

Detecting Outliers in High-Dimensional Neuroimaging Datasets with Robust Covariance Estimators

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

▶ **To cite this version:**

Virgile Fritsch, Gaël Varoquaux, Benjamin Thyreau, Jean-Baptiste Poline, Bertrand Thirion. Detecting Outliers in High-Dimensional Neuroimaging Datasets with Robust Covariance Estimators. Medical Image Analysis, 2012, 16, pp.1359-1370. 10.1016/j.media.2012.05.002 . hal-00701225

HAL Id: hal-00701225
https://inria.hal.science/hal-00701225

Submitted on 24 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detecting Outliers in High-Dimensional Neuroimaging Datasets with Robust Covariance Estimators

Virgile Fritsch[a,c], Gaël Varoquaux[b,a,c], Benjamin Thyreau[c], Jean-Baptiste Poline[c,a], Bertrand Thirion[a,c]

[a] *Parietal Team, INRIA Saclay-Île-de-France, Saclay, France*
[b] *Inserm, U992, Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[c] *CEA, DSV, I^{2}BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France*

Abstract

Medical imaging datasets often contain deviant observations, the so-called *outliers*, due to acquisition or preprocessing artifacts or resulting from large intrinsic inter-subject variability. These can undermine the statistical procedures used in group studies as the latter assume that the cohorts are composed of homogeneous samples with anatomical or functional features clustered around a central mode. The effects of outlying subjects can be mitigated by detecting and removing them with explicit statistical control. With the emergence of large medical imaging databases, exhaustive data screening is no longer possible, and automated outlier detection methods are currently gaining interest. The datasets used in medical imaging are often high-dimensional and strongly correlated. The outlier detection procedure should therefore rely on high-dimensional statistical multivariate models. However, state-of-the-art procedures, based on the Minimum Covariance Determinant (MCD) estimator, are not well-suited for such high-dimensional settings. In this work, we introduce regularization in the MCD framework and investigate different regularization schemes. We carry out extensive simulations to provide backing for practical choices in absence of ground truth knowledge. We demonstrate on functional neuroimaging datasets that outlier detection can be performed with small sample sizes and improves group studies.

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

1. Introduction

Medical image acquisitions are prone to a wide variety of errors such as scanner instabilities, acquisition artifacts, or issues in the underlying bio-medical experimental protocol. In addition, due to the high variability observed in populations of interest, these datasets may also contain uncommon, yet technically correct, observations. The inclusion of overly noisy or aberrant images in medical datasets typically results in additional analysis and interpretation challenges. In both cases, images deviating from normality are called *outliers*. Outliers may be numerous, especially in neuroimaging, where the between-subjects variability of anatomical and functional features is very high and images can have a low signal-to-noise ratio. Considering the dramatic influence of outliers in standard statistical procedures such as Ordinary Least Squares regression [12, 26], clustering [3, 7], manifold learning [37] or neuroimaging group analyzes [16, 18], it is crucial to detect outliers as a preprocessing step of any statistical study. Once labeled as such, outlying observations can be down-weighted or even dis-

Email address: virgile.fritsch@inria.fr (Virgile Fritsch)

carded in group studies. Down-weighting requires a measure of the deviation from normality for each observation and the rejection of observations must be grounded on a statistical control. Furthermore, an automatic procedure for data screening is also necessary in the case of a large number of observations and to avoid subjective bias.

An intrinsic difficulty of outlier detection in medical imaging lies in the lack of formal definition for abnormal data; in particular, no generative model for outliers might be sufficient to model the variety of situations where such data are observed in practice. Moreover, in high-dimensional settings, i.e. when the number of observations is less than five times the number of data descriptors (or *features*) [9], the problem of outlier detection is ill-posed since it becomes very difficult to characterize deviations from normality. From a practical perspective, manual outlier detection is impossible in such a situation. Current methods dealing with outliers in a high-dimensional context are essentially univariate methods, i.e. they consider different dimensions one by one [22, 39]. These methods may fail to tag as outliers observations that are deviant with respect to a combination of several of their characteristics, but for which each descriptor considered individually does not reveal deviation from normality. Medical imaging data, and in particular neuroimaging data, are high dimensional, the underlying dimension being the number of degrees of freedom in their variance, which can be of the order of the number of image voxels. This is typically much larger than the number of available samples. In functional MRI studies, neuroscientists often screen the data manually (see e.g. [24]), because of the lack of an adapted outlier detection framework. The criteria for discarding data are not always quantitatively defined. For instance, images may be discarded if, upon visual inspection, they are not reflecting the expected brain activation pattern (e.g. in a so called contrast map). Such a process is tedious and unreliable, but most importantly it makes the statistical analysis of the group data invalid for that pattern – as it implies that the variance of this pattern will be underestimated.

In this contribution, we explore several extensions to the state-of-the-art outlier detection framework of [26] to high-dimensional settings. In particular, we introduce and compare three procedures, all based on a robust and regularized covariance estimate. Using a covariance estimate relies on the assumption that regular observations, called the *inliers*, are Gaussian distributed, and that outliers are characterized by some distance to the standard model. We simulate various scenarii that result in outliers by using mixture models, where the location and covariance parameters of the outliers component are chosen according to three different outlier models. We focus mostly on the accuracy of outlier detection procedures, and only address the challenging question of exact statistical control through simulation procedures; importantly, our choice of a parametric approach facilitates this control. Eventually, we compare our parametric approach to a non-parametric procedure, the One-Class Support Vector Machine (One-Class SVM) algorithm, since this method has raised much interest recently [8, 19].

The layout of the paper is the following. In Section 2, we present the state-of-the-art outlier detection framework, comprising a robust estimator of location and covariance, the *Minimum Covariance Determinant (MCD)* [25], and we point out its limitations in our context. We then introduce in Section 3 three new robust covariance estimators derived from the MCD but suitable for high-dimensional settings. In Section 4, we present experiments on simulations that we use to assess the performance of our new outlier detection methods, together with the corresponding results. Experiments on anatomical and functional neuroimaging datasets are then described in Section 5. Finally, in Section 6, we discuss the results in the context of statistical inference in functional neuroimaging group studies.

Notations. We write vectors with bold letters, $\boldsymbol{a} \in \mathbb{R}^n$, matrices with capital bold letters, $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. $\boldsymbol{I}$ is the identity matrix. $\mathbf{0}$ and $\mathbf{1}$ are constant matrices filled respectively with 0 and 1. Quantities estimated from the data at hand are written with a hat, e.g. $\widehat{\boldsymbol{A}}$. $\boldsymbol{A}^\mathsf{T}$ is the transposed matrix and $\boldsymbol{A}^{-1}$ is the matrix inverse of $\boldsymbol{A}$. We call the inverse of a covariance matrix the *precision matrix*. $\boldsymbol{X}$ refers to an $n \times p$ matrix representing a dataset of n observations rep-

resented by p features. Considering a set of integers H, we denote $\boldsymbol{X}_H$ the matrix $(x_{i\,j})_{i\in H,\,j\in[1..p]}$. $\kappa(\boldsymbol{A})$ is the condition number of the matrix $\boldsymbol{A}$, defined by $\frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$, $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ being respectively the largest and the smallest eigenvalues of $\boldsymbol{A}$. chol$(\boldsymbol{A})$ is the lower triangular matrix obtained by the Cholesky decomposition of $\boldsymbol{A}$.

2. Detecting outlying observations: state of art procedures

2.1. MCD estimator: robust location and covariance estimates

Assuming a high-dimensional Gaussian model, an observation $\boldsymbol{x}_i \in \mathbb{R}^p$ within a set $\boldsymbol{X}$ can be characterized as outlier whenever it has a large Mahalanobis distance to the mean of the data distribution, defined as $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ being respectively the dataset location and covariance. Crucially, robust estimators of location and covariance must be used to compute these distances [2, 23].

The state-of-the-art robust covariance estimator for multidimensional Gaussian data is Rousseeuw's Minimum Covariance Determinant (MCD) estimator [25]. Given a dataset with n p-dimensional observations, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, MCD aims at finding h observations considered as inliers, by minimizing the determinant of their scatter matrix. We refer to these observations as the *support* of the MCD.

The core procedure commonly used to compute the MCD estimate of the covariance of a population is given in Algorithm 1. It consists of alternatively choosing a subset $\boldsymbol{X}_H$ of h observations to minimize a Mahalanobis distance, and updating the covariance matrix $\hat{\boldsymbol{\Sigma}}_H$ used to compute the Mahalanobis distance. $|\hat{\boldsymbol{\Sigma}}_H|$ decreases at each update of $\boldsymbol{X}_H$. Standard algorithms such as the Fast-MCD algorithm [27] perform this simple procedure several times from different initial subsets $\boldsymbol{X}_H$ and retain only the solution with the minimal determinant. The MCD can be understood as an alternated optimization of the following problem:

$$(\hat{H}, \hat{\boldsymbol{\mu}}_{\mathrm{h}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{h}}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma},H}{\operatorname{argmin}} \Bigg(\log|\boldsymbol{\Sigma}| + \frac{1}{h}\sum_{i\in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \Bigg), \tag{1}$$

The limitations of the MCD come from the fact that the scatter matrix must be full rank, as it is used to define a Mahalanobis distance. As a consequence, h must be greater than $h_{\min} = \frac{n+p+1}{2}$: the MCD cannot learn the inlier distribution if there are less than $h_{\min}$ inliers. In high-dimensional settings, as $\frac{p}{n}$ becomes large, $h_{\min}$ increases and 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 address 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, referred to as Regularized MCD in the remainder of the text.

2.2. One class SVM: a non-parametric procedure

Medical imaging data is not necessarily well described by a Gaussian distribution. Thus it might be profitable to seek decision rules not based on Mahalanobis distances to screen deviant data. For this, we use the *One-Class Support Vector Machine (One-Class SVM)* [29], which is not limited by any prior shape of the separation between in- and outlying observations. This choice was motivated by the fact that other robust, high-dimensional, non-parametric tools such as Robust PCA [13] or Local Component Analysis [28] have not yet been considered in practical applications. The One-Class SVM algorithm solves the following quadratic program:

$$\min_{w\in F, \boldsymbol{\xi}\in\mathbb{R}^n, \rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu\, n}\sum_i \xi_i - \rho \tag{2}$$

$$\text{subject to} \quad (w\cdot\Phi(\boldsymbol{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \tag{3}$$

where Φ is a feature map $\mathbb{R}^p \to F$ verifying $K(\boldsymbol{x},\boldsymbol{y}) = \Phi(\boldsymbol{x})\cdot\Phi(\boldsymbol{y})$ for any observations $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and a given

Algorithm 1 MCD estimation algorithm

1. Select h observations (call the corresponding dataset $\boldsymbol{X}_H$);
2. Compute the empirical covariance $\hat{\boldsymbol{\Sigma}}|H$ and mean $\hat{\boldsymbol{\mu}}|H$;
3. Compute the Mahalanobis distances $d^2_{\boldsymbol{\mu}|H,\boldsymbol{\Sigma}|H}(\boldsymbol{x}_i)$, $i = 1..n$;
4. Select the h observations having the smallest Mahalanobis distance;
5. Update $\boldsymbol{X}_H$ and repeat steps 2 to 5 until $|\hat{\boldsymbol{\Sigma}}|H|$ no longer decreases.

kernel K. The important parameter of the One-Class SVM is the margin parameter ν, which is both an upper bound on the proportion of observations that will lie outside the frontier learned by the algorithm and a lower bound on the number of support vectors of the model [29].

In our experiments on simulated data, we set ν to the amount of contamination (i.e. the proportion of outliers in the dataset). Note that this choice favors the One-Class SVM compared to methods that ignore the ratio of outliers. For real data experiments, we set $\nu = 0.5$ as we work with at most 50% contamination. We use a *Radial Basis Function* (RBF) kernel and select its inverse bandwidth σ with an heuristic inspired by [32]: $\sigma = \frac{0.01}{\Delta}$, where Δ is the 10^{th} percentile of the pairwise distances histogram of the observations. We verified that this heuristic is close to the optimum parameter on simulations, although the results are not very sensitive to mild variations of σ around this value.

3. Regularized MCD estimators

To improve upon the classical MCD estimator, we introduce regularization in the MCD estimation procedure.

3.1. ℓ_2 regularization (RMCD-ℓ_2)

We first investigate outlier detection with estimators resulting from a penalized versions of the likelihood in Equation 1. This corresponds to replacing the step 2 of the MCD Algorithm 1 by a penalized maximum-likelihood estimate of the covariance matrix.

We consider ℓ_2 *regularization* (or *ridge regularization*): let $\lambda \in \mathbb{R}^+$ be the amount of regularization, and $\hat{\boldsymbol{\Sigma}}_{\text{r}}|H$ the covariance estimate of a $n \times p$ dataset $\boldsymbol{X}_H$ that maximizes the penalized negative log-likelihood:

$$(\hat{\boldsymbol{\mu}}_{\text{r}}, \hat{\boldsymbol{\Sigma}}_{\text{r}}|H) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}} \Bigg(\log|\boldsymbol{\Sigma}| + \frac{1}{h}\sum_{i\in H}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda \operatorname{Tr}\boldsymbol{\Sigma}^{-1} \Bigg), \tag{4}$$

yielding $\hat{\boldsymbol{\Sigma}}_{\text{r}}|H = \frac{\boldsymbol{X}_H^{\mathsf{T}}\boldsymbol{X}_H}{h} + \lambda \mathrm{I}$ and $\hat{\boldsymbol{\mu}}_r|H = \frac{\boldsymbol{X}_H^{\mathsf{T}}\mathbf{1}}{h}$. The covariance estimate is biased toward a spherical covariance matrix. This bias corresponds to an underlying assumption of isotropy. If the inlier distribution violates strongly this prior, the bias may introduce outliers in the estimator's support.

The λ parameter has to be chosen carefully to obtain the right trade-off between ensuring the invertibility of the estimated covariance matrix and not introducing too much bias in the estimator. If $\lambda = 0$ we recover the MCD estimator and its limitations. On the contrary, if λ is very large, the data structure is not taken into account since the distance becomes then the Euclidean distance to the data median. We proceed as follows: starting with an initial guess for $\lambda = \frac{1}{np}\operatorname{Tr}(\hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is the unbiased empirical covariance matrix of the whole dataset, we isolate an uncontaminated set of $\frac{n}{2}$ observations that provides the RMCD support. Using convex shrinkage, the estimated covariance matrix $\boldsymbol{\Sigma}_{\text{lw}}$ can be expressed as $(1-\alpha)\hat{\boldsymbol{\Sigma}} + \frac{\alpha}{p}\operatorname{Tr}(\hat{\boldsymbol{\Sigma}})\boldsymbol{I}$. We use the formula in [17] for the shrinkage coefficient α that gives the optimal solution $\alpha^\star$ in terms of Mean Squared Error (MSE) between the real covariance matrix to be estimated and the shrunk covariance matrix, yielding $\lambda = \frac{\alpha^\star}{p\,(1-\alpha^\star)}\operatorname{Tr}(\hat{\boldsymbol{\Sigma}})$. Alternative strategies for the choice of λ are discussed in Appendix B.

We also considered a ℓ_1-regularized version of the MCD (RMCD-ℓ_1, see Appendix A) that we found not to be competitive with alternative approaches, in terms of outlier detection accuracy as well as computation time.

3.2. Random Projections (RMCD-RP)

Another way to regularize the MCD estimator in a high-dimensional context is to run it on datasets of reduced dimensionality via random projections. This dimensionality reduction is done by projecting to a randomly selected subspace of dimension $k < p$. Outlier detection can be performed with the MCD on the projected data if k/n ratio is small enough. Since the choice of the projection subspace is crucial for detection accuracy, the procedure has to be repeated several times in order not to miss the most discriminating subspaces. In our experiments, the results of the detections were averaged using the geometric mean of the p-values obtained in the different projections.

Setting the subspace dimension. The choice of the dimension k of the projection subspace is crucial. A too small value of k results in a large loss of information during the projection step and thus raises the issues encountered with the univariate method. On the other hand, for large values of k, the geometry is preserved but the method might suffer the same issues as the MCD, even though the dimensionality reduction should make RMCD-RP more robust. We performed several outlier detection experiments with various choices for the value of k between $p/10$ and p. Our observation was that taking $k = p/5$ was a good trade-off. This choice furthermore ensures that the RMCD-RP-based outlier detection method will be applicable for p/n ratios up to 1, since the underlying MCD-based outlier detections take place in a context where the MCD is computationally stable ($k/n < 0.2$).

Setting the number of projections. While too many random projections is computationally costly for a limited gain, too few projections may miss a good *angle* of the dataset. Outlier detection experiments convinced us that a number of projections equal to the number of dimensions is enough to explore the whole working space while being computationally tractable: further increase of this parameter does not improve the performance of the RMCD-RP method.

3.3. Regularized MCD computation

As we have seen in subsection 2.1, the computation of the MCD estimate is performed several times from different initial subsets of observations. In order to emphasize the effect of the random selection of initial subsets, we first project our dataset to a random one-dimensional subspace and use the h observations closest to the median of the projected dataset, as a starting subset for the MCD computation. In our high-dimensional context, this strategy consistently yielded a better solution than Algorithm 1.

We also modify the convergence criterion of Algorithm 1 to compute regularized estimates of the MCD. The penalized negative log-likelihood corresponding to the model is now the function minimized by the procedure. We can therefore use the expression of the penalized log-likelihood as a criterion for the convergence of the algorithm.

3.4. Statistical control of the outlier detection procedure

In the absence of closed-formula, we propose to control the specificity of the outlier detection procedures using simulations, as described in Appendix C. Note that this is only possible with the parametric approach used here, since the Mahalanobis distance can easily be compared between real data and adapted simulations.

4. Experiments on simulated data

We compare the outlier detection accuracy obtained from the Mahalanobis distances of simulated datasets, using the MCD and its regularized versions. We also include the One-Class SVM in our comparisons since it is more robust to deviations from the Gaussian distribution hypothesis. One potential drawback of the One-Class SVM is that it has many parameters to tune, which is difficult in the absence of the ground truth. In particular, we choose the threshold on the decision function that sets the number of

outliers detected to control the specificity/sensitivity trade-off.

Since the methods investigated are location invariant, we can make the assumption that the inliers are centered ($\boldsymbol{\mu} = \mathbf{0}$) without loss of generality.

4.1. Simulations description

Let $\boldsymbol{\Sigma} = \text{chol}(\boldsymbol{C}\boldsymbol{Z}\boldsymbol{C}^\intercal)$ be the covariance matrix for the inliers, where $c_{i,j} \sim \mathcal{U}(0,1)\ \forall i,j$ and $\boldsymbol{Z}$ is a diagonal matrix with uniformly distributed diagonal elements, rescaled so that the smallest is 1 and the largest is $\kappa(\boldsymbol{\Sigma})$. Let $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ be the location and covariance matrix for the outliers. we simulate three outliers types using mixture models (see Figure 1):

Variance outliers are obtained by setting $\boldsymbol{\Sigma}_q = a\boldsymbol{\Sigma}$, $a > 1$ and $\boldsymbol{\mu}_q = \mathbf{0}$. This situation models signal normalization issues or aberrant data, where the amount of variance in outlier observations is abnormally large. This type of outlier only requires the accurate estimation of the covariance up to a multiplicative factor, so that performance drops hint at numerical stability issues. Indeed, even if some outliers are included into the MCD computation, the location and covariance estimates are not shifted towards those outliers because of the global symmetry of the whole dataset. The ranking of the Mahalanobis distances are thus the same as with the real covariance and only computational issues would explain a drop in the MCD-based outlier detection accuracy.

Multimodal outliers are obtained by setting $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_q = b\mathbf{1}$. This simulates the study of an heterogeneous population. Multimodal outliers challenge the outlier detection methods in terms of robustness in high-dimensional settings. For instance, when $\frac{p}{n} \leq \frac{(1-2\gamma)p}{p+1}$ (where γ is the rate of contamination), detecting outliers with the MCD estimator theoretically yields perfect results; but when the modes are distant enough, it suffices to include only one outlier in the support H to bias the MCD estimate. Therefore, we expect the MCD performance to drop when p increases.

Multivariate outliers are obtained by setting $\boldsymbol{\mu}_q = \mathbf{0}$, $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma} + c\,\sigma_{\max}(\boldsymbol{\Sigma})\,\boldsymbol{a}\,\boldsymbol{a}^\intercal$ where $\boldsymbol{a}$ is a p-dimensional vector drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ distribution. This model simulates outliers as sets of points having potentially abnormally high values in some random direction. While variance and multimodal outliers are useful to empirically demonstrate the theoretical limitations of the MCD, multivariate outliers's random shape offer a good framework for testing the accuracy of the new RMCD-based outlier detection methods that we propose.

In each case, we relied on the theoretical result[1] $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}) \sim \chi^2_p$ to generate the outlier observations in such a way that with a probability of 99%, they do not fall in the inliers support. This was done to ensure that we can distinguish between in- and outliers if we know the real covariance matrix of the former.

4.1.1. Relevant models parameters.

Beyond the outlier type, we investigated the influence of various model parameters impacting the global configuration of the outlier detection problem. To compare the robustness of the methods investigated, and estimate their breakdown points, we look at their behavior under various *amounts of contamination* γ, which is the proportion of outliers in the dataset. Second, because regularized estimators of covariance are biased toward a spherical covariance model, we evaluate the methods performance for inliers covariance matrix having a *condition number* $\kappa(\boldsymbol{\Sigma})$ comprised between 1 and 10,000. Finally, as the ℓ_1 regularization is known to benefit from the *sparsity* of the original inliers precision matrix, we also look at this parameter's influence.

4.1.2. Deviation from Gaussian distribution.

Real-world data, and in particular, medical imaging data, are often not Gaussian distributed [5, 15, 35]. Yet, in absence of a better model, assuming that the observations are Gaussian distributed is a very popular choice in many fields of applied statistics and within the neuroimaging community, as it amounts to reducing data models to the specification of location and covariance parameters.

In order to address deviations from normality, we simulate neuroimaging real data as data coming from

[1] Note that this result only holds for $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X})$, as discussed in Appendix C.

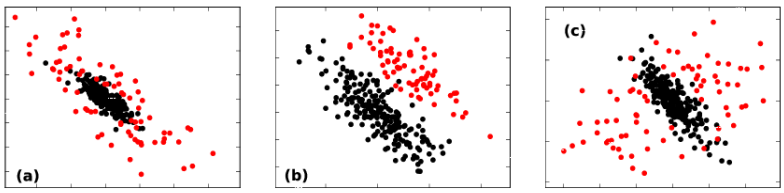


Figure 1: Three different ways to generate multivariate outliers for Gaussian data with $p = 2$. **(a) Variance outliers** ($a = 3$). **(b) Multi-modal outliers** ($b = 3$). **(c) Multivariate outliers** ($c = 5$). The contamination rate is 25%.

a mixture of m Gaussian distributions, the modes of which are randomly drawn from a $\mathcal{N}(\mathbf{0}, \frac{1}{\beta^2}\boldsymbol{I})$ distribution. The β parameter controls the expected distance between the modes. Each component of the model is affected by a given number of variance outliers ($a = 1.15$, see Section 4.1). We also generated outliers so that they do not lie within the 99% support of any of the components. We choose the β parameter in such a way that the different components overlap. To quantify the deviation from gaussianity of our simulated dataset, we look at the distribution of the p-values of a thousand normality tests (Shapiro test [33]) performed on random one-dimensional projections of the data, and report how frequently these p-values are below .05.

4.1.3. Success metrics.

For a given outlier model and a fixed p/n ratio, we call an *experiment* 100 outlier detection runs, using MCD and its regularized versions. We average the results of these runs to build a unique ROC curve [40] per method, and the Area Under the Curve (AUC) [10] is computed. AUC values obtained for various p/n ratios provide a measure of each method's accuracy for outlier detection.

4.2. Experimental results

We first give general results for simulated data according to outliers types before investigating how these results can be influenced by the amount of contamination or the inliers covariance matrix condition number and sparsity. All our results are given for a number of features p equal to 100, similar to the real setting ($p = 113$). They hold for greater or lower dimensions (data not shown), although small dimensions are of no interest and computation time becomes a burden for very high dimensions. When reporting results, we denote by *oracle* the best possible decision, knowing the underlying distributions of inliers and outliers.

4.2.1. Variance outliers.

As illustrated in Figure 2, we also observe a significant drop of the MCD accuracy as p/n increases. The MCD ℓ_1- and ℓ_2-regularized versions always give an accuracy above 0.9, RMCD-ℓ_1 performing a bit better. RMCD-RP does not perform well and the OCSVM method can break down if the condition number is very large (not shown). RMCD-ℓ_1 and RMCD-ℓ_2 performance show that the regularization parameter selection is adapted to our problem, i.e. that we do not introduce too much bias by regularizing the covariance estimate. Indeed, both methods achieve almost perfect outlier detection performance for all values of the covariance matrix condition number.

4.2.2. Multimodal outliers.

When dealing with multimodal outliers, we observe the expected drop of accuracy of the MCD accuracy. This demonstrates empirically MCD's theoretical limitations. All the regularized versions of the MCD estimator yielded a perfect outlier detection accuracy, even for $p/n > 1$. Shortening the distance between the modes only impacted the performance of the RMCD-RP-based method, especially when the amount of contamination was high, as shown in Table 1: When projecting to a k-dimensional subspace,

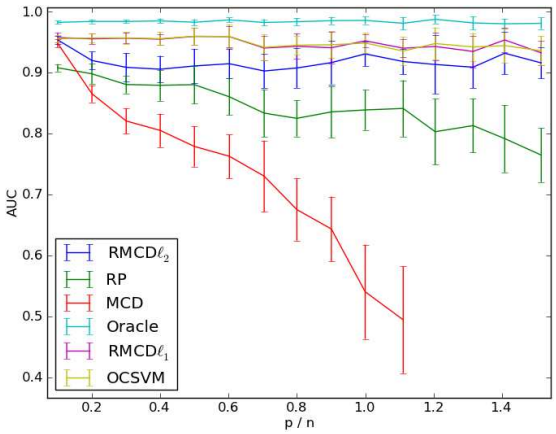


Figure 2: AUC for various outlier detection methods in the case of variance outliers ($p = 100$, $\kappa(\mathbf{\Sigma}) = 100$, $a = 1.15$, $\gamma = 40\%$). ℓ_1- and ℓ_2-regularized versions of the MCD outperform by far the standard MCD, benefiting from the isotropic distribution of the outliers. RP-RMCD's accuracy slightly decrease with p/n, which makes it not suitable for our problem. OCSVM also give good performance but can drastically break down when the condition number is too large (not shown).

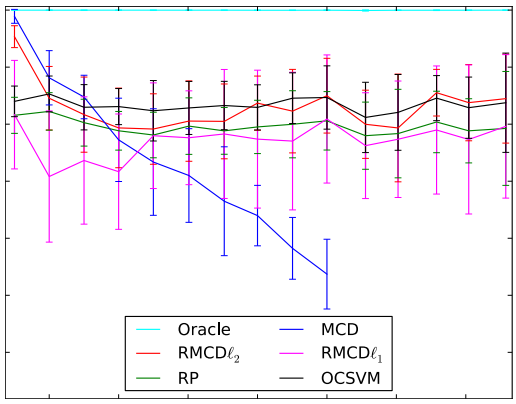


Figure 3: AUC for various outlier detection methods in the case of multivariate outliers ($p = 100$, $\kappa(\mathbf{\Sigma}) = 50$, $c = 20$, $\gamma = 30\%$). While MCD's accuracy drops, the regularized versions of the MCD almost give the same detection accuracy for each p/n ratio. RP and RMCD-ℓ_1 have a lower AUC than RMCD-ℓ_2.

the expected distance between two observations decreases by a factor $\sqrt{k/n}$ [14] so there is a weaker chance to randomly draw a subspace which preserves the separability between the two modes. Finally, the One-Class SVM is not adapted to this outliers model because it considers every densely populated region as composed of inliers. So in the presence of several clusters, the One-Class SVM would only detect outliers as abnormal subjects *with respect to their closest cluster*. This does not correspond to our assumptions of a single main cluster containing inliers.

4.2.3. Multivariate outliers.

Provided the outliers are strong enough (i.e. $c \geq 10$), the MCD estimator is well adapted to the case of multivariate outliers for $p/n < 0.2$, since its AUC is almost always above 0.9. Yet, the latter drops as the p/n ratio increases. Since RMCD-ℓ_1 and -ℓ_2 have stable performance, they outperform the MCD for large p/n values. In-between, depending on the condition number of the inliers covariance matrix and on the amount of contamination, the relative performance may vary in favor of one method or another. Figure 3 gives a general picture of the results obtained with the different methods confronted with multivariate outliers.

For $c < 10$, none of the methods can distinguish between in- and outliers and the AUC of each method increases with c, the strongest outliers being detected first. Even though the regularization parameter selection was adapted in the case of variance outliers, RMCD-ℓ_1 and RMCD-ℓ_2 confuse in- and outliers when confronted with multivariate outliers, because of the difficulty to choose an adapted regularization parameter in that case: the *most concentrated set of observations* depends on a prior knowledge about the shape of the global data set. The (R)MCD support is thus difficult to define, and so is the (R)MCD.

4.3. Influence of the simulation parameters

4.3.1. Covariance matrix condition number.

The covariance matrix condition number only has an influence in the case of multivariate and variance outliers. In both cases, we observe an improved accuracy for RMCD-ℓ_1, RMCD-ℓ_2 and OCSVM when

p/n	0.1	0.4	0.6	0.8	1.
MCD	**1.** ±0.008	**1.** ±0.065	0.8 ±0.052	0.65 ±0.058	0.55 ±0.067
RMCD-RP	**1.** ±0.008	0.98 ±0.035	0.95 ±0.031	0.90 ±0.056	0.8 ±0.057
One-Class SVM	0.76 ±0.009	0.76 ±0.020	0.76 ±0.016	0.75 ±0.025	0.76 ±0.028
others	**1.** ±0.	**1.** ±0.	**1.** ±0.	**1.** ±0.	**1.** ±0.

Table 1: AUC for MCD and RMCD-RP confronted with multimodal outliers ($p = 100$, $b = 3$, $\kappa(\mathbf{\Sigma}) = 10$, $\gamma = 30\%$). MCD breaks down for $p/n > 0.4$, which is the theoretical breakdown point. RMCD-RP's AUC stays above 0.8, which indicates good performance although it decreases when p/n increases. Other regularized methods achieve perfect outlier detection (AUC= 1) and the One-Class SVM's AUC remains constant at a low level.

the condition number is small. On the other hand, OCSVM systematically breaks down when $\kappa(\mathbf{\Sigma}) \geq 1000$, which is not the case for RMCD-ℓ_1, RMCD-ℓ_2 and RMCD-RP that keep the same AUC for every $\kappa(\mathbf{\Sigma}) > 100$. MCD is not affected by this parameter. An increase of the inliers covariance matrix condition number causes the accuracy of the three methods to decrease when confronted with multivariate outliers. This phenomenon is depicted in Figure 4.

4.3.2. Contamination rate.

Outlier detection accuracy remains similar for each method and amount of contamination, except for the RMCD-RP-based method that is very sensitive to the number of outliers when these are of the multimodal type (see Table 2).

4.3.3. Sparsity coefficient.

Sparsity of the precision matrix does not have a strong influence on the methods AUC: only the RMCD-ℓ_1 has a slightly improved AUC when the inverse covariance is very sparse. Yet, this method is not more accurate than the others, so we did not report the results.

4.4. Non-Gaussian models

Under deviations from normality, RMCD-RP and One-Class SVM outperform RMCD-ℓ_2, as shown in Figure 5. RMCD-ℓ_1 results are not reported since RMCD-ℓ_2 always yields better performance. All methods but MCD have similar and stable performance for $p/n > 0.4$. Interestingly, all methods have an AUC close to 1 for $p/n < 0.1$, which justifies the use of MCD on the complete database to build a reference labeling in our real-data experiments.

A stronger deviation from normality yields poorer performance as well as a larger variability of the outlier detection accuracy. RMCD-RP remains the best method for detecting outliers with an AUC above 0.85. MCD and RMCD-ℓ_2 still achieve almost perfect outlier detection for $p/n < 0.1$ with an AUC close to 1.

RMCD-RP performance is explained by the fact that in high-dimension, the distribution of randomly projected observations is closer to normal than the original data [4]. Therefore, applying the MCD on projected data yields a more accurate detection since the outlier detection threshold can be set exactly.

5. Outlier detection in neuroimaging

5.1. Real data description

We used data from a large functional neuroimaging database [30] containing functional Magnetic Resonance Images (fMRI) associated with 99 different contrast images in more than 1500 subjects.

Eight different 3T scanners from multiple manufacturers (GE, Siemens, Philips) were used to acquire the data. Standard preprocessing, including slice timing correction, spike and motion correction, temporal detrending (functional data), and spatial normalization (anatomical and functional data), were performed on the data using the SPM8 software and its default parameters. All images were warped in the MNI152 coordinate space. Gross outliers easily detected using simple rules such as large registration or segmentation errors or very large motion parameters were removed before hand. BOLD time series was recorded using Echo-Planar Imaging, with TR = 2200 ms, TE = 30 ms, flip angle = 75° and spatial

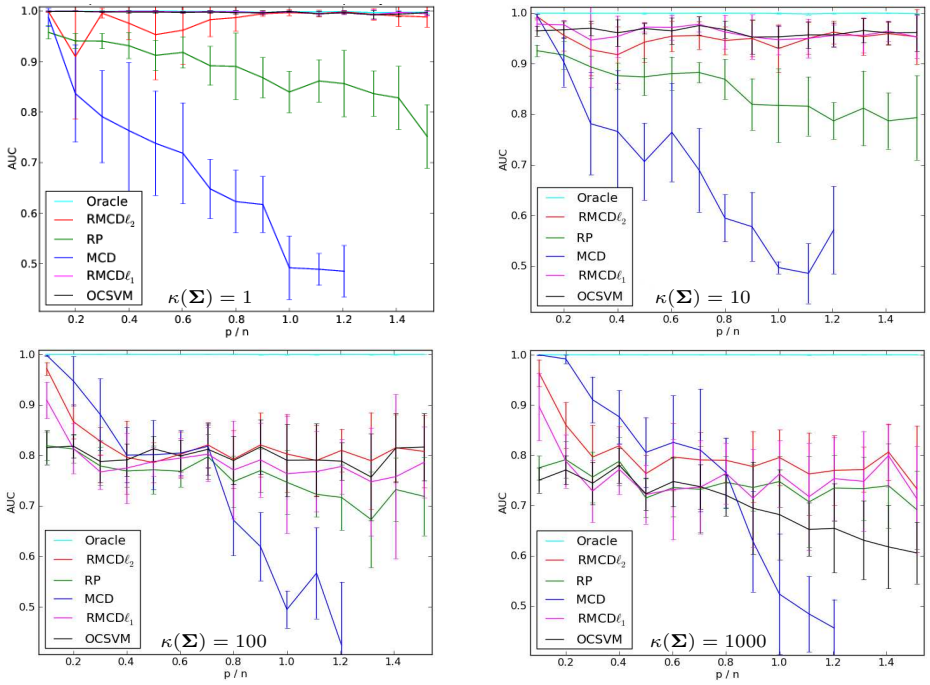


Figure 4: AUC for various outlier detection methods in the case of multivariate outliers ($p = 100$, $\kappa(\mathbf{\Sigma}) = \{1, 10, 100, 1000\}$, $c = 10$, $\gamma = 20\%$). A small condition number give advantage to the RMCD-ℓ_1 and -ℓ_2 methods, as well as the OCSVM. For $\kappa(\mathbf{\Sigma}) > 100$, all RMCD approaches perform similarly.

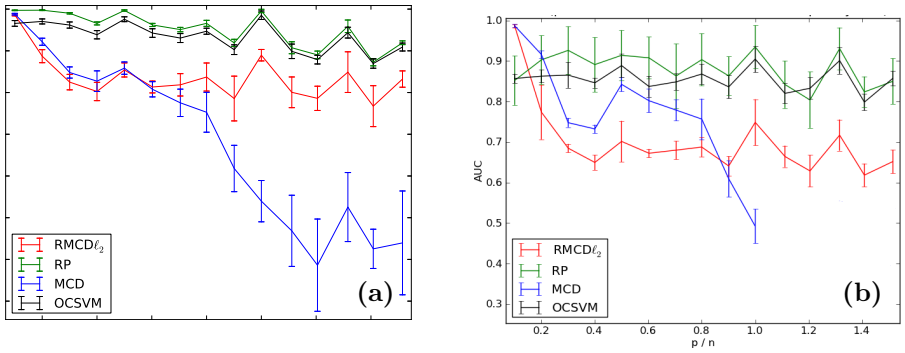


Figure 5: AUC curves of the methods on datasets generated by a mixture of Gaussian distributions. Observations were equally distributed between the components. $\gamma = 0.4$. **(a) Mild deviation from normality.** $m = 4$, $\beta = 1.1$. RMCD-ℓ_2 AUC is stable for large p/n ratios but is roughly 0.1 below RMCD-RP and One-Class SVM AUC. RMCD-RP has the best accuracy. **(b) Strong deviation from normality.** $m = 4$, $\beta = 0.7$. The different modes are observable in two- or one-dimensional projections of the data. RMCD-ℓ_2's performance is poor compared to OCSVM and RMCD-RP.

p/n	0.1	0.4	0.6	0.8	1.
$\gamma = 20\%$	**1.** ± 0	**1.** ± 0.005	**1.** ± 0.008	**0.99** ± 0.019	**0.98** ± 0.027
$\gamma = 30\%$	**1.** ± 0	0.98 ± 0.023	0.95 ± 0.051	0.90 ± 0.104	0.8 ± 0.187
$\gamma = 40\%$	0.65 ± 0.198	0.6 ± 0.164	0.59 ± 0.075	0.55 ± 0.063	0.58 ± 0.084

Table 2: Illustration of the drop of the RMCD-RP-based outlier detection method AUC with the amount of contamination γ. Multimodal outliers ($p = 100$, $b = 3$, $\kappa(\mathbf{\Sigma}) = 10$).

resolution 3mm × 3mm × 3mm. Gaussian smoothing at 5mm-FWHM was finally added. T1-weighted MPRAGE [38] anatomical images were acquired with spatial resolution 1mm × 1mm × 1mm, and gray matter probability maps were available for 1986 subjects as outputs of the SPM8 "New Segmentation" algorithm applied to the anatomical images.

5.2. Real data experiments

In a first experiment, we work with five different contrasts images (i.e. linear combination of parameter images associated with different experimental conditions) that show brain regions implied in simple cognitive tasks (computed on more than 1500 subjects):

- an auditory task as opposed to a visual task;
- a left motor task as opposed to a right motor task;
- a right motor task as opposed to a left motor task;
- a computation task as opposed to a sentences reading task;
- an angry faces viewing task;

For outlier detection, we extracted 113 features by computing on each contrast image the average activation intensity value from 113 regions of interest. These regions were given by the Harvard-Oxford cortical and sub-cortical structural atlases[2]. We removed the regions covering more than 1% of the whole brain volume, because the mean signal within such large regions did not summarize well the functional signal. We removed the effect of gender, handedness and acquisition center by using a robust regression based on M-estimators [12], using the scikit.statsmodels Python package [31] implementation. We then performed an initial outlier detection at a p-value $P < 0.1$ family-wise corrected, including all subjects ($n > 1500$). With such a small $\frac{p}{n}$ value, a statistically controlled outlier detection can be achieved using the MCD estimate. The outlier list obtained from this first outlier detection was then held as a reference labeling for further outlier detection experiments performed on reduced sample sizes, using MCD and all the Regularized MCD estimators. Note that for very small samples, we could not use the MCD-based outlier detection method. The outlier lists were compared to the reference labeling and ROC curves were constructed. For each sample size, we repeated the detection 10 times with 10 different, randomly selected samples.

We perform a second experiment using the gray matter probability maps available in this database. We use 120 regions of interest defined as 4mm-radius balls centered around locations of highly variable gray matter probability value trough subjects: we used the watershed algorithm [20] to segment the voxel-wise variability map into homogeneous regions, and the signal peak locations of the 120 regions of highest mean signal were retained as regions of interest. We limited the number of regions to 120 in order to keep an accurate statistical control of outlier detection with the full dataset. However, the choice and the size of the regions as well as the different type of data used in this second experiment should demonstrate how well regularized covariance-based outlier detection methods generalize to different contexts encountered in medical imaging. For the sake of completeness, we also tried outlier detection using the Harvard-Oxford atlas regions of interest on the gray matter probability maps.

[2] http://www.cma.mgh.harvard.edu/fsl_atlas.html

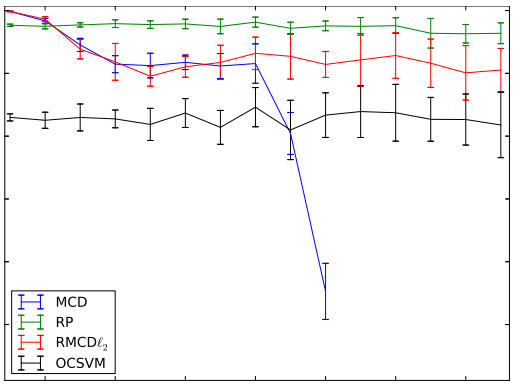


Figure 6: Results on functional MRI data after removal of the effect of gender, handedness and acquisition center. AUC curve illustrating the ability of each method to find back a reference labeling from randomly selected subsamples corresponding to various p/n ratios. Reference labeling was constructed with the MCD from $n = 1995$ observations ($p = 113$).

5.3. Results

5.3.1. Functional neuroimaging.

Figure 6 shows the outlier detection performance obtained on a dataset constructed from an fMRI contrast reflecting the brain activity related to angry faces viewing. The RMCD-RP method's curve dominates the others methods' curves for $p/n > 0.2$. RMCD-ℓ_2's accuracy is always above 0.9 while MCD-based outlier detection breaks down when p/n becomes large.

Results obtained without removing the effect of gender, handedness and acquisition center are similar to our first results, although the difference between RMCD-RP and RMCD-ℓ_2 is a bit larger (not shown).

Results obtained with others functional contrasts are similar to those of Figure 6. This suggests that the general structure of observations distribution does not depend on the contrast. Figure 7 shows activity maps (thresholded at $P < 0.01$ family-wise corrected) of out- and inliers subjects in a plot of the first two components of a Principal Components Analysis performed on the full, outlier-free data set. Outlier observations were projected to the same low-dimensional space. Outliers found by RMCD-ℓ_2-based method stand far from the central cluster, which illustrates the accuracy of the method. State-of-the-art MCD finds only three outliers, missing strongly abnormal observations. It is clear from the figure that some observations would be tagged as abnormal because the global activation pattern deviates from the standard ones (e.g. too much activity for subjects *(a)*, *(b)* and *(c)*). Yet manual screening may not be sufficient to detect some subtleties in the pattern differences. For instance, the dissimilarity between subjects *(d)* and *(e)* (both were yet in the RMCD-RP support) is not apparent in the low-dimensional projection. Note that some outliers seem to fall amongst inliers due to an artifact of projection since the original data lie in a 100-dimensional space. Indeed, only 70% of the variance is fit by the first two components.

Figure 8 shows the results of a group analysis performed on a dataset including 100 subjects drawn from the full data set MCD's support and the 20 strongest outliers found from the full dataset. The analysis was also performed on the same dataset after outliers removal using the RMCD-ℓ_2-based method. Results of both analyses were compared to a group analysis performed on the whole inliers set (1414 subjects). Activation in the left Globus Pallidus was missed in the contaminated set, but was detected after outlier removal. Also, activation in the right occipital cortex was only found from the latter dataset. Although it was obtained from less subjects (resulting in a statistical power loss), the group activation pattern for the "cleaned" group better reflects the activity pattern of the whole dataset, showing a stronger effect in every activated regions than the group map obtained from the contaminated set.

5.3.2. Anatomical brain images.

Figure 9 gives the outlier detection accuracy of the RMCD-ℓ_2, RMCD-RP, MCD and OCSVM methods on gray matter probability maps. Despite the use of a different imaging modality and ROI selection procedure, the relative performance of the methods is very similar to the performance obtained in our experiment with functional data. The number of outliers is much smaller in the reference labeling ($\simeq 3\%$). The

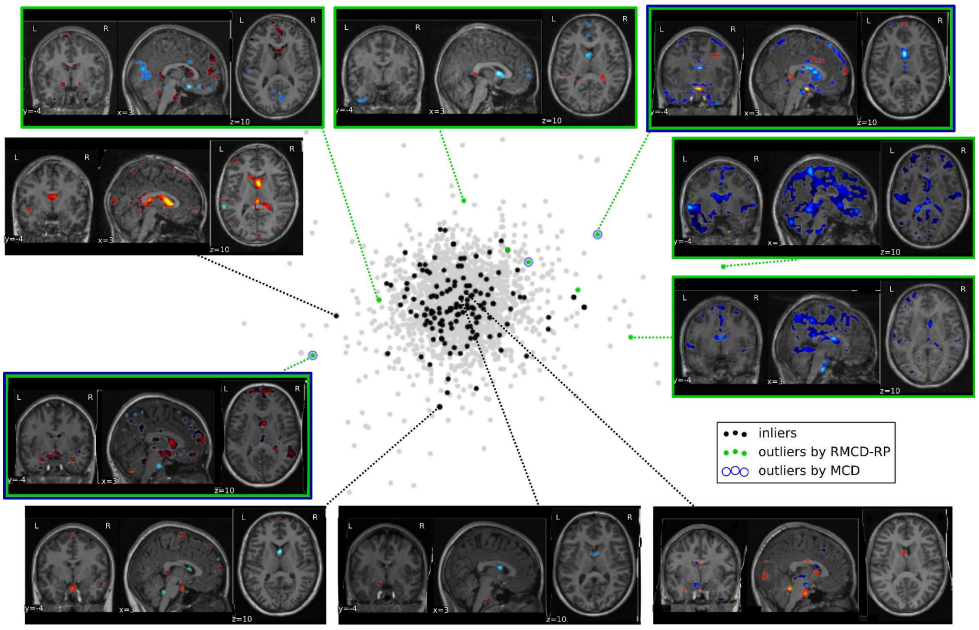
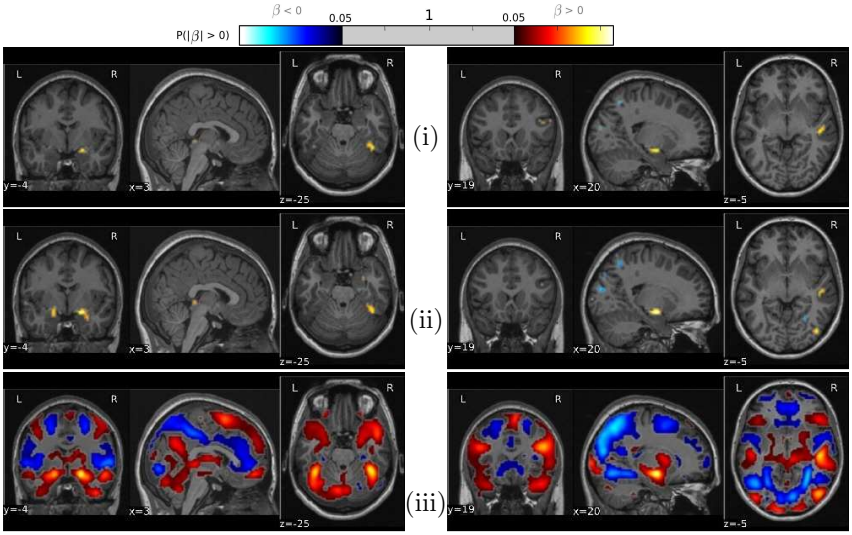


Figure 7: Neuroimaging data projection on the space spanned by the two principal components of the full, cleaned dataset. Observations tagged as outliers by the RMCD-RP method are indeed outliers at least along the two first PCA components. MCD-based outlier detection method only finds three outliers and misses strong ones. This figure illustrates the difficulty of manual outlier detection: the deviation from normality can result in unusual patterns that are not easily compared to the others.



(a) ($x = 3$mm, $y = -4$mm, $z = -25$mm) cut.

(b) ($x = 20$mm, $y = 19$mm, $z = -5$mm) cut.

Figure 8: Illustration of the benefit of removing outliers. Group activity map (two-sided test for a null intercept hypothesis $\beta = 0$, rejected at $P < 0.05$ level, family-wise corrected) for the angry faces viewing task on *(i)* a reduced dataset containing 100 inlier subjects and the 20 strongest outlier subjects, *(ii)* the same dataset with outliers removed according to RMCD-ℓ_2 method, *(iii)* the full dataset with outliers removed according to RMCD-ℓ_2 method. The results of the second row, obtained after removal of the outliers, are closer to the full dataset group analysis than the results of the first row. This illustrates the adverse consequences of including outliers in group-level inference.

MCD drops faster and breaks down for $p/n > 0.5$. The variability of all methods but RMCD-RP is much larger, which may be related to the deviation from the Gaussian distribution hypothesis that can be observed in the PCA plot given in Figure 9.

Using the Harvard-Oxford atlas's regions of interest mean signal as a descriptive feature of the gray matter images, we obtained similar results, confirming the RMCD-RP's more accurate performance for outlier detection on real datasets (not shown).

6. Discussion

Different models of outliers. The concept of outlier is ill-defined when dealing with medical data. They can be the result of a acquisition issues as well as poor preprocessing. They can also correspond to a patient or a subject with uncommon characteristics. In general, there is no good generative model that produces realistic outliers, so we use three different models to simulate contaminated neuroimaging datasets, assuming that observations are Gaussian distributed. We also investigate deviations from normality by generating inliers according to a mixture of Gaussian distributions contaminated by variance outliers (see Section 4.1.2). The relative performance of the outlier detection methods that we compare depends on the statistical characterization of outliers in the simulations. We demonstrated theoretically and empirically that the state-of-the-art method, based on the MCD estimator, breaks down in every model, as soon as the number of dimension approaches the number of observations. Our experiments also demonstrated that under each outlier model, it was possible to find a method that outperforms the RMCD-ℓ_1-based outlier detection method. Each of the three others methods have pros and cons depending on the outlier type under consideration.

RMCD-ℓ_2 detects clusters of outliers and is a good compromise. If the outliers are grouped in clusters separated from the main cluster of inliers, the only method that achieves a perfect outlier detection is the RMCD-ℓ_2 (see 4.2.2). The One-Class SVM was not adapted to this case because it takes densely-populated regions as being composed by inliers and so considers the clusters of outliers as being valid observations. RMCD-RP's accuracy drops if the outlier mode gets closer to main mode and if the contamination rate is high, mainly because the projection tends to reduce the separation between these clusters [4]. RMCD-ℓ_2 can focus on the inliers cluster (i.e. the biggest one) which is consistent with its definition. Because the RMCD-ℓ_2's accuracy is always close to the accuracy of the best method for every outlier type and any amount of contamination or inliers' covariance matrix condition number, we recommend to use this method by default because it does not require any parameter tuning, it yields interpretable results, and it is faster to compute than RMCD-ℓ_1 or RMCD-RP.

RMCD-RP for non-Gaussian distributed data. As most outlier-detection procedures, the RMCD-RP's accuracy slightly drops as p/n increases. Yet, except for extreme cases such as *multivariate outliers and large condition number* (4.2.3) or *multimodal outliers and large amount of contamination* (4.2.2), the method's AUC is higher than 0.8, which makes it attractive in practice. RMCD-RP was shown to have the best accuracy for non-Gaussian distributed data sets (see 4.4) under mild or strong deviation from normality. While the performance of RMCD-ℓ_2 breaks with stronger deviations from normality, RMCD-RP performances dominates with a gain in AUC of 0.2 or more in non-Gaussian settings. In medical imaging settings, RMCD-RP can be considered as useful, due to its robustness to deviations from normality. A procedure for the explicit control of false detections with RMCD-RP is presented in Appendix C.

One-Class SVM works well on unimodal datasets. One-Class SVM has been shown to have the best accuracy for variance and multivariate outliers, provided the condition number of the inliers covariance matrix is not too large ($\kappa(\mathbf{\Sigma}) \leq 100$, see 4.2.3). Otherwise, the number of support vectors required for spanning the whole *inliers space* and defining a frontier around has to be very large. This does not correspond to our heuristic to set the ν parameter, which is a lower bound on the number of support vectors. The remaining issue is the choice of the One-

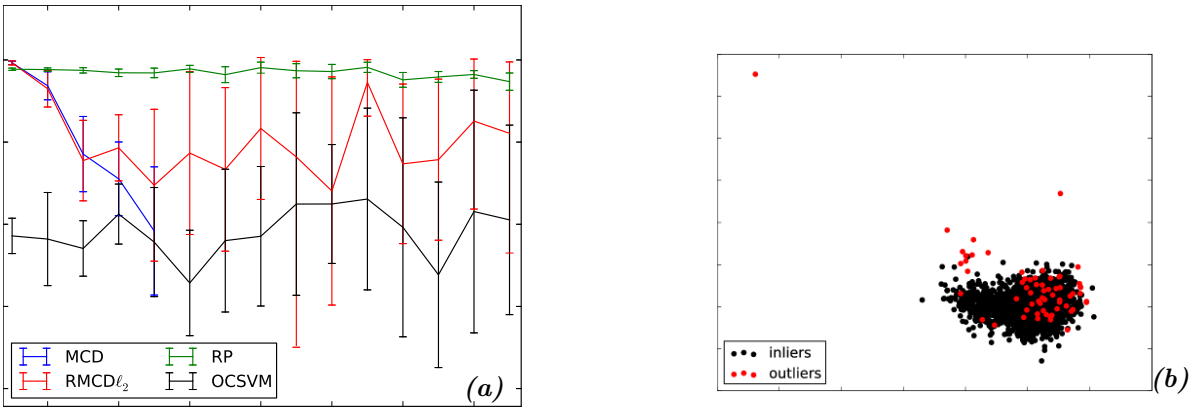


Figure 9: Outlier detection accuracy of the RMCD-ℓ_2, RMCD-RP, MCD and OCSVM methods on gray matter probability maps and representation of the corresponding dataset. **(a)** The relative performance is very similar to the performance obtained with functional data, although MCD drops faster. RMCD-RP still outperforms with an AUC above 0.95. **(b)** Projection of the dataset according to the first two components of a PCA decomposition. Outliers (in red) and inliers (in black) of the reference labeling are represented.

Class SVM parameters, especially when the inliers are much more variable in certain directions than in others. We plan to investigate in further work the use of robust covariance estimate to compute a distance to improve the performance of the One-class SVM.

Use of non-parametric tools. Non-parametric outlier detection tool as the One-Class SVM have a strong potential, provided we can set their parameters correctly. Indeed, non-parametric methods do not rely on any Gaussian nor symmetry assumption, and should therefore be more sensitive. Using a semi-supervised approach, Mourão-Miranda et al. [19] recently demonstrated the ability of the One-Class SVM to capture the shape of an homogeneous part of the data (i.e. the inliers), which makes outlier detection possible as the distance to the One-Class SVM frontier can be used as a measure of abnormality. Our experiments confirm these findings. However, One-Class SVM is ill-suitable to a medical context as the current heuristics used to tune the parameters prevent good statistical control, and its lack of simple decision frontier renders its decisions hard to interpret.

Performance on neuroimaging datasets. Functional neuroimaging datasets we used appeared to be non-Gaussian distributed. We showed in subsection 4.4 that using regularized versions of the MCD was still relevant to detect outliers. The RMCD-RP estimator is particularly adapted to that context (see Figure 5 and Figure 9) since the actual outlier detection is made on projected subspaces that appear *more Gaussian* than in the native space. Even on small datasets ($p/n > 0.2$), the new outlier detection methods that we propose can detect outliers that would not be detected by hand.

7. Conclusion

We modified the Minimum Covariance Determinant (MCD), a robust estimator of location and covariance part of the state-of-the-art outlier detection framework, in order to make it usable for outlier detection when the number of observations is small compared to the number of features describing them. Our main contribution is to introduce regularization in the definition of the MCD. We give algorithms to actually compute the regularized es-

timates and we propose a method to set the regularization parameters. ℓ_2 regularization was shown to perform generally well in simulations, but random projections outperform the latter in practice on non-Gaussian, and more importantly, on real neuroimaging data. Outlier detection using Regularized MCD can be performed in medical image processing before any group study, and was shown to advantageously replace widely-used manual screening of the data. Stabilizing group analysis is of broad interest in medical applications, such as pharmaceutic studies. Indeed, patient populations often present large heterogeneities and current studies often rely on an objective assessment of inclusion criteria.

Acknowledgments. This work was supported by a Digiteo DIM-Lsc grant (HiDiNim project, N°2010-42D). The data were acquired within the Imagen project. JBP was partly funded by the Imagen project, which receives research funding from the E.U. Community's FP6, LSHM-CT-2007-037286. This manuscript reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

Appendix A. ℓ_1 regularization (RMCD-ℓ_1)

We build a regularized version of the MCD using the ℓ_1 penalty $\|\boldsymbol{A}\|_{\text{off}} = \sum_{i \neq j} |a_{i\,j}|$ that corresponds to the ℓ_1 norm of the off-diagonal coefficients of the matrix $\boldsymbol{A}$ (note that this is not a matrix norm) in the expression of the penalized negative log-likelihood at step 2 of Algorithm 1:

$$(\hat{\boldsymbol{\mu}}_{\ell_1}, \hat{\boldsymbol{\Sigma}}_{\ell_1} | H) = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmin}} \Big(\log |\boldsymbol{\Sigma}| + \frac{1}{h} \sum_{i \in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda \, \|\Sigma^{-1}\|_{\text{off}} \Big). \quad \text{(A.1)}$$

We denote the corresponding estimator RMCD-ℓ_1.

The solution of this problem is known to have a sparse inverse [34]. This sparsity property is useful for interpretation of the solution in terms of graphical models. For instance in the functional neuroimaging context, not all brain regions are statistically related to each other [36]. The choice of the regularization parameter λ is particularly important, as the estimate is very sensitive to this value. When $\lambda \to \infty$ this converges to a diagonal matrix. We choose the regularization parameter through an approach using cross-validation (see Appendix B).

Since no closed form solution exists for the problem (A.1), we use the GLasso algorithm [6], implemented in the scikit-learn package [21].

Appendix B. Alternative strategies to set RMCD's shrinkage

We report here three strategies that we investigated to choose the RMCD-ℓ_2 shrinkage parameter:

i) The first strategy is based on *likelihood* maximization under the Gaussian distribution model for the inliers. Starting with an initial guess for $\lambda = \frac{1}{n\,p}\text{Tr}(\hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is the unbiased empirical covariance matrix of the whole dataset, we isolate an uncontaminated set of $\frac{n}{2}$ observations that correspond to the RMCD's support. Let $\lambda = \frac{\delta}{n\,p}\text{Tr}(\hat{\boldsymbol{\Sigma}}_{\text{pure}})$, where $\hat{\boldsymbol{\Sigma}}_{\text{pure}}$ is the empirical covariance matrix of the uncontaminated dataset. We choose δ so that it maximizes the ten-fold cross-validated log-likelihood of the uncontaminated dataset. Since we use cross-validation, we refer to the ℓ_2-regularized version of the MCD by *RMCD-$\ell_{2(\text{cv})}$*. We also used this strategy for the choice of the RMCD-ℓ_1 shrinkage parameter, since the subsequent strategies are not adapted to the ℓ_1 case.

The two other strategies are based on convex shrinkage, where the estimated covariance matrix $\boldsymbol{\Sigma}_{\text{lw}}$ can be expressed as $(1-\alpha)\hat{\boldsymbol{\Sigma}} + \frac{\alpha}{p}\text{Tr}(\hat{\boldsymbol{\Sigma}})\boldsymbol{I}$. *ii)* O. Ledoit and M. Wolf [17] derived a closed formula for the shrinkage coefficient α that gives the optimal solution in terms of Mean Squared Error (MSE) between the real covariance matrix to be estimated and the shrunk covariance matrix. *iii)* In a recent work, Chen et al. [1] derived another closed formula that gives a smaller MSE than Ledoit-Wolf formula under the assumption that the data are Gaussian distributed. They called it the *Oracle Approximating*

Shrinkage estimator (OAS). We adapt these results to set the regularization parameter of our MCD ℓ_2-regularized version by taking $\lambda = \frac{\alpha^\star}{p\,(1-\alpha^\star)}\mathrm{Tr}(\hat{\boldsymbol{\Sigma}})$ for $\alpha^\star$ obtained by Ledoit-Wolf and OAS formulas applied to the uncontaminated set, respectively yielding estimators that we refer to as RMCD-$\ell_{2\,(\mathrm{lw})}$ and RMCD-$\ell_{2\,(\mathrm{oas})}$ estimators.

We did not report the results for RMCD-$\ell_{2\,(\mathrm{cv})}$- and RMCD-$\ell_{2\,(\mathrm{oas})}$-based outlier detection methods since they systematically yielded an accuracy lower than or equal to RMCD-$\ell_{2\,(\mathrm{lw})}$. This is explained by the additional hypothesis required by OAS and cross-validation with respect to Ledoit-Wolf approach, and by the suboptimal cross-validation scheme. This finding suggests that the cross-validated likelihood may not be optimal as a criterion for choosing the RMCD-ℓ_1's shrinkage parameter and that we do not know how to set this parameter in practice.

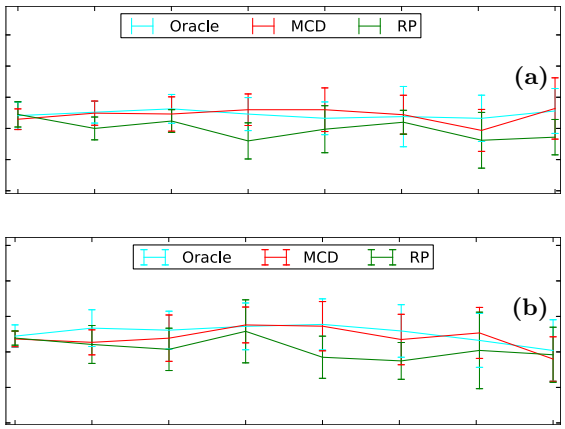


Figure C.10: Proportion of detected outliers on a clean Gaussian distributed dataset at $P < 0.05$ uncorrected. **(a)** $\kappa(\boldsymbol{\Sigma}) = 1$. **(b)** $\kappa(\boldsymbol{\Sigma}) = 1000$. Type I error rate of RMCD-ℓ_2 and RMCD-RP is close to the nominal value of 0.05 uncorrected chosen in this example.

Appendix C. Mahalanobis distance and statistical control

A crucial part of the covariance-based outlier detection is the derivation of a threshold on the Mahalanobis distances that helps performing a statistically controlled decision at the τ type I error maximum level. For any random variable $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is a well known result that $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}) \sim \chi^2_p$. Similar result exists for the distribution of $d^2_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}}(\boldsymbol{X})$, and [11] derived a theoretical formula approaching the distribution of the MCD-based Mahalanobis distances $d^2_{\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h}(\boldsymbol{X})$ for the observations that were not part of the MCD's support (the one within are distributed according to the second result we mentioned). But since the latter approximation only holds for large sample sizes, performing Monte-Carlo simulations remains the reference method to assess the distribution of $d^2_{\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h}(\boldsymbol{X})$: considering a $n \times p$ dataset on which outlier detection has to be performed, the MCD covariance estimate $\hat{\boldsymbol{\Sigma}}_h$ can be used to generate Gaussian distributed data from which a new $\hat{\boldsymbol{\Sigma}}_h$ can be estimated, together with the distribution of the ensuing Mahalanobis distances. Repeating this scheme several times, we obtain a tabulation $\hat{F_X} : x \mapsto P(X < x)$ of the MCD Mahalanobis distance distribution function under the current setting.

The same framework can be applied to RMCD-ℓ_2, and we adapted it to RMCD-RP in the following manner:

1. we tabulate the distribution F_{X_k} of the MCD-based Mahalanobis distance under $n \times k$ settings (k is the dimension of the projection subspaces);

2. we take τ/p as the new accepted error level as the number of random projections is equal to p;

3. Taking $d^* = F^{-1}(1 - \tau/p)$, define every observations with Mahalanobis distance greater than d^* in at least one subspace as outlier.

Despite the approximation made at step 2 of the previous procedure, Figure C.10 shows the proportion of type I errors made by the RMCD-RP for a desired theoretical value of $\tau = 0.05$ under various p/n settings. The final decision is a bit conservative but still relevant.

References

[1] Chen, Y., Wiesel, A., Eldar, Y., Hero, A., 2010. Shrinkage algorithms for MMSE covariance estimation. Signal Processing, IEEE Transactions on 58, 5016–5029.

[2] Daszykowski, M., Kaczmarek, K., Heyden, Y.V., Walczak, B., 2007. Robust statistics in data analysis – A review: Basic concepts. Chemometrics and Intelligent Laboratory Systems 85, 203–219.

[3] Dave, R., Krishnapuram, R., 1997. Robust clustering methods: A unified view. Fuzzy Systems, IEEE Transactions on 5, 270–293.

[4] Diaconis, P., Freedman, D., 1984. Asymptotics of graphical projections. The Annals of Statistics 12, 793–815.

[5] Falangola, M., Jensen, J., Babb, J., Hu, C., Castellanos, F., Di Martino, A., Ferris, S., Helpern, J., 2008. Age-related non-Gaussian diffusion patterns in the prefrontal brain. Journal of Magnetic Resonance Imaging 28, 1345–1350.

[6] Friedman, J., Hastie, T., Tibshirani, R., 2007. Sparse inverse covariance estimation with the lasso. ArXiv e-prints .

[7] Garcia-Escudero, L., Gordaliza, A., 1999. Robustness properties of K-Means and trimmed K-Means. Journal of the American Statistical Association 94, 956–969.

[8] Gardner, A., Krieger, A., Vachtsevanos, G., Litt, B., 2006. One-class novelty detection for seizure analysis from intracranial EEG. J. Mach Learn Res 7, 1025–1044.

[9] Hamilton, W.C., 1970. The revolution in crystallography. Science 169, 133–141.

[10] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143, 29–36.

[11] Hardin, J., Rocke, D.M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14, 928–946.

[12] Huber, P.J., 2005. Robust Statistics. John Wiley & Sons, Inc.. chapter 7. p. 149.

[13] Hubert, M., Engelen, S., 2004. Robust PCA and classification in biosciences. Bioinformatics 20, 1728–1736.

[14] Johnson, W., Lindenstrauss, J., Schechtman, G., 1986. Extensions of Lipschitz maps into Banach spaces. Israel Journal of Mathematics 54, 129–138.

[15] Joshi, S., Bowman, I., Toga, A., Van Horn, J., 2011. Brain pattern analysis of cortical valued distributions. Proc IEEE Int Symp Biomed Imaging , 1117–1120.

[16] Kherif, F., Flandin, G., Ciuciu, P., Benali, H., Simon, O., Poline, J.B., 2002. Model based spatial and temporal similarity measures between series of functional magnetic resonance images. Med Image Comput Comput Assist Interv , 509–516.

[17] Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

[18] Mériaux, S., Roche, A., Thirion, B., Dehaene-Lambertz, G., 2006. Robust statistics for nonparametric group analysis in fMRI, in: Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on, pp. 936–939.

[19] Mouro-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M., 2011. Patient classification as an outlier detection problem: An application of the one-class support vector machine. NeuroImage 58, 793–804.

[20] Najman, L., Schmitt, M., 1994. Watershed of a continuous function. Signal Processing 38, 99–112.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research .

[22] Penny, W.D., Kilner, J., Blankenburg, F., 2007. Robust bayesian general linear models. Neuroimage 36, 661–671.

[23] Pea, D., Prieto, F.J., 2001. Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43, 286–310.

[24] Pinel, P., Dehaene, S., Rivière, D., LeBihan, D., 2001. Modulation of parietal activation by semantic distance in a number comparison task. NeuroImage 14, 1013–1026.

[25] Rousseeuw, P.J., 1984. Least median of squares regression. J. Am Stat Ass 79, 871–880.

[26] Rousseeuw, P.J., Leroy, A.M., 2005. Robust Regression and Outlier Detection. John Wiley & Sons, Inc.. chapter 1. pp. 4–5.

[27] Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

[28] Roux, N.L., Bach, F., 2011. Local component analysis. CoRR abs/1109.0093.

[29] Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471.

[30] Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P., Dalley, J., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D., Ströhle, A., Struve, M., 2010. The IMAGEN study: Reinforcement-related behaviour in normal brain function and psychopathology. Molecular psychiatry 15, 1128–39.

[31] Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python, in: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference, pp. 57–61.

[32] Segata, N., Blanzieri, E., 2009. Fast and scalable local kernel machines. J. Mach Learn Res 11, 1883–1926.

[33] Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.

[34] Tibshirani, R., 1994. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

[35] Upadhyaya, A., Rieu, J., Glazier, J., Sawada, Y., 2001. Anomalous diffusion and non-Gaussian velocity distribution of hydra cells in cellular aggregates. Physica A: Statistical Mechanics and its Applications 293, 549–558.

[36] Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J., Thirion, B., 2010. A group model for stable multi-subject ICA on fMRI datasets. NeuroImage 51, 288–299.

[37] Wang, J., Saligrama, V., Castañón, D.A., 2011. Structural similarity and distance in learning. ArXiv e-prints .

[38] Wetzel, S.G., Johnson, G., Tan, A.G.S., Cha, S., Knopp, E.A., Lee, V.S., Thomasson, D., Rofsky, N.M., 2002. Three-dimensional, t1-weighted gradient-echo imaging of the brain with a volumetric interpolated examination. American Journal of Neuroradiology 23, 995–1002.

[39] Woolrich, M., 2008. Robust group analysis using outlier inference. Neuroimage 41, 286–301.

[40] Zweig, M., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 39, 561–577.