Principal Components Analysis and Extensions

Xiaogang Su, Ph.D.
Department of Mathematical Sciences
University of Texas at El Paso (UTEP)
xsu@utep.edu

March 2, 2018

Contents

Principal components analysis (PCA) is a classic topic and has various applications in dimension reduction, regression, data visualization, clustering, and so on. In this short note, I will provide a

brief coverage of methodological essentials.

1 PCA

PCA can be motivated and understood in several different (yet related) ways: 1) The first is to (linearly) transform correlated variables into a set of uncorrelated ones. Having uncorrelated predictors is useful in dealing with multicollinearity in regression. 2) The second is to seek mutually orthogonal directions (linear combinations) along which data show most variation; 3) The third is to find a lower-dimensional approximation to the original data matrix; 4) Factor analysis is yet another route leading to PCA via latent variables. PCA is also closely related to matrix approximation and distanced-based low-dimensional representation of data (referred to multidimensional scaling or MDS). It is interesting to note that different approaches have been taken in different fields to arrive at PCA; as a result, it may be referred to differently under several alternative names. Yet another important reason that PCA is first introduced here is to expose the singular value decomposition (SVD) of matrices. SVD will be used in the computation of PCA and is an indispensable tool for several other unsupervised and supervised methods in the ensuing topics.

The first approach of introducing PCA is the simplest one to understand, though we often take the second route in statistical convention. The idea, again, is to extract new orthogonal features or variables that preserve as much variation or information as possible by forming (normalized) linear combinations of original variables. Both approaches 2 and 3 involves optimization. In particular, approach 3 naturally leads to sparse PCA, for which more details will be presented in Section 3.

1.1 Transformation to Uncorrelated Components

Given a random vector $\mathbf{x}$ with mean vector $\boldsymbol{\mu}_x$ and variance-covariance matrix $\boldsymbol{\Sigma}_x$, or simply denoted as $\mathbf{x} \sim (\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, where WLOG we assume $\boldsymbol{\mu}_x = \mathbf{0}$ and hence $\boldsymbol{\Sigma}_x = \mathrm{E}(\mathbf{x}\mathbf{x}^T) \succ 0$, we want to seek a transformation matrix $\mathbf{V}'$ such that the linearly transformed random vector $\mathbf{x}' = \mathbf{V}'^T\mathbf{x}$ has uncorrelated components. Namely, $\mathrm{cov}(\mathbf{V}'^T\mathbf{x}) = \mathbf{V}'^T\boldsymbol{\Sigma}_x\mathbf{V}'$ is diagonal. This transform is often termed as the Karhunen-Loeve transform.

Since matrix $\boldsymbol{\Sigma}_x$ is symmetric and positive definite, it must have an eigen-decomposition. Denote it as

$$\boldsymbol{\Sigma}_x = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathbf{T},$$

where matrix $\mathbf{V}$ is orthogonal or unitary such that $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_j)$ is diagonal with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

We can see that setting $\mathbf{V}' = \mathbf{V}$ yields the solution since

$$\mathrm{cov}(\mathbf{V}^T\mathbf{x}) = \mathbf{V}^T\boldsymbol{\Sigma}_x\mathbf{V} = \mathbf{V}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{V} = \boldsymbol{\Lambda}.$$

To gain more insight, rewrite $\mathbf{x}' = \left(x'_1, \ldots, x'_p\right)^T$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$ so that $\mathbf{v}_j$ is the j-th column of $\mathbf{V}$. Then it follows that

$$x'_j = \mathbf{v}_j^T\mathbf{x} = \langle \mathbf{v}_j, \mathbf{x} \rangle \quad \text{and} \quad \mathbf{x} = \mathbf{v}\mathbf{x}' = \sum_{j=1}^{p} x'_j\mathbf{v}_j.$$

Clearly, this is a special case of the general (inner product) Hilbert space results. Furthermore, $\mathrm{var}(x'_j) = \mathbf{v}_j^T\boldsymbol{\Sigma}_x\mathbf{v}_j = \lambda_j\mathbf{v}_j^T\mathbf{v}_j = \lambda_j$. In other words, x'_1 has the largest variance λ_1; x'_2 has the second largest variance λ_2 and so on.

In addition, define

$$\mathbf{x}'_m = \sum_{j=1}^{m} x'_j \mathbf{V}_j = \sum_{j=1}^{m} \langle \mathbf{V}_j, \mathbf{x} \rangle \, \mathbf{V}_j,$$

for some $m \leq p$, which is the projection of $\mathbf{x}$ on the first m bases $\{\mathbf{V}_j : j = 1, \ldots, m\}$. Random vector $\mathbf{x}'_m$ can be viewed as an approximation of $\mathbf{x}$. Then, in terms of the approximation quality, it can be easily seen (via $\mathbf{V}_j^T \mathbf{V}_{j'} = 1$ for $j = j'$ and 0 otherwise) that the mean square error (MSE)

$$\mathrm{E}\|\mathbf{x} - \mathbf{x}'_m\|^2 = \sum_{j=m+1}^{p} \lambda_j.$$

It can be further established that this is the minimum MSE among any approximation of $\mathbf{x}$ by any (random) vector from the subspace spanned by $\{\mathbf{V}_j : j = 1, \ldots, m\}$.

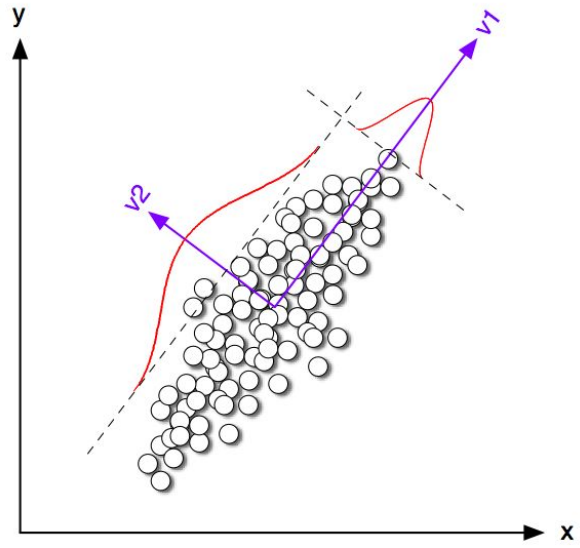

Figure 1: Illustration of PCA in the 2-D scenario. Note that the straight line that the first PC form corresponds to the total LS line that minimized the total perpendicular distances from each point to the straight line, which can be contrasted with the LS fitted line.

Figure 1 provides an illustration of PCA in the two-dimensional case. Clearly, for the decorrelation purpose only, the Karhunen-Loeve transform is not unique; there are many other transform that can do the same job. However, the Karhunen-Loeve transform based on $\mathbf{V}$ will be made unique through an optimization statement (maximizing explained variance) in the next section.

1.2 Explaining Variations in Data

Why do we want to seek mutually orthogonal directions along which data show most variation? Gentle (2009) explains it nicely. First of all, the information in data is presented as variation. That is why many statistical methods such as analysis of variance (ANOVA) are design to analyze the variation. Secondly, correlation among variables reduces the amount of information that the

variables contain. Thus, PCA is aimed to seek transforms of variables that are uncorrelated to each other and at the same time explain the maximum variation in the original variables.

1.2.1 The Theoretical Version

Suppose that random vector $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that no distributional assumptions is assumed. Here is the problem statement for finding the first PC of $\mathbf{x}$. We want a vector $\mathbf{v}$ s.t. $\|\mathbf{v}\| = 1$ that the linear combination $c = \mathbf{v}^T\mathbf{x} \in \mathbb{R}$ maximizes $\text{var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$. In other words, the solution $\mathbf{v}^\star$ solves

$$\mathbf{v}^\star = \arg\max_{\mathbf{v}} \left(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}\right), \quad \text{subject to } \mathbf{v}^T\mathbf{v} = 1. \tag{1}$$

Using the technique of Lagrange (or undetermined) multiplier, consider the Lagrangian function

$$\phi(\mathbf{v}; \alpha) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \alpha\left(\mathbf{v}^T\mathbf{v} - 1\right), \tag{2}$$

where α is the Lagrangian multiplier and the minus sign could be made positive alternatively. Then

$$\frac{\partial\phi}{\partial\mathbf{v}} = 2\boldsymbol{\Sigma}\mathbf{v} - 2\alpha\mathbf{v} \overset{\text{set}}{\equiv} \mathbf{0},$$

which amounts to

$$\boldsymbol{\Sigma}\mathbf{v} = \alpha\mathbf{v}$$

or the characteristic (polynomial) equation of $|\boldsymbol{\Sigma} - \alpha\mathbf{I}| = 0$. This is because $\|\mathbf{v}\| = 1$ or $\mathbf{v} \neq \mathbf{0}$, yet $(\boldsymbol{\Sigma} - \alpha\mathbf{I})\mathbf{v} = \mathbf{0}$, hence $\boldsymbol{\Sigma} - \alpha\mathbf{I}$ must be singular.

Furthermore,

$$\text{var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T(\boldsymbol{\Sigma}\mathbf{v}) = \mathbf{v}^T(\alpha\mathbf{v}) = \alpha.$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ denote the eigenvalues of $\boldsymbol{\Sigma}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$ denote the corresponding normalized eigenvectors. We must have $\mathbf{v}^\star = \mathbf{v}_1$ with $\text{var}(\mathbf{v}_1^T\mathbf{x}) = \lambda_1$. The scalar quantity $c_1 = \mathbf{v}_1^T\mathbf{x}$ is the first principal component, denote it as PC1.

Next, we want vector $\mathbf{v}$ s.t. $\|\mathbf{v}\| = 1$ and $\mathbf{v}^T\mathbf{v}_1 = 0$ that maximizes $\text{var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$. Note that the added constraint $\boxed{\mathbf{v}^T\mathbf{v}_1 = 0}$ can be geometrically interpreted as orthogonality (or being perpendicular to each other) of $\mathbf{v}$ and $\mathbf{v}_1$. Statistically, this means that

$$\text{cov}(\mathbf{v}^T\mathbf{x}, \mathbf{v}_1^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}_1 = \mathbf{v}^T\lambda_1\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}^T\mathbf{v}_1 = 0,$$

i.e., $\mathbf{v}^T\mathbf{x}$ is uncorrelated with PC1 $\mathbf{v}_1^T\mathbf{x}$. Clearly, the constraint $\mathbf{v}^T\mathbf{v}_1 = 0$ is equivalent to $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}_1 = 0$.

The associated Lagrangian becomes

$$\phi(\mathbf{v}; \alpha_0, \alpha_1) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \alpha_0\left(\mathbf{v}^T\mathbf{v} - 1\right) - 2\alpha_1\left(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}_1\right), \tag{3}$$

where the constant 2 really does not matter here as α_0 and α_1 have constraints and is added just for the convenience of simplification. Its partial derivatives with respect to $\mathbf{v}$ are

$$\frac{\partial\phi}{\partial\mathbf{v}} = 2\boldsymbol{\Sigma}\mathbf{v} - 2\alpha_0\mathbf{v} - 2\,\alpha_1 \cdot \boldsymbol{\Sigma}\mathbf{v}_1 \overset{\text{set}}{\equiv} \mathbf{0}. \tag{4}$$

Multiplying $\mathbf{v}_1$ to the left side on both sides of (4) yields

$$0 \equiv= 2\mathbf{v}_1^T\boldsymbol{\Sigma}\mathbf{v} - 2\alpha_0\mathbf{v}_1^T\mathbf{v} - 2\,\alpha_1 \cdot \mathbf{v}_1^T\boldsymbol{\Sigma}\mathbf{v}_1 = -2\,\alpha_1 \cdot \mathbf{v}_1^T\boldsymbol{\Sigma}\mathbf{v}_1 = -2\alpha_1 \cdot \lambda_1,$$

which implies that $\alpha_1 = 0$. Bringing this back into (4) leads to $\boldsymbol{\Sigma}\mathbf{v} - \alpha_0\mathbf{v} = 0$. In other words, $\mathbf{v}^\star$ must be an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue $\mathbf{v}$. Plus the observations that $\text{var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} = \mathbf{v}^T\alpha_0\mathbf{v} = \alpha_0$ and $\mathbf{v}^T\mathbf{v}_1 = 0$, we must have the second PC direction is given by $\mathbf{v}^\star = \mathbf{v}_2$ with $\text{var}(\mathbf{v}_2^T\mathbf{x}) = \lambda_2$. Denote the corresponding second principal component, PC2, as $c_2 = \mathbf{v}_2^T\mathbf{x}$.

In general, the j-th PC problems becomes seeking $\mathbf{v}$ s.t. $\|\mathbf{v}\| = 1$ and $\mathbf{v}^T\mathbf{v}_{j'} = 0$ for $j' = 1, \ldots, (j-1)$ that maximizes $\text{var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$. The Lagrangian is

$$\phi(\mathbf{v}; \alpha_0, \alpha_1, \ldots, \alpha_{j-1}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \alpha_0\left(\mathbf{v}^T\mathbf{v} - 1\right) - 2\sum_{j'=1}^{j-1}\alpha_{j'}\left(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}_{j'}\right). \tag{5}$$

And the solution would be $\mathbf{v}^\star = \mathbf{v}_j$ with $\text{var}(\mathbf{v}_j^T\mathbf{x}) = \lambda_j$.

Let $\mathbf{V} = (\mathbf{v}_1,\ \mathbf{v}_2,\ \cdots\ \mathbf{v}_p)$ be the matrix with $\mathbf{v}_j$'s as columns. It follows that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, i.e., $\mathbf{V}$ is orthogonal. Also we have

$$\text{cov}(\mathbf{V}^T\mathbf{x}) = \mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V} = \Lambda = \text{diag}(\lambda_j).$$

Or, equivalently,

$$\boldsymbol{\Sigma} = \mathbf{V}\Lambda\mathbf{V}^T,$$

which gives the spectral decomposition or eigen-decomposition of $\boldsymbol{\Sigma}$. Therefore, the variance of the jth PC c_j is the j-th eigenvalue λ_j.

1.2.2 The Sample Version

What's available in reality is a data matrix $\mathbf{X}$ of dimension $n \times p$, whose row $\mathbf{x}_i$'s are i.i.d. copies of the random vector $\mathbf{x}$ we discussed in the preceding section. WLOG, let's assume that $\mathbf{X}$ has been centered so that $\bar{\mathbf{x}} = 0$. The sample PCA is executed on the $p \times p$ variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{6}$$

or the sample correlation matrix. The $\hat{\boldsymbol{\Sigma}}$ given above is the maximum likelihood estimator of $\boldsymbol{\Sigma}$ if we assume $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Alternatively, the $1/n$ factor in (6) is replaced by $1/(n-1)$ in the more commonly used unbiased estimator of $\boldsymbol{\Sigma}$.

The following theorem (Theorem 11.3.1; Anderson, 1984) justifies the PCA based on $\hat{\boldsymbol{\Sigma}}$ by saying that the PCs $\mathbf{C} = \mathbf{V}\mathbf{X}$ obtained from the spectral decomposition of $\hat{\boldsymbol{\Sigma}}$ in (6) provides MLE of the true PCs based on $\boldsymbol{\Sigma}$.

Theorem 1.1. *Assume $\mathbf{x}_i \in \mathbb{R}^p \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$ and data matrix $\mathbf{X}_{n\times p}$ has rows $\mathbf{x}_i^T$. Further assume that $\boldsymbol{\Sigma}$ has p different characteristic roots. Let $\hat{\lambda}_1 \geq \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ be the eigenvalues of the MLE $\hat{\boldsymbol{\Sigma}}$ and $\{\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p\}$ be the corresponding eigenvectors. Then $(\hat{\lambda}_j, \hat{\mathbf{v}}_j)$ is the maximum likelihood estimator (MLE) of $(\lambda_j, \mathbf{v}_j)$.*

With slight abuse of notations, we shall use $\{\mathbf{V}, \mathbf{\Lambda} = \text{diag}(\lambda_j), \mathbf{C} = (\mathbf{c}_j)\}$ to denote both their population and sample versions for the sake of simplicity. Therefore, Analogous to the population version, the eigen-decomposition of $\hat{\mathbf{\Sigma}}$ is given by

$$\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{7}$$

The sample variance of the jth PC $\mathbf{c}_j \in \mathbb{R}^n$ is the j-th estimated eigenvalue λ_j. In order to determine the number of PCs for practical purposes, the cumulative proportion of explained variance is often computed as a percentage $\lambda_j / \sum_{j'} \lambda_{j'}$ for $j = 1, \ldots, p$. On this basis, a scree plot can be made to inspect for the desired number of PCs. See Figure 2 for an example.

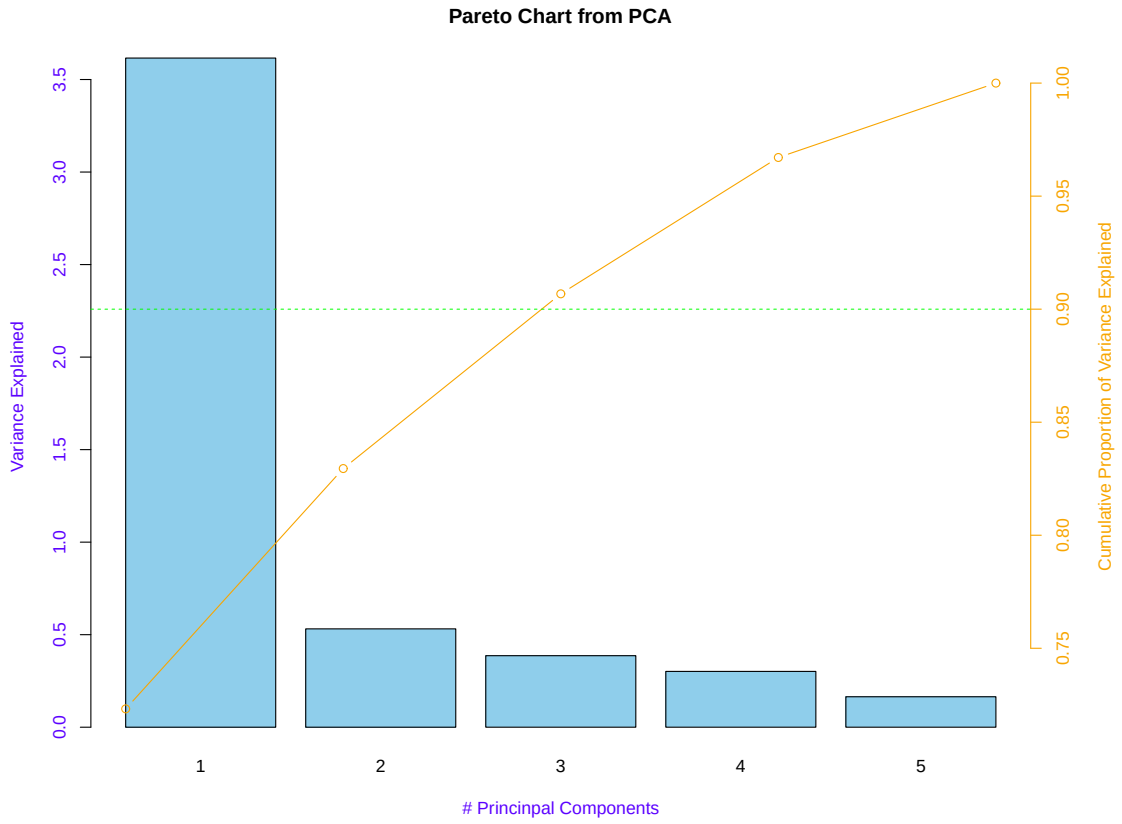

Figure 2: Scree plot from principla components analysis (PCA) with the `bumpus` data.

2 Computation of PCA via SVD of X

The singular value decomposition (SVD) of $\mathbf{X}_{n\times p}$ with $n \geq p$ has the following form

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}V^T_{p\times p}, \tag{8}$$

where both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal with $\mathbf{U^T U} = \mathbf{V^T V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ (note that $\mathbf{U}\mathbf{U}^T \neq \mathbf{I}_n$); the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{U})$; the columns of $\mathbf{V}$ span the row space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X}^T) = \mathcal{C}(\mathbf{V})$; $\mathbf{D} = \text{diag}(\sigma_j)$ is diagonal with *singular values* $\sigma_1 \geq \sigma_j \geq \cdots \geq \sigma_p \geq 0$.

Different forms of SVD are available depending on the relative magnitude of quantities $\{n, p, r\}$, where $r = \text{rank}(\mathbf{X})$. It is often further assumed that $\sigma_j > 0$ strictly so that there are r singular values σ_j in total. The Schatten family of matrix norms are defined on the basis of the singular values σ_j.

Definition 2.1. For a matrix $\mathbf{A}$, its Schatten p-norm is defined by

$$\|\mathbf{A}\|_p = \left(\sum_{j=1}^{r} \sigma_j^p \right)^{1/p} . \tag{9}$$

When $p = 1$, $\|\mathbf{A}\|_1 = \sum_{j=1}^{r} \sigma_j$ is called the nuclear norm; the case $p = 2$ with $\|\mathbf{A}\|_2 = \sum_{j=1}^{r} \sigma_j^2$ matches with the Frobenius norm; when $p = \infty$, $\|\mathbf{A}\|_\infty = \sigma_1$ reduces to the spectral norm, i.e., the maximum singular value σ_1 of $\mathbf{X}$.

Define the column space $\mathcal{C}(\mathbf{X})$ as

$$\mathcal{C}(\mathbf{X}) = \{\mathbf{c} \in \mathbb{R}^n : \ \mathbf{c} = \mathbf{X}\mathbf{b} \text{ for some vector } \mathbf{c} \in \mathbb{R}^p\} .$$

This concept plays a critical role in the geometric interpretation of linear regression. It can be shown that $\mathcal{C}(\mathbf{X}) = \mathbf{C}(\mathbf{X}\mathbf{X}^T)$ and $\mathcal{C}(\mathbf{X}^T) = \mathcal{C}(\mathbf{X}^T\mathbf{X})$. The symmetric and nonnegative definite (n.n.d.) matrix $\mathbf{X}^T\mathbf{X}$, termed as the Gram matrix, is frequently encountered in regression and many other problems. For example, the Frobenius matrix norm is given by

$$\|\mathbf{X}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} X_{ij}^2 = \text{trace}(\mathbf{X}^T\mathbf{X}) = \sum_{j=1}^{p} \sigma_j^2.$$

Given the SVD of $\mathbf{X}$ in (8), it follows that the eigen-decomposition of $\mathbf{X}^T\mathbf{X}$ is given by

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^\mathbf{T}\mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T} = \mathbf{V}\mathbf{D}^\mathbf{2}\mathbf{V}^\mathbf{T}.$$

Referring to equation (7), it follows that $\mathbf{D}^2 = \text{diag}(\sigma_j^2) = \text{diag}(\lambda_j \cdot n)$. Similarly, $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}^\mathbf{2}\mathbf{U}^T \in \mathbb{R}^{n \times n}$ is referred to as the kernel matrix consisting of inner products $\mathbf{x}_i^T\mathbf{x}_{i'}$ as elements.

Denote

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_p) \quad \text{and} \quad \mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_p) ,$$

where $\mathbf{u}_j \in \mathbb{R}^n$ and $\mathbf{v}_j \in \mathbb{R}^p$ denotes the columns of $\mathbf{U}$ and $\mathbf{V}$ respectively. Then we have

$$\mathbf{X} = \sum_{j=1}^{p} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^T .$$

In terms of matrix approximation, we have the following theorem.

Theorem 2.1. *Fixing $k \in \mathbb{N}$ such that $k \leq p$, let $\mathbf{X}_k = \sum_{j=1}^{k} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^T$. Then, for any matrix $\mathbf{A}$ of rank at most k, we have*

$$\|\mathbf{X} - \mathbf{X}_k\|_F \ \leq \ \|\mathbf{X} - \mathbf{A}\|_F .$$

The above inequality also holds for the spectral matrix norm $\|\mathbf{X}\|_\infty = \sigma_1$, i.e., the largest singular value σ_1 of $\mathbf{X}$. In particular, $\|\mathbf{X} - \mathbf{X}_k\|_\infty = \sigma_{k+1}$.

In other words, matrix $\mathbf{X}_k$ is the solution of the following optimization problem for matrix approximation:

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\|_F, \quad \text{s.t. } \operatorname{rank}(\mathbf{A}) \le k. \tag{10}$$

On this basis, the robust PCA make convex relaxation of $\operatorname{rank}(\mathbf{A})$ to its nuclear norm (or ℓ_1) norm. The connection between SVD and PCA has already been manifested by the eigen-decomposition of $\mathbf{X}^T\mathbf{X}$.

To summarize, assume that the columns of $\mathbf{X}$ all have been centered or standardized so that $\bar{\mathbf{x}} = \mathbf{0}$. In this case, we have

$$\mathbf{X}^T\mathbf{X} = n \cdot \hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}^\mathbf{2}\mathbf{V}^\mathbf{T},$$

where the factor n could be $(n-1)$ alternatively. The spectral decomposition of $\hat{\mathbf{\Sigma}}$ is given by $\hat{\mathbf{\Sigma}} = \mathbf{V}(\mathbf{D}^\mathbf{2}/\mathbf{n})\mathbf{V}^\mathbf{T}$. It follows that $\sigma_j^2/n = \lambda_j$ for $j = 1, \ldots, p$. This implies that matrix $\mathbf{V}$ in SVD (8) of $\mathbf{X}$ contains the loading factors for PCA based on $\hat{\mathbf{\Sigma}}$ and the matrix $\mathbf{C}$ of resultant sample PC's (as columns) is

$$\mathbf{C} = \mathbf{X}\mathbf{V} = (\mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T})\mathbf{V} = \mathbf{U}\mathbf{D}.$$

It is clear that $\mathbf{c}_j = \mathbf{u}_j\sigma_j$ is the *j-th principal component* of $\mathbf{X}$ and $\mathbf{v}_j$ is the *j-th principal component direction* (or the loadings) with $\mathbf{X}\mathbf{v}_j = \mathbf{u}_j\sigma_j$ for $j = 1, \ldots, p$. Usually, $\mathbf{u}_j$ is called the j-th normalized principal component (i.e., with length $\|\mathbf{u}_j\| = 1$). We have

$$\operatorname{var}(\mathbf{u}_j\sigma_j) = \operatorname{var}(\mathbf{X}\mathbf{v}_j) = \sigma_j^2/n = \lambda_j.$$

3 PCA as a Regression Problem

Since principal components account for variation in the original data matrix $\mathbf{X}$, it is not surprising that PCA can be formulated as a regression problem in view of analysis of variance. This formulation is closely related to factor analysis.

PCA has a minimum distance property in the following sense. Given data $\{\mathbf{x}_i : i = 1, \ldots n\}$ as n IID copies of $\mathbf{x}$, we assume WLOG that $\bar{x} = \mathbf{0}$. Consider approximating or modeling $\mathbf{x}_i$ linearly as $\mathbf{V}_q\mathbf{c}_i$ where $\mathbf{V}_q \in \mathbb{R}^{p\times q}$ has q orthonormal columns and $q \le p$. With square loss, this leads to optimization problem

$$\min_{\mathbf{V}_q, \mathbf{c}_i} \quad \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{V}_q\mathbf{c}_i\|^2. \tag{11}$$

Plugging the solution for $\hat{\mathbf{c}}_i = \mathbf{V}_q^T\mathbf{x}_i$, the problem becomes

$$\min_{\mathbf{V}_q} \quad \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{V}_q\mathbf{V}_q^T\mathbf{x}_i\|^2, \tag{12}$$

where $\mathbf{V}_q\mathbf{V}_q^T = \mathbf{V}_q(\mathbf{V}_q^T\mathbf{V}_q)^{-1}\mathbf{V}_q^T$ is a projection matrix. In view of the SVD decomposition of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, the solutions $\mathbf{V}_q$ to (12) turns out to be the the first q right-singular vectors or the first q columns of $\mathbf{V}$; matrix $\mathbf{C}_q = (\mathbf{c}_i^T, \ldots, \mathbf{c}_i^T)^T \in \mathbb{R}^{n\times q}$ (i.e., having $\mathbf{c}_i^T$ as ith row) is the first q columns of matrix $\mathbf{U}\mathbf{D}$. Namely, $\mathbf{C}_q = (\sigma_1\mathbf{u}_1, \ldots, \sigma_q\mathbf{u}_q)$. Recall that columns of $\mathbf{U}\mathbf{D}$ are the principal components (p in total). The minimum of the objective function in (12) becomes $\|\mathbf{X} - \mathbf{C}_q\mathbf{V}_q^T\|_F$, as also manifested by Theorem 2.1.

The PC directions can also be recovered by regressing PC components on the p variables, which essentially corresponds to the Lagrangian form of (11). See Proposition 3.1 below.

Proposition 3.1. *For each j, denote $\mathbf{c}_j = \mathbf{u}_j\sigma_j$ the j-th principal component. Consider the following ridge (or ℓ_2-regularized) regression problem*

$$\min_{\boldsymbol{\beta}} \|\mathbf{c}_j - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2,$$

for any $\lambda > 0$. Let $\tilde{\boldsymbol{\beta}}_j$ denote the solution. Then we have $\tilde{\boldsymbol{\beta}}_j /\!/ \mathbf{v}_j$, i.e.,

$$\tilde{\boldsymbol{\beta}}_j/\|\tilde{\boldsymbol{\beta}}\| = \mathbf{v}_j$$

gives the j-th principal component direction.

The above proposition holds for any $\lambda > 0$ and it holds, independent of the specific value of λ. The PC directions can also be recovered by regressing $\mathbf{x}_i$ (the i-th row vector of $\mathbf{X}$) on its linear combinations. First consider the first PC direction.

Proposition 3.2. *For any $\lambda > 0$, let*

$$(\tilde{\mathbf{c}}, \tilde{\boldsymbol{\beta}}) \;=\; \arg\min_{\mathbf{c},\boldsymbol{\beta}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{c}\,\left(\boldsymbol{\beta}^T\mathbf{x}_i\right)\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \quad s.t. \|\mathbf{c}\|^2 = 1.$$

Then $\tilde{\boldsymbol{\beta}} \;\propto\; \mathbf{v}_1$ provides the first or leading PC direction.

The whole sequence of PCs can be derived in a similar manner.

Proposition 3.3. *Suppose we are considering the first k principal components. Let matrix*

$$\mathbf{A}_{p\times k} = [\mathbf{c}_1, \ldots, \mathbf{c}_k] \quad \textit{and} \quad \mathbf{B}_{p\times k} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k]\,.$$

For any $\lambda > 0$, let

$$(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \;=\; \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\left(\mathbf{B}^T\mathbf{x}_i\right)\|^2 + \lambda\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2, \quad s.t.\ \mathbf{A}^T\mathbf{A} = \mathbf{I}_{k\times k}.$$

Then $\tilde{\boldsymbol{\beta}}_j = \mathbf{v}_j$ for $j = 1, 2, \ldots, k$ provide the first k PC directions.

On the basis of this connection, Zou, Hastie, and Tibshirani (2006) proposed sparse PCA by adding another ℓ_1 penalty term (in this case, they call the penalty *elastic net*).

4 Extensions of PCA

PCA can be extended in various ways. These include

- Robust PCA seeks to recover $\mathbf{L}$ and $\mathbf{S}$, if the data matrix $\mathbf{X} = \mathbf{L} + \mathbf{S}$ for a low-rank matrix $\mathbf{L}$ contaminated by a sparse matrix $\mathbf{S}$. In conventional settings, robust PCA executes PCA on a *robust* estimate of the variance-covariance or correlation matrix that are free of outliers.
- Regularized PCA or Sparse PCA makes available variable selection by having some 0 loading values via ℓ_1 regularization. The main advantage of sparse PCA is its enhanced interpretability. Another classical method for improving interpretability is the oblique PCA via rotations by allowing for correlated components.

- Correspondence analysis (CA) (also called reciprocal averaging) applies performs dimension reduction that is conceptually similar to PCA on categorical data.
- Principal curves and surfaces for finding nonlinear principal directions.
- Sliced Inverse Regression (SIR) for (sufficient) dimension reduction in the supervised learning setting. Variants of similar flavor to PCA for supervised learning also include ridge regression, partial least squares regression, and continuum regression.
- Independent components analysis (ICA) seeks independent components for non-Gaussian data as exemplified by the 'cock tail party' problem.
- Kernel PCA also faciliates nonlinear directions by first projecting $\mathbf{X}$ to the hidden features in the reproducing kernel Hilbert space (RKHS).
- Canonical correlation analysis (CCA) seeks maximally correlated linear combinations between two sets of variables.

In the ensuing sections, we will briefly discuss principal curve and canonical correlation analysis here. Several other extensions will be discussed in future topics.

4.1 Principal Curves

Given a random vector $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we want to find a smooth curve $\mathbf{h}(s) = (h_1(s), \ldots, h_p(s))^T$ (parameterized by one single parameter $s \in \mathbf{R}$) that passes through the 'middle" of its distribution.

Definition 4.1. The *projection index* $s_h(\mathbf{x})$ of a point $\mathbf{x}$ on a curve $\mathbf{h}(s)$ in the p-dimensional space is defined as the value of s corresponding to the point on $\mathbf{h}(s)$ that is closest to $\mathbf{x}$. In other words, point $\mathbf{h}(s_h(\mathbf{x}))$ is the projection of $\mathbf{x}$ on the curve $\mathbf{h}(s)$.

Definition 4.2. A *principal curve* is a curve satisfying the self-consistency property

$$\mathbf{h}(s) \;=\; \mathrm{E}\left(\mathbf{x} | s_h(\mathbf{x}) = s\right). \tag{13}$$

If we project each point $\mathbf{x}$ to the curve $\mathbf{h}(s)$, then its project point $\mathbf{T} = \mathbf{h}(s_h(\mathbf{x}))$ is the expectation or average of all points that project to it, as illustrated in Figure 3 for the 2D scenario.

Hastie and Stuetzle (1989) showed that a principal curve is a critical point of the squared distance $\mathbf{E}\sum_{j=1}^{p}\{X_j - h_j(s)\}^2$. Thus the principal curve generalizes the minimum distance property of linear principal components.

Theoretically the principal curve can be obtained using Algorithm 1.

With sample data, the conditional expectation step in the algorithm is replaced by any one-dimensional nonparametric smoothing. Figure 4 provides an empirical illustration.

It is worth noting that there is one unsettling property with the principal curve. Suppose that $\mathbf{X}$ satisfies

$$X_j = h_j(s) + \varepsilon_j, \quad \text{for } j = 1, 2, \ldots, p$$

where s and ε_j's are independent with $\mathrm{E}(\varepsilon_j) = 0$. Then $\mathbf{h} = (h_1, \ldots, h_p)^T$ is not in general a principal curve of the distribution of $\mathbf{X}$.

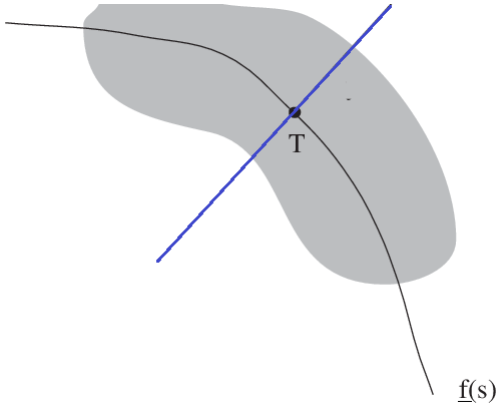

Figure 3: Two-Dimensional Illustration of How Principal Curve is Defined: Reproduced from Tibshirani (1992). The grey area is the distribution of $\mathbf{x}$. Given a point $\mathbf{T}$ on the principal curve $\mathbf{h}(s)$, the average of all points $\mathbf{x}$'s that project to T (the blue line) is point $\mathbf{T}$.

Algorithm 1 Principal Curve Algorithm

Start with $\mathbf{h}(s) = \mathrm{E}(\mathbf{x}) + \mathbf{d}s$, where $\mathbf{d}$ is the first eigenvector of $\mathbf{\Sigma}$ and $s = s_{\mathbf{h}}(\mathbf{x})$ for each $\mathbf{x}$.
repeat
  Smoothing: Fixing s, minimize $\mathrm{E}\|\mathbf{x} - \mathbf{h}(s)\|^2$ to obtain $h_j(s) = \mathrm{E}\{X_j | s_h(\mathbf{x}) = s\}$ for each j.
  Projection: Fixing $\mathbf{h}(s)$, obtain $s = s_{\mathbf{h}}(\mathbf{x})$ for each $\mathbf{x}$.
until the change in $\mathrm{E}\|\mathbf{x} - \mathbf{h}(s)\|^2 < \varepsilon_0$
return $\mathbf{h}(s)$.

4.2 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) focuses on the correlations (thus linear association) between linear combinations of variables in one set and linear combinations of variables in another set. CCA is useful in multivariate regression by treating variables in one set as responses.

We consider the theoretical version. To set up, let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$. WLOG, let's assume $p \leq q$; otherwise, we switch $\mathbf{x}$ with $\mathbf{y}$. Suppose that

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \left\{ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \right\}$$

Let $c = \mathbf{a}^T\mathbf{x}$ and $d = \mathbf{b}^T\mathbf{y}$, both being scalars, denote linear combinations of $\mathbf{x}$ and $\mathbf{y}$, respectively. Note that

(i) $\mathrm{var}(c) = \mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a}$ and $\mathrm{var}(d) = \mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b}$.

(ii) $\mathrm{cov}(c, d) = \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}$.

(iii) $\mathrm{corr}(c, d) = \dfrac{\mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a}} \cdot \sqrt{\mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b}}}$.

For the first pair of canonical variables (or first canonical variate pair), we seek (c_1, d_1) having

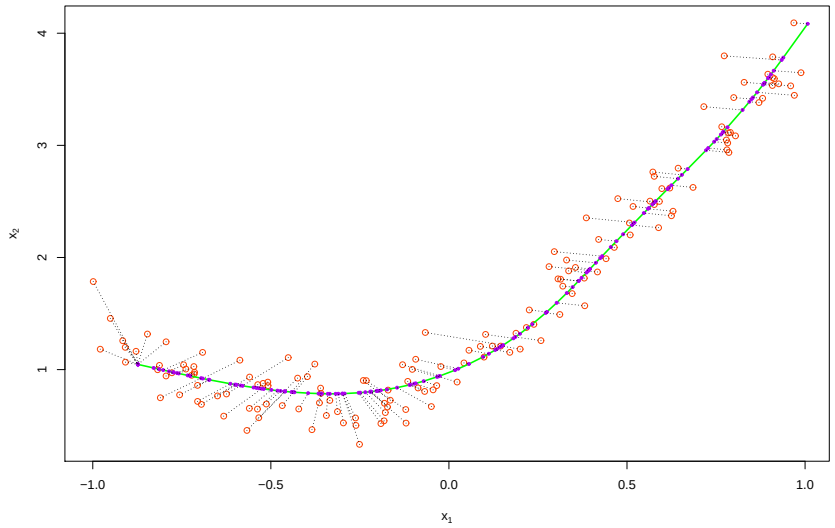

Figure 4: Illustration of Principal Curves with Function `principal.curve()` in R Package **princurve**.

unit variance, which maximizes the correlation by solving

$$\max_{\mathbf{a},\mathbf{b}} \quad \text{corr}(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{y}), \quad \text{s.t. } \mathbf{a}^T\boldsymbol{\Sigma}_{11}\mathbf{a} = \mathbf{b}^T\boldsymbol{\Sigma}_{22}\mathbf{b} = 1.$$

The constraints correspond to the unit variance conditions on c and d. See Figure 5 for illustration.

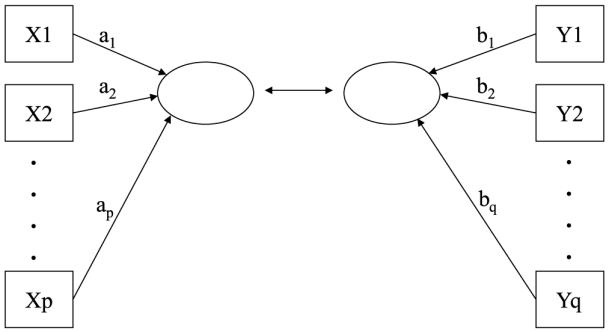

Figure 5: Illustration of CCA: the canoncal pair.

Next, we seek (c_2, d_2) having unit variances, which maximizes $\text{corr}(c, d)$, among all choices that are uncorrelated with the first pair of canonical variables.

And so on, the k-th pair of canonical variables (c_k, d_k) having unit variances, which maximize $\text{corr}(c, d)$, among all the choices that are uncorrelated with all the previous $(k-1)$ canonical variable pairs.

Note that at most p canonical pairs can be found. They are explicitly given by the following theorem. The derivation essentially involves applying the Cauchy-Schwarz inequality to $\text{corr}(c, d)$.

Theorem 4.1. *Suppose that $p \leq q$ and $\mathbf{\Sigma}$ has full rank. Then*

$$\begin{cases} c_k &= \mathbf{e}_k \mathbf{\Sigma}_{11}^{-1/2} \mathbf{x} \\ d_k &= \mathbf{h}_k \mathbf{\Sigma}_{22}^{-1/2} \mathbf{y} \end{cases}$$

for $k = 1, \ldots, p$ are the p canonical pairs and let $\rho_k = corr(c_k, d_k)$ be the maximized correlation, where $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_p$ are the eigenvalues of matrix $\mathbf{\Sigma}_{11}^{-1/2} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1/2}$ with associated eigenvectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_p\} \in \mathbb{R}^p$; in fact, $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_p$ are also the eigenvalues of matrix $\mathbf{\Sigma}_{22}^{-1/2} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1/2}$ with associated eigenvectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_p\} \in \mathbb{R}^q$; and each $\mathbf{h}_k$ is proportional to $\mathbf{\Sigma}_{22}^{-1/2} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1/2} \mathbf{e}_k$. It also can be verified that

$$corr(c_k, c_l) = corr(d_k, d_l) = corr(c_k, d_l) = 0, \quad for\ k \neq l.$$

The sample version of CCA can be carried out in a similar manner. The computation of CCA can be done via SVD of the sample covariance matrix. It turns out that CCA can also be obtained via a restricted multivariate regression problem (see p.85; Hastie, Tibshirani, and Friedman, 2009), a fact that can be of immediate use for proposing sparse CCA. Clearly, CCA can have robust, kernel, and other variants.

In R, CCA is implemented by the `cancor()` function in the **basic** package. The package **CCA** provides other functionality, including missing value handling and the penalized CCA.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Edition. Wiley. ISBN-13: 978-0471889878.

Gentle, J. E. (2009). *Computational Statistics*. Fairfax, VA: Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Section 14.5–14.7. Springer.

Hastie, T., and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, **84**: 502–516.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*, 2nd edition. San Diego, CA: Academic Press.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing*, **2**(4): 183–190.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2): 265–286.