Lecture 10

Jussi Klemelä

March 31, 2009

1 Transformations

In regression function estimation it is often useful to transform the explanatory variables. We discuss two transformations: data sphering and copula preprocessing.

Copula preprocessing. The copula transformation changes the marginal distributions but keeps the copula (the joint distribution) the same. The copula preprocessing of data matrix $\mathbb{X}_n = (x_i^j)$, $i = 1, \ldots, n$, $j = 1, \ldots, d$, is defined in two steps.

1. We make each margin approximately uniformly distributed by the following transformation: let z_i^j, $i = 1, \ldots, n$, $j = 1, \ldots, d$, be the number of observations smaller or equal to x_i^j, divided by n (z_i^j is the rank of x_i^j, divided by n):

$$z_i^j = n^{-1} \# \{x_l^j : x_l^j \leq x_i^j, l = 1, \ldots, n\}.$$

2. When $X \sim \text{Unif}([0,1])$, then $F^{-1}(X) \sim F$, where F is any continuous distribution function. We have made in step 1 each margin approximately uniformly distributed, and next we can make the margins to be approximately normally distributed by defining

$$y_i^j = \Phi^{-1}(z_i^j), \qquad i = 1, \ldots, n, \ j = 1, \ldots, d,$$

where Φ is the distribution function of the standard Gaussian distribution. The copula preprocessed data matrix is $\mathbb{Y}_n = (y_i^j)$, $i = 1, \ldots, n$, $j = 1, \ldots, d$.

Data sphering. We can make the scales of variables compatible by normalizing each column of the data matrix to have unit variance. Data sphering is more extensive transformation; we make such linear transformation of data that the covariance matrix becomes the identity matrix. The sphering is almost the same as the principal component transformation. In the principal component transformation the covariance matrix is diagonalized but it is not made the identity matrix.

1. Sphering of a random vector $X \in \mathbf{R}^d$ means that we make a linear transform of X so that the new random variable has expectation zero and the identity covariance matrix. Let

$$\Sigma = E(X - EX)^T(X - EX)$$

be the covariance matrix and make the spectral representation of Σ:

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

where A is orthogonal and Λ is diagonal. Then

$$Y = \Lambda^{-1/2}A^T(X - EX)$$

is the sphered random vector. Indeed,

$$\mathrm{Cov}(Y) = \Lambda^{-1/2}A^T\mathrm{Cov}(X)A\Lambda^{-1/2} = \Lambda^{-1/2}A^T\Sigma A\Lambda^{-1/2} = I_d.$$

2. Data sphering of the data means that the data matrix is transformed so that the arithmetic mean of each column is zero and the empirical covariance matrix is the unit matrix. Let Σ_n be the empirical covariance matrix,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)^T,$$

where $\bar{X}_n$ is the $d \times 1$ column vector of arithmetic means. We find the spectral representation of Σ_n,

$$\Sigma_n = A_n\Lambda_n A_n^T,$$

where Λ_n is a diagonal matrix. The sphered data matrix is

$$\mathbb{Y}_n = \left(\mathbb{X}_n - 1_{n\times 1}\bar{X}_n^T\right) \times A_n\Lambda_n^{-1/2},$$

where $\mathbb{X}_n$ is the original $n \times d$ data matrix, and $1_{n\times 1}$ is the $n \times 1$ column vector of ones.

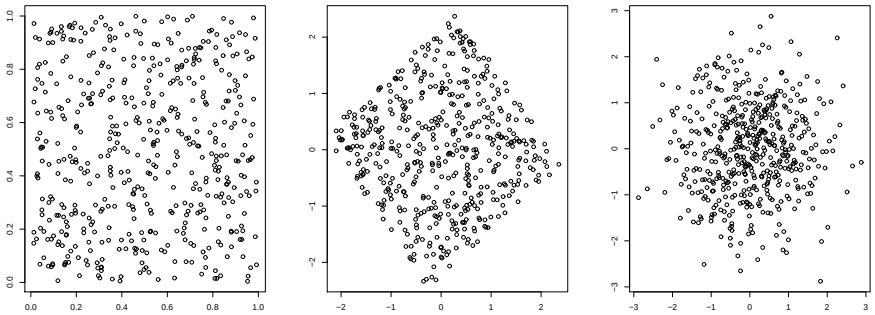

Figure 1: *(Independent uniform marginals.)* The figure illustrates data preprocessing with data whose marginals are independent and uniformly distributed. Frame a) shows a scatter plot of the data, frame b) shows the sphered data, and frame c) shows the copula transformed data.

Illustrations We illustrate data sphering and the copula transform.

1. Figure 1(a) shows a scatter plot of a simulated sample of size 500, where the marginals are uniform on $[0, 1]$ and independent from each other. Frame b) shows sphered data and frame c) shows copula transformed data, where the marginals are approximately standard Gaussian. The data in frame c) is distributed as standard 2D Gaussian.

2. Figure 2(a) shows a scatter plot of exchange rates of Brazilian Real and Mexican new Peso between 1995-01-05 and 2007-09-26. The rates are with respect to one U.S. Dollar and transformed to returns $(r_i \mapsto (r_i - r_{i-1})/r_{i-i})$. There are 3197 observations. Frame b) shows sphered data and frame c) shows copula transformed data, where the marginals are approximately standard Gaussian. The data is provided by Federal Reserve Economic Data (http://research.stlouisfed.org).

3. Figure 3(a) shows a scatter plot of the returns of the German stock index DAX and the French stock index CAC between 1990-01-05 and 2008-01-14. There are 4277 observations. Frame b) shows sphered data and frame c) shows copula transformed data, where the marginals are approximately standard Gaussian. The marginals appear to be almost independent. The data is provided by Yahoo.

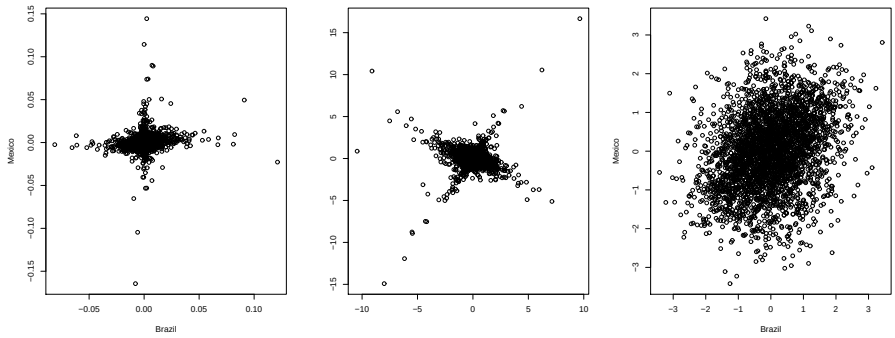

Figure 2: *(Exchange rates.)* The figure illustrates data preprocessing with data of exchange rates of Brazilian Real and Mexican Peso (n=3197). Frame a) shows a scatter plot of the data, frame b) shows the sphered data, and frame c) shows the copula transformed data.

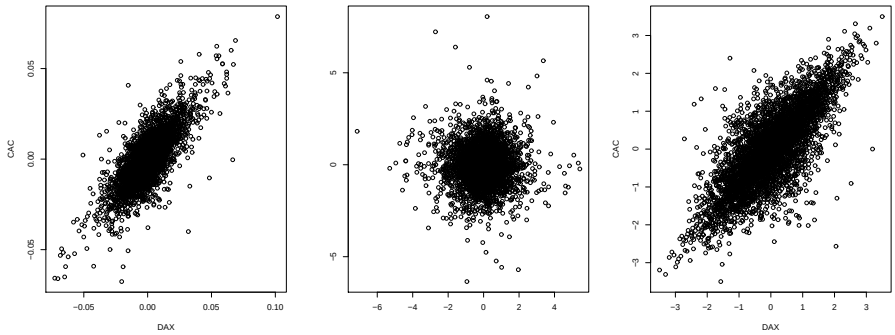

Figure 3: *(Stock indexes.)* The figure illustrates data preprocessing with data of the German stock index DAX and the French stock index CAC (n=4277). Frame a) shows a scatter plot of the data, frame b) shows the sphered data, and frame c) shows the copula transformed data.

2 Empirical risk minimization

Let $\mathcal{F}$ be a class of functions $\mathbf{R}^d \to \mathbf{R}$ and let $\epsilon > 0$. We define the empirical risk minimizer $\hat{f} : \mathbf{R}^d \to \mathbf{R}$ to be such that

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon,$$

where the most common definition for γ_n is the sum of squared errors:

$$\gamma_n(g) = \sum_{i=1}^{n} (Y_i - g(X_i))^2, \qquad g : \mathbf{R}^d \to \mathbf{R}.$$

More generally, we can define

$$\gamma_n(g) = \sum_{i=1}^{n} \gamma(Y_i, g(X_i)).$$

Examples of the contrast function γ include

1. The power functions:
$$\gamma(y, z) = |y - z|^p,$$
for $p \geq 1$.

2. The ϵ-sensitive loss function
$$\gamma(y, z) = I_{[\epsilon, \infty)}(|y - z|)(|y - z)| - \epsilon),$$
for $\epsilon > 0$.

The linear least squares estimator is obtained by choosing

$$\mathcal{F} = \{\beta_0 + \beta^T x : \beta_0 \in \mathbf{R}, \;\; \beta_1 \in \mathbf{R}^d\}.$$

2.1 Local empirical risk

Local constant estimator We can define the weighted empirical risk by

$$\gamma_n(\theta, x) = \sum_{i=1}^{n} p_i(x)(Y_i - \theta)^2, \qquad \theta \in \mathbf{R},$$

where the weights $p_i(x)$ can be chosen as the kernel weights

$$p_i(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)},$$

where $K : \mathbf{R}^d \to \mathbf{R}$ is a kernel function, $K_h(x) = K(x/h)/h^d$, and $h > 0$ is the smoothing parameter. Let

$$\hat{f}(x) = \text{argmin}_\theta \gamma_n(\theta, x), \qquad x \in \mathbf{R}^d.$$

The solution to the minimization problem is

$$\hat{f}(x) = \sum_{i=1}^n p_i(x) Y_i, \qquad x \in \mathbf{R}^d.$$

The is estimator is identical to the kernel estimator.

Local linear estimator A local linear estimator is

$$\hat{f}(x) = \hat{\alpha}(x) + \hat{\beta}(x)^T x, \qquad x \in \mathbf{R}^d,$$

where $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ are defined by

$$(\hat{\alpha}(x), \hat{\beta}(x)) = \text{argmin}_{\alpha \in \mathbf{R}, \beta \in \mathbf{R}^d} \gamma_n(\alpha, \beta, x),$$

where

$$\gamma_n(\alpha, \beta, x) = \sum_{i=1}^n p_i(x) \left[Y_i - \alpha - \beta^T X_i \right]^2, \;\; \alpha \in \mathbf{R}, \;\; \beta \in \mathbf{R}^d, \;\; x \in \mathbf{R}^d.$$

We can find an explicite expression for $\hat{\alpha}(x)$ and $\hat{\beta}(x)$, similarily as in the case of linear regression. Let us denote by $\mathbf{X}$ the $n \times (d+1)$-matrix whose ith row is $(1, X_i^T)$, where X_i is interpreted as a column vector of length d. Let $\mathbf{y}$ be the column vector of length n whose ith element is Y_i. Let $W(x)$ be the $n \times n$ diagonal matrix with diagonal elements $p_i(x)$. Then

$$\hat{b}(x) = (\hat{\alpha}(x), \hat{\beta}(x)^T)^T = (\mathbf{X}^T W(x) \mathbf{X})^{-1} \mathbf{X}^T W(x) \mathbf{y}.$$

Then we can write

$$\hat{f}(x) = \sum_{i=1}^n q_i(x) Y_i,$$

for certain weights $q_i(x) \geq 1$, $\sum_{i=1}^n q_i(x) = 1$. Indeed,

$$q_i(x) = p_i(x) \, \frac{s_2(x) - s_1(x) X_i}{s_2(x) - s_1^2(x)},$$

where

$$s_k(x) = \sum_{i=1}^n p_i(x) X_i^k, \qquad k = 0, 1, 2, \ldots.$$

3 Examination

Possible questions in the examination:

10) Define the copula transform.

11) Define the locally linear estimator.