Robust Principal Component Analysis

Boris Polyak

Institute for Control Sciences, Moscow

Berlin, May 16, 2013

Workshop "Advances in predictive modeling and optimization"

Outline

- Low-dimensional approximations
- Principal Component Analysis
- Robust PCA - known approaches
- Model 1
- Model 2
- Optimization problems and solution methods
- Results of calculation
- Open problems

Low-dimensional approximations

Main idea of modern data analysis — low-dimensional approximations. Lasso regression, sparse parameters, compressed sensing, L_1 techniques etc.

One of the problems: given a cluster of points $x_i \in R^n, i = 1, \dots, N$, approximate them with low-dimensional affine subspace. In statistics it is done by use of Principal Component Analysis (PCA).

PCA: calculate

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad H = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T,$$

find eigenvectors and eigenvalues of H:

$He_i = \lambda_i e_i, 0 \le \lambda_1 \le \lambda_2 \le \dots \le \lambda_n$, then $e_n, e_{n-1}, \dots, e_m$ are principal components (i.e. $n - m + 1$-dimensional approximation of the cluster).

There are 2 validations of this choice, we consider both of them and their robust counterparts.

PCA — Pearson 1901

560 Prof. K. Pearson *on Lines and Planes of*

of leg-lengths; but a point at a given time will have one position only, although our observations of *both* time and position may be in error, and vary from experiment to experiment. In the case we are about to deal with, we suppose the observed variables—all subject to error—to be plotted in plane, three-dimensioned or higher space, and we endeavour to take a line (or plane) which will be the "best fit" to such a system of points.

Of course the term "best fit" is really arbitrary; but a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or plane a minimum.

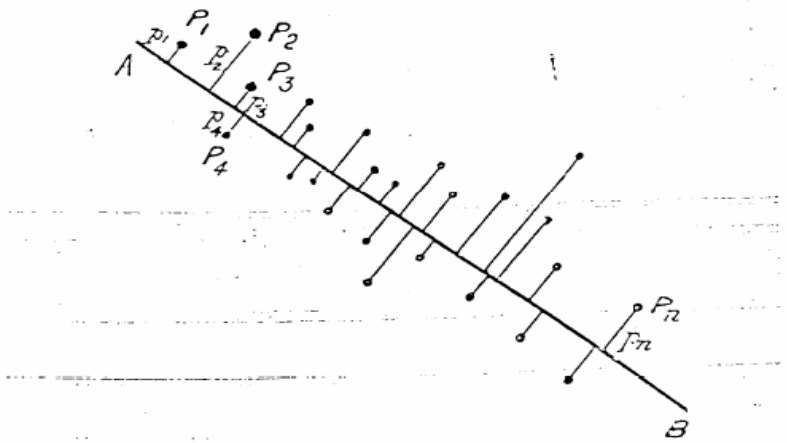
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y' - y)^2 = \text{a minimum}$$

(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



Now clearly $U = S(p^2)$ is the moment of momentum, the second moment of the system of points, supposed equally loaded, about the line A B. But the second moment of a system about a series of parallel lines is always least for the

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2:559-572. *http://pbil.univ-lyon1.fr/R/pearson1901.pdf*

Best fit by a hyperplane

Pearson — no statistical validation, just best fit!
Problem: given $x_i \in R^n, i = 1, \dots, N$, find a hyperplane $(c, x) = 0$ which provides the best (in mean-square sense) approximation to them. We normalize $c : ||c|| = 1$. Then the distance r_i from x_i to the hyperplane equals $|(c, x_i)|$ and the best hyperplane solves

$$\min_{||c||=1} \sum_i (c, x_i)^2$$

If we denote $H = \frac{1}{N} \sum x_i x_i^T, He_i = \lambda_i e_i, ||e_i|| = 1, 0 \le \lambda_1 \le \dots \le \lambda_n$ then the obvious solution is

$$c^* = e_1$$

that is the eigenvector corresponding to the least eigenvalue of H. It is the solution obtained by Pearson.

Non-centered sample

We assumed above that the hyperplane contains the origin. More typical situation is an affine hyperplane $(c, x) = \beta$. Then optimization problem becomes

$$\min_{\beta, ||c||=1} \sum_i ((c, x_i) - \beta)^2$$

Its solution is

$$c^* = e_1, \beta^* = (c^*, \bar{x})$$

where

$$\bar{x} = \frac{1}{N} \sum_i x_i, H = \frac{1}{N} \sum (x_i - \bar{x})(x_i - \bar{x})^T, He_i = \lambda_i e_i, 0 \leq \lambda_1 \leq \cdots \leq \lambda_n$$

This solution is also due to Pearson.

$(n-m)D$ subspace

The solution above is the best fit of points by $n-1$ dimensional affine subspace. Now let's extend it to arbitrary dimension. We are looking for a linear manifold defined by m linear equalities $Cx = b$ where rows $c_i, i = 1, \dots m$ of matrix C are orthogonal and scaled $(c_i, c_j) = \delta_{ij}$, that is $CC^T = I_m$. Then the distance from x_i to this manifold equals $||Cx_i - b||$ and we arrive to optimization problem

$$\min_{b, CC^T = I} \sum_i ||Cx_i - b||^2$$

LemmaThe solution is given by

$$\bar{x} = \frac{1}{N} \sum_i x_i, H = \frac{1}{N} \sum (x_i - \bar{x})(x_i - \bar{x})^T,$$

$$He_i = \lambda_i e_i, ||e_i|| = 1, 0 \leq \lambda_1 \leq \dots \leq \lambda_n,$$

$$C^* = \begin{pmatrix} e_1 & e_2 & \dots & e_m \end{pmatrix}^T, b^* = C^* \bar{x}.$$

Comments

The proof can be obtained by Lagrange multipliers method. Similar results are known (Brockett 1991, Chu, Driessel 1994). Pearson provided the solution for $m = 1, n - 1$ only (hyperplane and straight line).
It is interesting that we are looking for matrix C subject to condition $rankC = m$. Such constraints are nonconvex, and there are special relaxation approaches to approximate the optimization problem with a convex one. Here we avoid the relaxation.
Until now we had no statistical model of data. We just constructed the least square approximation.

Statistical Model

Principal Component Analysis has very natural statistical validation. It is assumed that $x_i = z_i + \xi_i$ where $Cz_i = b$ while ξ_i are i.i.d. with common distribution $N(0, I_n)$. Then $Cx_i - b$ are i.i.d. with common distribution $N(0, I_m)$. Maximum Likelihood Estimate (MLE) for b and C are given by the above formulas. The first m Principal Components are the largest half-axes of the ellipsoid $(H(x - \bar{x}), x - \bar{x}) \leq 1$, this is standard result in PCA.

Robustness in statistics

Standard PCA is very sensitive to *outliers*. If among x_i there are few points which are outliers (i.e. they have distribution which differs from the common one and their distances from $\bar{x}$ are large enough), they can strongly imply the solution.
This effect is well known in statistics; the word "*robust*" was introduced by famous statistician Box (1952). David Bernoulli (1777) discussed the problem of outliers, while Mendeleev (1895) proposed *trimming* procedure to get robust estimates.

Robust PCA

There are numerous attempts to robustify PCA , most of them are based on sparsity models and l_1 techniques. Typical example is E. J. Candes, X. Li, Y. Ma, J. Wright, Robust Principal Component Analysis, Journal of ACM, 2009, 58(1), 1-37.
Our approach relies on ideas of the monograph P.Huber, "Robust statistics", 1981 (second edition 1996 has > 10000 citations), it had great influence on modern statistics. The "nearest relative" to our technique is: R.Maronna, *Robust M-estimators of multivariate location and scatter*, Annals of Stat., 1976. However numerical algorithms are different.

Robust estimation of location parameter

Huber 1964. Given sample $x_i \in R^1, i = 1, \ldots, N$, find their center. If $x_i \sim N(a, 1)$ i.i.d., then the best estimate is arithmetic mean $\hat{a} = \bar{x} = \frac{1}{N} \sum x_i$. However it is not robust w.r.t. outliers. Median $\hat{a} = \arg\min_x \sum |x_i - x|$ is much more robust. Huber's function

$$h(t) = \left\{ \begin{array}{ccc} t^2/2 & if & |t| \leq 1 \\ |t| - 1/2 & if & |t| > 1 \end{array} \right\}$$

provides robust estimate $x^* = \arg\min_x \sum h(x_i - x)$ which is in between median and arithmetic mean. It is optimal for *contaminated* Gaussian families of distributions in some minimax asymptotic sense.

Statistical Model 1 for Robust PCA

Suppose that

$$x_i = z_i + \xi_i, \quad Cz_i = b, \xi_i \sim N(0, I) i.i.d.$$

$CC^T = I$, rank $C = m$. Then MLE for C, b is given by optimization problem

$$\min_{b, CC^T = I} \sum_i ||Cx_i - b||^2$$

considered above. Now assume that ξ_i's are contaminated Gaussian. Then optimization problem becomes

$$\min_{b, CC^T = I} \sum_i h(||Cx_i - b||).$$

This is our first *robust PCA* estimate. Other statistical models will be addressed later.

Discussion

1. The optimization problem is matrix one and it is nonconvex. However its solution can be found explicitly if objective function is quadratic.
2. Our constraint can be written in the form rank $C = m$. Such constraints are hard in optimization, and usually convex relaxation is exploited to treat them. We'll solve optimization problem avoiding convex relaxation.
3. The only analog of such approach without convexification for different problem is found in Teboulle, 2011. At each iteration he solves non-convex problem

$$\min(a, x) \quad ||x||_0 = m, ||x||_2 = 1$$

($||x||_0$ is the number of nonzero entries of x). Its explicit solution is $x_i^* = -\gamma a_i, i = 1, \dots, m, x_i^* = 0, i > m, |a_1| \leq |a_2| \leq \dots \leq |a_n|$.

How to solve the nonconvex problem?

The idea is to apply *weighted least squares method*, proposed for Steiner-Weber problem by Weiszfeld, 1937 (see F.Plastira, *Ann. Op. Res.* 2009).

$$\min_x \sum ||a_i - x||, x \in R^n$$

Let x^k be k-th iteration. Approximate $||x - a_i||$ with quadratic approximation at $t_{ik} = x^k - a_i$: $1/2(||x - a_i||^2/|t_{ik}| + |t_{ik}|)$ thus minimization problem becomes

$$\min_x \sum w_{ik}||a_i - x||^2, w_{ik} = 1/||x^k - a_i||,$$

and its solution is

$$x^{k+1} = \frac{\sum_i w_{ik} a_i}{\sum_i w_{ik}}$$

Similar idea of upper quadratic approximation can be applied in our case.

Optimization problem

$$\min_{b, CC^T = I} \sum_i h(||Cx_i - b||)$$

where $b \in R^m, C \in R^{m \times n}, x_i \in R^n, i = 1, \ldots, N$ and

$$h(t) = \left\{ \begin{array}{ccc} t^2/2 & if & |t| \leq 1 \\ |t| - 1/2 & if & |t| > 1 \end{array} \right\}$$

Algorithm for Model 1

Begin

$$\bar{x}^0 = \frac{1}{N}\sum_i x_i, H^0 = \frac{1}{N}\sum_i (x_i - \bar{x}^0)(x_i - \bar{x}^0)^T,$$

$$H^0 e_i = \lambda_i e_i, ||e_i|| = 1, 0 \le \lambda_1 \le \cdots \le \lambda_n,$$

$$C^1 = \begin{pmatrix} e_1 & e_2 & \dots & e_m \end{pmatrix}^T, b^1 = C^1 \bar{x}^0.$$

k-th iteration $t_{ik} = ||C^k x_i - b^k||, w_{ik} = \left\{ \begin{array}{ccc} 1 & if & |t_{ik}| \le 1 \\ 1/|t_{ik}| & if & |t_{ik}| > 1 \end{array} \right\}$

$$\bar{x}^k = \frac{\sum_i w_{ik} x_i}{\sum_i w_{ik}}, H^k = \sum_i w_{ik}(x_i - \bar{x}^k)(x_i - \bar{x}^k)^T,$$

$$H^k e_i = \lambda_i e_i, ||e_i|| = 1, 0 \le \lambda_1 \le \cdots \le \lambda_n,$$

$$C^{k+1} = \begin{pmatrix} e_1 & e_2 & \dots & e_m \end{pmatrix}^T, b^{k+1} = C^{k+1} \bar{x}^k.$$

Implementation details

1. *Choice of threshold Δ in Huber's function*

$$h(t) = \left\{ \begin{array}{ccc} t^2/2 & if & |t| \leq \Delta \\ \Delta|t| - \Delta^2/2 & if & |t| > \Delta \end{array} \right\}$$

   Larger is contamination smaller is Δ.

2. *Detection of outliers*. Points x_i with small w_{ik} are outliers. Notice that outliers in affine subspace are not detected.

3. *Choice of m* when it is not fixed in advance. Notice that the solution depends on m in contrast with quadratic case.

4. Main question: *convergence* and rate of convergence of the algorithm. Easy to prove: objective function is monotonically decreasing. Unfortunately there exist examples where convergence to local minima holds for some initial approximations. However in most cases global convergence is met.

Gaussian sample — Model 2

We start with another validation of PCA as Maximum Likelihood technique for Gaussian distribution. Now we do not assume points lie in a subspace, they are $x_i = a + C^{-1}z_i; z_i \sim N(0, I)$ are i.i.d., C is $n \times n$ non-degenerate matrix, $i = 1, \dots, N$. Our goal is to estimate a, C. Points x_i have common density

$$p(x) = \frac{\det C}{(2\pi)^{n/2}} \exp(-1/2||C(x-a)||^2).$$

For $Q = CC^T$ the likelihood function is

$$\phi(Q, a) = -\sum \log p(x_i) = -\frac{N}{2} \log\det V + \frac{1}{2}\sum_i ||V^{1/2}(x_i - a)||^2.$$

Thus MLE for a is $\hat{a} = \arg\min_a \phi(Q, a) = \bar{x} = \frac{1}{N}\sum x_i$ and likelihood function becomes

$$J(Q) = -\frac{N}{2} \log\det Q + \frac{N}{2} < H, Q >)$$

Gaussian sample — Model 2 Contd

where $H = (1/N)\sum_i (x_i - \bar{x})(x_i - \bar{x})^T$ and $< ., . >$ stands for scalar product of matrices in the space of symmetric matrices equipped with Frobenius norm. Thus we get the MLE

$$\hat{Q} = \arg\min_{Q>0}(-\log\det Q + < H, Q >)$$

This convex optimization problem has the explicit solution $\hat{Q} = H^{-1}$. We conclude that MLE for $x_i \sim N(a, V), V = Q^{-1}$ is sample mean $\bar{x} = \frac{1}{N}\sum x_i$ and sample covariance $\bar{V} = (1/N)\sum_i (x_i - \bar{x})(x_i - \bar{x})^T$

Now principal components of the sample x_i are eigenvectors of $\bar{V}$ corresponding to few smallest eigenvalues. Their number can be chosen either a priory or with some heuristics.

Robust PCA — Model 2

We address the same model as above $x_i = a + C^{-1}z_i$, but now z_i have contaminated Gaussian distribution with zero mean and radial-depending density. We assume that this density corresponds to "the worst" contamination in scalar model, that is common density for x_i is

$$p(x) = \alpha \det C \exp(-h(||C(x-a)||))$$

where $h(t)$ is Huber's function and α is scaling parameter. Then the likelihood function is

$$\phi(V,a) = -\sum \log p(x_i) = -\frac{N}{2}\log\det V + \sum_i \frac{1}{2}h(||V^{1/2}(x_i-a)||).$$

Finding MLE is a convex optimization problem in V, a; however in contrast with Gaussian case its explicit solution is not available. To minimize this function we exploit the same idea of upper quadratic approximation as above. Thus we arrive to following algorithm.

Algorithm for Model 2

Begin

$$\bar{x}^0 = \frac{1}{N}\sum_i x_i, V^0 = \frac{1}{N}\sum_i (x_i - \bar{x}^0)(x_i - \bar{x}^0)^T, C^0 = (V^0)^{1/2}$$

k-th iteration $t_{ik} = ||C^k(x_i - \bar{x}^k)||, w_{ik} = \left\{ \begin{array}{ccc} 1 & if & |t_{ik}| \leq 1 \\ 1/|t_{ik}| & if & |t_{ik}| > 1 \end{array} \right\}$

$$\bar{x}^{k+1} = \frac{\sum_i w_{ik} x_i}{\sum_i w_{ik}}, V^{k+1} = \frac{\sum_i w_{ik}(x_i - \bar{x}^{k+1})(x_i - \bar{x}^{k+1})^T}{\sum_i w_{ik}},$$

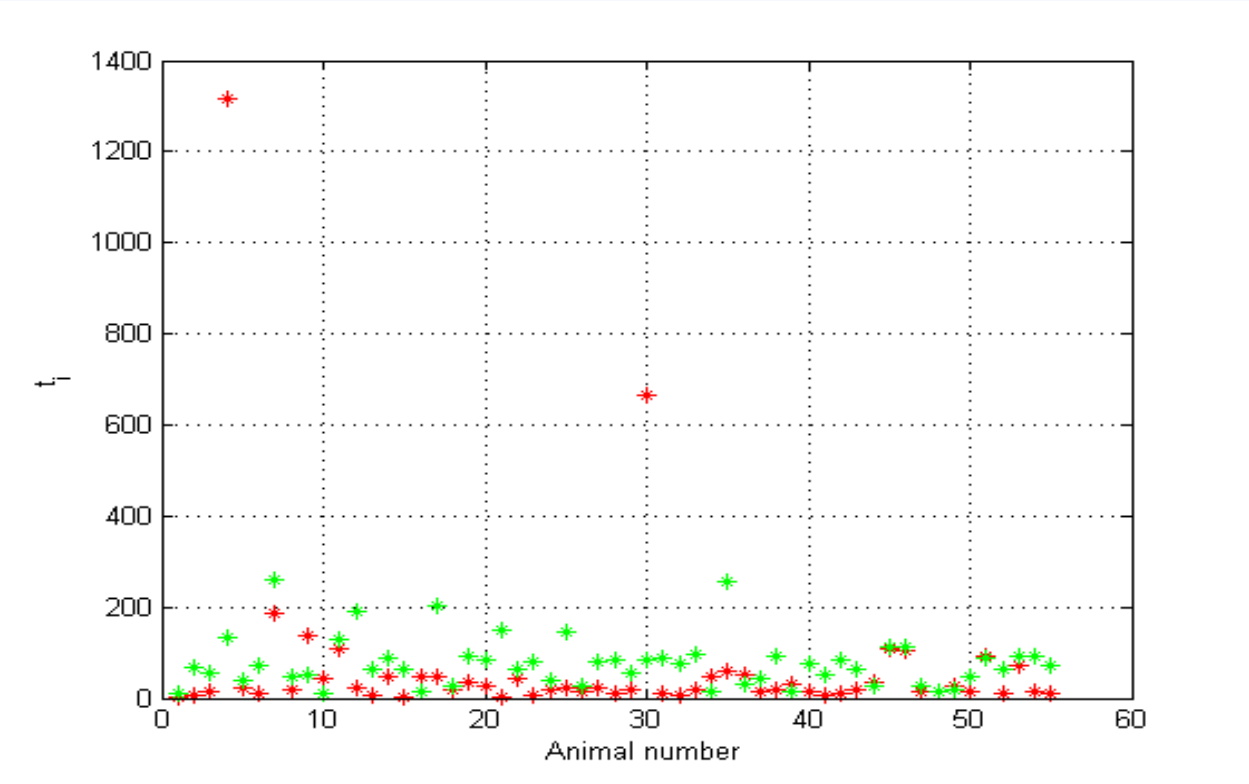
$$C^{k+1} = (V^{k+1})^{1/2}$$

Comments

1. The optimization problem under consideration is a convex one. It is possible to prove global convergence of the algorithm.
2. The rate of convergence is fast. Probably it is linear.
3. Choice of threshold Δ in Huber's function depends on contamination level.

Some examples. Mammals 1

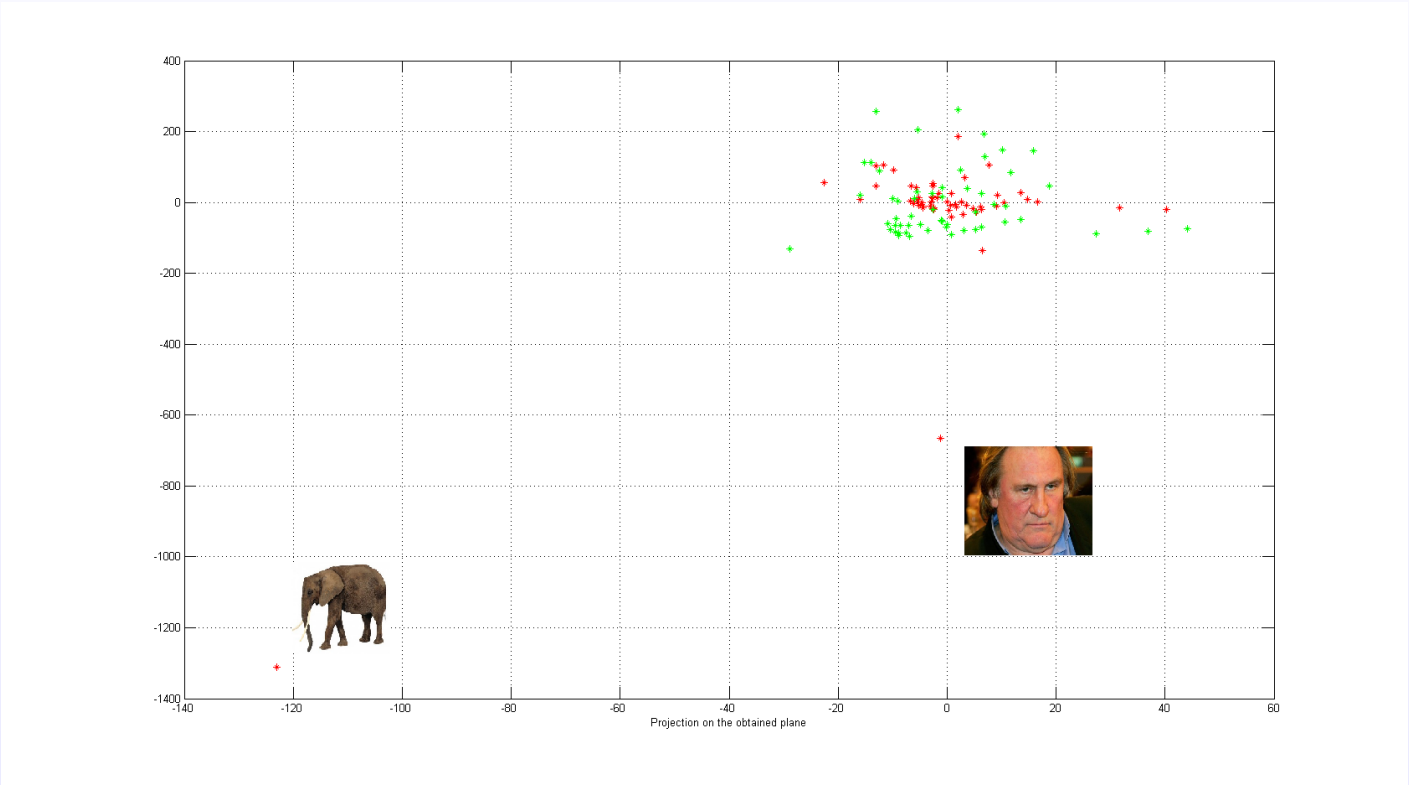
Well known test. 55 mammals, 4 factors (weight, brain weight etc.), goal - to visualize data. Results for 2D approximation via PCA and robust PCA. t_i — distance from x_i to this 2D approximation.



Values r_i are smaller for PCA than for RPCA, however RPCA indicates 2 outliers.

Mammals 2

Same data. Results for 2D approximation via PCA and robust robust PCA. Projections of x_i onto 2D plane orthogonal to this 2D approximation.



Points for RPCA lie more compact than for PCA, however RPCA indicates 2 outliers.

Other examples

Many other examples from databases
http://archive.ics.uci.edu/ml/
http://lib.stat.cmu.edu/
have been tested.

Open problems

1. *Statistical validation.* Asymptotic behavior of the Maximum Likelihood Estimates, their asymptotic consistency and asymptotic normality.
2. *Numerical validation.* Rigorous proof of convergence and rate of convergence. Situation with local and global convergence for model 1.
3. *Application for real-life data.*
4. *Application to classification problems etc.*

Acknowledgements

Helpful discussions with S.Boyd, A.Nemirovski, V.Spokoiny, Yu.Nesterov.
Simulation results by V.Khlebnikov and P.Shcerbakov.