Robust regularized M-estimators of regression parameters and covariance matrix

Esa Ollila, Hyon-Jung Kim and Visa Koivunen
Department of Signal Processing and Acoustics, SMARAD CoE
Aalto University, Finland

Abstract: High dimension low sample size (HD-LSS) data are becoming increasingly present in a variety of fields, including chemometrics and medical imaging. Especially problems with \( n < p \) (more variables than measurements) present a challenge to data analysts since the classical techniques can not be used. In this paper, we consider HD-LSS data in regression parameter and covariance matrix estimation problems. In particular, we consider and extend convex relaxation (or shrinkage regularization, diagonal loading) approach for \( M \)-estimation of regression coefficients and covariance (scatter) matrix. We demonstrate the utility of the methods in beamforming and tensor decomposition applications

Keywords: Covariance matrix estimation, Diagonal loading, High-dimensional data, M-estimator, Regularization, Ridge regression, Shrinkage

1 Introduction

High dimensional data sets are challenging for data analyst. In regression one often resorts to convex-relaxation methods such as ridge regression which seeks a balance with bias-variance trade-off. HD-LSS data provides a challenge to classical multivariate analysis as well. For example, principal component analysis (PCA) is a common pre-processing and standardization step which cannot be performed due to possibly rank deficient sample covariance matrix (SCM). Also impulsive measurement environments and outliers are commonly encountered in many practical applications. In this paper we tackle these two important issues and consider robust shrinkage approaches for regression parameter and covariance (scatter) matrix estimation problems in case of HD-LSS data. Recently robust shrinkage approaches for covariance matrix estimation were addressed in [1, 7] and for regression setting in [4].

2 Shrinkage M-estimates of regression

We consider the multiple linear regression model \( y_i = \phi_i^\top s + \varepsilon_i, \ i = 1, \ldots, n, \) which can be more conveniently expressed in matrix form as \( y = \Phi s + \varepsilon, \) where \( y = (y_1, \ldots, y_n)^\top \) is the observed data vector (measurements), \( \Phi = (\phi_1, \ldots, \phi_n)^\top \) is the known \( n \times p \) measurement matrix, \( s = (s_1, \ldots, s_p)^\top \) is the unobserved signal vector (or regression coefficients) and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \) is the (unobserved) noise vector. The problem is then of estimating the unknown \( s. \) All measurements and parameters are assumed to be real-valued. Suppose \( \varepsilon_i \)‘s are i.i.d. from a continuous symmetric distribution with p.d.f. \( f_\varepsilon(e) = \frac{1}{2} f_0\left(\frac{e}{\sigma}\right), \) where \( \sigma > 0 \) denotes the scale parameter and \( f_0(\cdot) \) the standard form of the p.d.f.

2.1 Ridge M-estimates with preliminary scale

We first assume that scale parameter \( \sigma \) is known or replaced by its estimate. In either case, we replace hereafter \( \sigma \) by \( \hat{\sigma} \). First recall that ridge regression (RR) estimator [2] is defined as the (unique) minimizer of the penalized residual sum of squares objective function \( J_{RR}(s) = \sum_{i=1}^n(y_i - \phi_i^\top s)^2 + \lambda \|s\|_2^2 \) where \( \lambda > 0 \) is the ridge (shrinkage or regularization) parameter. The bigger the \( \lambda \), the greater is the amount of shrinkage of coefficients toward zero. Let us denote residuals for a given (candidate) \( s \) by \( e_i(s) = y_i - \phi_i^\top s \) and write \( e(s) = (e_1(s), \ldots, e_n(s))^\top \). We define the ridge regression \( M \)-estimator \( \hat{s}_{RR} \) as the minimizer of

\[
J(s) = \sigma^2 \sum_{i=1}^n \rho\left(\frac{e_i(s)}{\sigma}\right) + \lambda \|s\|_2^2 \quad (1)
\]

where \( \rho \) is continuous and differentiable even function \( (\rho(e) = \rho(-e)) \) and increasing for \( e \geq 0 \). Note that the multiplying factor \( \sigma^2 \) is used so that the objective function coincide with \( J_{RR}(s) \) when \( \rho(e) = e^2 \).

Let us write \( \psi(e) = \rho'(e) \) and \( w(e) = \psi(e)/e \) with convention that \( w(0) = 0 \) for \( e = 0 \). To obtain robust RR \( M \)-estimates, we need \( \rho \)-functions that give small or zero weights for large residuals. In the conventional regression model \( (n > p) \), the maximum likelihood (ML) estimator of \( s \) is found by choosing \( \rho(e) = -\log f_0(e) + c \) in (1) with \( \lambda = 0 \). In case of Cauchy error terms, \( \rho_C(e) = \frac{1}{2} \log(1 + e^2) \) and \( \psi_C(e) = \rho_C'(e) = e/(1 + e^2) \) whereas Huber’s \( \rho \) function is

\[
\rho_H(e) = \begin{cases} 
\frac{1}{2} e^2, & \text{for } |e| \leq k \\
|e| - \frac{1}{2} k^2, & \text{for } |e| > k 
\end{cases}
\]

and the corresponding \( \psi \)-function is \( \psi_H(e) = \psi_H(|e|) \).
max[−k, min(k, c)]. Above k is a user-defined tuning constant that affects robustness and efficiency of the method. With Huber’s ρ-function, the objective function in (1) is convex but due to the non-convexity of Cauchy ρ, also the associated optimization problem is non-convex.

**Computation:** By setting the derivatives of (1) to zero shows that \( \hat{s}_\lambda \) solves the following estimating equation:

\[
(\Phi^T W \Phi + 2\lambda I) \hat{s}_\lambda = \Phi^T W y
\]

(2)

where \( W = \text{diag}(\{w_i\}_1^n) \) with \( w_i = w(e_i(\hat{s}_\lambda)/\hat{\sigma}) \). This suggest the computation of the estimator by "iteratively (re)weighted RR (IWRR) algorithm. which iterates

\[
s_{t+1} = (\Phi^T W \Phi + 2\lambda I)^{-1} \Phi^T W y
\]

(3)

until convergence. Note that an initial estimate \( s_0 \) of \( s \) and an estimate \( \hat{\sigma} \) of the scale of the residuals is needed. Following [4] it can be shown that the objective function (1) descends at each iteration. Thus for convex problems the IWRR algorithm can be used to find the global minimum.

### 2.2 Ridge M-estimates of regression and scale

Next we consider joint estimation of regression parameter \( s \) and scale \( \sigma \). We define the joint ridge M-estimates of regression \( \hat{s}_\lambda \) and scale \( \hat{\sigma} \) as the minimizers of

\[
J(s, \sigma) = \sum_{i=1}^n \rho \left( \frac{e_i(s)/\sigma}{\sigma} \right) + \lambda |s|^2 + \alpha(n\sigma).
\]

(4)

where \( \rho \) and \( \lambda \) are as earlier and \( \alpha \geq 0 \) is a tuning parameter. Note that if \( \lambda = 0 \), then the equation reduces to approach proposed by Huber. By setting the derivatives of (4) w.r.t. \( (s, \sigma^2) \) to zero shows that \( (\hat{s}_\lambda, \hat{\sigma}) \) solves the estimating equation (2) jointly with the estimating equation

\[
\frac{1}{n} \sum_{i=1}^n \chi \left( \frac{e_i(\hat{s}_\lambda)}{\sigma} \right) = \alpha
\]

(5)

where \( \chi(e) = \psi(e) - e - \rho(e) \) and \( \psi = \rho' \) as earlier.

### 3 Shrinkage M-estimators of covariance

Here we consider the complex-valued case. Recall that a random vector \( (r.v.) z \in \mathbb{C}^p \) is said to have a p-variate complex elliptically symmetric (CES) distribution [5] if its p.d.f. is of the form \( f(z) = C_{p,g}(\Sigma^{-1} g(z^H \Sigma^{-1} z)) \), for some positive definite Hermitian (PDH) \( p \times p \) scatter matrix parameter \( \Sigma \) and function \( g : \mathbb{R}_+^p \rightarrow \mathbb{R}_+^p \), called the density generator; We shall write \( z \sim \text{CE}_{p,g}(\Sigma, g) \).

Above \( C_{p,g} \) is a normalizing constant and \( (\cdot)^H \) denotes Hermitian transpose. Note that the covariance matrix of \( z \) (when exists) is equal to \( \mathbb{E}[zz^H] = c \cdot \Sigma \). Consider \( n \) i.i.d. samples from a CES distribution. Let us add a regularization term \( \lambda \text{Tr}(\Sigma^{-1}) \) to \(-1 \times \log\text{-likelihood function and minimize}

\[
L(\Sigma) = \sum_{i=1}^n \rho(z_i^H \Sigma^{-1} z_i) + n \ln|\Sigma| + \lambda \text{Tr}(\Sigma^{-1}).
\]

(6)

where \( \rho(t) = -\ln(g(t)) + \lambda > 0 \) is the fixed regularization parameter. Thus we impose a bound on \( \text{Tr}(\Sigma^{-1}) = \sum_{i=1}^n \frac{1}{\gamma_i} \), where \( \gamma_i \)'s denote the eigenvalues of \( \Sigma \). As a consequence, the solution will not be ill-conditioned when \( n < p \). In the Gaussian case, \( z_i \sim \mathcal{CN}_p(0, \Sigma) \), we have that \( \rho(t) = t \) and the solution to (6) is easily shown to be \( \Sigma = S + \lambda I \), where \( S = \frac{1}{n} \sum_{i=1}^n z_i z_i^H \) denotes the SCM. We shall consider generalization of the methodology in [1, 7] and consider shrinkage M-estimators of \( \Sigma \) for general \( \rho \) functions.

**Applications:** Conventional beamforming cannot be used in HD-LSS underdetermined problems, e.g., the minimum variance distortionless response (MVDR) beamformer weight vector requires the inverse of the SCM. Applying the diagonal loading, i.e., using \( S + \lambda I \) in place of \( S \) is the commonly used approach; see [3]. Our simulations show that robust shrinkage covariance matrix estimators (e.g., shrinkage Tyler’s M-estimator [1, 7] or Huber’s M-estimator proposed here) provide superior performance in non-Gaussian impulsive noise.

### References


[3] J. Li, P. Stoica, and Z. Wang. On robust Capon beamformer weight vector requires the inverse of the SCM. Applying the diagonal loading, i.e., using \( S + \lambda I \) in place of \( S \) is the commonly used approach; see [3].


