Non-parametric Density Modeling and Outlier Detection in Medical Imaging Datasets

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

Non-parametric Density Modeling and Outlier Detection in Medical Imaging Datasets

Virgile Fritsch[1,2], Gael Varoquaux[3,1,2], 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, I^{2}BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[3] Inserm, U992, Neurospin bât 145, 91191 Gif-Sur-Yvette, France

Abstract. The statistical analysis of medical images is challenging because of the high dimensionality and low signal-to-noise ratio of the data. Simple parametric statistical models, such as Gaussian distributions, are well-suited to high-dimensional settings. In practice, on medical data made of heterogeneous subjects, the Gaussian hypothesis seldom holds. In addition, alternative parametric models of the data tend to break down due to the presence of outliers that are usually removed manually from studies. Here we focus on interactive detection of these outlying observations, to guide the practitioner through the data inclusion process. Our contribution is to use *Local Component Analysis* as a non-parametric density estimator for this purpose. Experiments on real and simulated data show that our procedure separates well deviant observations from the relevant and representative ones. We show that it outperforms state-of-the-art approaches, in particular those involving a Gaussian assumption.

Keywords: Outlier detection, non-parametric density estimation, One-Class SVM, Parzen, Local Component Analysis, neuroimaging, fMRI

1 Introduction

Group studies based on medical images often attempt to extract representative samples from a given dataset in order to summarize the whole population to one or a few data prototypes. A stronger yet standard assumption is to consider that the data are Gaussian distributed around these prototypes. In the case of neuroimaging, this choice seems convenient because of the high dimensionality of the datasets, but a simple univariate Shapiro-Wilk normality test [9] demonstrates that the Gaussian hypothesis is not correct (see e.g. [10]). Furthermore, a unimodal distribution hypothesis is inconsistent with approaches that emphasize the impact of population stratification such as genome-wise association studies, or diagnosis settings that imply a separation between patients and healthy subjects. More generally, any parametric characterization of the population statistical structure is challenged by the presence of many outliers related to acquisition

or processing issues. Outlier detection and subsequent data cleansing is therefore a first step towards a better understanding of the statistical structure (including between-subjects variability) of medical imaging datasets, which in turn would be of broad interest regarding group analyses. Manifold learning is an alternative solution [2] useful for visualization, but lacking statistical guarantees.

In a recent contribution [1], a regularized version of the standard Mahalanobis distances-based outlier detection method was introduced in this context; it relies on the assumption that inliers are Gaussian distributed. Under mild deviations from this assumption, the approach has been shown to be accurate, making it possible to point out outliers in both high-dimensional and highly polluted neuroimaging datasets. However, although Mahalanobis distances-based approaches can rank the observations accurately (a property that we refer to as *accuracy*), the choice of a threshold on this ranking is strongly related to the actual data distribution.

There has also been interest in non-parametric algorithms such as *One-Class Support Vector Machine (SVM)* [8] for subjects versus patients discrimination [3,5]. One-Class SVM is well suited for medical image models since the algorithm is computationally efficient and does not rely on any prior distribution assumption. However, it defers part of the work to the practitioner, who has to build a training set with already labeled observations, or directly indicate the amount of contamination in the dataset.

The focus of our work is the analysis of the statistical structure of medical imaging datasets through high-dimensional non-parametric density estimation algorithms. We use an algorithm derived from *Parzen windows density estimation* (or *Kernel Density Estimation (KDE)*), that estimate the parameter θ of a given kernel $K(.;.,\theta)$ so that, given a learning set $(x_i)_{i=1,..,n}$, the density probability of an observations $x \in \mathbb{R}^p$ can be written $p(x) = \frac{1}{n}\sum_{i=1}^{n} K(x; x_i, \theta)$. As a new contribution, we subsequently embed this density estimator into a mode-seeking procedure to build a simple representation of the data and thus discard potential outliers. This framework provides an easy and efficient way of checking data homogeneity, a feature often required when performing further data analyzes or clinical studies [5].

In Section 2 and 3, we briefly describe the different tools that we use in this work and our contributions, that adapt these tools to medical imaging settings. In Section 4, we present some experiments on both simulated and real data. We show that the accuracy of density-based outlier detection is greater than the accuracy of state-of-the-art outlier detection methods. We also demonstrate that, starting from the Local Component Analysis density estimator, we can obtain a simple, yet relevant, differentiation between inliers and outliers. Finally, we present the results of our experiments in Section 5 and discuss them in Section 6.

2 State-of-the-art methods

Regularized Minimum Covariance Determinant. We consider datasets of n observations $x_1, \ldots, x_n$ in $\mathbb{R}^p$. Based on given mean $\boldsymbol{\mu}$ and covariance

$\boldsymbol{\Sigma}$ parameters, we define the Mahalanobis distance $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)$ of an observation $\boldsymbol{x}_i$ by $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$. The larger this quantity, the more likely $\boldsymbol{x}_i$ is an outlier. We use the *Regularized Minimum Covariance Determinant (RMCD)* [1] as the reference method for outlier detection and also to whiten the data in distance-based approaches (see Section 4).

One-Class SVM. The One-Class SVM *novelty detection* algorithm [8] is a supervised clustering algorithm that relies on a thresholded Parzen windows density estimator to define a frontier between two populations. It can be adapted to the unsupervised problem of outlier detection, but remains a descriptive model that only provides a deviation index for each observation.

Local Component Analysis. Local Component Analysis is another extension of Parzen windows density estimation where the isotropic assumption inherent in most kernels is relaxed to anisotropic covariance parameters. The θ parameter of the kernel hence becomes the local data covariance matrix $\boldsymbol{\Sigma}$, which we estimate using a leave-one-out cross-validation scheme as in [7]:

$$\boldsymbol{\Sigma}^* = \arg\min_{\boldsymbol{\Sigma}} \left[-\sum_{i=1}^{n} \log \left(|\boldsymbol{\Sigma}|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \frac{1}{n-1} \sum_{j \neq i} \exp\left(-\frac{1}{2} d^2_{\boldsymbol{x}_j,\boldsymbol{\Sigma}}(\boldsymbol{x}_i) \right) \right) \right]. \tag{1}$$

3 Theoretical contributions

Setting the LCA regularization term $\boldsymbol{\lambda}$. In [7], an internal regularization term is used to ensure LCA computation stability. The proposed default value is set to $\lambda = 10^{-4}$ [7]. In our work, we choose λ so that it models properly the central mode of the data. Since outliers may have a large influence on the observed variance of the dataset along some dimensions, we use a robust heuristic: we select the 50% most concentrated observations according to a Parzen windows density estimation, compute the Ledoit-Wolf [4] coefficient shrinkage α from this subsample, and set $\lambda = \frac{\alpha}{1-\alpha}$.

Building an interactive outlier detection framework. We propose an efficient procedure to summarize the necessary information about the data structure so that the practitioner can find how many observations to discard: Within the LCA computation, proximity measures of each observation from another are computed as $k_{ij} = \exp\left(-\frac{1}{2}(x_i - x_j)^\mathsf{T} \boldsymbol{\Sigma}^{*-1} (x_i - x_j)\right)$, thus providing a kernel-based representation of the data as a symmetric positive definite matrix $\boldsymbol{K} = (k_{ij})_{i,j \in [1..n]^2}$, that summarizes the whole data set structure. Let $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}$, where $\boldsymbol{U}$ and $\boldsymbol{D}$ are the matrix of the eigenvectors and diagonal matrix of eigenvalues $(\sigma_i, i \in [1..n])$ of $\boldsymbol{K}$. Let $\boldsymbol{D}_\delta$ be the diagonal matrix obtained by shrinking the elements of $\boldsymbol{D}$ by a factor $\delta \geq 0$ like in [11]: $\boldsymbol{D}_\delta(i,i) = 1 - \frac{\delta}{\sigma_i}$ if $\sigma_i > \delta$,

$\boldsymbol{D}_\delta(i,i) = 0$ otherwise. D_δ yields a shrunk density estimate at each observation $g_\delta(x_i) = \boldsymbol{e}_i^\mathsf{T}\boldsymbol{U}\boldsymbol{D}_\delta\boldsymbol{U}^\mathsf{T}\boldsymbol{e}$, where $\boldsymbol{e}_i$ is a vector whose entries are 0 except its i-th element which is 1 and $\boldsymbol{e}$ is a vector of ones; note that the normalization constant is omitted as it plays no role in our analysis. We finally define $\Delta(x_i) = \min_\delta\{\delta : g_\delta(x_i) < 0.5\}$, which associates each observation with the minimal shrinkage value δ that –almost– cancels it. Δ can be further used to identify different levels of homogeneity amongst the data. Typically, outliers would correspond to a group of observations that vanish with the smallest values of δ, whereas larger δ also trim off regular observations as in Fig. 2.

We define the *disappearance function*, a ranked version of Δ, as: $\Delta_{\text{rank}}(i) = \Delta(x_i)_{i:n}$, where $\Delta(x_i)_{i:n}$ is the i-th order value of $\Delta(x_i)$. Working with simulated datasets and various values of p, p/n ratios and contamination amount γ, we show that the first knee in the variation of Δ_{sort} provides a reliable estimation of the number of outliers in the dataset, while no such estimation can be made from the LCA's ranked density function $g_{\text{rank}}(i) = g(x_i)_{i:n}$, where $g(x_i) = \frac{1}{Z}\boldsymbol{e}_i^\mathsf{T}\boldsymbol{K}\boldsymbol{e}$ and Z is a normalization factor. As we will show in Section 4, Δ_{rank} better characterizes data structure than g_{rank}.

4 Experiments

In our experiments, we first compare One-Class SVM, Parzen density estimation and LCA outlier detection accuracy on both simulated and real data. Parzen density estimation and One-Class SVM are also applied to whitened data (see Section 2), which we refer to as One-Class SVM_w and Parzen_w. In a second set of experiments, we demonstrate that the subsequent interactive outlier detection framework described in Section 3 provides a usable representation of the data distribution that helps the user to isolate a set of homogeneous samples.

Data description. We generate a γ-polluted $n \times p$ dataset by drawing $(1-\gamma)n$ observations from a $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution and $\gamma\, n$ observations from a $\mathcal{N}(\boldsymbol{\mu}, \alpha\boldsymbol{\Sigma})$ distribution, $\alpha > 1$. We also consider Student distributed datasets. In both cases, μ can be set to zero without loss of generality. In all our simulations, we generated the outlier observations so that they can be distinguished from the inliers data.

Our real data were functional MRI contrast maps that were acquired with 3T scanners from multiple manufacturers. Each contrast was available for a number of subject n comprised between 1700 and 2000. BOLD time series was recorded using Echo-Planar Imaging, with TR = 2200 ms, TE = 30 ms, flip angle = 75° and spatial resolution 3mm isotropic. Standard pre-processing, including slice timing correction, spike and motion correction, temporal detrending, Gaussian smoothing at 5mm-FWHM, and spatial normalization, were performed on the data using the SPM8 software and its default parameters. Functional contrast maps were obtained by fitting a General Linear Model to this data using SPM8 also. For each available subject, we extract the mean signal of every region of interest defined by an anatomical atlas [6] and concatenate the extracted values so that every subject is represented by a p-dimensional vector in an $n \times p$ matrix.

Outlier detection accuracy measures. With simulated data, for various values of p/n, γ, $\boldsymbol{\Sigma}$, and α, we build Receiver-Operating Characteristic (ROC) curves [12] of outlier detection accuracy for each method. ROC curves were averaged over 10 runs for each method and each experiment. We compute Area Under Curve values (AUC) that reflect the general outlier detection accuracy of a method for a realistic range of p/n ratios.

Our real datasets are composed of $n \sim 1900$ (the exact value depends on the contrast) observations, each described by 113 features ($p = 113$). With such a p/n ratio, we can construct a fair approximation of the ground truth with covariance-based outlier detection. Computing AUCs as we did with simulated data, and sub-sampling the original dataset, we assess the ability of the different methods to accurately rank the observations by their *degree of abnormality*.

5 Results

5.1 Quality of outlier detection with density estimators

Gaussian and Student distributed data. On both Gaussian and Student distributed data, the accuracy of outlier detection with LCA dominates the accuracy of the other methods. Generally, all methods perform well with an AUC above 0.9, except when the condition number of the covariance matrix $\kappa(\boldsymbol{\Sigma})$ is above 100. Table 1 illustrates this phenomenon. In the latter case, one has to whiten the data previously to using One-Class SVM and Parzen. The main advantage of LCA is that such a transformation is part of the algorithm.

Real dataset. Fig. 1 shows the accuracy of the different methods on a real neuroimaging dataset, using a contrast related to the perception of angry versus neutral faces. All methods perform well with an AUC above 0.8. LCA achieves the highest accuracy yet, which remains above 0.95 for all p/n ratios. Whitening the data prior to outlier detection with One-Class SVM or Parzen density estimation is relevant since it increases the accuracy of the latter methods by roughly 0.1. Similar results were obtained in five other functional contrasts.

p/n	0.1	0.5	0.8	1.0
LCA	**0.99** ±0.0017	**0.99** ±0.0054	**0.99** ±0.0080	**0.98** ±0.0078
Parzen	0.98 ±0.0043	0.98 ±0.0103	0.98 ±0.0091	0.96 ±0.0094
Parzen_w	**0.99** ±0.0022	0.97 ±0.0055	0.97 ±0.0082	0.97 ±0.0095
One-Class SVM	**0.99** ±0.0037	0.91 ±0.0296	0.77 ±0.0795	0.64 ±0.0593
$\text{One-Class SVM}_\text{w}$	**0.99** ±0.0022	0.96 ±0.0061	0.97 ±0.0095	0.97 ±0.0104
RMCD	**0.99** ±0.0023	0.95 ±0.0055	0.97 ±0.0083	0.97 ±0.0090

Table 1. AUC values of the different outlier detection methods confronted with variance outliers (Gaussian distributed data, $p = 100$, $\gamma = 0.4$, $\kappa(\boldsymbol{\Sigma}) = 1000$, $\alpha = 1.15$).

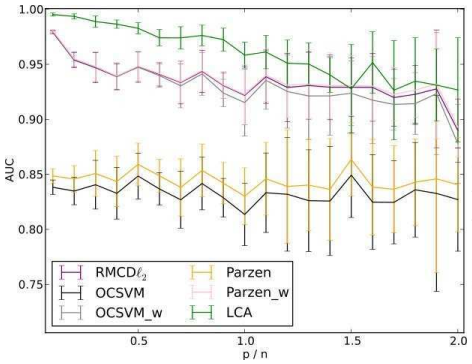
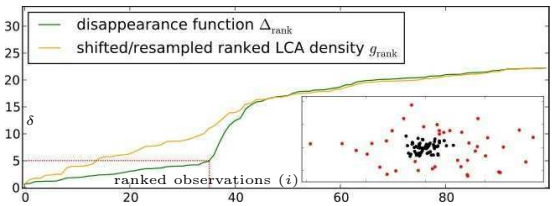


Fig. 1. Outlier detection accuracy of non-parametric density estimation algorithms, represented by their AUC (real data). LCA outperforms both Parzen density estimation and One-Class SVM, even applied on whitened data. RMCD parametric method has the same accuracy than the latter. LCA seems to be sensitive to the p/n ratio as its performance decreases with this ratio.

5.2 Relevance of an interactive outlier detection procedure

Finding outliers on simulated datasets. We verified with extensive simulations that the first knee of the disappearance function directly provides an estimate of the number of outliers. Fig. 2 illustrates this statement. This result holds for various p, p/n and α values, even though the decision showed to be a bit conservative. This behavior is yet required to guarantee a low false detections rate on heavy tailed distributions such as the Student distribution. Fig. 3 shows that our procedure does not encourage discarding observations when applied to pure Student distributed data.

Investigating the statistical structure of real data. Fig. 4 gives the spectrum of real neuroimaging datasets as obtained from LCA-learned density transformations. Knees can be easily identified in this curve, indicating that two or more relevant groups of observations are present. This observation rank property could not be inferred from the standard decision function. It is noticeable that many observations (about half of the dataset) seem to be suggested as outliers, while looking at a standard bidimensional PCA plot (not shown for the sake of place) would have suggested a much lower number.



Gaussian-distributed data, $p = 2$, $n = 100$, $\alpha = 4$, $\gamma = 0.4$

Fig. 2. Functions summarizing the data structure from their density. Difference between outliers (red dots) density and inliers (black dots) density only appears in the disappearance function. Choosing $\delta \simeq 5$ yields an outlier detection corresponding to estimating $\hat{\gamma} \simeq 35\%$ for a real value of $\gamma = 40\%$.

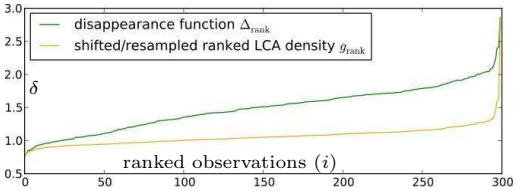


Fig. 3. Data statistical structure investigation in an uncontaminated Student-distributed data. No hard decision seems to be suggested. $p = 100$, $n = 300$, $\gamma = 0$.

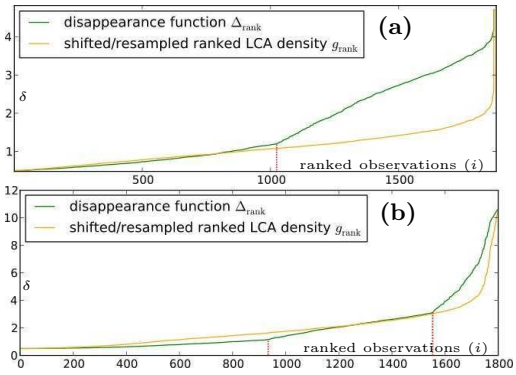


Fig. 4. Dataset structure spectrum obtained by density analysis on real neuroimaging datasets. **(a)** Viewing Angry faces - viewing neutral faces. A slope breakdown is observed at $i \simeq 1000$, suggesting that half the observations should be removed to obtain an heterogeneous set. **(b)** Rewarding task. The procedure suggests that three observations scales are present in the data. The first one may be composed of outliers. In both cases, g_{rank} does not reveal any structure.

6 Discussion

Statistical modelling of medical images is challenging because of the dimensionality of the data. Most current approaches rely on a Gaussian assumption. Here we use density estimation to rank medical images according to their degree of abnormality in a group study. We demonstrate that outlier detection with Local Component Analysis (LCA) achieves higher accuracy than state-of-the-art methods. This was shown for various p/n settings on both Gaussian and Student distributed data contaminated with up to 40% outliers. Real data experiments showed that LCA accuracy is generally above 0.9, although it seems to slightly decrease in high-dimension. Our choice of the LCA regularization parameter seemed to be optimal in that regard and our experiments demonstrated that LCA should be preferred to other non-parametric methods.

Non-parametric methods do not rely on distributional assumptions, but at the expense of explicit statistical control. We propose a simple way to perform outlier detection in an unsupervised framework that does not require any prior knowledge and guides the user in his final decision about how many observations to discard. Because it uses an internal cross-validation scheme, LCA adapts to the data local structure and comes with a natural kernel-based representation of the data. We apply a trace norm penalization to capture the information carried in the kernel matrix which reveals important features on the structure of the data [11]. As this penalization is the convex relaxation of principal components analysis-based truncation of the kernel matrix, it results in a stable criterion. We used it to characterize the difference between outliers and inliers in a robust procedure. This is meant to provide practitioners a faithful representation of possible inhomogeneities in the population under study. We verified on sev-

eral simulated and real functional neuroimaging datasets that this heuristic to chose the regularization of LCA does not yield spurious outlier detections. An attractive generalization of the LCA approach for high-dimensional settings is a mixed model, in which some dimensions are simply modeled as a Gaussian, while others are modeled through equation (1) [7].

Conclusion. Local Component Analysis was shown to have a good outlier detection accuracy under more general settings than parametric approaches. The interactive outlier detection framework presented in this contribution is of broad interest in medical imaging where manual data screening is impossible because of the high-dimensionality and sample size of the data, and yet essential due to the poor quality of the datasets used in many clinical studies.

This work was supported by a Digiteo DIM-Lsc grant (HiDiNim project, N°2010-42D). 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.

References

1. Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Medical Image Analysis in press (2012)
2. 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)
3. Kalatzis, I., Piliouras, N., Ventouras, E., Papageorgiou, C., Rabavilas, A., Cavouras, D.: Design and implementation of an SVM-based computer classification system for discriminating depressive patients from healthy controls using the P600 component of ERP signals. Comp Meth Prog Bio 75(1), 11 – 22 (2004)
4. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2), 365–411 (2004)
5. Mourao-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M.: Patient classification as an outlier detection problem: An application of the one-class SVM. NeuroImage 58(3), 793–804 (2011)
6. 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)
7. Roux, N.L., Bach, F.: Local component analysis. ArXiv e-prints (2011)
8. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471 (July 2001)
9. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika 52(3/4), 591–611 (Dec 1965)
10. Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B.: Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 35(1), 105 – 120 (2007)
11. Wang, J., Saligrama, V., Castañón, D.A.: Structural similarity and distance in learning. ArXiv e-prints (oct 2011)
12. Zweig, M., Campbell, G.: Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 39(4), 561–577 (1993)