Extended CIL Summary
FS 2013

Pascal Spörri
pascal@spoerri.io

July 8, 2013

Part I.
Dimensionality Reduction

Select the *most interesting* dimensions.

1. Intrinsic Dimensionality

Pairwise Distances

Assume components of data $x = (x_1, \dots, x_D)^T \in \mathbb{R}^D$ are i.i.d. Gaussian distributed:

$$x_d \sim \mathcal{N}(0,1) \implies x_d - y_d \sim \mathcal{N}(0,2).$$

Using χ^2-distribution:

$$\frac{1}{2}(x_d - y_d)^2 \sim \chi^2(1),$$

and extending to D dimensions:

$$\frac{1}{2}\sum_{d=1}^{D}(x_d - y_d)^2 \sim \chi^2(D) = \Gamma\left(\frac{D}{2}, 2\right)$$

$$\text{Recall: } \forall z, k, \theta > 0,\ \Gamma(z; k, \theta) = \frac{\theta^k}{\Gamma(k)} y^{k-1} e^{-\theta z}$$

Hence, the dimension-normalised squared distance is:

$$\frac{1}{D}\sum_{d=1}^{D}(x_d - y_d)^2 \sim \Gamma\left(\frac{D}{2}, \frac{4}{D}\right)$$

is Gamma distributed with mean 2 and variance $\frac{8}{D}$.
$\Gamma\left(\frac{D}{2}, \frac{4}{D}\right)$ tends towards normality with shrinking width for large D. Therefore, most points have *constant* pairwise distances in this limit.

2. Principal Component Analysis

Objectives of PCA:

1. Minimise error $\|x_n - \tilde{x}_n\|$ of point x_n and its approximation $\tilde{x}_n$.
2. Reveal "interesting" information: maximise *variance*.

Both objectives are show to be formally equivalent.
Consider a set of observations $\{x_n\},\ n = 1, \ldots, N$ and $x_n \in \mathbb{R}^D$.

Goal Project data onto $K < D$ dimensional space while maximising variance of the projected data.

For K = 1 Define direction of projection as u_1. Set $\|u_1\|_2 = 1$ (only the direction of the projection is important.

2.1. Statistics of Projected Data

Original Data

Mean is given by the sample mean $\bar{x}$.

Covariance of the Data:

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$$

Projected Data

Mean is given by: $u_1^T\bar{x}$.

Variance is given by:

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\left\{u_1^T x_n - u_1^T\bar{x}\right\}^2 &= \frac{1}{N}\sum_{n=1}^{N}\left\{u_1^T(x_n - \bar{x})\right\}^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}u_1^T(x_n - \bar{x})(x_n - \bar{x})^T u_1 \\ &= u_1^T\Sigma u_1.\end{aligned}$$

2.2. Maximisation Problem

These statistics now can be fed into a maximisation problem:

$$\max_{u_1} \; u_1^T \Sigma u_1$$

such that $\|u_1\|_2 = 1$.
Writing the Lagrangian results in in:

$$\mathcal{L} := u_1^T \Sigma u_1 + \lambda_1 (1 - u_1^T u_1).$$

Setting $\frac{\delta}{\delta u_1} \mathcal{L} \stackrel{!}{=} 0$ results in:

$$\Sigma u_1 = \lambda_1 u_1$$

We observe that u_1 is an *eigenvector* of Σ and λ_1 it's associated *eigenvalue*. Furthermore λ_1 is also the variance of the projected data:

$$\lambda_1 = u_1^T \Sigma u_1$$

2.2.1. Second principal direction

The second principal direction can be obtained by maximising the variance $u_2^T \Sigma u_2$, subject to $\|u_2\|_2 = 1$ and $u_2^T u_1 = 0$:

$$\mathcal{L} = u_2^T \Sigma u_2 + \lambda_2 \left(1 - u_2^T u_2\right) + \nu \left(u_2^T u_1\right).$$

The maximum i found by setting $\frac{\delta \mathcal{L}}{\delta u_2} \stackrel{!}{=} 0$:

$$2\Sigma u_2 - 2\lambda_2 u_2 + \nu u_1 = 0.$$

Because of the orthogonality between u_2 and u_1 we observe that u_2 contains no component of u_1 and hence $\nu = 0$. We get:

$$\Sigma u_2 = \lambda_2 u_2.$$

We observe that u_2 is an eigenvector of Σ with the second largest eigenvalue of λ_2.

2.3. Solution: Eigenvalue Decomposition

Hence we see that the eigenvalue decomposition of the covariance matrix

$$\Sigma = U \Lambda U^T$$

contains all relevant information.
For a projection space of size $K \leq D$ we choose the K eigenvectors $\{u_1, \ldots, u_k\}$ with the largest associated eigenvalues $\{\lambda_1, \ldots, \lambda_2\}$.

2.4. Error Formulation

We define an *orthonormal* basis $\{u_d\}$, $d = 1, \ldots, D$ of $\mathbb{R}^D$. The scalar projection of x_n onto u_d (magnitude) is given by:

$$z_{n,d} = x_n^T u_d.$$

The associated projection onto u_d amounts to $z_{n,d}u_d$. Therefore, each data point can be represented in the basis by:

$$x_n = \sum_{d=1}^{D} z_{n,d}u_d = \sum_{d=1}^{D} \left(x_n^T u_d\right) u_d.$$

Restricted representation using $K < D$ basis vectors can be written as:

$$\tilde{x}_n = \sum_{d=1}^{K} a_{n,d}u_d + \sum_{d=K+1}^{D} b_d u_d,$$

where b_d does not depend on the data point x_n.
The approximation error can be represented by:

$$J(\{a_{n,d}\}, \{b_d\}) = \frac{1}{N}\sum_{n=1}^{N} \|x_n - \tilde{x}_n\|_2^2$$

$$\text{Minimisation of } J \text{ w.r.t.} a_{n,d} = x_n^T$$

$$\text{Minimisation of } J \text{ w.r.t.} b_d = \bar{x}^T u_d$$

The displacement can be obtained by resubsittuing $a_{n,d}$ and b_d:

$$x_n - \tilde{x}_n = \sum_{d=K+1}^{D} \left\{ \left(x_n - \bar{x}\right)^T u_d \right\} u_d.$$

We observe that the displacement vector is orthogonal to the principal space!
Resubstituting the displacement into the error criterion leads to:

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=K+1}^{D} \left(x_n^T u_d - \bar{x}^T u_d\right)^2 = \sum_{d=K+1}^{D} u_d^T \Sigma u_d$$

2.5. Matrix viewpoint

The data can be represented as matrix:

$$X = [x_1, \ldots, x_n, \ldots, x_N]$$

The corresponding zero-centered data is:

$$\bar{X} = X - M,$$

where $M = \underbrace{[\bar{x}, \ldots, \bar{x}]}_{N \text{ times}}$.
Compute the projection of $\bar{X}$ on $U_k = [u_1, \ldots, u_K]$ with:

$$\underbrace{\bar{Z}_K}_{K \times N} = \underbrace{U_K^T}_{K \times D} \cdot \underbrace{\bar{X}}_{D \times N}.$$

To approximate $\bar{X}$, we return to the original basis:

$$\tilde{\bar{X}} = U_K \cdot \bar{Z}_K.$$

For $K = D$ we obtain a perfect reconstruction.

2.6. Computation

First compute the *empirical mean*:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Then *center the data* by subtracting the mean from each sample:

$$\bar{X} = X - M,$$

where $M = \underbrace{[\bar{x}, \ldots, \bar{x}]}_{N \text{ times}}$. Now compute the *Covariance matrix*:

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N} \underbrace{\bar{X}\bar{X}^T}_{\text{Scatter Matrix } \mathbf{S}}.$$

Σ is *symmetric*.
Now the *Eigenvalue decomposition* can be computed:

$$\Sigma = U \Lambda U^T,$$

where $\Lambda = diag[\lambda_1, \ldots, \lambda_D]$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$ with orthonormal eigenvectors.

Transformation the data can be transformed on to the new basis of K dimensions:

$$\tilde{\bar{Z}} = U_K^T \bar{X},$$

$\bar{Z} \in \mathbb{R}^{K \times N}$: We obtain a dimension reduction of the data.

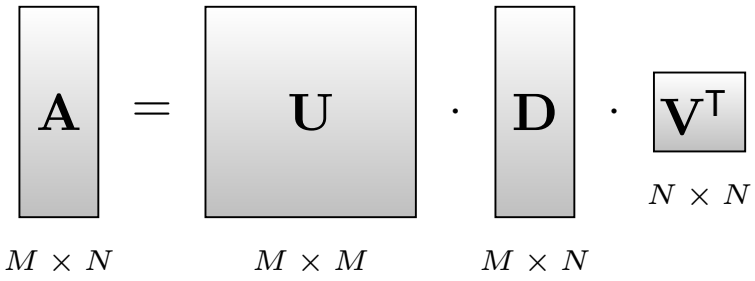
Reconstruction Go back to the original basis by computing

$$\tilde{\bar{X}} = U_K \bar{Z}$$
$$\tilde{X} = \tilde{\bar{X}} + M$$

3. Singular Value Decomposition

3.1. Introduction

The *Singular Value Decomposition* (SVD) is a widely used technique to decompose a matrix into several component matrices exposing many of the useful and interesting properties of the original matrix like rank, null-space, orthogonal basis of column and row space.
Every rectangular, real or complex matrix S has an SVD decomposition into a set of three matrix factors.
Let A be any real M by N matrix, $A \in \mathbb{R}^{M \times N}$, then A can be decomposed as $A = UDV^T$:



- U is an $M \times M$ orthogonal matrix, $U^T U = I$
- D is an $M \times N$ diagonal matrix
- V^T is an $N \times N$ orthogonal matrix, $V^T V = I$

3.2. Singular values

The elements of D are only non-zero on the diagonal and are called the *singular values*. By convention, the order of the singular vectors is determined by the *high-to-low* sorting of singular values, with the highest singular value in the upper left index of the D matrix. The first r columns of U are called *left singular vectors*, they form an orthogonal basis for the space spanned by the columns of the original matrix A.
Similarly the first r rows of V^T are the *right singular vectors*, they form an orthonormal basis for the row space of A.
SVD provides an explicit representation of the range and null-space of a matrix A.

- The right side singular vectors corresponding to vanishing singular values of A, span the null space of A:

$$d_i = 0 \quad \Longrightarrow \quad Av_i = 0 \quad \Longrightarrow \quad v_i \in Null(A).$$

- The left singular vectors corresponding to the non-zero singular values of A span the range of A.

As a consequence, the rank of A equals the number of non-zero singular values (= the number of non-zero elements in D).

$$Rank(A) = \#d_i > 0.$$

3.3. Closest Rank-k Matrix

Let the SVD of $A \in \mathbb{R}^{M \times N}$ be given by $A = UDV^T$. If $k < r = Rank(A)$ and

$$A_k = \sum_{i=1}^{k} d_i u_i v_i^T.$$

Then

$$\min_{Rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2.$$

This means that A_k is the closest $Rank(k)$ approximation to A in the Eculidean matrix norm sense hence:

$$\|A - A_k\|_2 = d_{k+1}.$$

3.4. Properties

The columns of U are the eigenvectors of AA^T. This claim can be verified using the SVD decomposition:

$$AA^T = UDV^T VDU^T = UD^2U^T.$$

Similarly the rows of V^T (or columns of V) are the eigenvectors of $A^T A$ as:

$$A^T A = VDU^T UDV^T = VD^2V^T.$$

Part II.
Appendix

A. Matrix Definitions and Theorems

A.1. Norms

A *norm* is a function $\|\circ\| : V \mapsto \mathbb{R}$ quantifying the size of a vector. It must satisfy

- *Positive scalability:*

$$\|a \cdot x\| = |a| \cdot \|x\|.$$

- *Triangle inequality*

$$\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in V.$$

- *Separability*:

$$\|x\| = 0 \implies x = 0.$$

A.1.1. Vector norms

p-norms The most commonly used matrix norms are p-norms.

$$\|x\|_p := \left(\sum_{i-1}^{n} |x_i|^p \right)^{1/p}$$

for $p \in [1, \infty]$, where $|x_i|$ denotes the absolute value of coordinate x_i.

A special case of the p norm is the *Eclidean norm*:

$$\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}.$$

0-**norm** technically not really a norm is defined by:

$$\|x\|_0 := \text{number of nonzero coordinates in } x.$$

A.1.2. Matrix norms

p-**norm** for matrices:

$$\|X\|_p := \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

A special case is the Euclidean or *spectral norm*:

$$\|X\|_2 = \sigma_{\max}(X),$$

the largest singular value of X.

Frobenius norm is defined as:

$$\|X\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2} = \sum_{i=1}^{\min(m,n)} \sigma_i^2,$$

where σ_i are the singular values of X.

A.2. Orthogonality

Orthogonal vectors Two vectors in an inner product are orthogonal if their inner product is zero.

Orthonormal vectors Orthogonal vectors that have unit length 1

Orthogonal matrix An orthogonal matrix is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e. orthonormal vectors). For orthogonal matrices it also holds that

$$A^T A = I \quad \Longrightarrow \quad A^T = A^{-1} \text{ since,}$$

$$(A^T A)_{i,j} = a_i^T a_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$