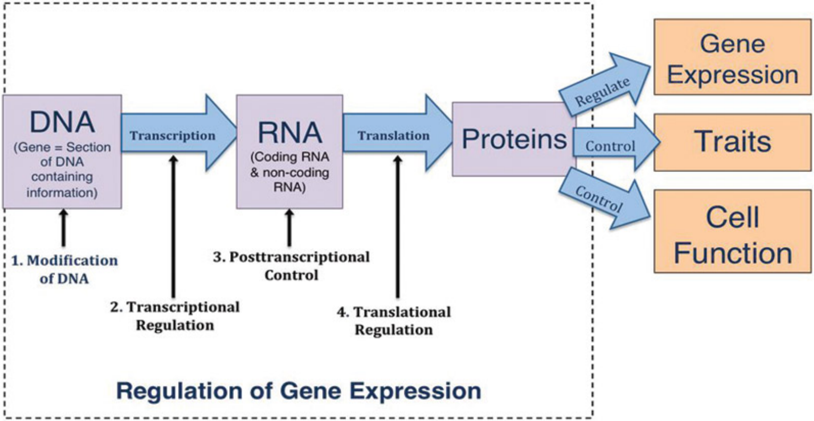
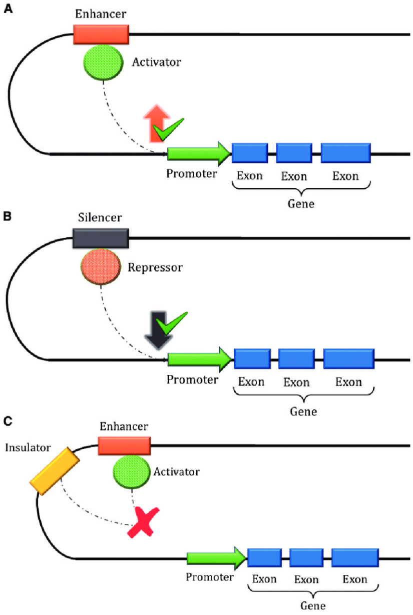
SVD and PCA in HD

December 5, 2023

Motivation - Gene Expression


DNA
(Gene = Section of DNA containing information)
Transcription
RNA
(Coding RNA & non-coding RNA)
Translation
Proteins
Regulate
Control
Control
Gene Expression
Traits
Cell Function
1. Modification of DNA
2. Transcriptional Regulation
3. Posttranscriptional Control
4. Translational Regulation
Regulation of Gene Expression

Transcription factors



PCA - reduction of dimensionality of ”omics” data

$X_{n \times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes

PCA - reduction of dimensionality of ”omics” data

$X_{n \times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes

PCA - reduction of dimensionality of ”omics” data

$X_{n \times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes

Assumptions : $X = M + E$, where M is of a low rank and E is a random noise

PCA - reduction of dimensionality of ”omics” data

$X_{n\times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$,
$p \approx 20000$ - number of genes

Assumptions : $X = M + E$, where M is of a low rank and E is a random noise

We usually assume that $e_{ij} \sim N(0, \sigma)$

PCA - reduction of dimensionality of ”omics” data

- $X_{n\times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes
- Assumptions : $X = M + E$, where M is of a low rank and E is a random noise
- We usually assume that $e_{ij} \sim N(0, \sigma)$
- Mathematical goal - recovering M, separation of the signal from noise

PCA - reduction of dimensionality of ”omics” data

$X_{n \times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes

Assumptions : $X = M + E$, where M is of a low rank and E is a random noise

We usually assume that $e_{ij} \sim N(0, \sigma)$

Mathematical goal - recovering M, separation of the signal from noise

Practical goal - data compression, several basis vectors [Principal Components] may contain most of the information and be applied for prediction (of the patient’s response to the therapy)

Principal Components Analysis (2)

Method - Singular Value Decomposition:

$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \quad ,$$

$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \;\; l = min\{n,p\},$ and D is the diagonal matrix

Principal Components Analysis (2)

Method - Singular Value Decomposition:

$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \quad ,$$

$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \;\; l = min\{n, p\},$ and D is the diagonal matrix

Eckhart-Young theorem: SVD truncated to k largest singular values (i.e. setting $D_{k+1,k+1} = \ldots = D_{l,l} = 0$) is the best rank k approximation to X in the Frobenius norm, i.e.

$$\hat{M}_k = U_{n\times k} D_{k\times k} V^T_{k\times p} = \underset{M:rank(M)=k}{\arg\min} \sum_{ij} (X_{ij} - M_{ij})^2$$

Principal Components Analysis (2)

Method - Singular Value Decomposition:

$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \quad ,$$

$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \quad l = min\{n, p\},$ and D is the diagonal matrix

Eckhart-Young theorem: SVD truncated to k largest singular values (i.e. setting $D_{k+1,k+1} = \ldots = D_{l,l} = 0$) is the best rank k approximation to X in the Frobenius norm, i.e.

$$\hat{M}_k = U_{n\times k} D_{k\times k} V^T_{k\times p} = \underset{M:rank(M)=k}{\arg\min} \sum_{ij} (X_{ij} - M_{ij})^2$$

$\hat{M}_k$ is the maximum likelihood estimator of M under the restriction $rank(M) = k$

Principal Components Analysis (2)

Method - Singular Value Decomposition:

$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \quad ,$$

$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \quad l = min\{n, p\},$ and D is the diagonal matrix

Eckhart-Young theorem: SVD truncated to k largest singular values (i.e. setting $D_{k+1,k+1} = \ldots = D_{l,l} = 0$) is the best rank k approximation to X in the Frobenius norm, i.e.

$$\hat{M}_k = U_{n\times k} D_{k\times k} V^T_{k\times p} = \underset{M:rank(M)=k}{\arg\min} \sum_{ij} (X_{ij} - M_{ij})^2$$

$\hat{M}_k$ is the maximum likelihood estimator of M under the restriction $rank(M) = k$

Statistical Goal - determining rank k of matrix M

PESEL (PEnalized SEmi-integrated Likelihood)

Sobczyk, Bogdan, Josse, Journal of Computational Graphical Statistics, 2017

Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - nested sequence of statistical models

Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - nested sequence of statistical models

Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - nested sequence of statistical models

In our example A_k - $rank(M) \leq k$

Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - nested sequence of statistical models

In our example A_k - $rank(M) \leq k$

θ - vector of parameters of A_k:

eleements of $U_k \in S_{k,n}$, $V_k \in S_{k,p}$, D_k, i σ

$S_{k,n}$ - Stiefel manifold of orthonormal matrices of dimension $n \times k$

Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - nested sequence of statistical models

In our example A_k - $rank(M) \leq k$

θ - vector of parameters of A_k:

eleements of $U_k \in S_{k,n}$, $V_k \in S_{k,p}$, D_k, i σ

$S_{k,n}$ - Stiefel manifold of orthonormal matrices of dimension $n \times k$

$l(X, \theta)$ - likelihood function (density of the distribution describing the data)

Bayesian Information Criterion (BIC) (2)

In general situation BIC suggests selecting the model maximizing

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

where N is the number of independent observations.

Bayesian Information Criterion (BIC) (2)

In general situation BIC suggests selecting the model maximizing

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

where N is the number of independent observations.

BIC is justified (consistent) where $dim(A_k) = const$ when $N \to \infty$

Bayesian Information Criterion (BIC) (2)

In general situation BIC suggests selecting the model maximizing

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

where N is the number of independent observations.

BIC is justified (consistent) where $dim(A_k) = const$ when $N \to \infty$

In our case $N = np$, so $dim(A_k)$ increases with n and p

Bayesian Information Criterion (BIC) (2)

In general situation BIC suggests selecting the model maximizing

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

where N is the number of independent observations.

BIC is justified (consistent) where $dim(A_k) = const$ when $N \to \infty$

In our case $N = np$, so $dim(A_k)$ increases with n and p

Idea - reduction of the number of parameters by integrating them out with respect to some prior distribution

PESEL for large p

Assume that $M = TW^T$, where
$T = [t_{i,l}]_{n \times k}$ is the matrix of ”hidden factors”,
$W = [w_{i,l}]_{p \times k}$ is the matrix of coefficients

PESEL for large p

Assume that $M = TW^T$, where
$T = [t_{i,l}]_{n \times k}$ is the matrix of ”hidden factors”,
$W = [w_{i,l}]_{p \times k}$ is the matrix of coefficients

prior distribution -

$$w_{j\cdot} \sim N(0, I_k) \quad ,$$

which implies, that $x_{\cdot 1}, \ldots, x_{\cdot p}$ są are iid random vectors from the distribution

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

PESEL for large p

Assume that $M = TW^T$, where
$T = [t_{i,l}]_{n \times k}$ is the matrix of ”hidden factors”,
$W = [w_{i,l}]_{p \times k}$ is the matrix of coefficients

prior distribution -

$$w_{j\cdot} \sim N(0, I_k) \quad ,$$

which implies, that $x_{\cdot 1}, \ldots, x_{\cdot p}$ są are iid random vectors from the distribution

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

Now we have p independent vectors and the number of parameters does not depend on p - we can apply BIC if only $p >> n$

PESEL for large p

Assume that $M = TW^T$, where
$T = [t_{i,l}]_{n \times k}$ is the matrix of ”hidden factors”,
$W = [w_{i,l}]_{p \times k}$ is the matrix of coefficients

prior distribution -

$$w_{j\cdot} \sim N(0, I_k) \quad ,$$

which implies, that $x_{\cdot 1}, \ldots, x_{\cdot p}$ są are iid random vectors from the distribution

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

Now we have p independent vectors and the number of parameters does not depend on p - we can apply BIC if only $p >> n$

The related criterion is called PESEL (Penalized SEmi-integrated Likelihood).

Consistency of PESEL

Let $\hat{k}_0(p, n)$ be the PESEL estimator of the rank of M.

Consistency of PESEL

Let $\hat{k}_0(p, n)$ be the PESEL estimator of the rank of M.

Let $\frac{1}{p}MM' \to L$ with $rank(L) = k_0$.

Consistency of PESEL

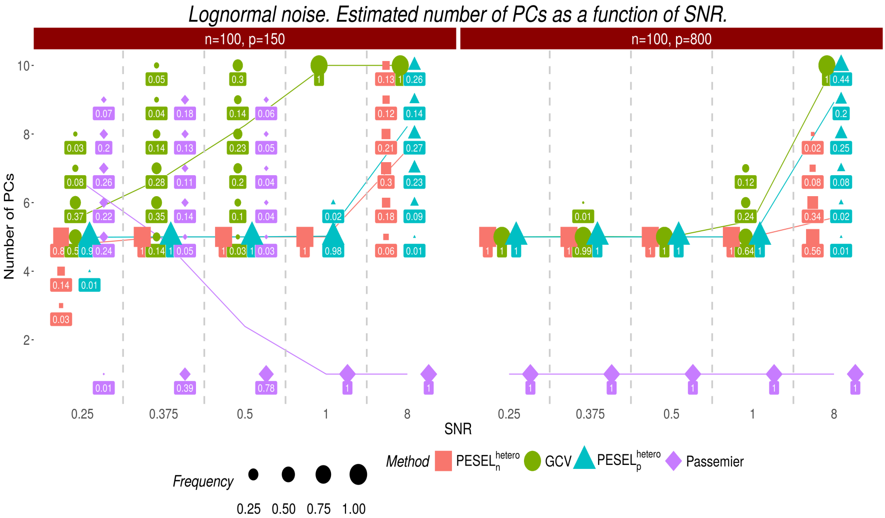
Let $\hat{k}_0(p, n)$ be the PESEL estimator of the rank of M.

Let $\frac{1}{p}MM' \to L$ with $rank(L) = k_0$.

Then, for n fixed and under mild regularity conditions it holds

$$P(\exists p_0 \; \forall p > p_0 \quad \hat{k}_0(p, n) = k_0) = 1.$$

Errors from the log-normal distribution



Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Graczyk, Josse, Panloup, Seegers, 2020

Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Graczyk, Josse, Panloup, Seegers, 2020

Awards for young scientists P. Sobczyk (Vienna workshop on simulation, 2015) i S. Wilczyński (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Graczyk, Josse, Panloup, Seegers, 2020

Awards for young scientists P. Sobczyk (Vienna workshop on simulation, 2015) i S. Wilczyński (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Goal: Identification of groups of co-regulated variables (genetic pathways) and selection of appropriate Principal Components.

.

Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Graczyk, Josse, Panloup, Seegers, 2020

Awards for young scientists P. Sobczyk (Vienna workshop on simulation, 2015) i S. Wilczyński (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Goal: Identification of groups of co-regulated variables (genetic pathways) and selection of appropriate Principal Components.

Mathematics: clustering of variables into groups, such that each of them is spanned by just few of ”hidden” variables.

Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Graczyk, Josse, Panloup, Seegers, 2020

Awards for young scientists P. Sobczyk (Vienna workshop on simulation, 2015) i S. Wilczyński (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Goal: Identification of groups of co-regulated variables (genetic pathways) and selection of appropriate Principal Components.

Mathematics: clustering of variables into groups, such that each of them is spanned by just few of ”hidden” variables.

Package *varclust* by P. Sobczyk and S. Wilczyński- Algorithm K-centroids around PCs. Estimation of the number of clusters and their dimensions by modifications of BIC.

Methodology

K-centroids algorithm

Methodology

K-centroids algorithm

Centers - PCs, distance - BIC

Methodology

K-centroids algorithm

Centers - PCs, distance - BIC

Estimation of clusters dimensions by PESEL

Methodology

K-centroids algorithm

Centers - PCs, distance - BIC

Estimation of clusters dimensions by PESEL

Repeat for different K and estimate K by mBIC

Overview of the algorithm

Algorithm: Multiple Latent Clustering Components

```
Input: $n$ - number of individuals, $p$ - number of variables,
  $X_{n \times p} = (x_1, \ldots, x_p)$ - data set, $d$ - maximal subspace
  dimension, $N$ - number of runs of the algorithm
  Scale $X$ to have columns with mean 0 and unit variance
  for $i \in \{1, \ldots, N\}$ do
     Find the model using K-means and store its value of mBIC
  end for
  Choose the model with the highest value of mBIC and return
  the model (segmentation, mBIC, factors) as the result. =0
```

K-means step

1. Initialize clusters’ centres

K-means step

1. Initialize clusters' centres
2. Until convergence or maximal number of iterations is reached repeat:

K-means step

1. Initialize clusters' centres
2. Until convergence or maximal number of iterations is reached repeat:
   - For every variable x_j and every cluster factors $F_{j'}$ fit a linear regression model without intercept $lm(x_j \sim F_{j'})$ and store BIC value as $BIC_{jj'}$
   - Assign variable x_j to the cluster M_q where

$$q = \underset{j' \in \{1,\ldots,K\}}{\arg\max}\, BIC_{jj'}$$

K-means step

1. Initialize clusters' centres
2. Until convergence or maximal number of iterations is reached repeat:
   - For every variable x_j and every cluster factors $F_{j'}$ fit a linear regression model without intercept $lm(x_j \sim F_{j'})$ and store BIC value as $BIC_{jj'}$
   - Assign variable x_j to the cluster M_q where
     $$q = \underset{j' \in \{1,\ldots,K\}}{\arg\max}\ BIC_{jj'}$$
   - For every cluster M_i use PESEL to estimate its dimensionality k_i with an upper bound of d. Use PCA to compute the first k_i principal components and store them in F_i

Convergence (1)

Both the partition (BIC) and the center's selection (PESEL) steps lead to increase of Laplace approximations to the model posterior probability.

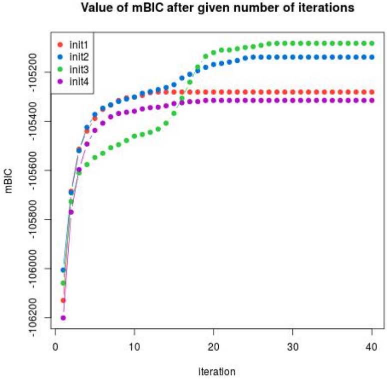
Convergence (1)

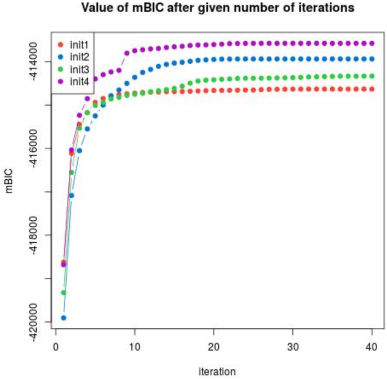
Both the partition (BIC) and the center's selection (PESEL) steps lead to increase of Laplace approximations to the model posterior probability.

For sufficiently large *n* and *p* each step of the algorithm leads to increase of mBIC.

Convergence (2)

$p = 750 p = 3000$





Informative prior distribution and mBIC

- The problem with BIC (non-informative prior)
- Prior distribution taking into account the number of clusters and maximal dimension of the subspace

$$P(M) = \frac{1}{K^p}\frac{1}{d^K}$$

$$mBIC = \sum_{i=1}^{K} \ln\left(\widehat{P}(X_i|M_i)\right) - p\ln(K) - K\ln(d)$$

Informative prior distribution and mBIC

- The problem with BIC (non-informative prior)
- Prior distribution taking into account the number of clusters and maximal dimension of the subspace

$$P(M) = \frac{1}{K^p} \frac{1}{d^K}$$

$$mBIC = \sum_{i=1}^{K} \ln\left(\widehat{P}(X_i|M_i)\right) - p\ln(K) - K\ln(d)$$

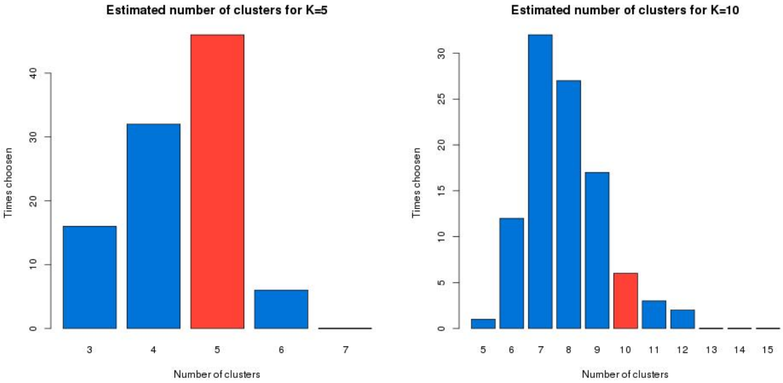
Application

mBIC can be used to compare different models and to choose the number of clusters in the data.

mBIC

Figure: Estimation of the number of clusters. Simulation parameters: $n = 100,\ p = 600,\ d = 3,\ SNR = 1\ mode : not\ shared$.

$K \qquad = \qquad 5K \qquad = \qquad 10$



Compared methods

1. **Sparse Subspace Clustering** (SSC)
2. Low Rank Subspace Clustering (LRSC)
3. **MLCC with random initialization** (MLCC)
4. **MLCC with initialization by the result of SSC** (MLCC_{aSSC})
5. MLCC with initialization by sparse PCA (MLCC_{sPCA})
6. ClustOfVar (COV)

Data generation - shared factors

```
**Input:** $n$, $SNR$, $K$, $p$, $d$
  Number of factors $m \leftarrow K\frac{d}{2}$
  Factors $F = (f_1, \ldots, f_m)$, $f_i \sim N(0, I_n)$
  Draw subspaces' dimension $d_1, \ldots d_K$ uniformly from $\{1, \ldots, d\}$
  **for** $i = 1, \ldots, K$ **do**
    $F_i \leftarrow$ sample of size $d_i$ from columns of $F$
    Draw matrix of coefficients $C_i$ from $U(0.1, 1) \cdot sgn(U(-1, 1))$
    Variables in the $i$-th subspace are $X_i \leftarrow F_i C_i$
  **end for**
  Scale matrix $X = (X_1, \ldots, X_K)$ (columns with unit variance)
  return $X + Z$ where $Z \sim N(0, \frac{1}{SNR} I_n)$ =0
```

Data generation - independent subspaces

Remark

To generate data without shared factors we draw independently i-th subspaces basis F_i as sample of size d_i from standard multivariate normal distribution

Measures of effectiveness

Compare two partitions $A = (A_1, \ldots A_n),\ B = (B_1, \ldots, B_m)$

- Adjusted Rand Index (ARI)
- Integration
- Acontamination
- ARI $\in [-1, 1]$, Integration, Acontamination $\in [0, 1]$.

Measures of effectiveness

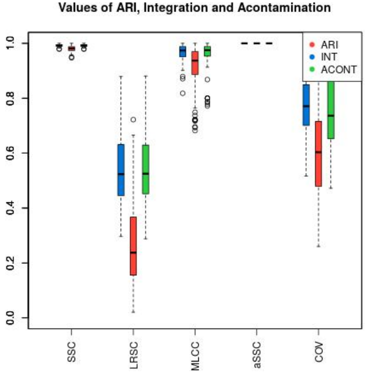
Compare two partitions $A = (A_1, \dots A_n),\ B = (B_1, \dots, B_m)$

- Adjusted Rand Index (ARI)
- Integration
- Acontamination
- ARI $\in [-1, 1]$, Integration, Acontamination $\in [0, 1]$.

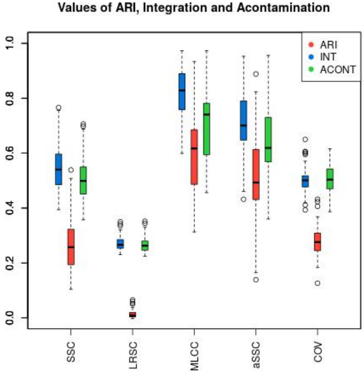
Remark

The bigger the indices, the better the clustering.

Mode

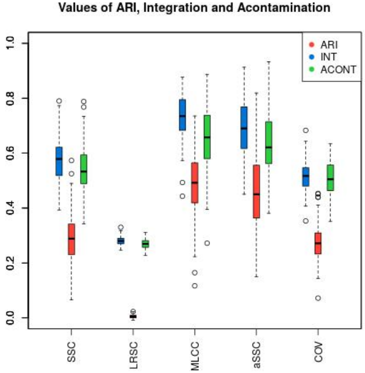


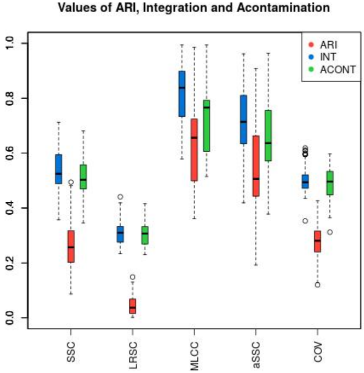
Values of ARI, Integration and Acontamination



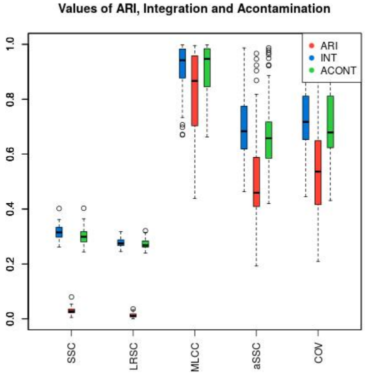
Values of ARI, Integration and Acontamination

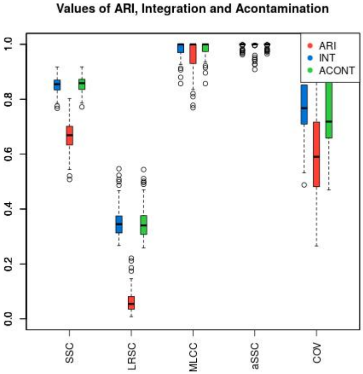
Number of variables





Signal to noise ratio

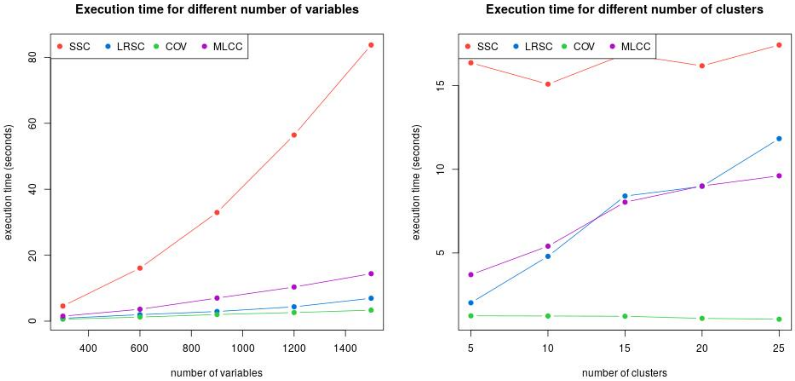




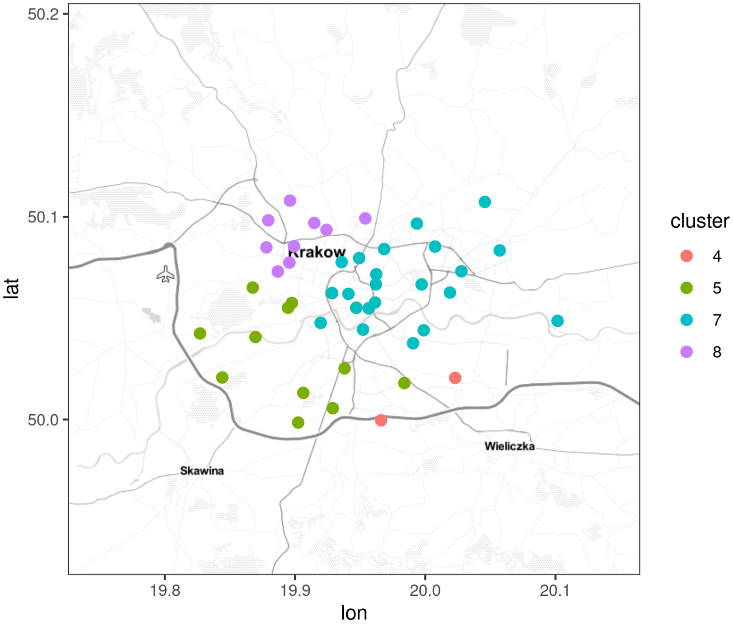
Execution time

Figure: Comparison of the execution time of the methods with respect to p and K. Simulation parameters:$n = 100,\ d = 3,\ SNR = 1\ mode : shared$.

With respect to theWith respect to the
number of variablesnumber of clusters



Particle matter measurements clustered using varclust



Sparse PCA (1)

$$\text{maximize} \;\; a'\Sigma a, \;\text{subject to} \;\; ||a|| = 1 \;\; \text{and}$$

1. $||a||_0 \leq k$
2. $||a||_1 \leq t$ (ScoTLASS)

Sparse PCA in R (*sparsepca*)

$$argmin_B \;\; \frac{1}{2}||X - XBA'|| + \phi(B) \;\; ,$$

where $AA' = I$ and $\phi(B)$ is the sparsity inducing penalty (LASSO or elastic net)