Detecting Outlying Subjects in High-Dimensional Neuroimaging Datasets with Regularized Minimum Covariance Determinant

Virgile Fritsch, Gaël Varoquaux, Thyreau Benjamin, Jean-Baptiste Poline, Bertrand Thirion

▶ To cite this version:

Detecting Outlying Subjects in High-Dimensional Neuroimaging Datasets with Regularized Minimum Covariance Determinant

Virgile Fritsch[1,2], Gael Varoquaux[3,1,2], Benjamin Thyreau[2], Jean-Baptiste Poline[2,1], and Bertrand Thirion[1,2]

[1] Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
virgile.fritsch@inria.fr,
WWW home page: http://parietal.saclay.inria.fr
[2] CEA, DSV, I2BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[3] Inserm, U992, Neurospin bât 145, 91191 Gif-Sur-Yvette, France

Abstract. Medical imaging datasets used in clinical studies or basic research often comprise highly variable multi-subject data. Statistically-controlled inclusion of a subject in a group study, i.e. deciding whether its images should be considered as samples from a given population or whether they should be rejected as outlier data, is a challenging issue. While the informal approaches often used do not provide any statistical assessment that a given dataset is indeed an outlier, traditional statistical procedures are not well-suited to the noisy, high-dimensional, settings encountered in medical imaging, *e.g.* with functional brain images. In this work, we modify the classical *Minimum Covariance Determinant* approach by adding a regularization term, that ensures that the estimation is well-posed in high-dimensional settings and in the presence of many outliers. We show on simulated and real data that outliers can be detected satisfactorily, even in situations where the number of dimensions of the data exceeds the number of observations.

Keywords: Outlier detection, Minimum Covariance Determinant, regularization, robust estimation, neuroimaging, fMRI

1 Introduction

Between-subject variability is a prominent effect in many fields of medical imaging, and particularly in brain imaging. While part of this variability can be viewed as normal fluctuations within a population or across repeated measurements, and can be considered as an effect of interest for diagnosis problems, part of it may be a confound, related to scanner instabilities, experimental issues, or acquisition artifacts. Such confounding factors can be much larger than the effects of interest: for instance, in functional neuroimaging, the variability related to acquisition issues (motion, defective experimental setup, scanner spikes) can mask the true effect of interest, which is the variability in brain functional organization related to diseases, psychological or genetic factors.

The detection of abnormal data, or outlier detection, is important in order to ensure that the ensuing statistical analysis will be robust to such undesired effects. This detection should be automated for the sake of reproducibility and to be time efficient, as cohorts can now encompass up to several hundreds of subjects. This detection is challenging because *i)* images, in particular brain images, are *complex, high-dimensional objects* with some unknown latent structure; *ii)* the problem is *unsupervised*, in the sense that outlier detection procedures can in general not be calibrated on training data; and *iii)* in many cases, it is impossible to *normalize* the signal or its variability.

So far, high-dimensional analysis procedures have been confined to high SNR data, such as anatomical images, e.g. with the use of manifold learning techniques [1, 3]. These, however, are not robust to outlier data, and are not applicable to functional Magnetic Resonance Imaging (fMRI) since they may easily be confounded by noise. As a first step to alleviate this issue, univariate outlier detection methods have been proposed for fMRI, in which one particular image feature is studied, and compared to other data [6, 12]. *Kherif et al.* [5] point out the need of homogeneous datasets in fMRI studies and propose a model-based multivariate framework as a solution. However, their work is restricted to small cohorts and does not discuss statistical control.

While the robust statistics literature generally considers that problems with a number of dimensions comparable to the number of observations cannot be addressed in model-based approaches, we investigate whether outlier detection is still possible in that setting. Specifically, we modify the Minimal Covariance Determinant method [8] so that its performance approaches the level of non-parametric methods, such as one-class Support Vector Classification [2]. We describe the new robust estimator in the next section and show its well-posedness. We then perform some experiments on simulated data and assess the behaviour of the proposed method with respect to state-of-the-art techniques. Finally, we describe the application of our approach to an fMRI dataset, where we show that outliers can still be detected on medium-sized groups of subjects.

2   Robust location and covariance estimates

We focus on a model-based approach, as it yields more interpretable results as well as a probabilistic control of false detections: Assuming a high dimensional Gaussian model, an observation $x_i \in \mathbb{R}^p$ within a set X can be characterized as outlier whenever it has a large Mahalanobis distance to the mean of the data distribution, defined as $d^2_{\hat{\mu},\hat{\Sigma}}(x_i) = (x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})$, $\hat{\mu}$ and $\hat{\Sigma}$ being respectively estimates of the dataset location and covariance. Crucially, robust estimators of location and covariance have to be used for the computation of these distances.

MCD estimator and FastMCD algorithm   The state-of-the-art robust covariance estimator for multidimensional Gaussian data is Rousseeuw's Minimum Covariance Determinant (MCD) estimator [8], which can be computed using the Fast-MCD algorithm [10]. Given a dataset with n p-dimensional observations, MCD

aims at finding h observations (referred to as the *support*), the scatter matrix of which has a minimal determinant. For the scatter matrix to be well-conditioned, h must be greater than $h_{\min} = \frac{n+p+1}{2}$. As $\frac{p}{n}$ becomes large, $h_{\min}$ increases so outliers are potentially included in the covariance estimation if there are more than $\frac{n-p-1}{2}$ of them. When $p = n - 1$, the MCD estimator is equivalent to the unbiased maximum likelihood estimator, which is not robust. Finally, if $p \geq n$, the MCD estimator is not defined. To alleviate these issues we propose to use half of the observations in the support ($h = \frac{n}{2}$) and compensate the shortage of data for covariance estimation with regularization.

Regularized MCD estimator (R-MCD) We consider ridge regularization: let $\lambda \in \mathbb{R}^+$ be the amount of regularization, $(\mu_{\mathrm{r}}, \Sigma_{\mathrm{r}})$ the location and covariance estimates of a $n \times p$ dataset X that maximize the penalized negative log-likelihood:

$$(\mu_{\mathrm{r}}, \Sigma_{\mathrm{r}}) = \operatorname*{argmin}_{\mu, \Sigma} \left(\log |\Sigma| + \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \lambda \operatorname{Tr} \Sigma^{-1} \right), \quad (1)$$

yielding $\Sigma_{\mathrm{r}} = \frac{X^T X}{n-1} + \lambda \operatorname{Id}_p$ and $\mu_{\mathrm{r}} = \frac{1}{n} X^T \mathbf{1}_n$.

Convergence of the Fast R-MCD algorithm Fast-MCD is an iterative algorithm that successively culls out outliers using Mahalanobis distances defined with the covariance of the most homogeneous fraction of the data. In our new algorithm, *Fast-R-MCD*, we replace the sample covariance matrix used in MCD to define the Mahalanobis distance by the ridge estimate. The convergence of Fast-R-MCD stems from the following lemma, that generalizes the proof of convergence of Fast-MCD [4]:

$$\forall \eta > 0, \quad (\mu_{\mathrm{r}}, \Sigma_{\mathrm{r}}) = \begin{cases} \operatorname{argmin}_{\mu, \Sigma} |\Sigma|, \\ \text{s.t.} \quad \mathbb{E}\left[(X - \mu)^{\mathrm{T}} \Sigma^{-1} (X - \mu)\right] + \lambda \operatorname{tr} \Sigma^{-1} = \eta \end{cases} \quad (2)$$

which straightforwardly implies that the determinant of Σ_{r} will decrease at each iteration of the Fast-R-MCD algorithm.

Setting the regularization parameter λ Starting with an initial guess for $\lambda = \frac{\operatorname{tr}(\hat{\Sigma})}{n\,p}$ where $\hat{\Sigma}$ is the unbiased empirical covariance matrix of the whole dataset, we isolate an uncontaminated set of $\frac{n}{2}$ observations, as in the Fast-MCD approach. Let $\lambda = \delta \frac{\operatorname{tr}(\hat{\Sigma}_{\mathrm{pure}})}{n\,p}$, where $\hat{\Sigma}_{\mathrm{pure}}$ is the empirical covariance matrix of the uncontaminated dataset. We choose δ so as to maximize the ten-fold cross-validated log-likelihood of the uncontaminated dataset.

3 Experiments

We compared the outlier detection accuracy that can be obtained from the Mahalanobis distances of the samples, using respectively MCD and R-MCD.

3.1 Simulations

Data generation In our simulations, we sample a core set of $n - q$ $(q < \frac{n}{2})$ observations from a $\mathcal{N}(0_p, \Sigma)$ distribution corresponding to regular observations (also called *inliers*). We add q outliers from a $\mathcal{N}(\mu_q, \Sigma_q)$ distribution ($\mu_q \in \mathbb{R}^p, \Sigma_q \in S_n^+(p)$), thus generating a total of n observations with p features. We use three outliers types (see Fig. 1):

Variance outliers are obtained by setting $\Sigma_q = \alpha\Sigma, \alpha > 1$ and $\mu_q = 0_p$. This situation models signal normalization issues, where the amount of variance in outlier observations is abnormally large.

Multi-modal outliers are obtained by setting $\Sigma_q = \Sigma$ and $\mu_q \neq 0_{\mathbb{R}^p}$, which simulates the presence of an heterogeneous population.

Multivariate outliers are obtained by setting $\mu_q = 0_p$, $\Sigma_q = \Sigma + \alpha\,\mathbf{a}\,\mathbf{a}^T$ where $\mathbf{a} = \frac{\mathbf{a}_{\mathrm{rand}}}{||\mathbf{a}_{\mathrm{rand}}||_2}$ and $\mathbf{a}_{\mathrm{rand}}$ is a vector p-dimensional vector with coordinates drawn from a Bernoulli distribution $\mathcal{B}(\frac{1}{2})$. This model simulates outliers as sets of points having potentially abnormally high values in some random directions.

In our experiments, we also investigated the influence of Σ's condition number $\kappa(\Sigma) = ||\Sigma||_2 \cdot ||\Sigma^{-1}||_2$ and contamination rate $\frac{q}{n}$.

Methods comparison Given a simulated dataset, we estimated the location and covariance of the data using MCD and R-MCD estimators. Both were computed with the Fast-MCD (or Fast-R-MCD) algorithm without consistency and re-weighting steps (see [10]), leading to what we call *raw estimates*. The parameter that influences the most the relative performances of MCD- and R-MCD-based outlier detection methods is the $\frac{p}{n}$ ratio. Every other parameter being fixed, we averaged 100 ROC curves for each value of $\frac{p}{n}$ in a given range, and finally expressed Area Under Curve (AUC) as a function of $\frac{p}{n}$.

We also compare the R-MCD sensitivity with the One-class SVM sensitivity, holding the latter as a reference since it is not limited by any prior shape of the separation between in- and outlying observations. We used a *RBF* kernel and selected its bandwidth γ with an heuristic inspired by [11]: $\gamma = \frac{0.01}{\Delta}$, where Δ is the 10th percentile of the pairwise distances histogram of the observations.
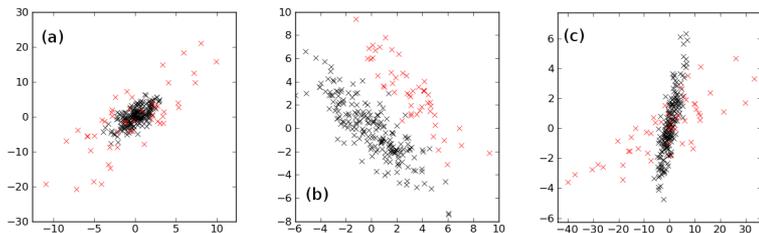

Fig. 1. Three different ways to generate multivariate outliers for Gaussian data. **(a)** all directions ($\alpha = 3$). **(b)** second cluster ($\mu_q = 3 \times \mathbf{1}_{\mathbb{R}^2}$). **(c)** multivariate ($\alpha = 5$). Outliers are represented in red and inliers in black. $\kappa(\Sigma) = 10$. Contamination is 40%.

3.2 Outliers identification in functional neuroimaging

We used data from a large functional neuroimaging database containing several fMRI contrast images in more than 1500 subjects. 3T scanners from multiple manufacturers were used for acquiring the data with TR = 2200 ms, TE = 30 ms, and flip angle = 75°. Standard preprocessing were performed on the data using the SPM8 software. Here we focus on a control contrast that shows brain regions implied in auditory tasks as opposed to visual tasks.

We used a probabilistic brain atlas [7] to extract an average activation intensity value from 145 regions of interest in all the contrast images. We then performed an initial outlier detection at $P < 0.1$ familywise corrected, including more than 1000 subjects. With such a small $\frac{p}{n}$ value, a statistically controlled outlier detection could be done using the MCD estimate. The outliers list obtained from this first outlier detection was then held as a ground truth for further outlier detection experiments performed on reduced sample, using MCD and R-MCD estimators. Note that for very small samples, we could not use the MCD-based outliers detection method. The outliers lists were compared to the ground truth and ROC curves were hence constructed. For each sample size, we repeated the detection 10 times with 10 different, randomly selected samples.

4 Results

4.1 Simulation results

We first give the results on a 30-dimensional dataset with a 40% contamination rate ($\frac{q}{n} = 0.4$), generated from the multivariate outliers model. We show the case where $\kappa(\Sigma) = 1000$. The accuracy of the R-MCD-based method is much higher than the accuracy of the MCD-based method as soon as $\frac{p}{n} > 0.3$ (Fig 2). Using regularization, it is possible to go beyond the $p = n$ limit, keeping an AUC greater than 0.70.
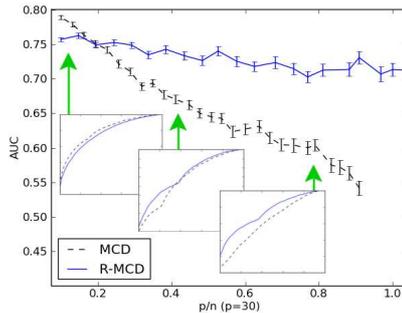

Fig. 2. AUC for MCD- and R-MCD-based outliers detection methods. 40% multivariate outliers are generated ($\alpha = 5$, $\kappa(\Sigma) = 1000$). The R-MCD-based method keeps an AUC of $\simeq 0.70$ up to $\frac{p}{n} = 3.5$ (not shown) while the MCD-based method breaks down.

Outliers type Table 1.a summarizes the AUC results for experiments with the variance outliers model. In this case, the MCD performance drops when $\frac{p}{n} > 0.5$ while the R-MCD-based method always achieves at least a 80% accuracy. In the multimodal case (Table 1.b), the R-MCD-based method is also more successful at detecting outliers.

Covariance matrix condition number and contamination rate The methods' performance depends weakly on the condition number. A small condition number yields better results with R-MCD while a high condition number gives advantage to the use of the MCD estimator when $\frac{p}{n}$ is small. The higher the contamination rate, the more MCD estimator is likely to break. On the other hand, the performance of a R-MCD-based detection method is stable (Table 2).

Comparison to One-class SVM In both cases of variance and multivariate outliers, One-class SVM achieves a better specificity/sensitivity compromise than R-MCD-based outlier detection method. Yet, for a $\frac{p}{n}$ ratio of the order of .5, the R-MCD performance remains comparable to that of the One-class SVM, with an asymptotic score that remains below (0.05 difference in the AUC).

4.2 Application on a real dataset

We give the averaged ROC curves for detecting outliers on an auditory task in Fig 4. The reference outlier detection was performed on 1118 subjects, each being described by 145 features. The results shown correspond to 10 random sets of 290 subjects. In the useful range (FP $\leq$ 5%), R-MCD outperforms MCD. Even with only 100 samples ($\frac{p}{n} = 1.5$), R-MCD can still be used to find outliers, as Fig 4 and Fig 5 demonstrate. On this latter figure, outlying subjects 1, 2 and 3 indeed exhibit much variable activity patterns than subjects A or B, despite the presence of a few mistakes (subject 4).

5 Discussion

In high-dimensional Gaussian datasets, our results show that Regularized MCD can reach a significantly higher sensitivity in outlier detection than the standard MCD. We assumed that neuroimaging data are distributed according to a multivariate Gaussian distribution. This strong hypothesis lead us to focus on Mahalanobis-distances-based approaches since they can exploit the assumed shape of the dataset to *estimate its covariance matrix*. Since R-MCD systematically deals with half of the observations, it is not subject to the known masking and swamping effects [9]. We plan to investigate a ℓ_1 norm for covariance regularization, as it may fit with standard hypotheses on brain covariance structure.

Under the Gaussian assumption we made, outlier detection with the R-MCD estimator is the only method so far that both holds in high-dimension and allows a probabilistic control of the false detection rate. Although the One-class

Table 1. AUC values: **a: left** variance outliers model ($p = 30$, $q/n = 40\%$, $\alpha = 1.25$), **b: right** multimodal outliers model ($p = 30$, $q/n = 20\%$, $\mu_q = 2 \cdot \mathbf{1}_p$).

p/n	0.1	0.2	0.3	0.4	0.5	0.7	0.8
MCD	0.86	0.82	0.77	0.73	0.70	0.66	0.63
R-MCD	**0.87**	**0.86**	**0.85**	**0.85**	**0.84**	**0.82**	**0.82**

p/n	0.1	0.2	0.3	0.4	0.5	0.7	0.8
MCD	0.62	0.60	0.58	0.57	0.55	0.55	0.51
R-MCD	**0.76**	**0.77**	**0.78**	**0.81**	**0.78**	**0.75**	**0.77**

Table 2. Influence of the contamination rate. $p = 30$, multivariate outliers ($\alpha = 5$). Unlike MCD, the R-MCD performances are independent of the contamination value.

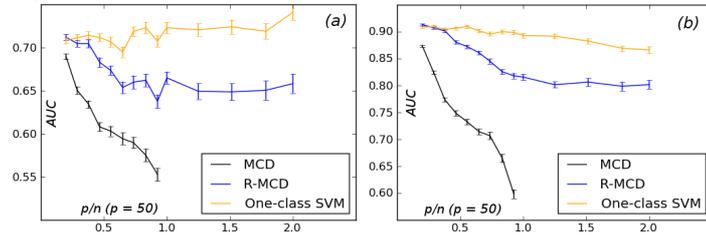| q/n	10 %				20 %				30 %				40 %			
p/n	0.1	0.3	0.6	0.9	0.1	0.3	0.6	0.9	0.1	0.3	0.6	0.9	0.1	0.3	0.6	0.9
MCD	**0.83**	**0.77**	0.68	0.56	**0.82**	0.74	0.65	0.57	**0.81**	0.72	0.65	0.55	**0.79**	0.71	0.64	0.54
R-MCD	0.77	0.74	**0.72**	**0.68**	0.76	**0.75**	**0.72**	**0.71**	0.76	**0.74**	**0.72**	**0.72**	0.76	**0.75**	**0.74**	**0.71**

Fig. 3. One-class SVM comparison with 40% contamination and $\kappa(\Sigma) = 100$. **(a)** AUC for multivariate outliers ($\alpha = 5$). **(b)** AUC for variance outliers ($\alpha = 1.25$).
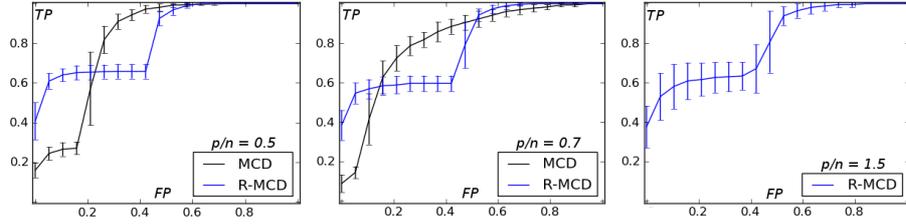

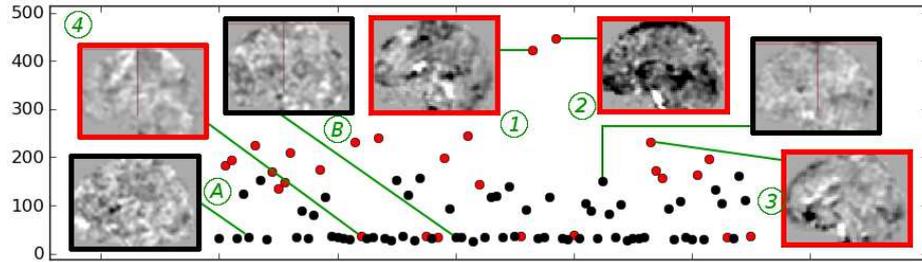**Fig. 4.** ROC curves showing that R-MCD outperforms MCD on real fMRI data.

Fig. 5. R-MCD-based Mahalanobis distances of a small sample. The higher the Mahalanobis distance, the higher the probability for an observation to be tagged as outlying. Points in red are outliers subjects according to the whole population.

SVM non-parametric algorithm achieves a better sensitivity/specificity compromise and is still applicable with non-Gaussian data, its lack of interpretability and statistical control, as well as the difficulty to tune its parameters, makes it unsuitable in a medical context.

Conclusion We introduced the R-MCD estimator, a regularized version of a robust covariance estimator commonly used to detect outlying observations on the basis of their Mahalanobis distances. We showed that the Fast-MCD algorithm is still valid to compute this new estimator. Our application to neuroimaging, where studies have a high exclusion rate, shows that it is possible to build automatic procedures to detect outliers even though the number of descriptors is higher than the number of available subjects. This property is of broad interest in medical applications where heterogeneous populations have to be considered and relies on an objective assessment of normal variability.

References

1. Aljabar, P., Wolz, R., Srinivasan, L., Counsell, S., Boardman, J.P., Murgasova, M., Doria, V., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D.: Combining morphological information in a manifold learning framework: application to neonatal MRI. Med Image Comput Comput Assist Interv 13, 1 (2010)
2. Gardner, A., Krieger, A., Vachtsevanos, G., Litt, B.: One-class novelty detection for seizure analysis from intracranial EEG. J. Mach Learn Res 7, 1025 (2006)
3. Gerber, S., Tasdizen, T., Joshi, S., Whitaker, R.: On the manifold structure of the space of brain images. Med Image Comput Comput Assist Interv 12, 305 (2009)
4. Grubel, R.: A minimal characterization of the covariance matrix. Metrika 35, 49 (1988)
5. Kherif, F., Flandin, G., Ciuciu, P., Benali, H., Simon, O., Poline, J.B.: Model based spatial and temporal similarity measures between series of functional magnetic resonance images. Med Image Comput Comput Assist Interv p. 509 (2002)
6. Penny, W.D., Kilner, J., Blankenburg, F.: Robust bayesian general linear models. Neuroimage 36, 661 (2007)
7. Perrot, M., Rivière, D., Tucholka, A., Mangin, J.F.: Joint bayesian cortical sulci recognition and spatial normalization. Inf Process Med Imaging 21, 176–187 (2009)
8. Rousseeuw, P.J.: Least median of squares regression. J. Am Stat Ass 79, 871 (1984)
9. Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. WIREs Data Mining Knowl Discov 1, 73 (2011)
10. Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3), 212 (1999)
11. Segata, N., Blanzieri, E.: Fast and scalable local kernel machines. J. Mach Learn Res 11, 1883 (2009)
12. Woolrich, M.: Robust group analysis using outlier inference. Neuroimage 41, 286 (2008)