been demonstrated that for the same value of s, the PLSE of $\mathbf{b}$ has better predictability than the corresponding PCR estimator.

Note 4.13 The PLSE of $\mathbf{b}$ can be considered a special kind of constrained LS estimator (Section 2.2.9), in which $\mathbf{b}$ is constrained to lie in the Krylov subspace of dimensionality s defined by $\mathcal{K}_s(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) = \text{Sp}(\mathbf{K}_s)$, where $\mathbf{K}_s = [\mathbf{G}'\mathbf{z}, (\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{z}, \cdots, (\mathbf{G}'\mathbf{G})^{s-1}\mathbf{G}'\mathbf{z}])$ is called the Krylov matrix. Since $\text{Sp}(\mathbf{W}_s) = \mathcal{K}_s(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$, $\mathbf{b}$ can be reparameterized as $\mathbf{b} = \mathbf{W}_s\mathbf{a}$ for some $\mathbf{a}$. Then (4.310) can be rewritten as

$$\mathbf{z} = \mathbf{G}\mathbf{W}_s\mathbf{a} + \mathbf{e}. \tag{4.312}$$

The OLSE of $\mathbf{a}$ is given by $\hat{\mathbf{a}} = (\mathbf{W}'_s\mathbf{G}'\mathbf{G}\mathbf{W}_s)^{-1}\mathbf{W}'_s\mathbf{G}'\mathbf{z}$, from which the constrained LS estimate of $\mathbf{b}$ is found by

$$\hat{\mathbf{b}}^{(s)}_{CLSE} = \mathbf{W}_s\hat{\mathbf{a}} = \mathbf{W}_s(\mathbf{W}'_s\mathbf{G}'\mathbf{G}\mathbf{W}_s)^{-1}\mathbf{W}'_s\mathbf{G}'\mathbf{z}. \tag{4.313}$$

To show that (4.313) is indeed equivalent to (4.311), we need several well known results in the PLS literature (e.g., Bro and Eldén 2009; de Jong 1993; Eldén 2004; Phatak and de Hoog 2002). First of all, $\mathbf{W}_s$ is columnwise orthogonal, that is,

$$\mathbf{W}'_s\mathbf{W}_s = \mathbf{I}_s. \tag{4.314}$$

In fact, $\mathbf{W}_s$ is identical to the matrix of orthogonal basis vectors generated by the Arnoldi orthogonalization of $\mathbf{K}_s$. This orthogonalization method finds, starting from $\mathbf{w}_1 = \mathbf{G}'\mathbf{z}/\|\mathbf{G}'\mathbf{z}\|$, $\mathbf{w}_{i+1}$ ($i = 1, \cdots, s-1$) by successively orthogonalizing $\mathbf{G}'\mathbf{G}\mathbf{w}_i$ ($i = 1, \cdots, s-1$) to all previous $\mathbf{w}_i$'s by a procedure similar to the Schmidt orthogonalization method (Section 2.1.11), yielding $\mathbf{W}_s$ such that $\mathbf{G}'\mathbf{G}\mathbf{W}_s = \mathbf{W}_s\mathbf{H}_s$, or

$$\mathbf{W}'_s\mathbf{G}'\mathbf{G}\mathbf{W}_s = \mathbf{H}_s, \tag{4.315}$$

where $\mathbf{H}_s$ is tridiagonal. Secondly, $\mathbf{T}_s$ is also columnwise orthogonal, i.e.,

$$\mathbf{T}'_s\mathbf{T}_s = \mathbf{I}_s, \tag{4.316}$$

and

$$\mathbf{T}_s\mathbf{L}_s = \mathbf{G}\mathbf{W}_s, \tag{4.317}$$

where $\mathbf{L}_s$ is the Lanczos upper bidiagonal matrix (Bro and Eldén 2009). Relations (4.315) through (4.317) imply that

$$\mathbf{W}_s'\mathbf{G}'\mathbf{G}\mathbf{W}_s = \mathbf{L}_s'\mathbf{L}_s = \mathbf{H}_s, \tag{4.318}$$

where $\mathbf{H}_s$ is tridiagonal as noted above. Thirdly,

$$\mathbf{V}_s' = \mathbf{T}_s'\mathbf{G}, \tag{4.319}$$

so that

$$\mathbf{L}_s = \mathbf{T}_s'\mathbf{G}\mathbf{W}_s = \mathbf{V}_s'\mathbf{W}_s. \tag{4.320}$$

Now it is straightforward to show that

$$\begin{aligned}\hat{\mathbf{b}}_{CLSE}^{(s)} &= \mathbf{W}_s\mathbf{H}_s^{-1}\mathbf{L}_s'\mathbf{T}_s'\mathbf{z} \\ &= \mathbf{W}_s(\mathbf{L}_s'\mathbf{L}_s)^{-1}\mathbf{L}_s'\mathbf{T}_s'\mathbf{z} \\ &= \mathbf{W}_s\mathbf{L}_s^{-1}\mathbf{T}_s'\mathbf{z} \\ &= \mathbf{W}_s(\mathbf{V}_s'\mathbf{W}_s)^{-1}\mathbf{T}_s'\mathbf{z} = \hat{\mathbf{b}}_{PLSE}^{(s)}.\end{aligned} \tag{4.321}$$

Furthermore, let $\mathbf{D}_s = [\mathbf{d}_0, \cdots, \mathbf{d}_{s-1}]$ denote the set of the first s direction vectors obtained by the conjugate gradient method to solve the normal equation $\mathbf{G}'\mathbf{G}\mathbf{b} = \mathbf{G}'\mathbf{z}$. Then

$$\hat{\mathbf{b}}_{CG}^{(s)} = \mathbf{D}_s(\mathbf{D}_s'\mathbf{G}'\mathbf{G}\mathbf{D}_s)^{-1}\mathbf{D}_s'\mathbf{G}'\mathbf{z} \tag{4.322}$$

is also identical to $\hat{\mathbf{b}}_{PLSE}^{(s)}$ (Phatak and de Hoog 2002). The equivalence is clear from $\mathbf{D}_s = \mathbf{W}_s\mathbf{A}$ for some square nonsingular matrix $\mathbf{A}$. (The $\mathbf{D}_s$ and $\mathbf{W}_s$ span the same Krylov subspace $\mathcal{K}_s(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$.) One notable fact about $\hat{\mathbf{b}}_{CG}^{(s)}$ is that $\mathbf{D}_s'\mathbf{G}'\mathbf{G}\mathbf{D}_s$ is diagonal, and hence easily invertible, due to the conjugacy of $\mathbf{D}_s$ with respect to $\mathbf{G}'\mathbf{G}$. The PLSE of regression parameters reduces to the OLSE if $s = \text{rank}(\mathbf{G})$.

We now turn to the multivariate case (PLS2), for which PLS algorithms are not as well established as in the univariate case. There are two representative algorithms. One is the original algorithm (e.g., Abdi 2007) called the NIPALS algorithm, and the other is the SIMPLS algorithm (de Jong 1993). The two algorithms give slightly different results for $s > 1$. In what follows, $\mathbf{Z}$ denotes the $n \times m$ matrix of criterion variables.

PLS2 (NIPALS) Algorithm

Step 1. Set $\mathbf{G}_0 = \mathbf{G}$.

Step 2. For $i = 1, \cdots, s$, obtain $\mathbf{w}_i$, $\mathbf{t}_i$, $\mathbf{c}_i$, and $\mathbf{u}_i$ by iterating the following substeps until convergence:

Step 2.1. Set $\mathbf{w}_i = \mathbf{G}_{i-1}'\mathbf{u}_i / \|\mathbf{G}_{i-1}'\mathbf{u}_i\|$. (An initial value of $\mathbf{u}_i$ may be generated randomly.)

Step 2.2. Set $\mathbf{t}_i^* = \mathbf{G}_{i-1}\mathbf{w}_i$.

Step 2.3. Set $\mathbf{c}_i = \mathbf{Z}'\mathbf{t}_i^* / \mathbf{t}_i^{*'}\mathbf{t}_i^*$.

Step 2.4. Set $\mathbf{u}_i = \mathbf{Z}\mathbf{c}_i / \mathbf{c}_i'\mathbf{c}_i$.

Step 3. If $i < s$, set $\mathbf{G}_i = \mathbf{G}_{i-1} - \mathbf{t}_i^*\mathbf{v}_i^{*'}$ (deflation), where $\mathbf{v}_i^* = \mathbf{G}'\mathbf{t}_i^* / \mathbf{t}_i^{*'}\mathbf{t}_i^*$, and go back

to Step 2. Otherwise, stop.

The vectors $\mathbf{w}_i$, $\mathbf{t}_i^*$, $\mathbf{v}_i^*$, and $\mathbf{c}_i$ ($i = 1, \cdots, r$) are collected in matrices $\mathbf{W}_s$, $\mathbf{T}_s^*$, $\mathbf{V}_s^*$, and $\mathbf{C}_s$. Then the PLS2 (NIPALS) estimate of the matrix of regression coefficients, $\hat{\mathbf{B}}_{PLSE}^{(s)}$, is given by

$$\hat{\mathbf{B}}_{NIPALS}^{(s)} = \mathbf{W}_s(\mathbf{V}_s^{*'}\mathbf{W}_s)^{-1}\mathbf{C}_s'. \tag{4.323}$$

Note that this formula is essentially the same as (4.311), which can be seen from $\mathbf{T}_s = \mathbf{T}_s^*(\mathbf{T}_s^{*'}\mathbf{T}_s^*)^{-1/2}$, $\mathbf{V}_s^* = (\mathbf{T}_s^{*'}\mathbf{T}_s^*)^{-1}\mathbf{T}_s^{*'}\mathbf{G}$, and $\mathbf{C}_s' = (\mathbf{T}_s^{*'}\mathbf{T}_s^*)^{-1}\mathbf{T}_s^{*'}\mathbf{Z}$. Note also that at convergence Step 2 obtains the left singular vector $\mathbf{w}_i$ of $\mathbf{G}_{i-1}'\mathbf{Z}$ corresponding to the largest singular value, which is also equal to the first eigenvector of $\mathbf{G}_{i-1}'\mathbf{Z}\mathbf{Z}'\mathbf{G}_{i-1}$. Lindgren et al. (1993) proposed a more efficient algorithm than the one given above. This algorithm bypasses the estimation of $\mathbf{t}_i^*$ and $\mathbf{u}_i$, and deflates $\mathbf{G}'\mathbf{Z}$ and $\mathbf{G}'\mathbf{G}$ (rather than $\mathbf{G}$).

The SIMPLS algorithm (de Jong 1993) also partially uses a trick similar to Lindgren et al. (1993) to speed up the algorithm:

PLS2 (SIMPLS) Algorithm

Step 1. Set $\mathbf{S} = \mathbf{G}'\mathbf{Z}$.
Step 2. For $i = 1, \cdots, s$, repeat the following substeps:

Step 2.1. When $i = 1$, obtain the SVD of $\mathbf{S}$, or when $i > 1$, obtain the SVD of $(\mathbf{I} - \tilde{\mathbf{V}}_{i-1}(\tilde{\mathbf{V}}_{i-1}'\tilde{\mathbf{V}}_{i-1})^{-1}\tilde{\mathbf{V}}_{i-1}')\mathbf{S}$.

Step 2.2. Set $\mathbf{r}_i$ equal to the first left singular vector from Step 2.1 above.

Step 2.3. Set $\tilde{\mathbf{t}}_i = \mathbf{G}\mathbf{r}_i$.

Step 2.4. Set $\tilde{\mathbf{v}}_i = \mathbf{G}'\tilde{\mathbf{t}}_i / \tilde{\mathbf{t}}_i'\tilde{\mathbf{t}}_i$.

Step 2.5. Append $\mathbf{r}_i$, $\tilde{\mathbf{t}}_i$, and $\tilde{\mathbf{v}}_i$ to $\mathbf{R}_{i-1}$, $\tilde{\mathbf{T}}_{i-1}$, and $\tilde{\mathbf{V}}_{i-1}$ to obtain $\mathbf{R}_i$, $\tilde{\mathbf{T}}_i$, and $\tilde{\mathbf{V}}_i$. (Assume that $\mathbf{R}_0$, $\tilde{\mathbf{T}}_0$, and $\tilde{\mathbf{V}}_0$ are null matrices.)

The SIMPLS estimate of $\mathbf{B}$ is then obtained by

$$\hat{\mathbf{B}}_{SIMPLS}^{(s)} = \mathbf{R}_s(\tilde{\mathbf{T}}_s'\tilde{\mathbf{T}}_s)^{-1}\tilde{\mathbf{T}}_s'\mathbf{Z}. \tag{4.324}$$

Again this formula is similar (but not identical) to (4.323) and (4.311). This may be seen from $\tilde{\mathbf{T}}_s \approx \mathbf{T}_s^*$, $\tilde{\mathbf{V}}_s \approx \mathbf{V}_s^*$, and $\mathbf{R}_s = \mathbf{W}_s(\mathbf{V}_s^{*'}\mathbf{W}_s)^{-1} \in \{(\mathbf{V}_s^{*'})^-\}$. Both NIPALS and SIMPLS algorithms reduce to PLS1 when there is only one criterion variable. When $s = 1$, both methods give identical results, which can also be obtained via SVD($\mathbf{G}'\mathbf{Z}$). McIntosh and his collaborators (McIntosh and Lobaugh 2004; McIntosh et al. 1996, 2004) proposed a procedure called spatiotemporal PLS for analysis of brain imaging data (e.g., fMRI data). This method calculates SVD($\mathbf{G}'\mathbf{Z}$), where $\mathbf{G}$ is the matrix of orthonormalized contrast vectors (i.e., $\mathbf{G}'\mathbf{G} = \mathbf{I}$), assuming that $s = 1$. This technique is clearly a special case of CPCA.

Note 4.14 Apart from PLS and $s = 1$, the SVD of $\mathbf{G}'\mathbf{Z}$ could be interesting in its own right. In factor analysis, there is a method of factor extraction called interbattery factor analysis (Tucker 1958). This method calculates SVD($\mathbf{G}'\mathbf{Z}$) to avoid estimation of communality in factor analysis of $[\mathbf{G}, \mathbf{Z}]$. Co-inertia analysis (Dolédec and Chessel 1994; Dray et al. 2003) obtains GSVD of a product of several matrices under suitable metric matrices. This method reduces to

SVD($\mathbf{G}'\mathbf{Z}$) in the special case of two data sets and identity metric matrices. See also Section 2.3.8 on product SVD.

There is a version of PLS called latent variables path analysis (LVPA; Lohmöller 1989; Wold 1982) to relate multiple sets of variables. It is a kind of structural equation model (SEM) similar to ERA and GSCA in Sections 6.7 and 6.8. Let $\mathbf{Z}_j$ ($j = 1, \cdots, J$) represent the jth block of variables, and define $\mathbf{S}_{jk} = \mathbf{Z}_j'\mathbf{Z}_k$ ($j, k = 1, \cdots, J$) to be the covariance matrix between the jth and kth blocks. Let $\mathbf{w}_j$ denote the vector of weights applied to $\mathbf{Z}_j$ to derive the latent variable (component) for the jth block, and let g_{jk} represent a variable such that $g_{jk} = 0$ if the variable sets j and k are unrelated, $g_{jk} = 1$ if they are positively related, and $g_{jk} = -1$ if they are negatively related. There are two representative algorithms for LVPA, one called Mode A, and the other called Mode B. The Mode A algorithm updates $\mathbf{w}_j$ ($j = 1, \cdots, J$) by

$$a_j \mathbf{w}_j = \sum_{k \neq j} g_{jk} \mathbf{S}_{jk} \mathbf{w}_k, \tag{4.325}$$

where a_j is a normalization factor to make $\mathbf{w}_j'\mathbf{w}_j = 1$. The Mode B algorithm, on the other hand, updates $\mathbf{w}_j$ by

$$a_j \mathbf{w}_j = \sum_{k \neq j} g_{jk} \mathbf{S}_{jj}^{-1} \mathbf{S}_{jk} \mathbf{w}_k, \tag{4.326}$$

where a_j is the normalization factor to make $\mathbf{w}_j'\mathbf{S}_{jj}\mathbf{w}_j = 1$. When $g_{jk} = 1$ for all j and $k \neq j$, The Mode B algorithm reduces to SUMCOR (Horst 1961; see also Kettenring 1971) for multiple-set canonical correlation analysis similar to GCANO (Section 4.10). However, SUMCOR does not reduce to SVD unlike GCANO. PLS in general has no explicit global criteria for optimization. McDonald (1996, Appendix) has shown, however, that under the same condition as above (i.e., $g_{jk} = 1$ for all j and $k \neq j$) both Mode A and Mode B algorithms can be derived from well-defined optimization criteria, assuming that only one LV (component) is extracted in each block of variables. More recently, Tenenhaus and Tenenhaus (2011) have proposed a very comprehensive algorithm which subsumes both modes as special cases.

Bibliography

Abdi, H. 2007. Partial least squares regression. 2007. In *Encyclopedia of measurement and statistics*, ed. N. J. Salkind, 740–54. Thousand Oaks, CA: Sage. []

Bro, R. and L. Eldén. 2009. PLS works. *Journal of Chemometrics* 23:69–71.

de Jong, S. 1993. SIMPLS: An alternative approach to partial least squares regression. *Journal of Chemometrics* 18:251–263.

Dolédec, S. and D. Chessel. 1994. Co-inertia analysis: An alternative method for studying species-environment relationships. *Freshwater Biology* 31:277–99.

Dray, S., D. Chessel, and J. Thioulouse. 2003. Co-inertia analysis and the linking ecological data tables. *Ecology* 84:3078–89.

Eldén, L. 2004. Partial least-squares vs Lanczos bidiagonalization – I: Analysis of a projection method for multiple regression. *Computational Statistics and Data Analysis* 46:11–31.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Clinical Psychology Special Monographs Supplement* 14:331–47.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika* 58:433–51.

LaMotte, L. R. 2007. A direct derivation of the REML likelihood function. *Statistical Papers* 48:321–7.

Lindgren, F., P. Geladi, and S. Wold. 1993. The kernel algorithm for PLS. *Journal of Chemometrics* 7:45–59.

Lohmöller, J. B. 1989. *Latent variables path modeling with partial least squares*. Heidelberg: Physica-Verlag.

McCulloch, C. E. and S. R. Searle. 2001. *Generalized, linear, and mixed models*. New York: Wiley.

McDonald, R. P. 1978. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 31:59–72.

McIntosh, A. R., W. K. Chau, and A. B. Protzner. 2004. Spatiotemporal analysis of event-related fMRI data using partial least squares. *NeuroImage* 23:764–75.

McIntosh, A. R. and N. J. Lobaugh. 2004. Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage* 23:S250–63.

McIntosh, A. R., F. L.Bookstein, J. V. Haxby, and C. L. Grady. 1996. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* 3:143–157.

Phatak, A. and F. de Hoog. 2002. Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics* 16:361–3367.

[1] Numbers in square brackets indicate page numbers on which the article is referenced.

Pinheiro, J. C. and D. M. Bates. 2000. *Mixed effects models in S and Splus*. New York: Springer.

Rao, C. R. 1967. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, ed. L. M. Le Cam and J. Neyman, 335–72. Berkeley: University of California Press.

Rosipal, R. and N. Krämer. 2006. Overview and recent advances in partial least squares. In *SLSFS 2005, LNCS 3940*, eds. C. Saunders et al., 34–51, Berlin: Springer.

Tenenhaus, A. and M. Tenenhaus, M. 2011. Regularized generalized canonical correlation analysis. *Psychometrika* 76:257–84.

Tucker, L. R. 1959. Intra-individual and inter-individual multidimensionality. In *Psychological Scaling*, ed. H. Gullicksen and S. Messick, 155–67. New York: Wiley.

Verbyla, A. P. 1990. A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* 32:227–30.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. In *Multivariate analysis*, (ed.) P. R. Krishnaiah, 391–420, New York: Academic Press.

Wold, H. 1982. Soft modeling: The basic design and some extensions. In *Systems under indirect observations, Part 2*, eds. K. G. Jöreskog and H. Wold, 1–54, Amsterdam: North-Holland.