# *USE OF SPARSE METHODS IN COSMETICS*

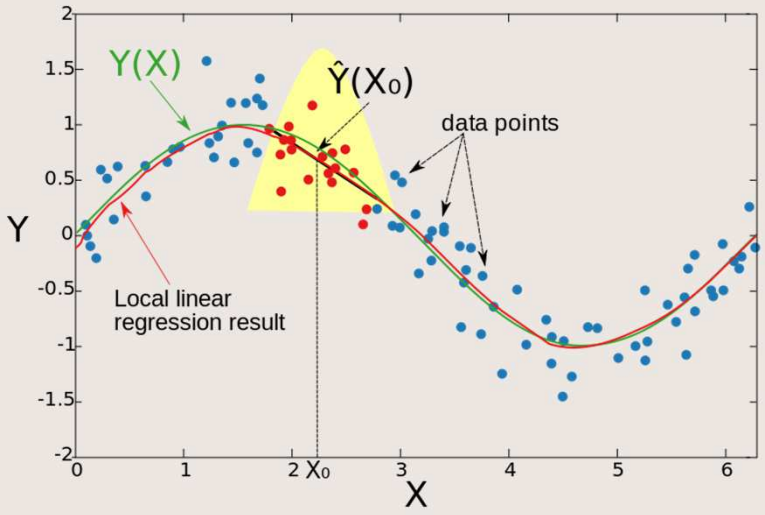
Philippe Bastien

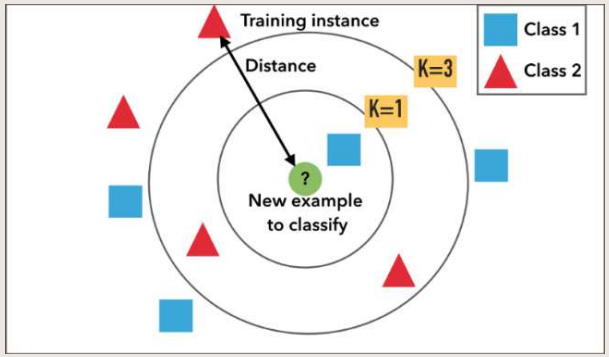
***L'Oreal R&I, Aulnay, France (pbastien@rd.loreal.com)***

# Curse of Dimensionality

Let $f$ a smooth function, a natural way to estimate $f(x_0)$ is by some average of the $y$ associated to the $x$ in the vicinity of $x_0$.

The most simple version of this idea is the k-nearest neighbours estimator using a local average of the data.



# Curse of Dimensionality

- **Unfortunately, when the number of independent variables *p* increases, the notion of « nearest points » vanishes.**

Let $x_1, \dots, x_p, p\ i.i.d.$ descriptive variables which follow a uniform distribution on [0,1] and $y \in \mathbb{R}$ a response variable.

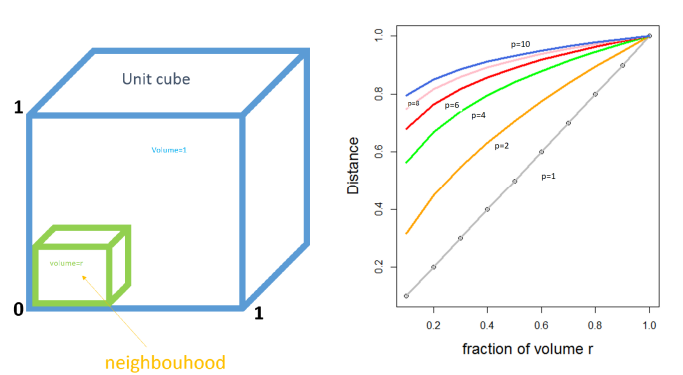
In order to fill the hypercube $[0,1]^p$, i.e. to have **at least one point at a distance less than 1** from x, we need at least:

$$n \geq \left(\frac{p}{2\pi e}\right)^{p/2} \sqrt{p\pi} \text{ observations.}$$

This number of points **grows more than exponentially fast** with p

| P | 20 | 30 | 50 | 100 | 200 |
|---|---|---|---|---|---|
| N | 39 | 45630 | $5.7*10^{12}$ | $42*10^{39}$ | *larger than the number of estimated particules in the observable universe* |

# Curse of Dimensionality

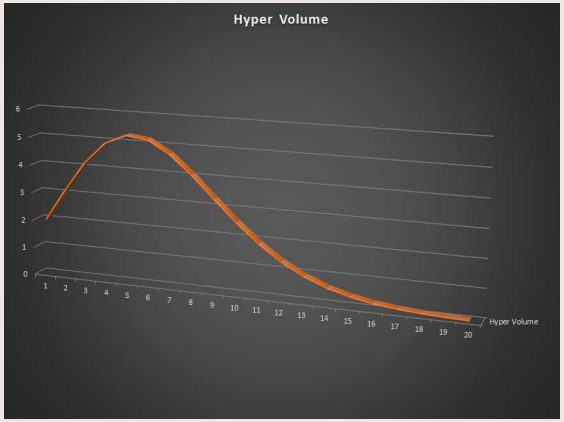



For $p$ = 10 , in order to capture **10 %** of the data one needs to cover **80 %** of the range of each dimension.

- The concept of « local » neighborhood becomes irrelevant.
- Any estimator based on local averaging will fail with such data.

The volume $V_p(r)$ of a p-dimensional sphere of radius **r goes to zero** with p and is **concentrated in its crust**.

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} r^p \sim (2\pi e^{r^2})^{p/2} (p\pi)^{-1/2}$$



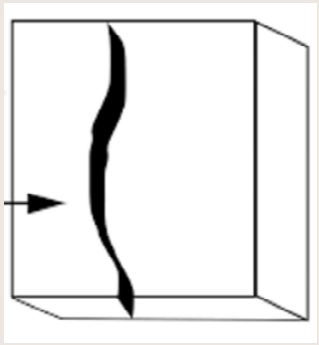
# Curse of Dimensionality

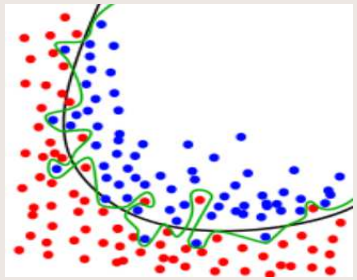
- Accumulation of small fluctuations in many different directions can produce a large global fluctuation (increasing variability in the estimations)
- Multicollinearity issues (incoherent and unstable models)
- Empirical covariance not reliable in high-dimensional settings
- Distance measures lose their effectiveness to measure dissimilarity in highly dimensional spaces : $lim_{p\to\infty} \frac{d_{max}-d_{min}}{d_{min}} \to 0$
- Need of interpretable model for knowledge
- Computational complexity.
- False discovery.

# Circumvent the Curse of Dimensionality

The high dimensionality of the data that seems at first to be a blessing is actually a major issue for the statistical analyses. **The situation may appear hopeless**.

**Fortunately**, **high dimensional data** are often much more low dimensional that they seem to be. Usually they are not uniformely spread in $R^p$, but rather **concentrated** around small **low-dimensional structures**. This is due to the relatively **small complexity** of the system producing the data.





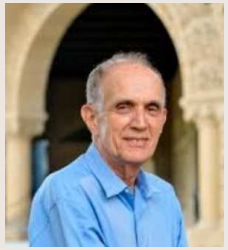
Ex: biological data are the outcome of a biological system which is strongly regulated and whose regulation network has a relatively small complexity.

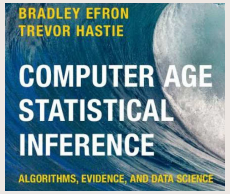
# Circumvent the Curse of Dimensionality

When the low-dimensional structures are known, we are back to some classical «low-dimensional statistics». The major issue with high-dimensional data is that these **structures are usually unknown** and the main task is to identify them.

Importance of bringing information a priori (**regularity**) to **reduce the size of the space**. Deep neural networks use a priori information through the network structure.

Beyond the importance of **regularization** for reliable predictions, **sparsity/parcimony** is important to have an **interpretable** model for knowledge.






*Bradley EFRON*

*« **Maximum likelihood** estimation has shown itself to be an **inadequate and dangerous tool** in many twenty-first century applications.*
***Unbiasedness** can be an **unaffordable luxury** when there are hundred or thousands of parameters to estimate at the same time. »*

# Regularization

***Need for regularization to obtain a stable predictive model***

- ***Dimension reduction****: PCR/PLS regression*
- ***Penalization*** *(linear and generalized linear models)*
  - *L1 (Lasso, group Lasso, sparse group Lasso)*
  - *L2 (Ridge)*
  - *L1&L2 (Elastic net)*
  - *non-convex penalties*
- ***Dimension reduction and Penalization***
  - *sparse PCR, sparse PLS regression, sparse group PLS*
  - *sparse PCA, sparse CCA, sparse GCCA*
- ***Clustering***
  - *sparse k-means*
  - *sparse hierarchical clustering*
  - *sparse bi-clustering*

# Sparse PCA

- Principal components are interpreted by examining the loadings $\{v_j\}_{j=1}^{r}$ in order to determine which of the variables play a **significant role**.
- With a large number of variables it is often desirable to select a smaller subset of relevant variables. Needs for **sparse loadings**.

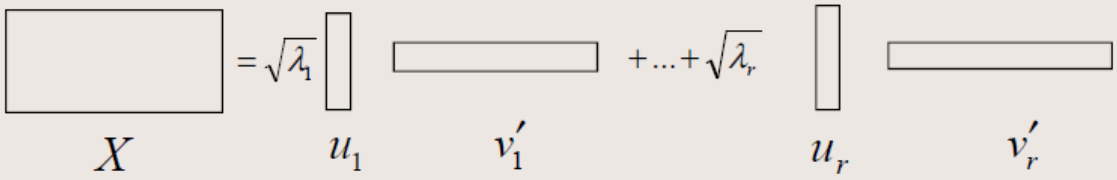
- From a **theoretical** point of view, when p>>N, **PCA** is known to **break down** very badly in that the eigenvectors of the sample covariance could be far from the population eigenvectors.
  Imposing **sparsity** on the PC makes the **problem well-posed** and is therefore essential.

- Sparse PCA for **unsupervised variables selection**

# Sparse PCA

**Singular Value Decomposition of X** (SVD):

$$\mathbf{X} = \mathbf{UDV}^T \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$$




SPC (based on PMD algorithm)

$$max_{u,v} u'Xv \quad s.t.\ \|\boldsymbol{v}\|_1 \leq \boldsymbol{c},\ \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1$$

*Witten, Tibshirani, and Hastie (2009) 'A penalized matrix decomposition, with applications to sparse canonical correlation analysis and principal components', Biostatistics.*

# Penalized Matrix decomposition

**(Witten, Tibshirani, Hastie 2009)**

**PMD algorithm** :

1 / Initialize $v$ to have $L_2$ norm 1

2 / Iterate until convergence

*a)* $\mathrm{u} \leftarrow \arg\max u'Xv \;\; \text{st.} \sum_i |u_i| \le c_1, \; \|u\|_2^2 \le 1$

*b)* $\mathrm{v} \leftarrow \arg\max u'Xv \;\; \text{st.} \sum_i |v_i| \le c_2, \; \|v\|_2^2 \le 1$

→ *Biconvex algorithm*

# Solution for the SPC algorithm

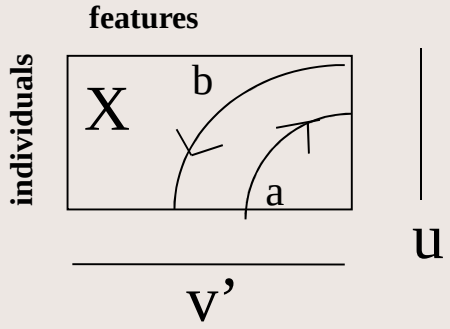
$let\ S_c(a)$=sign(a)$(|a| - c)_+$

1: initialize *v*, $\|v\|_2$=1
2: iterate until convergence

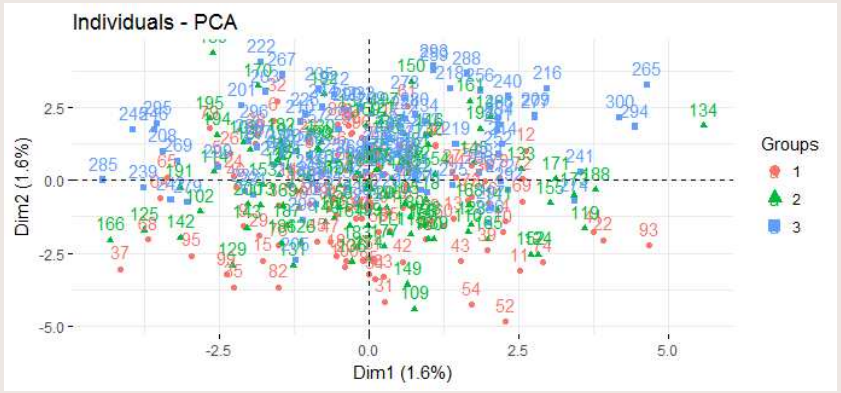
a) u $\leftarrow \frac{Xv}{\|Xv\|_2}$

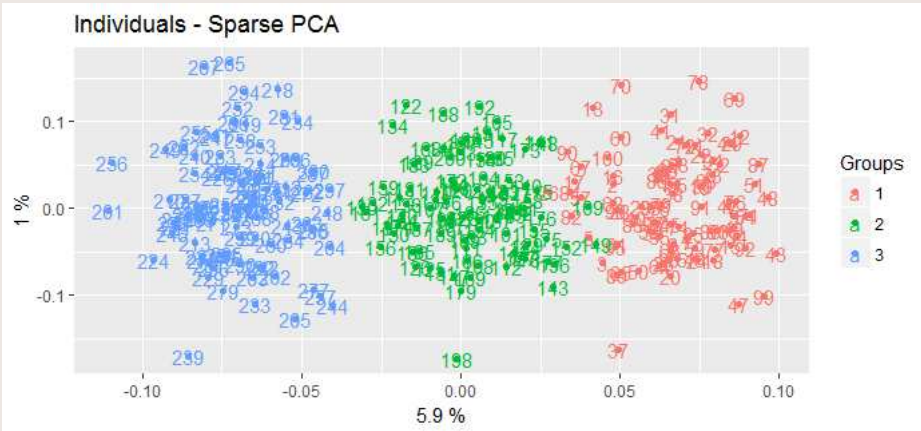
b) v $\leftarrow \frac{S_{\delta_2}(X'u)}{\left\|S_{\delta_2}(X'u)\right\|_2}$

$where\ \delta_2 smallest\ value\ such\ that\ \|v\|_1 \leq c_2$



# Sparse PCA (simulation)



PLOS | ONE

RESEARCH ARTICLE

# Craniofacial similarity analysis through sparse principal component analysis

Junli Zhao[1,2], Fuqing Duan[3,4]☯*, Zhenkuan Pan[5]☯*, Zhongke Wu[3,4], Jinhua Li[1], Qingqiong Deng[3,4], Xiaona Li[1], Mingquan Zhou[3,4]

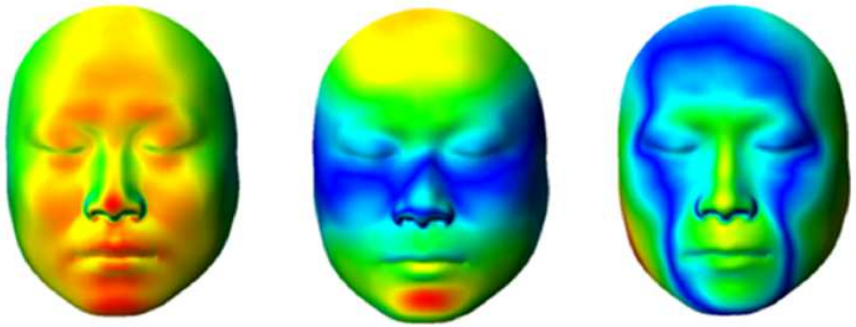


Fig 3. Reflected region by PCA principal component.

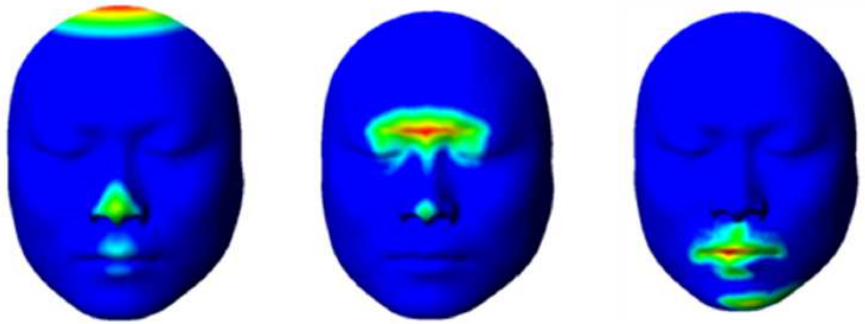
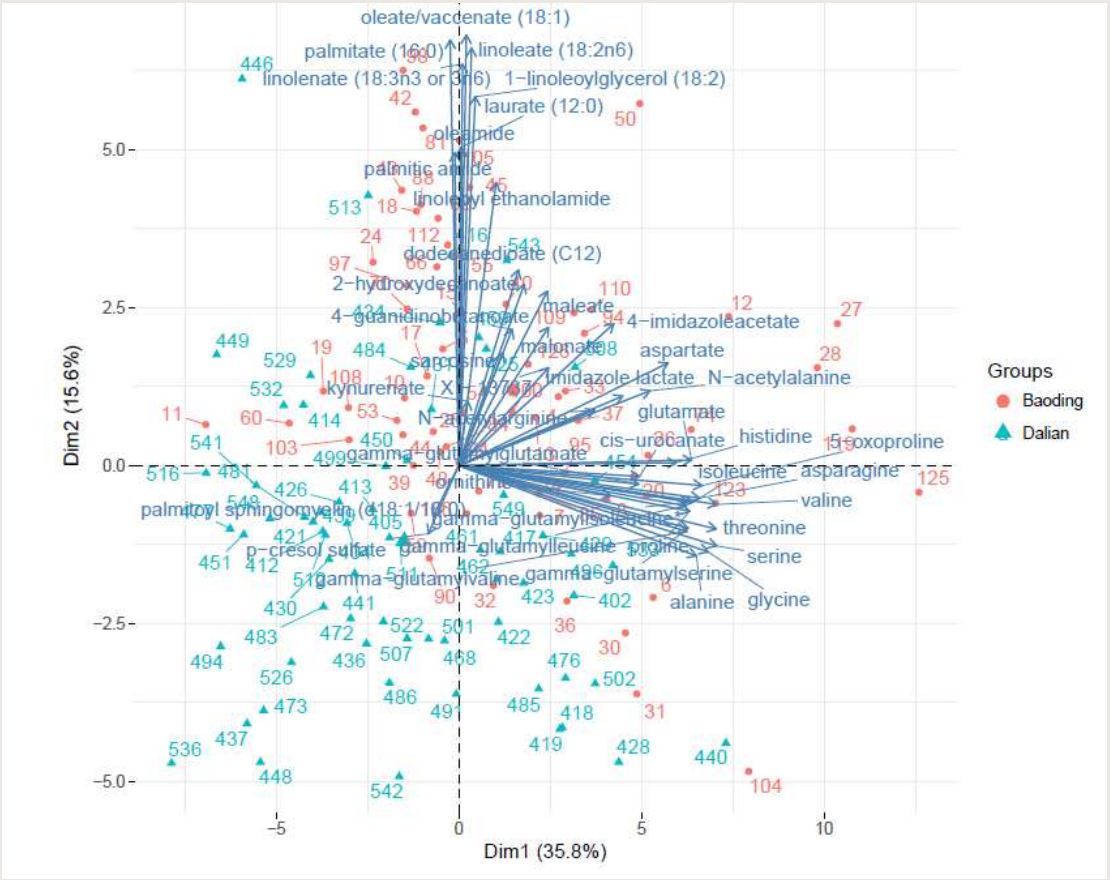


Fig 4. Reflected region by SPCA sparse principal component.

Each PCA component reflects the whole or a larger region of the craniofacial, whereas each sparse SPCA component reflects only a local part of the craniofacial, such as the mouth or nose. Thus, each sparse SPCA principal component reflects detailed areas.


# Sparse PCA: Metabolomics



# Sparse Clustering

(Witten, Tibshirani, 2010)

We wish to cluster the observations and suspect that the true underlying clusters differ only with respect to some features.

This results in more accurate identification of the groups and more interpretable that standard clustering.

*Witten and Tibshirani, A framework for feature selection in clustering, J Am Stat Assoc. 2010 Jun 1; 105(490): 713–726.*

# Sparse K-means clustering

$$max_{C_1,\ldots,C_k,w}\left\{\sum_{j=1}^{p} w_j\left(1/n\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} 1/n_k \sum_{i,i'\in C_k} d_{i,i',j}\right)\right\}$$

$$\|\boldsymbol{w}\|_{\boldsymbol{1}} \leq \boldsymbol{s}, \|w\|_2^2 \leq 1, w_j \geq 0$$

*Witten and Tibshirani, A framework for feature selection in clustering, J Am Stat Assoc. 2010 Jun 1; 105(490): 713–726.*

# Sparse K-means clustering

1: $Initialize\ w\ as\ w_1, \ldots, w_p = 1/\sqrt{p}$

2: $Iterate\ until\ convergence$

$a: Holding\ w\ fixed$

$$min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} 1/n_k \sum_{i,i' \in C_k} \sum_{j=1}^{p} w_j d_{i,i',j} \right\}$$

K-means on matrix $\sum_{j=1}^{p} w_j d_{i,i',j}$

$b: Holding\ C_1, \ldots, C_k fixed$

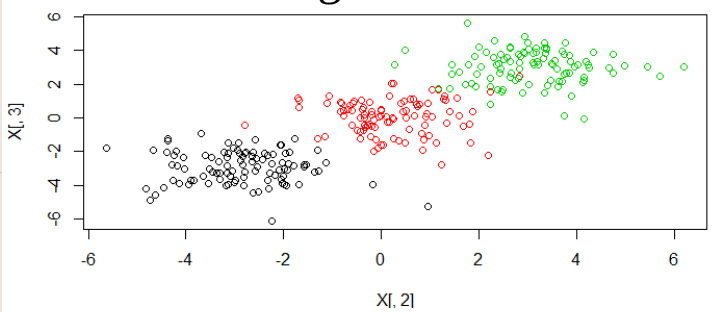
$$w = \frac{S(a_+, \Delta)}{\|S(a_+, \Delta)\|_2}$$

$$a_j = \left( 1/n \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} 1/n_k \sum_{i,i' \in C_k} d_{i,i',j} \right)$$

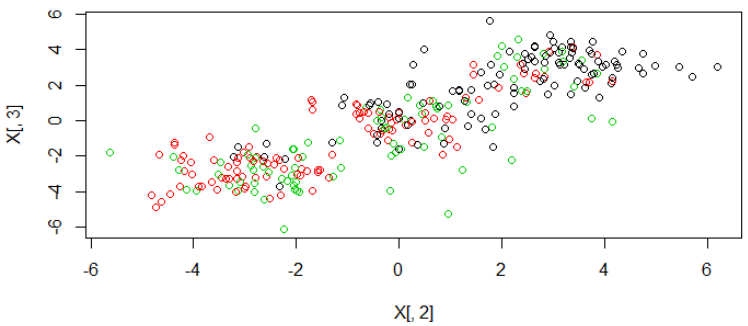
$\Delta = 0\ if\ that\ results\ in\ \|w\|_1 < s, else\ \Delta > 0, so\ that \|w\|_1 = s$

# Sparse K-means clustering

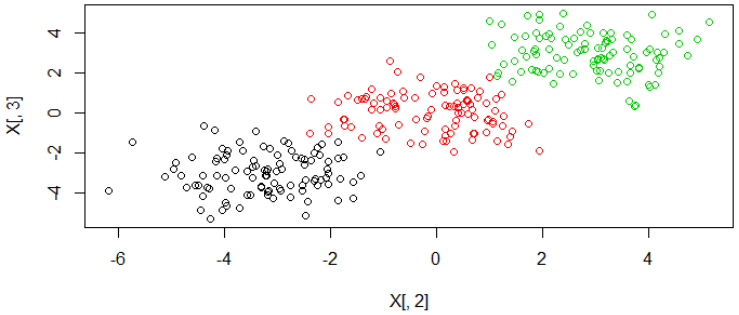
original data



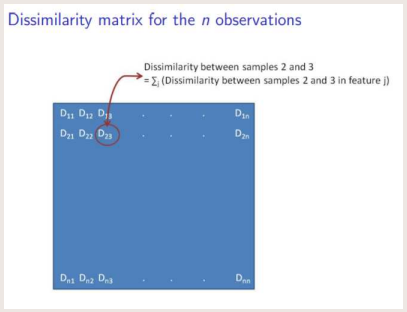

k-means

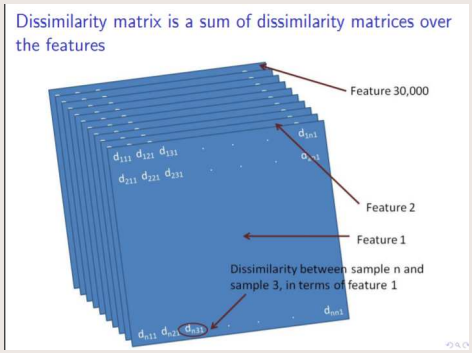


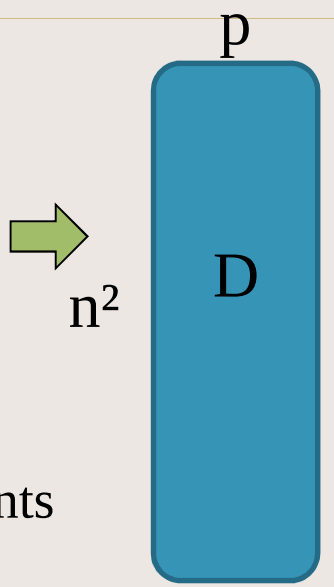

sparse k-means



# Sparse Hierarchical Clustering



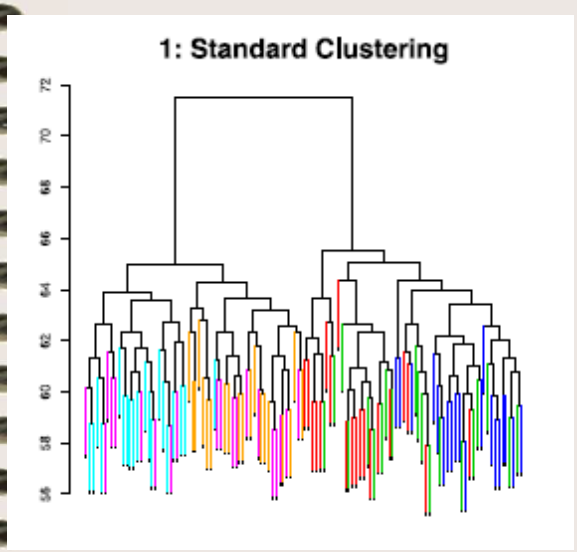



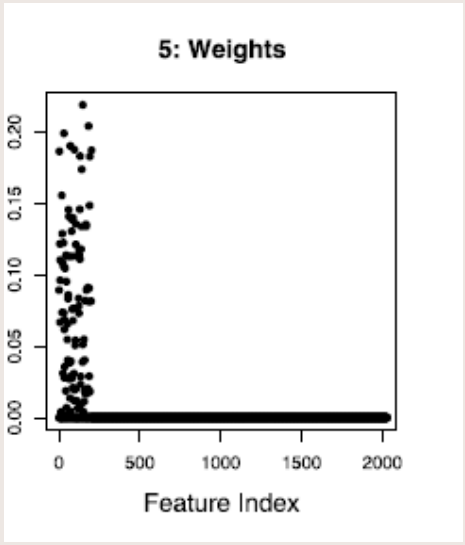

$Let\ D \in \mathbb{R}^{n^2 xp}$ with column j consists of the elements $\{d_{i,i',j}\}_{i,i'}$ unfolded into a vector

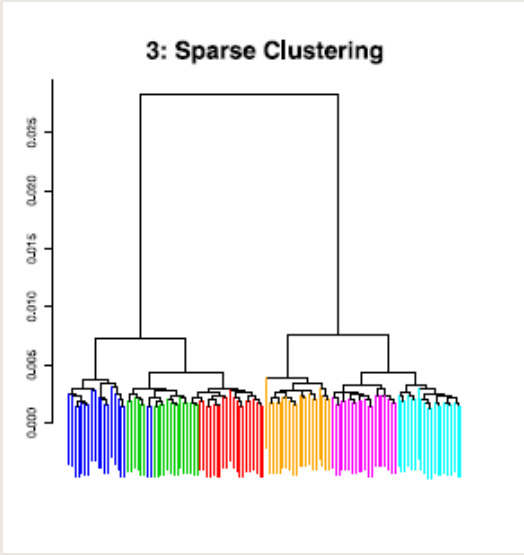
- Sparse PCA on D $max_{u,v}\{u'Dw\}$
  $with\ u \in \mathbb{R}^{n^2} s.t. \|u\|_2^2 \leq 1, \|w\|_2^2 \leq 1, \|w\|_1 \leq s, w_j \geq 0\ \forall j$
- Rewrite u as a n x n matrix U
- Perform a hierarchical clustering on U

*Daniela Witten , A penalized matrix decomposition, with application to sparse hierarchical clustering, PhD thesis 2009; Department of Statistics Stanford University*

# Sparse Hierarchical Clustering



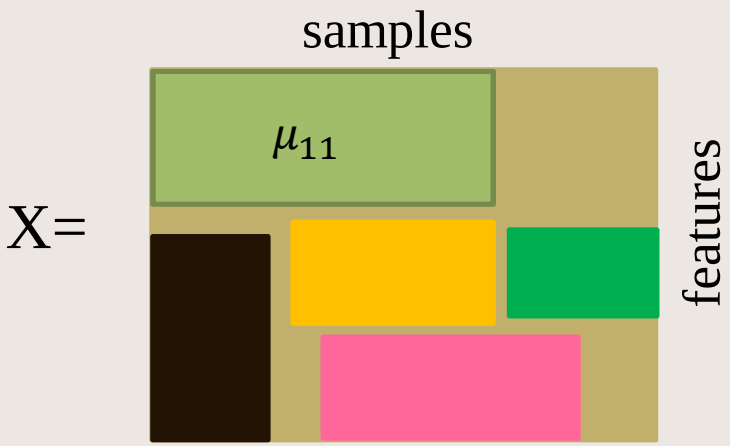




*Daniela Witten , A penalized matrix decomposition, with application to sparse hierarchical clustering, PhD thesis 2009; Department of Statistics Stanford University*

# Bi-clustering

Bi-clustering is of particular interest in situations when both rows and columns of the data matrix present certain scientific meaning and may contain clusters, such as for example gene expression data.

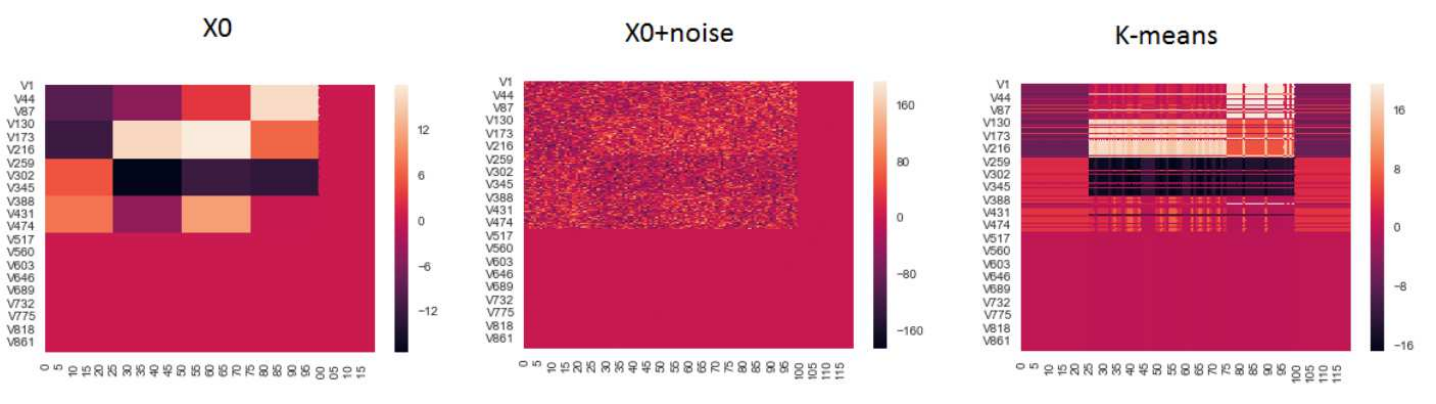
The subgroups of samples may be similar on a subsets of features and vice versa: the subgoups of features may behave similarly on a subsets of samples.

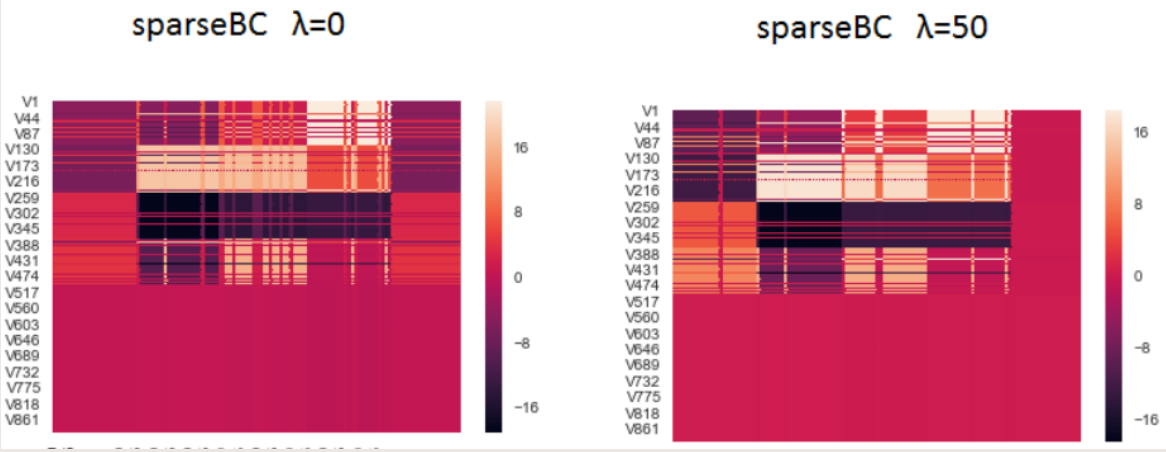



$$\min_{C_1, C_k, D_1, D_r, \mu} \frac{1}{2} \sum_k \sum_r \sum_{i \in C_k} \sum_{j \in D_r} \left( X_{ij} - \mu_{kr} \right)^2 + \lambda \sum_k \sum_r |\mu_{kr}|$$

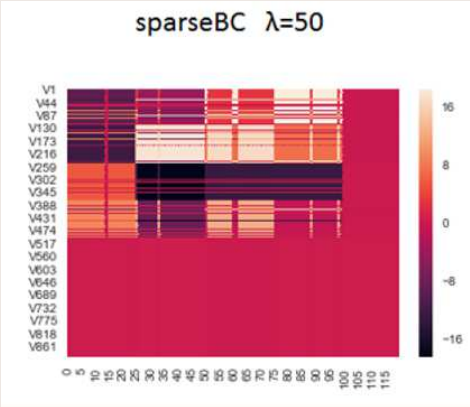
***Kean Ming Tan & Daniela M. Witten (2014) Sparse Biclustering of Transposable Data, Journal of Computational and Graphical Statistics, 23:4, 985-1008***

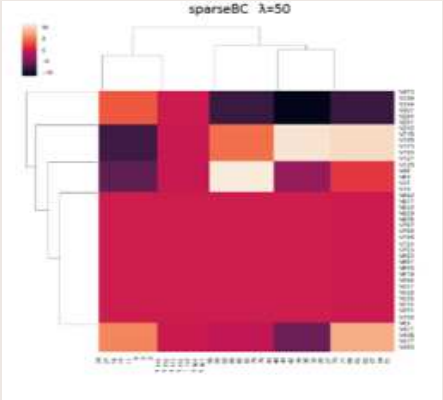
# Sparse bi-clustering

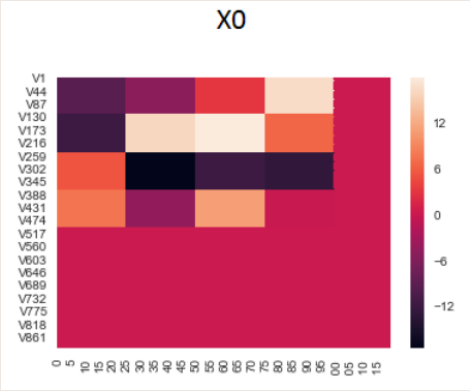


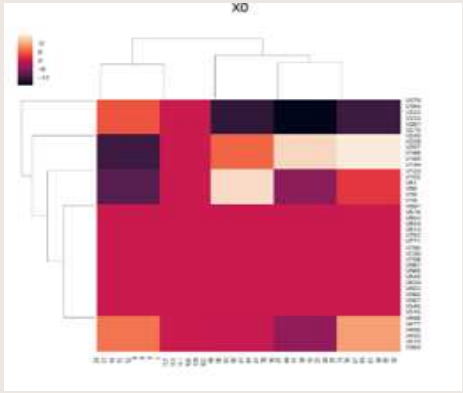



# Sparse bi-clustering

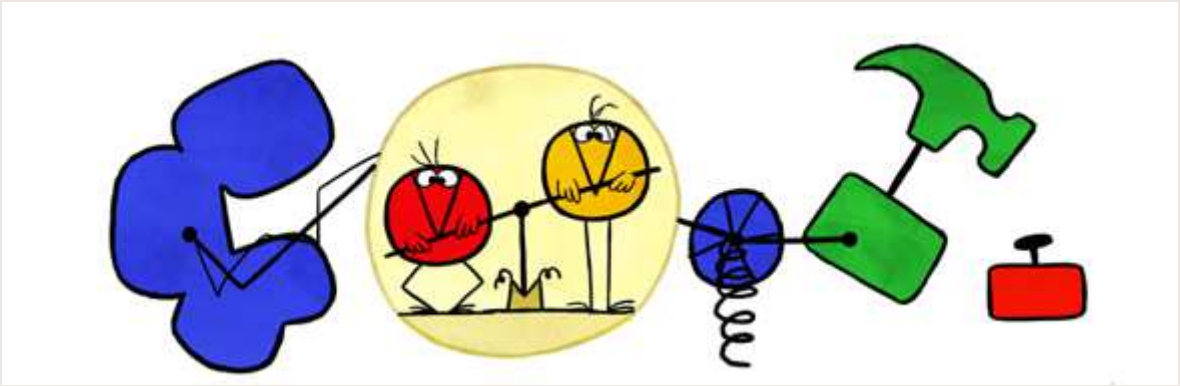


# Thanks for your attention



**Why make *sparse* when we can do complicated ?!**