STATS 235: Modern Data Analysis
Dimensionality Reduction

Babak Shahbaba

Department of Statistics, UCI

Outline

- In this lecture we discuss several dimensionality reduction methods such as principal component analysis (PCA), factor analysis (FA), and independent component analysis (ICA)
- We are mainly interested in unsupervised learning techniques for presenting high-dimensional data in low dimensional spaces, hoping to make the underlying structure in the data and patterns easier to see
- Very often, such unsupervised learning methods for dimensionality reduction are used a preprocessing step for supervised learning methods (discussed later) without taking the outcome variable into account
- Alternative supervised dimensionality reduction methods, such as partial least squares, are discussed later

Principal Component Analysis

Principal component analysis

- For a set of variables, we denote the centered matrix of of observed data as x
- The principal components are a set of orthonormal basis, $v_1, v_2, \ldots, v_p$, in the column space of x such that
  - v_1 is the basis with the largest sample variance
  - v_2 is the basis with the second largest sample variance that it is orthogonal to v_1
  - v_3 is the basis with the j^{th} largest sample variance that it is orthogonal to v_1, v_2
  - and so forth

Principal component analysis

- To find these principal components, we first find the eigenvectors of the covariance matrix $s = x^\top x/n$, and then order them based on the descending order of their corresponding eigenvalues, λ_j,

$$sv_j = \lambda_j v_j, \qquad j = 1, \ldots, p, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

- Recall that given a matrix s, eigenvectors are special vectors, v_j, such that we have $sv_j = \lambda v_j$ so v_j remains on the same line, but it either stretches, shrinks, reverses directions, or stays unchanged

- For each eigenvectors v, λ is the corresponding eigenvalue

Eigenvalues and eigenvectors (review)

- Suppose $s_{p\times p}$ has p independent eigenvectors, $v_1, \ldots, v_p$, which are the columns of an eigenvector matrix v
- The corresponding eigenvalues form a diagonal matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$
- Then, we can write them in a matrix form $sv = v\Lambda$, from which we can get

$$\begin{aligned} v^{-1}sv &= \Lambda \\ s &= v\Lambda v^{-1} \end{aligned}$$

- For symmetric matrices (such as covariance matrices), we have real eigenvalues and orthogonal eigenvector matrix. Therefore,

$$s = v\Lambda v^{\top}$$

Alternative view based on SVD

- Instead of explaining PCA based on finding eigenvectors of the square matrix $x^\top x$, we can explain it in terms of a singular value decomposition of x itself

$$x = u\Sigma v^\top$$

where u and v, called left and right singular vectors, are orthonormal and Σ, called singular values, is a diagonal matrix

- Then,

$$x^\top x = v\Sigma^\top u^\top u\Sigma v^\top = v\Sigma^\top \Sigma v^\top = v\Lambda v^\top$$

where Λ is a diagonal matrix that contains the eigenvalues of $x^\top x$

- Therefore, v contains the orthogonal eigenvectors of $x^\top x$, and the diagonal elements σ_i^2 are the positive eigenvalues of $x^\top x$

Singular value decomposition (SVD)

- Ordering $\sigma_1 \geq \ldots \geq \sigma_r > 0$, where r is the rank, we can write

$$x = u_1 \sigma_1 v_1^\top + \ldots + u_r \sigma_r v_r^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

- Left and right singular vectors are also known as Karhunen-Loève bases
- PCA is sometimes referred to as the Karhunen-Loève transformation

Principal component analysis

- After we find the eigenvectors v and order them according to their corresponding eigenvalues in decreasing order, we obtain a new set of derived variables, $z = (z_1, \ldots, z_p)$, as follows:

$$z = xv$$

- The columns of z are called *scores*, and the columns of v are called *loadings*

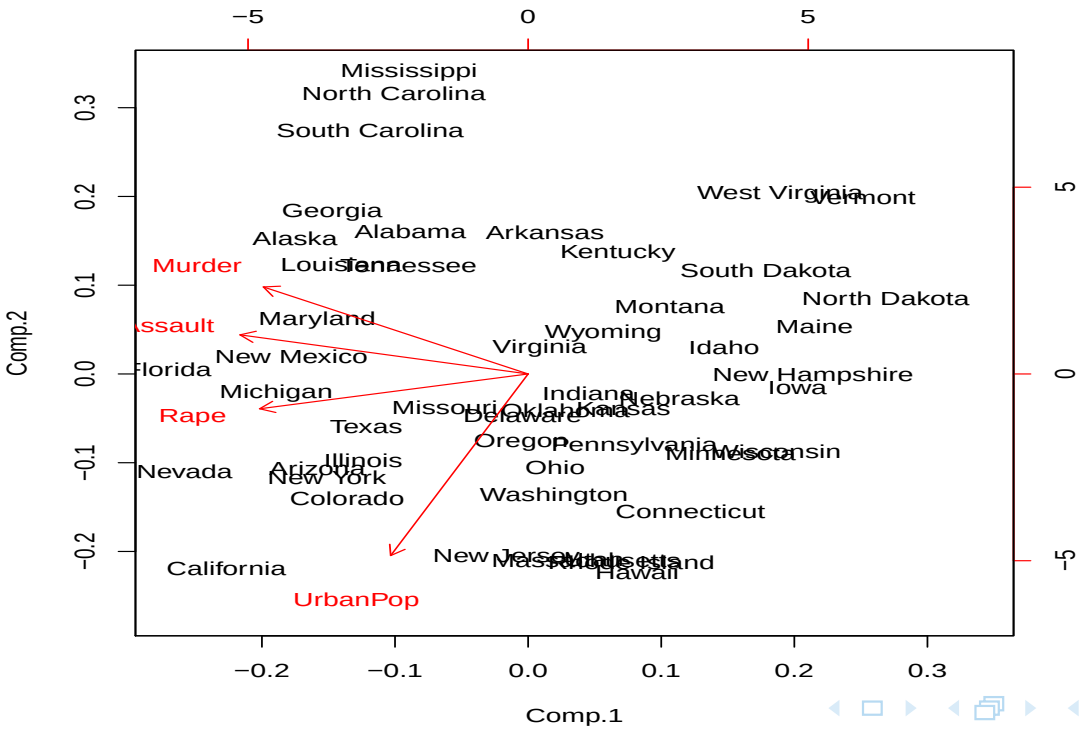
- Note that the sample variance of z_j, i.e., the j^{th} column of z, is

$$\mathrm{Var}(z_j) = z_j^T z_j / n = v_j^T x^T x v_j / n = v_j^T s v_j = \lambda_j$$

Therefore, the first column of z has the highest variance, the second column has the second highest variance, and so forth.

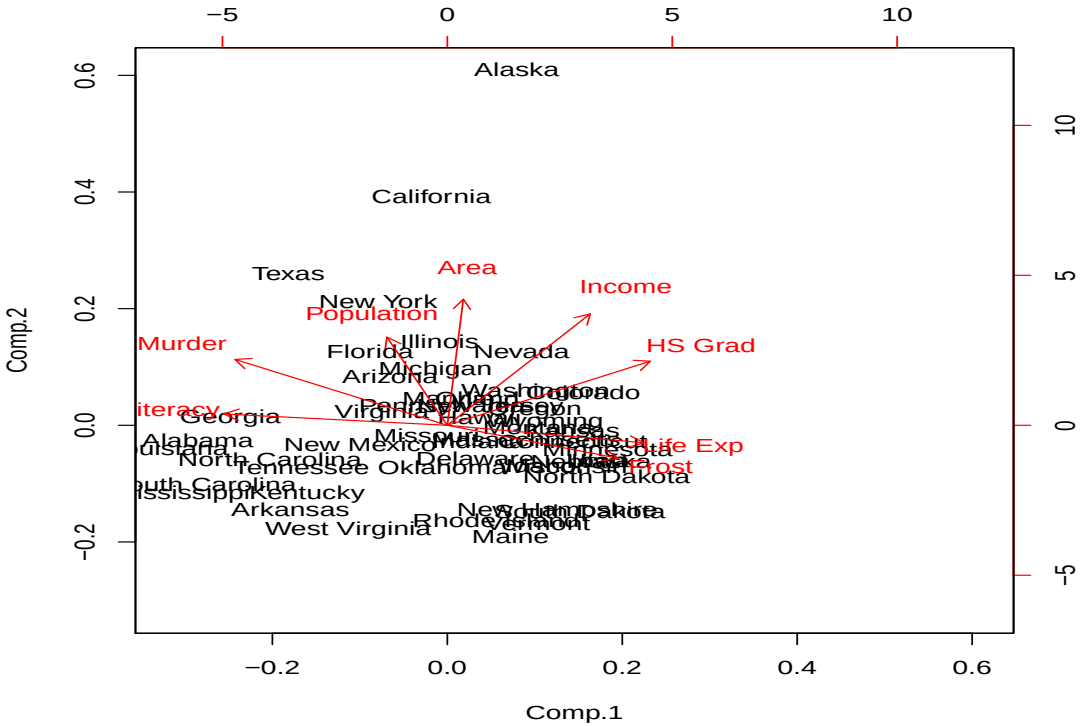
Example: US Arrests

- It is common to present the PCA results in a biplot (see my codes) using the first two principal components
- The following biplot shows the results based on the `USArrests` dataset, which is available in R, and according to its help file, it contains statistics in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states during 1973.



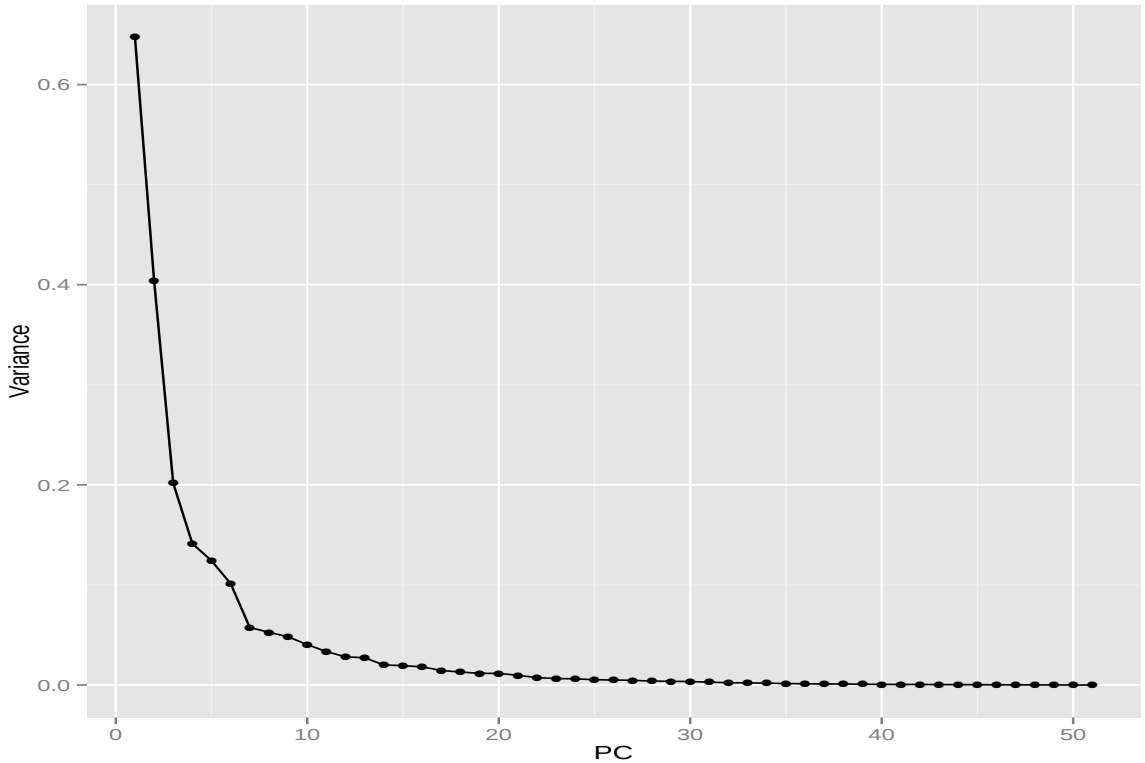
Example: States

- The following biplot is based on the `state.x77` dataset (also available in R)
- This dataset provides 8 different statistics (see the help file) on the 50 states of the United States of America



Example: Neurology data

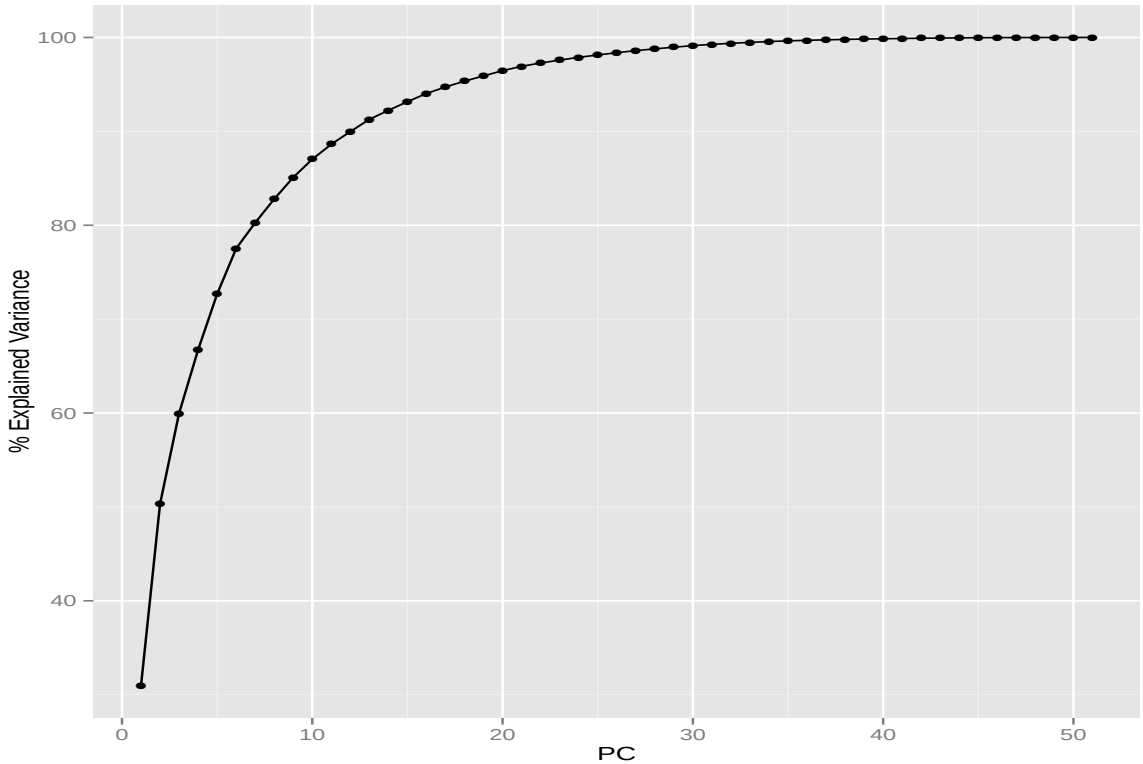
- To reduce dimensionality of x, we could choose the first q columns of z to represent the observed data
- For this, it is common to use the *scree* plot, which is the plot of all the eigenvalues (variances) in decreasing order
- The following is based on a neurology dataset (Burke et. al., 2014)



Example: Neurology data

- Alternatively, we can plot the cumulative percent of variance explained, which is calculated as follows:

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j'=1}^{p} \lambda_{j'}}$$



Factor Analysis

Factor Analysis (FA)

- PCA involves a standardized linear projection which maximizes the variance in the projected space (Hotteling, 1933; Tipping and Bishop, 1999)

- $v_1, \ldots, v_q$ are the top q dominant eigenvector (associated with the largest eigenvalues) of the sample covariance

- Using the corresponding scores z, we can approximate x,

$$\hat{x} = zv^\top$$

which minimizes the reconstruction error (error is zero if $q = p$)

$$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$$

- The main shortcoming of this approach is the lack of any underlying probabilistic model

Factor Analysis (FA)

- To address this issue, we use *Factor Analysis* (FA) models, which include *Probabilistic Principal Component Analysis* (PPCA) as a special case

- Instead of the above deterministic model, we use the following probabilistic model

$$\begin{aligned} x_i &= wz_i + \mu + \varepsilon_i \\ z_i &\sim N(0, I) \\ \varepsilon_i &\sim N(0, \Psi) \end{aligned}$$

- Here, z are q-dimensional latent variables

- w is a $p \times q$ matrix called the *factor loading matrix*

Factor Analysis (FA)

- Marginalizing over latent variables, we are in fact modeling the distribution of the observed data x as a multivariate normal

$$x_i \sim N(\mu, ww^\top + \Psi)$$

- That is, FA is a low rank representation of a multivariate normal distribution

- Note that this is similar to mixture models we discussed before, but instead of categorical latent factors we have continuous latent factors

- Therefore, we can use EM as before for parameter estimation

Factor Analysis (FA)

- For simplicity and identifiability, it is common to impose some structure (orthonormal, lower triangular, or sparse) on w
- Also, we typically make Ψ a diagonal matrix so given the latent variables, z, the original variables, x, become conditionally independent
- This way, the latent variables capture the correlation structure in the observed data
- That is, while PCA focuses on preserving the observed variance, FA focuses on preserving the observed correlation
- Finally, note that since this is a probabilistic model, we can make inference regarding the required number of latent factors using the usual χ^2 test for nested models based on likelihood ratio

Factor Analysis (FA)

- If we further simplify the model by setting $\Psi = \sigma^2 I$, we can write down the log-likelihood as follows:

$$\ell = -\frac{n}{2}[p \log(2\pi) + \log|c| + \mathrm{tr}(c^{-1}s)]$$

where s is the covariance matrix and

$$c = ww^\top + \sigma^2 I$$

- Tipping and Bishop (1999) showed that the MLE in this case is

$$\hat{w} = v_q(\Lambda_q - \sigma^2 I)^{1/2}$$

where $\Lambda_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ is a diagonal matrix of eigenvalues and the columns of v_q are the corresponding eigenvectors for s

Factor Analysis (FA)

- The MLE of σ is

$$\hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j$$

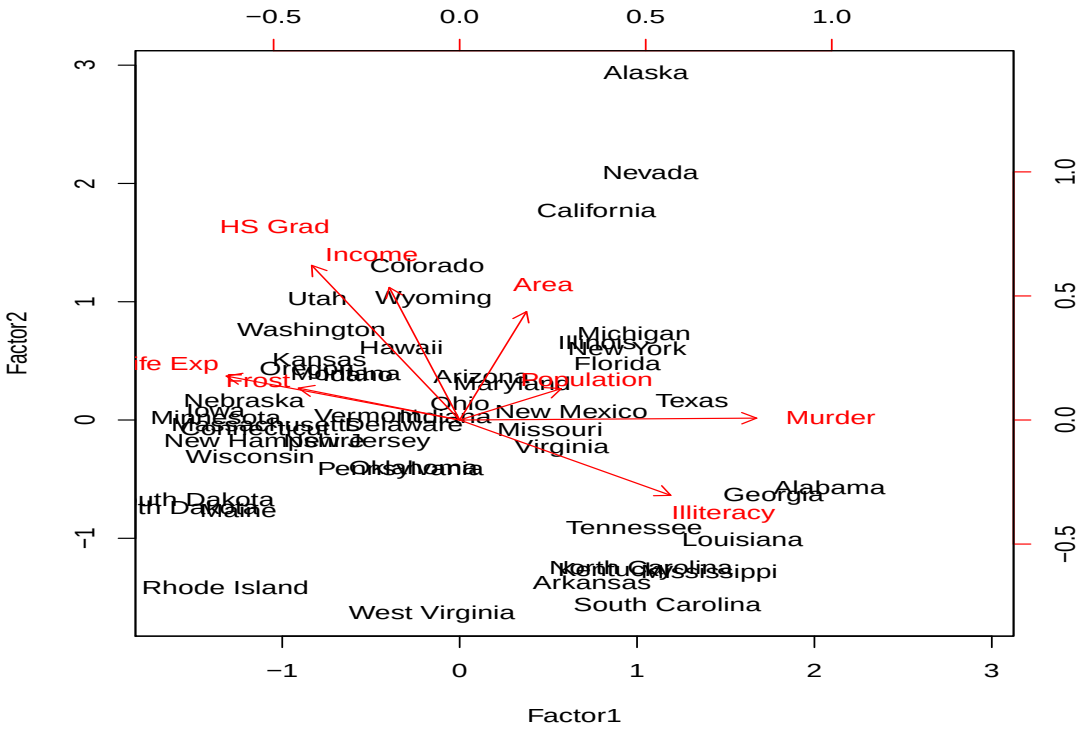
  which captures the average loss of variance through projection

- This approach is called PPCA

- As we can see, it reduces to standard PCA as $\sigma^2 \to 0$

Example: States

- The following biplot is based on factor analysis (two latent factors) of the `state.x77` dataset



Independent Component Analysis

Independent Component Analysis (ICA)

- The normality assumption for latent factors is very restrictive and could lead to unreasonable results
- We can relax this assumption and allow z to have any non-Gaussian distribution, but instead we assume that the components of z are independent so their joint distribution is separable,

$$P(z_i) = \prod_{j=1}^{q} P_j(z_{ij})$$

- Given z_i, we assume

$$x_i = wz_i + \varepsilon_i$$

- The resulting model is known as Independent Component Analysis
- This approach is commonly used in signal processing so z are usually called source variables and w called mixing matrix

Independent Component Analysis (ICA)

- We denote the centered (by subtracting the mean) and whitened (e.g., by using PCA) data as x
- Because the data are whitened, we have $\text{cov}(x) = I$; therefore, the mixing matrix is also orthogonal since $\text{cov}(z)$ is also I
- Our goal is to find the mixing matrix, w
- After we estimate w, assuming a noise-free model, we can use it to find the sources

$$z = vx$$

where $v = w^\top$

- To estimate w, we can use maximum likelihood estimation or EM (see section 12.6 of Murphy, 2012)
- However, it is more common to use entropy-based methods

Independent Component Analysis (ICA)

- For a continuous random variable y with density $f(y)$, the entropy (a.k.a., differential entropy) is defined as

$$H(y) = -\int f(y) \log f(y) dy$$

- It is well known that if we fix the variance, Gaussian variables have the highest entropy

- We can use the *mutual information* to measure the dependence among the components of z,

$$I(z) = \sum_j H(z_j) - H(z)$$

Independent Component Analysis (ICA)

- Our objective is to minimize $I(z)$
- We can show that this is equivalent to minimizing the sum of the entropies of individual components: $\sum_j H(z_j)$

Independent Component Analysis (ICA)

- Alternatively, we can measure the degree of non-Gaussianity of z_j in terms of the *negentropy* defined as follows (Hyvarinan and Oja, 2000):

$$J(z_j) = H(y_j) - H(z_j)$$

where y_j is a Gaussian random variable with the same mean and variance as z_j

- For highly non-Gaussian distributions, negentropy becomes large
- To find v, we maximize

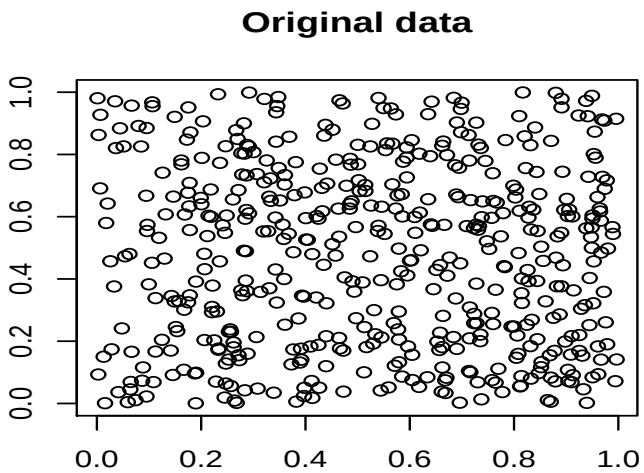
$$J(z) = \sum_j H(y_j) - H(z_j)$$

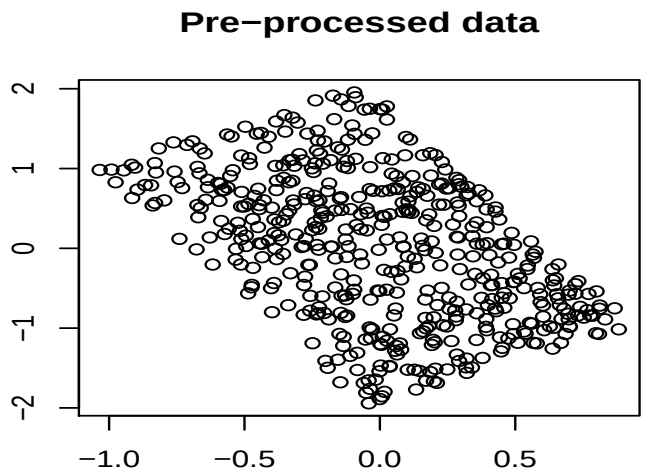
- We can show that this is equivalent to minimizing the mutual information and maximizing the likelihood (see Murphy, 2012)

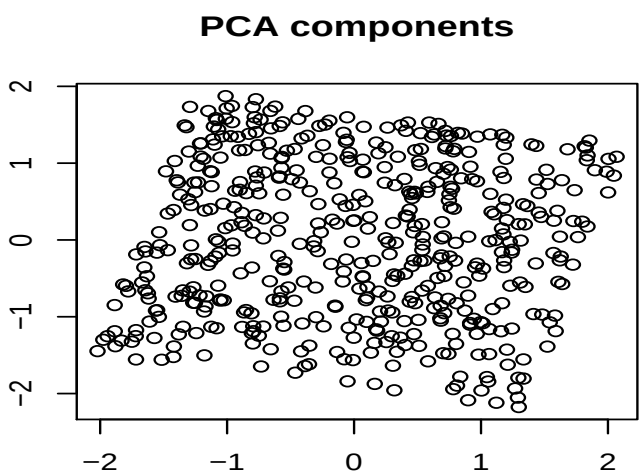
Independent Component Analysis (ICA)

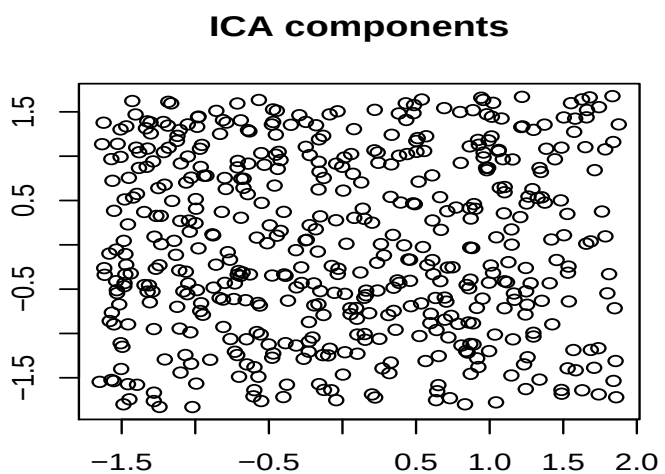
- Typical distributions for z_j are
  - *super-Gaussian*: with higher kurtosis than Gaussian, e.g., Laplace
  - *sub-Gaussian* with lower kurtosis than Gaussian, e.g., uniform
  - skewed distributions such as Gamma distribution
- However, we do not need to fully specify the shape of the distributions; we could instead specify how the distributions are deviating from Gaussian (e.g, sub-Gaussian or super-Gaussian)
- It is common to set $\log P(z)$ to $-\sqrt{z}$ or $-\log\cosh(z)$

Example: Two independent uniforms (from fastICA)









Example: Two independent signals (from fastICA)