

Robust covariance matrix estimation with regularization

Pietro Coretto and Christian Hennig

Abstract The sample covariance matrix performs poorly in situations with outliers and high dimensionality. Most high-breakdown point estimators of multivariate location and scatter cannot cope with datasets with small n/p ratio, where n is the sample size and p the dimension. In this paper we introduce and discuss an estimator of location and scatter defined as the solution of a system of nonlinear (implicit) equations. These equations are based on a pseudo-model representation of the underlying probability measure. The latter consists in an improper density function called “Black Hole Density”. The estimator smoothly trims data points far away from the center. The robustness of the procedure is based on tunings that have a straightforward interpretation. For datasets with small n/p ratio, regularization is provided by a constraint that bounds the condition number of the covariance matrix. Empirical evidence suggests that the proposed estimator compares well with other robust location-scatter estimators.

Key words: Robust estimation, multivariate location-scale estimates, improper density pseudo-model

1 Introduction

Location-scale statistics are basic ingredients of most analysis. Nevertheless estimation in multivariate settings can be tricky because of the strong influence of outlying observations and/or small n/p ratio, where n is the sample size and p is the

Pietro Coretto

Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Giovanni Paolo II no.132, 84084, Fisciano (SA), Italy, e-mail: pcoretto@unisa.it

Christian Hennig

Department of Statistical Science, University College London, 1-19 Torrington Place, London, WC1E 6BT, United Kingdom e-mail: c.hennig@ucl.ac.uk

dimension of the observed vector. Robust covariance matrix estimation has been a popular topic in the '80s and '90s, whereas recently there has been a growing interest in the high dimensional setting ($n \ll p$). Starting from Maronna's pioneering work [12], several robust estimators of multivariate location and scatter have been proposed with emphasis to symmetric elliptical populations. The list includes: MVE and MCD ([16, 15]), S-estimates and τ -estimates ([6, 10]), MM-estimates ([18]), SD-estimates ([17, 7]), OGK-estimates ([11]), NNVE estimator by [20], FS estimates by [13]. These methods offer more resistance to outliers than the classical sample covariance, some of them are well known for having high break down point. Unfortunately none of them can cope with small n/p . These days "high-dimensional" qualifies the case $n \ll p$. For covariance matrices, even when $n > p$, moderate values of n/p can cause troubles in practice. In particular the spectrum of the sample covariance is seriously biased unless $n \gg p$. Proposals to solve this problem includes: shrinkage ([8, 9]), thresholding ([2, 14, 3]), lasso-type penalized methods ([22, 1]), ML estimation under restrictions on eigenvalues ([21]). These contributions provide serious improvements over the sample covariance, however none of them considers the robustness issue. Some link between dimensionality and robustness has been established (see [6, 19]). In particular for small n/p there is no way to achieve good breakdown point behaviour if one is not willing to waive affine equivariance.

In this paper we outline an estimator of location and scatter defined as the solution of a system of nonlinear (implicit) equations. These equations are based on a pseudo-model representation of the underlying probability measure, that is an improper density that we call the "Black Hole Density". The pseudo density is used to construct a system of weights that reduce the influence of points away from the center in a data-dependent way. The robustness is controlled by tunings that have a simple interpretation. The user can decide whether to regularize the solution with a constraint that in practice bounds the condition number of the covariance matrix. Therefore the proposed estimator always provides well conditioned estimates. In the final section of this paper we report some empirical experience that suggests the potential of the proposed method.

The paper is organized as follows: in Section 2 we introduce the black hole pseudo model, in Section 3 we defined the proposed estimator final remarks are given.

2 The black hole pseudo density

We begin with some notations and definitions. Let $\mathbf{x}$ be a vector in $\mathbb{R}^p$ with $p > 1$, let x_k the generic element of $\mathbf{x}$ for $k = 1, 2, \dots, p$. The i th observed vector is $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ for $i = 1, 2, \dots, n$. $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the matrix with generic element $\mathbf{X}[i, k] = x_{ik}$ with $i = 1, 2, \dots, n$ and $k = 1, 2, \dots, p$. $\mathbf{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ denotes the observed sample. For a random vector $\mathbf{x} \in \mathbb{R}^p$, with expectation μ and covariance matrix Σ , the squared Mahalanobis distance of a point $\mathbf{x}$ to μ is given by

$$d^2(\mathbf{x}; \mu, \Sigma) := (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu).$$

Define the Gaussian β -ellipsoid as the set

$$\mathcal{E}_\beta(\mu, \Sigma) := \{\mathbf{y} \in \mathbb{R}^p : d^2(\mathbf{y}; \mu, \Sigma) \leq q_p(\beta)\},$$

where $q_p(\beta)$ is the quantile function of a χ_p^2 distribution computed at β . If the previously defined random vector $\mathbf{x}$ has a Gaussian law, then $\Pr(\mathcal{E}_\beta(\mu, \Sigma)) = 1 - \beta$. Finally $\phi(\mathbf{x}; \mu, \Sigma)$ denotes the density function of the p -variate Normal distribution at $\mathbf{x}$.

We want to handle situations where $\mathbf{X}_n$ is generated from a ‘‘Gross Error Model’’

$$P = (1 - \alpha)F + \alpha U, \quad (1)$$

where F is a symmetric elliptical distribution mainly responsible for the ‘‘regular’’ observations and U produces observations inconsistent with F called outliers or noise. α is the expected proportion of points sampled from U . Regular points are not the points sampled under F , but rather points that lie in a particular region defined relatively to F . We are interested in situations where the regular region has a shape that can be well approximated by a Gaussian ellipsoid. In other words the reference model for F is the Gaussian distribution. The following distinction between regular and outlier region is inspired to the work of [5]. Fix β , $\mathbf{x}$ is regular if $\mathbf{x} \in \mathcal{E}_\beta(\mu, \Sigma)$, a point $\mathbf{x} \in \mathbb{R}^p \setminus \mathcal{E}_\beta(\mu, \Sigma)$ is an outlier. The goal is to construct an estimator for μ , and Σ so that points in the outlier region do not affect estimates.

The proposed estimator in Section 3 is based on a pseudo-density representation of (1). Before to introduce it we fix additional notations. Fix $\beta_1 < \beta_2$ and take $q_p(\beta_1), q_p(\beta_2)$.

$$w(t) := \begin{cases} 0 & \text{if } t < q_p(\beta_1), \\ \frac{1}{2} \left[1 + \frac{2t - q_p(\beta_2) - q_p(\beta_1)}{q_p(\beta_2) - q_p(\beta_1)} + \frac{1}{\pi} \sin \left(\pi \frac{2t - q_p(\beta_2) - q_p(\beta_1)}{q_p(\beta_2) - q_p(\beta_1)} \right) \right] & \text{if } q_p(\beta_1) \leq t \leq q_p(\beta_2); \\ 1 & \text{if } t > q_p(\beta_2). \end{cases} \quad (2)$$

Define

$$\phi_{\beta_1}(\Sigma) := (2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp \left(-\frac{q_p(\beta_1)}{2} \right),$$

which gives the value of the Gaussian density on the border of $\mathcal{E}_{\beta_1}(\mu, \Sigma)$. Define the ‘‘improper black hole density’’ (IBHD) as

$$h(\mathbf{x}; \mu, \Sigma) := \phi_{\beta_1}(\Sigma) w(d^2(\mathbf{x}; \mu, \Sigma)), \quad (3)$$

Features of $h(\cdot)$ are: (i) $h(\cdot) = 0$ for $\mathbf{x} \in \text{int}(\mathcal{E}_{\beta_1}(\mu, \Sigma))$; (ii) $h(\cdot)$ is an improper density because $h(\cdot) = \phi_{\beta_1}(\Sigma) > 0$ for $\mathbf{x} \in \mathbb{R}^p \setminus \mathcal{E}_{\beta_2}(\mu, \Sigma)$; (iii) it increases smoothly from 0 to $\phi_{\beta_1}(\Sigma)$ going from the border of $\mathcal{E}_{\beta_2}(\mu, \Sigma)$ to the border of $\mathcal{E}_{\beta_1}(\mu, \Sigma)$. Once $h(\cdot)$ touches $\mathcal{E}_{\beta_1}(\mu, \Sigma)$ stays at $\phi_{\beta_1}(\Sigma)$ forever. Therefore $h(\cdot)$ puts zero mass on points in $\mathcal{E}_{\beta_1}(\mu, \Sigma)$ creating a ‘‘black hole’’ where we expect that points are typically

sampled under the regular component. Based on (3) define

$$\psi(\mathbf{x}; \theta) := (1 - \alpha)\phi(\mathbf{x}; \mu, \Sigma) + \alpha h(\mathbf{x}; \mu, \Sigma) \quad (4)$$

with $\theta = (\alpha, \mu, \text{vec}(\Sigma))$ and $0 \leq \alpha < 1$. $\psi(\cdot)$ is interpreted as an improper density representation of P where α is a measure of the “global” weight of the nonregular component. The pseudo model (4) is used to define a “local measure” of the weight of the regular component at $\mathbf{x}$:

$$\tau(\mathbf{x}; \theta) := \frac{(1 - \alpha)\phi(\mathbf{x}; \mu, \Sigma)}{\psi(\mathbf{x}; \mu, \Sigma)}. \quad (5)$$

At point $\mathbf{x}$ the quantity $1 - \tau(\mathbf{x}, \theta)$ measures the weight of the nonregular component relative to the overall density approximated by $\psi(\mathbf{x}; \theta)$.

3 Estimator

Define the matrix $\mathbf{S}(\theta) \in \mathbb{R}^{p \times p}$

$$\mathbf{S}(\theta) := \frac{1}{n(1 - \alpha)} \sum_{i=1}^n \tau(\mathbf{x}_i; \theta) (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top.$$

Let $l_1(\theta) \geq l_2(\theta), \dots, \geq l_p(\theta)$ be the ordered eigenvalues of $\mathbf{S}(\theta)$, and $\mathbf{L}(\theta) = \text{diag}(l_1(\theta), l_2(\theta), \dots, l_p(\theta))$. Let $\mathbf{G}(\theta)$ be the matrix of normalized eigenvectors associated to $\text{diag}(\mathbf{L}(\theta))$ so that $\mathbf{S}(\theta) = \mathbf{G}(\theta)\mathbf{L}(\theta)\mathbf{G}(\theta)^\top$ holds. Fix $\gamma \geq 1$ and define

$$l_k(m, \theta) := \begin{cases} m & \text{if } l_k(\theta) < m, \\ l_k(\theta) & \text{if } m \leq l_k(\theta) \leq \gamma m, \\ \gamma m & \text{if } l_k(\theta) > \gamma m. \end{cases} \quad (6)$$

for $k = 1, 2, \dots, p$ and $m > 0$. Furthermore, define the unique solution of the following one dimensional convex program

$$m_*(\theta) := \arg \min_{m > 0} \sum_{k=1}^p \frac{l_k(\theta)}{l_k(m, \theta)} + \log l_k(m, \theta). \quad (7)$$

Let $\mathbf{L}_*(\theta) = \text{diag}(l_1(m_*(\theta), \theta), l_2(m_*(\theta), \theta), \dots, l_p(m_*(\theta), \theta))$. We say that the vector of parameters $\theta = (\alpha, \mu, \text{vec}(\Sigma))$ solves the “Black Hole Equations” if the following system of equations holds:

$$\begin{cases} \alpha - \frac{1}{n} \sum_{i=1}^n (1 - \tau(\mathbf{x}_i; \theta)) & = 0, \\ \mu - \frac{1}{n(1 - \alpha)} \sum_{i=1}^n \tau(\mathbf{x}_i; \theta) \mathbf{x}_i & = 0, \\ \Sigma - \mathbf{G}(\theta)\mathbf{L}_*(\theta)\mathbf{G}(\theta)^\top & = 0. \end{cases} \quad (8)$$

Let Θ_r denote the set of solutions to (8). For $\alpha > 0$ members of Θ_r are *robust candidate approximations* to the μ and the Σ we are looking for. Observations outside $\mathcal{E}_{\beta_1}(\mu, \Sigma)$, will have less or no impact on μ and Σ since the corresponding $\tau(\cdot)$ will be small. The role of γ is to regularize the covariance matrix in high-dimensional settings or more generally when the ratio n/p is modest. In the *oracle case* suppose one knows that P is an uncontaminated normal distribution and sets $\alpha = 0$, then one can show that: (i) if $\gamma = +\infty$, then Θ_r contains a unique solution that coincides with MLE, moreover the resulting Σ is invertible for $n \geq p + 1$, (ii) if $\gamma < +\infty$ then Θ_r contains a unique solution that coincides with the MLE under a condition number constraint studied in [21] (similar constraints have been used in cluster analysis based on Gaussian mixtures, see [4] and references therein). The resulting Σ is invertible even for $n < p + 1$. However, apart from the *oracle case* one cannot generally expect that (8) has a unique solution. Therefore, within Θ_r we look for the solution that produces that best approximation to a Gaussian β_1 -ellipsoid. Define

$$\mathbb{W}_{t,n}(\theta) := \frac{1}{n(1-\alpha)} \sum_{i=1}^n \tau(\mathbf{x}_i; \theta) \mathbb{1} \{d^2(\mathbf{x}_i; \mu, \Sigma) \leq d^2(\mathbf{x}_i; \mu, \Sigma)\}, \quad (9)$$

where $\mathbb{W}_{t,n}(\theta)$ is the a weighted distribution function of the squared Mahalanobis distances according to weights $\tau(\cdot)$. Let $\mathcal{D}(P_1, P_2)$ be the Kuiper distance between distributions P_1 and P_2 , let $\chi_p^2(t)$ be the distribution function of a χ_p^2 at t , then our final estimator for θ is found by solving the following optimization problem:

$$\theta_n := \arg \min_{\theta \in \Theta_r, \alpha \leq \alpha} \mathcal{D}(\mathbb{W}_{t,n}(\theta), \chi_p^2(d^2(\mathbf{x}_i; \mu, \Sigma))) \quad (10)$$

The restriction $\alpha \leq \alpha_{\max}$ is needed in order to guarantee existence when $\gamma = +\infty$. However, $\alpha_{\max} = 0.5$ will just implement the familiar condition in robust statistics that we expect that no more than 50% of the observations are produced by the outlier component of P . The rational of θ_n can be understood by noting that $\mathcal{D}(\cdot)$ measures the lack of Gaussianity for the points in $\mathcal{E}_{\beta_1}(\cdot)$. Notice that the robustness of the estimator is controlled by β_1, β_2 . While β_2 does not impact results too much, β_1 is crucial and can be set by simple arguments inspired by [5]. In fact we may require that for an uncontaminated normal sample $\mathbf{X}_n^r$ $\Pr\{\mathbf{X}_n^r \in \mathcal{E}_{\beta_j}(\mu, \Sigma)\} = 1 - \pi_j$, the latter can be achieved by taking $\beta_j = (1 - \pi_j)^{\frac{1}{n}}$ for $j = 1, 2$. In our experiments $\pi_1 = 10\%$ and $\pi_2 = 5\%$ always produced satisfactory results.

Existence, robustness, computational issues and other theoretical properties of (10) will be discussed in future papers. Extensive numerical experiments have shown that the proposed estimator performs well compared with other robust alternatives. Numerical experiments have been conducted with observation sampled from highly correlated normal distributions with varying dimensions, various types and levels of contamination. Numerical results will be also discussed in future publications.

References

1. Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.
2. Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.
3. Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):672–684, 2011.
4. Pietro Coretto and Christian Hennig. Finding approximately Gaussian clusters via robust improper maximum likelihood, 2013. Url: <http://arxiv.org/abs/1309.6895>.
5. Laurie P Davies and Ursula Gather. The identification of multiple outliers. *J. Amer. Statist. Assoc.*, 88(423):782–801, 1993. With discussion and a reply by the authors.
6. P. L. Davies. Asymptotic behaviour of S -estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, 15(3):1269–1292, 1987.
7. David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston, 1982.
8. Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.
9. Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411, 2004.
10. Hendrik P. Lopuhaä. Multivariate τ -estimators for location and scatter. *Canad. J. Statist.*, 19(3):307–321, 1991.
11. Ricardo A. Maronna and Ruben H. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.
12. Ricardo Antonio Maronna. Robust M -estimators of multivariate location and scatter. *Ann. Statist.*, 4(1):51–67, 1976.
13. Marco Riani, Anthony C. Atkinson, and Andrea Cerioli. Finding an unknown number of multivariate outliers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):447–466, 2009.
14. Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186, 2009.
15. Peter Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht, 1985.
16. Peter J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880, 1984.
17. Werner A Stahel. Breakdown of covariance estimators. Technical report, Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1981.
18. Kay S. Tatsuoaka and David E. Tyler. On the uniqueness of S -functionals and M -functionals under nonelliptical distributions. *Ann. Statist.*, 28(4):1219–1243, 2000.
19. David E. Tyler. A note on multivariate location and scatter statistics for sparse data sets. *Statist. Probab. Lett.*, 80(17-18):1409–1413, 2010.
20. Naisyin Wang and Adrian E. Raftery. Nearest-neighbor variance estimation (NNVE): robust covariance estimation via nearest-neighbor cleaning. *J. Amer. Statist. Assoc.*, 97(460):994–1019, 2002. With comments and a rejoinder by the authors.
21. Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(3):427–450, 2013.
22. Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.