8

Principal component analysis

8.1 Introduction

When trying to understand the structure of a tabular dataset, a central question to ask is which features carry more information and whether some features can be recovered from other features.

In general, compactness of the data we work with is important not only for ease of interpretability, but also because working with high-dimensional representations can lead to various pathologies at the time of storing, organizing or analyzing the data. This is generically referred to as the *curse of dimensionality*.

Moreover, a point can be made that the ultimate goal of machine learning and scientific endeavour at large is to distill as compact models as possible to explain the natural phe-

nomena we read the data from.

The technique we are about to describe concerns the analysis of tabular data with numerical entries. From now on we will assume that an $n \times m$ matrix M encodes feature value information per sample of an experiment: each column of M will represent one feature and each row will represent a sample.

8.2 Sample covariance matrices

Given two numerical *random variables* X and Y that are jointly and randomly sampled n times we can define their *covariance* as the following expression:

$$\mathbf{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X))(Y_i - E(Y))$$

where $E(X)$ and $E(Y)$ are the respective means of the random variables. However theoretically convenient, this concept cannot be directly applied to any data unless we know exactly how the data is distributed, which happens rarely, if ever at all. Instead, we will employ a statistic the approximates the theoretical value, the *sample covariance*:

$$\Omega_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

where this time $\bar{X}$ and $\bar{Y}$ denote the respective sample means.

The reader may ask why there is a $1/(n-1)$ factor in the sample covariance expression. This is due to the fact that

with a $1/n$ factor the sample covariance would be a *biased* estimator of the covariance, i.e. the expected value of the sample covariance would miss the population covariance short by a $(n-1)/n$ factor. This is known as Bessel's correction. Although this correction is not obvious, we will not provide a proof here.

The sample covariance has a few interesting properties that make it a good statistic to assess the dependency between random variables which for convenience we can simply refer to as features, because this is the role random variables will have in our setting.

Proposition 31. *The following properties hold:*

1. $\boldsymbol{cov}(X, Y) = \boldsymbol{cov}(Y, X)$
2. $\boldsymbol{cov}(aX + bY, Z) = a\boldsymbol{cov}(X, Z) + b\boldsymbol{cov}(Y, Z)$
3. $\boldsymbol{cov}(X, X) = \sigma_X^2 \geq 0$, *the variance of* X
4. *If* X *and* Y *are independent random variables, then* $\boldsymbol{cov}(X, Y) = 0$

We define the *sample covariance matrix* of the data encoded in M as the $m \times m$ matrix Ω with entries Ω_{ij} containing the sample covariance of the i-th and j-th columns of M:

$$\Omega_{ij} = \Omega_{M_i, M_j} = \frac{1}{n-1} \sum_{k=1}^{n} (M_{ki} - \bar{M}_i)(M_{kj} - \bar{M}_j)$$

The sample covariance matrix is symmetric and contains the

sample variance of the columns of M as diagonal entries. By the very definition of the sample covariance, we can also see that Ω can be expressed as $A^t A$. Can you guess the matrix A that accomplishes this?

We can obtain A from M in two steps:

First we substract the respective column average from every element. This process is often referred to as *centering the data*, because the resulting columns have average equal to zero. Written with matrices, we obtain $\tilde{A}$

$$\tilde{A} = M - \frac{1}{n}\mathbf{1}_{n\times n}M,$$

where $\mathbf{1}_{n\times n}$ is the all ones $n \times n$ matrix.

Then we can rescale $\tilde{A}$ as

$$A = \frac{1}{\sqrt{n-1}}\tilde{A}$$

In other words, A is nothing less than the input data, albeit properly normalized, i.e., centered and rescaled.

8.3 PCA via eigendecomposition

The sample covariance matrix $\Omega = A^t A$ encapsulates information about the statistical dependence of the feature values across samples. This a positive-semidefinite (hence symmetric) matrix that admits an orthonormal basis $\mathcal{V} = \{v_1, \ldots, v_m\}$ of eigenvectors with eigenvalues

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_m^2.$$

Therefore the following matrix identity is satisfied,

$$V^t \Omega V = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2),$$

where V is the orthogonal matrix that has the eigenvectors v_i as columns. Taking into account the expression of Ω, note that the identity can be written as follows:

$$V^t A^t A V = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2),$$

meaning that if we use the matrix V to generate a new dataset $A_{\text{PCA}} = AV$ where the features are linear combinations of the orginal features, the new features will satisfy the following:

1. Pairs of distinct features have sample covariance equal to zero, i.e., the new features are statistically independent from each other.

2. The variance of the i-th feature is equal to σ_i^2.

We will say that the vectors $v_1, \ldots, v_m$ are the *principal components* or *principal directions* of our dataset. The entries of each principal component are referred to as its *loadings*: the loadings represent the weight that each original feature has in the principal component. With the principal components we have a new coordinate system in which we can express the data vectors corresponding to a sample in our dataset: the coordinates of the centered feature values in the basis of principal components are referred to as the principal component *scores*: there is one score value per sample and principal component.

Proposition 32. *The principal component score for a sample s and principal component v is the length of the orthogonal*

projections of the vector of centered feature values of the sample s onto v.

Let's denote $\tilde{w}$ the vector of centered feature values and w the vector of principal value scores, which according to the preceding argument can be seen as the column vectors of $\tilde{A}^t$ and $(\tilde{A}V)^t = V^t\tilde{A}^t$, respectively. Therefore, the following relation holds:

$$w = V^t\tilde{w}.$$

The i-th component of w can be expressed as

$$w_i = v_i^t\tilde{w} = v_i \cdot \tilde{w}.$$

In other words, the scores are the coordinates of $\tilde{w}$ in the new basis of principal components.