Generic Setup

Data: $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$.

Categorical variables are coded using dummy variables.

We collect the x-values in a big matrix

$$\mathbf{X} = \left\{ \begin{array}{c} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{array} \right\} = \left\{ \begin{array}{ccc} x_{1,1} & \ldots & x_{1,p} \\ x_{2,1} & \ldots & x_{2,p} \\ \vdots & & \vdots \\ x_{N,1} & \ldots & x_{N,p} \end{array} \right\}$$

with dimensions $N \times p$.

If a we need to work with a categorical x-coordinate that can occur on K different levels we can encode the variable as a $(K-1)$-dimensional vector of zero's and one's containing just a single one. This coding is known as dummy variables.

Figure 14.22 – Threes

In this example the resulting data matrix $\mathbf{X}$ is 130×256.

Linear Algebra - the Mean Value

Matrix computations and decompositions is the key to many theoretical results, and practical success relies heavily on efficient matrix computations.

With $\mathbf{1}$ the N-dimensional vector with one's at all positions, the *column means* can be computed as

$$\bar{x}^T = \frac{1}{N}\mathbf{1}^T\mathbf{X}$$

The *projection* in $\mathbb{R}^N$ onto $\mathbf{1}$ and the orthogonal complement $\mathbf{1}^\perp$ are given by the matrices

$$P = \frac{1}{N}\mathbf{1}\mathbf{1}^T, \quad I_N - P = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T,$$

respectively.

Linear Algebra - the Covariance Matrix

The empirical covariance matrix is

$$\begin{aligned}(N-1)\hat{\Sigma} &= (\mathbf{X}-\mathbf{1}\bar{x}^T)^T(\mathbf{X}-\mathbf{1}\bar{x}^T)\\ &= (\mathbf{X}-P\mathbf{X})^T(\mathbf{X}-P\mathbf{X})\\ &= ((I_N-P)\mathbf{X})^T(I_N-P)\mathbf{X}\\ &= \mathbf{X}^T(I_N-P)\mathbf{X}\end{aligned}$$

since $(I_N-P)^2 = I_N-P$.

Often we will use the augmented matrix $\{\mathbf{1}\ \mathbf{X}\}$ and often we will assume that $\mathbf{X}$ has then been orthogonalized with $\mathbf{1}$. This means that $\mathbf{X}$ has been replaced with $(I_N-P)\mathbf{X} = \mathbf{X}-\mathbf{1}\bar{x}^T$. This *does not change the column space* of $\{\mathbf{1}\ \mathbf{X}\}$.

Matrix decompositions

A core problem is to find useful decompositions

$$\mathbf{X} = \mathbf{AB}$$

for an $N \times p$ matrix $\mathbf{A}$ and a $p \times p$ matrix $\mathbf{B}$.

The *column space* of $\mathbf{A}$ and $\mathbf{X}$ is the same.

Objectives include:

- Computational benefits, e.g. efficient and reliable equation solving.
- Approximations: if $\mathbf{A}^q$ and $\mathbf{B}^q$ for $q < p$ denotes the first q rows and the first q columns, respectively, $\mathbf{A}^q\mathbf{B}^q$ provides an approximation.
- Projections: $\mathbf{A}^q$ holds the coefficients for the expansion of the x_i's in the first q rows $\mathbb{B}$, whose rows form a basis of $\mathbb{R}^p$.

Note that if $p > N$, or generally, if the rank of $\mathbf{X}$ is not p, the $\mathbf{B}$ matrix is not uniquely determined by $\mathbf{A}$. In that case it may be preferable to look for matrices of dimensions $N \times p'$ and $p' \times p$ with p' the rank of $\mathbf{X}$, that is, the dimension of the column space spanned by $\mathbf{X}$. Then $\mathbf{B}$ becomes uniquely determined by $\mathbf{A}$, and if needed, it can be arbitrarily supplemented by $p'-p$ additional rows to a full basis of $\mathbb{R}^p$.

Singular Value Decomposition

$p' = \min\{N, p\}$.

Theorem 1. *If $\mathbf{X}$ is an $N \times p$ matrix there exists an $N \times p'$ matrix U, a $p' \times p$ matrix V and a diagonal matrix*

$$D = \begin{Bmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_{p'} \end{Bmatrix}$$

such that $U^TU = I_{p'}$, $V^TV = I_{p'}$, $d_1 \geq \ldots \geq d_{p'} \geq 0$ *and*

$$\mathbf{X} = UDV^T.$$

We call $d_1, \ldots, d_{p'}$ the *singular values*. V is an orthogonal matrix with $V^{-1} = V^T$ if $p = p'$. The columns in U with corresponding $d_i > 0$ form an orthonormal basis for the column space of $\mathbf{X}$.

The proof of the singular value decomposition is based on the diagonalization of the symmetric, positive definite matrix $\mathbf{X}^T\mathbf{X}$, that is

$$\mathbf{X}^T\mathbf{X} = VD^2V^T$$

with D^2 a diagonal matrix with non-negative entries and V an orthogonal matrix. The U matrix is then found as

$$U = \mathbf{X}VD^{-1}$$

provided that all entries in D^2 are strictly positive, and we see that

$$\begin{aligned} U^TU &= D^{-1}V^T\mathbf{X}^T\mathbf{X}VD^{-1} \\ &= D^{-1}V^TVD^2V^TVD^{-1} \\ &= I_p. \end{aligned}$$

A suitable modification, solving only for the U-columns corresponding to diagonal entries > 0 and then supplementing the basis, works if some of the entries in D^2 are 0.

Figure 14.20 – Dimension Reduction

A one dimensional representation of 2D data points is sought.

The natural idea is to minimize the sum of squared distances from the line to the data points *perpendicular to the line.*

This differs from linear regression where we consider the sum of distances *parallel to the 2nd coordinate axis.*

Dimension Reduction and Projections

How can we visualize the data in $\mathbf{X}$? What is a good *low-dimensional projection* $P : \mathbb{R}^p \to \mathbb{R}^p$ with rank 1, 2 or 3?

With

$$V = \{V_q \; V_{p-q}\}$$

where V_q is $p \times q$, the projection onto the columns of V_q is

$$P_q = V_qV_q^T.$$

Then P_q *minimizes* among all rank q projections the *reconstruction error*

$$\sum_{i=1}^{N} ||x_i - P_q x_i||^2 = \text{trace}((\mathbf{X} - \mathbf{X}P_q)(\mathbf{X} - \mathbf{X}P_q)^T)$$

Note that a computation of the reconstruction error by computing the $N \times N$ matrix $(\mathbf{X} - \mathbf{X}P_q)(\mathbf{X} - \mathbf{X}P_q)^T$ and then computing the trace is a computational waste. All the non-diagonals in the matrix product are not needed.

We will generally always replace $\mathbf{X}$ by $\mathbf{X} - \mathbf{1}\bar{x}^T$ before attempting a projection onto a subspace. Because the p coordinates we measure by no means need to be measured on a common scale it is often also most relevant to normalize the columns to have unit length before we attempt a dimension reduction. That is, we divide each column by its empirical standard error. If there are other ways to bring all variables measured on a common scale that might be preferred. We should note that the projections obtained from the singular value decomposition are not invariant to marginal scaling of the columns in $\mathbf{X}$.

Figure 14.21 – Dimension Reduction and PC

The coordinates for the P_q projections of the data points in the V_q basis are called the q first principal components.

The coordinates are

$$\begin{aligned} XV_q &= UDV^TV_q \\ &= UD\text{diag}(1, \ldots, 1, 0, \ldots, 0) \\ &= U_qD_q \end{aligned}$$

with U_q and D_q the matrices with the q first columns from U and D, respectively.

Figure 14.23 – Two First Principal Components for Threes

The first principal component shows primarily the variation in how wide the hand written threes are. The second shows primarily the variation in how thick the drawn line is.

Figure 14.23 – Two First Principal Components for Threes

All pixel values are measured on the same scale so we would only centralize – not scale – the columns.

Factor Analysis

Let $X = \mathbf{A}S + \varepsilon$ with $\mathbf{A}$ a $p \times q$ matrix and S a q-vector – the *unobserved loadings* – with independent coordinates and ε a vector of i.i.d. noise variables.

With $\mathbf{X} = UDV^T$ the SVD, $\hat{A} = DV^T/\sqrt{N}$ and $\hat{S} = \sqrt{N}U$ we can interpret the first q columns of $\hat{S}$ as estimates of the unobserved loadings.

Unfortunately, any $q \times q$ orthogonal transformation of these columns qualify equally well.

Sparse PCA

Recent generalizations of PCA involve attempts to make *sparse* low-rank reconstructions, e.g. minimization of

$$\sum_{i=1}^{N} ||x_i - \Theta V^T x_i||^2 + \lambda \sum_{k=1}^{K} ||v_k||_2^2 + \sum_{k=1}^{K} \lambda_{1k} ||v_k||_1$$

subject to Θ being $p \times K$ with orthonormal columns and V being $p \times K$ with columns v_k. The penalization ensures that v_k has zeroes, thus for the reconstruction of x_i in terms of the K-basis in Θ each coefficient depends only on a subset of the coordinates in x_i.

Non-negative Matrix Factorization

Another recent idea is for positive matrices to look for factorizations WH such that

- the entries in W and H are all positive,
- W is $N \times q$ and H is $q \times p$ such that WH is a good approximation of $\mathbf{X}$.

The resulting basis columns in W may be interpretable.

But there is in general no unique positive matrix factorization ...

Figure 14.33 – Non-negative matrix factorization

Figure 14.33 – Non-negative matrix factorization

Computational Shortcuts

Suppose we consider the problem of minimizing

$$\sum_{i=1}^{N} L(y_i, \beta_0 + x_i^T \beta) + \lambda ||\beta||^2$$

over β_0 and $\beta \in \mathbb{R}^p$ with $p > N$. With $\mathbf{X} = UDV^T$ the singular value decomposition (when $p > N$, U is $N \times N$ orthogonal, D is $N \times N$ diagonal and V is $p \times N$ with $V^TV = I_N$), then $R = UD$ is an $N \times N$ matrix and

$$\mathbf{X}\beta = UDV^T\beta = RV^T\beta = R\theta$$

where $\theta = V^T\beta$ is N-dimensional. Writing $\beta = V\theta + \beta^\perp$ with $\beta^\perp$ orthogonal to the columns in V we see that

$$||\beta||^2 = ||V\theta||^2 + ||\beta^\perp||^2 \geq ||V\theta||^2 = \theta^T V^T V\theta = ||\theta||^2.$$

Since $\mathbf{X}\beta$ is unaffected by $\beta^\perp$ this term equals 0 and we need to minimize

$$\sum_{i=1}^{N} L(y_i, \theta_0 + r_i^T\theta) + \lambda||\theta||^2, \quad \theta \in \mathbb{R}^N.$$