Robust scatter regularization

G. Haesbroeck and C. Croux

University of Liège - University of Leuven

COMPSTAT 2010

Introduction

Let $X = (X_1, \ldots, X_p)^T$ be a p-dimensional random vector with

$$X_i \sim N_p(\mu, \Sigma)$$

where μ is the mean and Σ is the nonsingular covariance matrix.

Aim: Estimate, in a robust way, μ and $\Theta = \Sigma^{-1}$ (concentration matrix) using a sample of size n.

Maximum Likelihood estimator

The ML estimator of (μ, Θ) maximizes

$$\log(\det(\Theta)) - \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^T \Theta (x_i - \mu).$$

When the sample covariance matrix S is nonsingular,

$$(\hat{\mu}_{ML}, \hat{\Theta}_{ML}) = (\bar{x}, S^{-1}).$$

When S is singular (e.g. when $n < p$), the ML estimator does not exist.

Regularized Maximum Likelihood estimator

The Regularized ML estimator of (μ, Θ) maximizes

$$\log(\det(\Theta)) - \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^T\Theta(x_i - \mu) - \lambda J(\Theta),$$

where $\lambda \geq 0$ is the penalty parameter and J is a penalty function.

Typical choices:

- L_1-norm: $J(\Theta) = \sum_{i,j=1}^{p} |\Theta_{ij}|$
- L_2-norm: $J(\Theta) = \sum_{i,j=1}^{p} \Theta_{ij}^2$
- ...

Breakdown Point

Roughly speaking, the breakdown point is the smallest fraction of contamination that can drive the estimator over all bounds.

For a scatter estimator, breakdown can occur due to

explosion: $\lambda_1(\Theta) \to \infty$

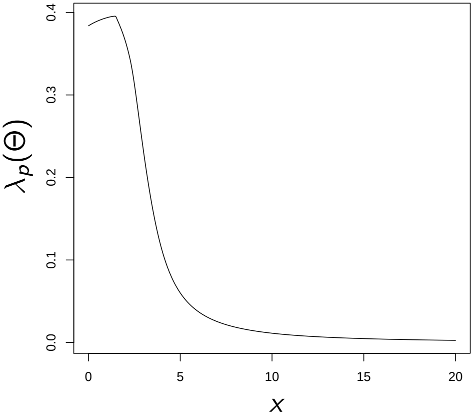
or

implosion: $\lambda_p(\Theta) \to 0$

with $\lambda_p(\Theta) \leq \ldots \leq \lambda_1(\Theta)$.

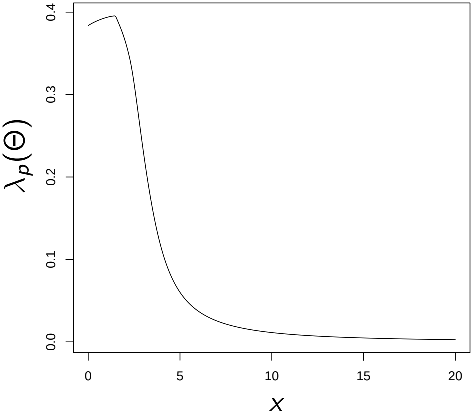
Breakdown of the Regularized ML procedure

$\mu = 0, \Sigma = I_p$ and $x'_n = x_n + x e_1$



Breakdown of the Regularized ML procedure

$\mu = 0, \Sigma = I_p$ and $x'_n = x_n + xe_1$



Robust alternatives are needed!

Minimum Covariance Determinant estimator

- Find a subsample H of size h (with $\frac{n}{2} \leq h \leq n$) minimizing the generalized variance
$$\log(\det(\Sigma_H))$$
(where Σ_H is the covariance matrix based on the h points).
- The location and scatter MCD estimates are given by the mean and covariance matrix of the optimal subsample.

Regularized MCD estimator

- Find a subsample H of size h maximizing

$$\log(\det(\Theta_H)) - \frac{1}{h}\sum_{i \in H}(x_i - \mu_H)^T \Theta_H (x_i - \mu_H) - \lambda J(\Theta_H)$$

- The regularized MCD estimator is given by the regularized ML estimator computed on the optimal subsample.

Properties of the Regularized MCD estimator

A. Robustness

The finite-sample breakdown point for joint location and scatter of the Regularized MCD estimator is equal to

$$\varepsilon^*((\hat{\mu}_{\mathrm{MCD}}, \hat{\Sigma}_{\mathrm{MCD}}); X) = \frac{\min(h, n-h+1)}{n}$$

where $\frac{n}{2} \leq h \leq n$ is the number of observations selected in the MCD solution.

In particular, for $h = n/2$, $\varepsilon^*((\hat{\mu}_{\mathrm{MCD}}, \hat{\Sigma}_{\mathrm{MCD}}); X) = 1/2$.

Properties of the Regularized MCD estimator

B. Computation

Iterative algorithm:

$$(\hat{\mu}_0, \hat{\Theta}_0) \rightarrow \ldots \rightarrow (\hat{\mu}_k, \hat{\Theta}_k) \rightarrow (\hat{\mu}_{k+1}, \hat{\Theta}_{k+1}) \rightarrow \ldots$$

- $(\hat{\mu}_0, \hat{\Theta}_0)$: Regularized ML estimator based on a random subset of 2 observations
- iteration k to $k+1$ by means of a $C-$step
- works for $n < p$

Simulations

Clean setting: $n = p = 50$, $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5I(i, j \leq 9)$ for all $i \neq j$.

Contaminated setting: 5% of shift and correlation outliers (intermediate or extreme)

L_1 penalty	ML		MCD	
	MSE($\hat{\mu}$)	KL($\widehat{\Theta}$)	MSE($\hat{\mu}$)	KL($\widehat{\Theta}$)
Clean	0.98	6.94	1.43	6.46
5% Intermediate	1.70	9.76	1.42	6.53
5% Extreme	200.89	17.58	1.41	6.53

where

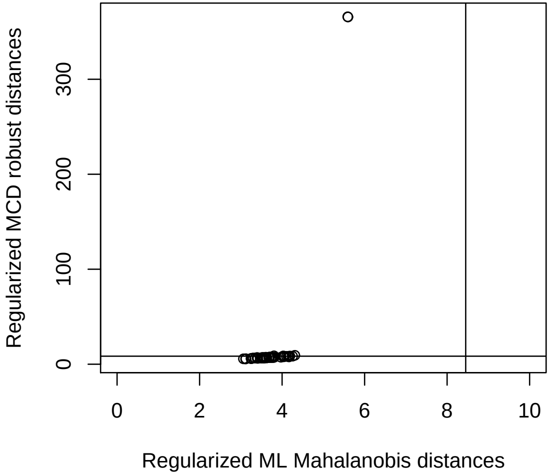
$$KL(\widehat{\Theta}) = -\log(\det(\widehat{\Theta})) + \mathrm{tr}(\widehat{\Theta}\Sigma) - (-\log(\det(\Sigma^{-1})) + p)$$

Applications

- Detection of outliers in high dimensional data (with $n < p$ or n/p small).
- Robust graphical modelling
- Robust regularized regression

Detection of outliers

$n = p = 50$, $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5I(i, j \leq 9)$ for all $i \neq j$,
5% of shift and correlation outliers



Conclusions

- Robust regularized scatter estimation is available.
- Other robust multivariate estimators can also be adapted to the penalized setting (e.g. M estimator,...).
- Still room for further research.