Statistical learning

Assignment 5

1. For $n = 50$ and $p \in \{100, 500, 1000, 5000\}$ generate 100 independent replicates of the design matrix $X_{n\times p}$ according to the formula

$$F_{n\times 5}: \quad \text{for } i \in \{1, \ldots, 5\} \;\; F_{,i} \sim N(0, I_{n\times n})$$

$$W_{p\times 5}: W_{,j} = (6-j)N(0, I_{p\times p)}$$

$$E_{n\times p}: E_{ij} \sim N(0, \sigma = 10)$$

$$X = FW^T + E$$

.

For each combination of n and p plot the histogram of the dimensions selected by PESEL.

2. For $n = 50$, $p = 800$ and $K = 4$ generate the design matrix $X_{n\times p}$ according to the formula

$$X = [X^1, \ldots, X^K] \;\; ,$$

where each X^i is of the dimension $n \times 200$ and is obtained as

$$X^i = F^i W^{iT} + E^i$$

$$F^i_{n\times 3}: \quad \text{for } j \in \{1, \ldots, 3\} \;\; F^i_{,j} \sim N(0, I_{n\times n})$$

$$W^i_{200\times 3}: W^i_{,j} = (4-j)N(0, I_{p\times p)}$$

$$E^i_{n\times 200}: E^i_{uj} \sim N(0, \sigma = 1) \;\; .$$

   a) Use PESEL to estimate the rank of X and calculate all important principal components of X. Critically analyze their loadings.

   b) Use spca function to calculate first k Sparse Principle Components of X, where k is the number of PCs selected by PESEL in point a). Limit the number of variables selected for each PCs to 10. Critically analyze the loadings of these sparse PCs and compare to the loadings obtained in point a).

   c) Use *varclust* to cluster the columns of X. Estimate the number of clusters, their dimensions, and calculate the values of ARI, Integration and Acontamination.

3. Generate $X_1, \ldots, X_{100}$ random vectors from the multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where the covariance matrix Σ is a 30×30 block-diagonal matrix consisting of three diagonal blocks of dimension 10×10. In each block, the diagonal elements are equal to 1, and all off-diagonal elements are equal to 0.7.

   Estimate the precision matrix using:

   1. The inverse of the sample covariance matrix.
   2. Graphical LASSO (gLASSO), with the tuning parameter adjusted to minimize the Mean Squared Error (MSE).

   Compare the MSE for these two approaches.

Malgorzata Bogdan