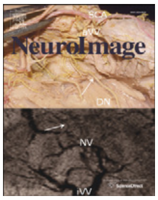
NeuroImage 111 (2015) 431–441

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Robust regression for large-scale neuroimaging studies[☆]

Virgile Fritsch [a,b,h,o,s,w,x,y,aa,*], Benoit Da Mota [a,b,h,o,s,w,x,y,aa], Eva Loth [d,h,o,s,w,x,y,aa], Gaël Varoquaux [a,b,h,o,s,w,x,y,aa], Tobias Banaschewski [e,h,o,s,w,x,y,aa], Gareth J. Barker [f,h,o,s,w,x,y,aa], Arun L.W. Bokde [g,h,o,s,w,x,y,aa], Rüdiger Brühl [h,o,p,s,w,x,y,aa], Brigitte Butzek [h,l,o,s,w,x,y,aa], Patricia Conrod [f,i,h,o,s,w,x,y,aa], Herta Flor [h,j,k,o,s,w,x,y,aa], Hugh Garavan [h,m,n,z,o,s,w,x,y,aa], Hervé Lemaitre [h,o,r,s,w,x,y,aa], Karl Mann [h,o,q,s,w,x,y,aa], Frauke Nees [h,j,k,o,s,w,x,y,aa], Tomas Paus [h,o,t,s,u,v,w,x,y,aa], Daniel J. Schad [h,l,o,s,w,x,y,aa], Gunter Schümann [f,h,o,s,w,x,y,aa], Vincent Frouin [b,h,o,s,w,x,y,aa], Jean-Baptiste Poline [c,h,o,s,w,x,y,aa], Bertrand Thirion [a,b,h,o,s,w,x,y,aa], the IMAGEN consortium [1]

[a] *Parietal Team, INRIA Saclay-Île-de-France, Saclay, France*
[b] *CEA, DSV, I2BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[c] *Helen Wills Neuroscience Institute, Henry H. Wheeler Jr. Brain Imaging Center, University of California at Berkeley, USA*
[d] *Department of Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, King's College London, London, United Kingdom*
[e] *Department of Child and Adolescent Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[f] *MRC Social, Genetic and Developmental Psychiatry (SGDP) Centre, London, United Kingdom*
[g] *Trinity College Institute of Neuroscience and Discipline of Psychiatry, School of Medicine, Trinity College Dublin, Dublin, Ireland*
[h] *Universitaetsklinikum Hamburg Eppendorf, Hamburg, Germany*
[i] *Department of Psychiatry, Universite de Montreal, CHU Ste Justine Hospital, Canada*
[j] *Central Institute of Mental Health, Mannheim, Germany*
[k] *Medical Faculty Mannheim, University of Heidelberg, Germany*
[l] *Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité Universitätsmedizin Berlin, Germany*
[m] *Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland*
[n] *Department of Psychiatry, University of VT, USA*
[o] *School of Physics and Astronomy, University of Nottingham, United Kingdom*
[p] *Physikalisch-Technische Bundesanstalt (PTB), Braunschweig, Berlin, Germany*
[q] *Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Department of Addictive Behaviour and Addiction Medicine, Germany*
[r] *Institut National de la Santé et de la Recherche Médicale, INSERM CEA Unit 1000 "Imaging & Psychiatry", University Paris Sud, Orsay, France*
[s] *AP-HP Department of Adolescent Psychopathology and Medicine, Maison de Solenn, University Paris Descartes, Paris, France*
[t] *Rotman Research Institute, University of Toronto, Toronto, Canada*
[u] *School of Psychology, University of Nottingham, United Kingdom*
[v] *Montreal Neurological Institute, McGill University, Canada*
[w] *The Hospital for Sick Children, University of Toronto, Toronto, Canada*
[x] *Behavioural and Clinical Neurosciences Institute, Department of Experimental Psychology, University of Cambridge, United Kingdom*
[y] *Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Germany*
[z] *Department of Psychology, University of VT, USA*
[aa] *Neuroimaging Center, Technische Universität Dresden, Germany*

ARTICLE INFO

Article history:
Accepted 19 February 2015
Available online 28 February 2015

Keywords:
Robust regression
Large cohorts
Neuroimaging genetics
fMRI
Outliers

ABSTRACT

Multi-subject datasets used in neuroimaging group studies have a complex structure, as they exhibit non-stationary statistical properties across regions and display various artifacts.

While studies with small sample sizes can rarely be shown to deviate from standard hypotheses (such as the normality of the residuals) due to the poor sensitivity of normality tests with low degrees of freedom, large-scale studies (e.g. >100 subjects) exhibit more obvious deviations from these hypotheses and call for more refined models for statistical inference. Here, we demonstrate the benefits of robust regression as a tool for analyzing large neuroimaging cohorts. First, we use an analytic test based on robust parameter estimates; based on simulations, this procedure is shown to provide an accurate statistical control without resorting to permutations. Second, we show that robust regression yields more detections than standard algorithms using as an example an

[☆] Authors 5–17 are listed in alphabetical order.

[*] Corresponding author at: Parietal Team, INRIA Saclay-Île-de-France, Saclay, France.
E-mail address: virgile.fritsch@inria.fr (V. Fritsch).

[1] URL: www.imagen-europe.com.

http://dx.doi.org/10.1016/j.neuroimage.2015.02.048

imaging genetics study with 392 subjects. Third, we show that robust regression can avoid false positives in a large-scale analysis of brain–behavior relationships with over 1500 subjects. Finally we embed robust regression in the Randomized Parcellation Based Inference (RPBI) method and demonstrate that this combination further improves the sensitivity of tests carried out across the whole brain. Altogether, our results show that robust procedures provide important advantages in large-scale neuroimaging group studies.

Introduction

Population-level inference or population comparison based on neuroimaging data most often rely on a mass univariate linear model, in which voxel-based brain measurements are modeled as a linear function of the variables of interest (e.g. age or sex) forming the so-called design matrix for a given group of subjects. Depending on the imaging modality, these measurements reflect tissue density or volume, neural activity (as measured by the BOLD signal) or (probabilistic) white-matter tract orientation through diffusion MRI. This mass-univariate framework is weakened by some well-known issues, such as *i*) the large number of statistical tests performed, which entails strong corrections for multiple comparisons to control for type I errors (Bennett et al., 2009); *ii*) the presence of correlations in the signal, that break the independence assumption (Friston et al., 1995); and *iii*) the presence of undesired effects or *artifacts* that substantially degrade the image quality, at a local or global spatial scale (Erasmus et al., 2004). Finally, inter-individual variability in brain anatomy, cognitive function and functional organization potentially results in mismatches in the image registration that degrade the sensitivity of statistical inference procedures.

Methods for neuroimaging studies

Neuroimaging group analyses aim at detecting the effect of a variable of interest by assessing the significance of its correlation with brain images. Many data processing and statistical analysis methods have been proposed in the literature to perform neuroimaging group analyses. These deal with the three main issues mentioned above: local averages within regions of interest (Flandin et al., 2002; Nieto-Castanon et al., 2003; Thirion et al., 2006) and feature selection (Hansen et al., 1999; Thirion and Faugeras, 2003; Spetsieris et al., 2009) are used to reduce the data dimension and the dependence between descriptors; prior smoothing of the images reduces registration mismatches (Worsley et al., 1996) and can be accounted for in standard multiple comparison corrections (Worsley et al., 1992); introducing noise regressors into the model aims at improving the sensitivity of the analyses (Lund et al., 2006); cluster-size analysis (Roland et al., 1993; Friston et al., 1993; Poline and Mazoyer, 1993), Threshold-Free Cluster Enhancement (TFCE) (Smith and Nichols, 2009; Salimi-Khorshidi et al., 2011) and Randomized Parcellation Based Inference (RPBI) (Da Mota et al., 2013) are state-of-the-art methods that combine several of the above-mentioned concepts to improve the statistical sensitivity of the analyses. For a more complete review, see Da Mota et al., 2013; Moorhead et al. (2005); and Petersson et al. (1999). All these methods rely on a set of assumptions about the statistical structure of the data (e.g. Gaussian-distributed data, "smooth-enough" images (Hayasaka et al., 2004), descriptors (in-)dependence), which are difficult to check in practice. Even though some tools have been designed to check whether the data exhibit artifacts, such as Luo and Nichols (2003), no guarantee is given that the images output by standard pre-processing pipelines will conform to these assumptions. In particular, most of the methods fit a linear model to the data with ordinary least squares (OLS) *regression*, a procedure that is optimal only if the noise is Gaussian-distributed with a given variance across samples (i.e. across individuals). Note that, by contrast, the variance can vary arbitrarily across voxels.

Large cohorts and the need for robust tools

Departure from normality has stronger effects in small sample settings than in large sample settings, where the central limit theorem leads to Gaussian errors on the estimated parameters. On the other hand, violation from standard hypotheses about the statistical structure of the data cannot be easily detected when 10 to 20 subjects are included in a neuroimaging experiment, while significant departure may be observable when larger groups of subjects are considered. Consequently, we can expect a much better model of the data, and some gains in sensitivity or specificity if we use a model that relaxes standard, simplistic assumptions such as Gaussian-distributed data, or homoscedastic noise. The need for such improved techniques becomes more apparent as more large-scale neuroimaging cohorts are now emerging (ADNI (Jack et al., 2008), IMAGEN (Schumann et al., 2010), Human Connectome (Van Essen et al., 2012) cohorts, Saguenay Youth Study (Pausova et al., 2007)). Using the simplest analysis scheme, i.e. the massively univariate voxel-wise inference, Wager et al. (2005) suggested to replace standard ordinary least squares regression by robust regression (Huber regression (Huber, 2005)), which has the advantage of *i*) relying on weak structural assumptions (symmetric, unimodal data) and *ii*) being robust to outliers. Wager and colleagues' work successfully showed sensitivity improvements for both inter- and intra-subject analyses, as well as better results stability in the presence of outliers. But this work was limited to the consideration of small groups of subjects (<20) and only the *outlier-resistant* property of the method seems to have been considered by the community (Poldrack, 2007; Ochsner et al., 2009; McRae et al., 2010; Kober et al., 2010; Atlas et al., 2010).

Robust regression schemes

Many robust regression settings have been proposed in the statistical literature to perform accurate detection in the context of non normally-distributed data. least absolute deviation (LAD) regression (or ℓ_1 regression) (Dodge, 1987) minimizes the sum of the absolute value of the model residuals. It is hard to compute in practice and the solution of the associated optimization problem may not be unique (Huber, 2005). The repeated median algorithm (Siegel, 1982) is a regression algorithm that targets a high level of outlier resistance, namely up to 50% of contamination (a property known as *high-breakdown point*). It is computationally expensive and the resulting estimate is not affine equivariant, because it is sensitive to a rotation of the data. The least median of squares (LMS) (Hampel, 1975) and least trimmed squares (LTS) (Rousseeuw, 1984) estimates also have a high breakdown point but can only be computed with algorithms for which there is no known global optimum. The efficiency of these methods for uncontaminated samples is generally poor. This can be easily understood if one conceptualizes robust methods as methods that reject the input samples that are most dissimilar to the others: the straightforward consequence is that the number of samples used in the estimation is smaller, resulting in more variable estimates and power loss. A compromise thus needs to be found between the amount of robustness to achieve and the estimation accuracy when there are no or few outliers.

Moreover, all of the above-mentioned regression estimators are difficult to apply in the context of neuroimaging problems, where thousands of correlated variables are involved. Huber's regression is the most convenient regression criterion to date. It offers a good compromise between interpretability, computational cost and robustness.

Contributions and outline

Here, we extend the work of Wager (Wager et al., 2005) by revisiting the testing framework based on Huber's robust regression and analyzing its impact on the results of large fMRI cohorts analyses, in particular in an imaging genetics study. Specifically, we show that Huber regression has some key advantages over two alternative robust regression schemes, least trimmed squares (LTS) and support vector regression (SVR) with the ϵ-insensitive loss, which is that it can be associated with an analytical test that controls type I error rate accurately; it is also ten times faster than LTS and, unlike LTS, it is guaranteed to converge to an optimal solution.

We first check the theoretical and empirical validity of robust regression in this context, and obtain additional detections on real data.

The impact of this approach on results for actual datasets is found to be important. This means that, while quality-check or data cleaning procedures are essential, the amount of data and their complexity make it impossible to set up optimal inclusion/exclusion criteria, so that robust inference remains necessary even after standard screening. Finally, we combine robust regression with state-of-the-art analysis methods to increase the analysis sensitivity. To this end we developed a fast algorithm for solving the robust regression problem, as well as a fast testing procedure. State-of-the-art analysis methods are indeed often used with permutation testing as they involve complex statistics. It is therefore crucial to reduce their computation cost.

The outline of the paper is the following. In Statistical inference for neuroimaging group analysis section we detail the neuroimaging group analysis setting. We introduce *Randomized Parcellations Based Inference* (*RPBI*), as this method will be used in combination with robust regression later on. We introduce robust regression, emphasize the associated test statistic and detail efficient procedures to perform a robust fit and the ensuing statistical tests in Robust regression section. In Robust analysis of simulated and real datasets section we introduce our simulations and real data experiments; the corresponding results are described in Results section.

Statistical inference for neuroimaging group analysis

Univariate analysis

We consider the linear model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\mathbf{y}$ is a set of n samples of an imaging feature (typically the signal value in a given voxel measured in n subjects), $\mathbf{X}$ is an $n \times p$ matrix of p variates describing the same n samples, and $\boldsymbol{\epsilon}$ represents noise. Some columns of $\mathbf{X}$ correspond to variates of no interest, while other columns are explanatory variates of interest, of which one wants to assess and test the influence on the data $\mathbf{y}$. The purpose of linear regression is to estimate the unknown coefficients β. This is generally done using OLS regression, which minimizes the sum of the squared residuals of the fitted model:

$$\hat{\beta}_{\text{OLS}} = \underset{\beta}{\text{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|^2. \tag{2}$$

In neuroimaging, the most common analysis method is a *voxel-wise inference*, that consists of fitting the model (1) independently at each voxel by using (2) (assuming that the images involved in the study have been spatially registered prior to analysis, an assumption that we rely on throughout this work). To reduce the number of statistical tests as well as the impact of registration mismatch, analysis methods based on regions of interest have been developed: Signal averages within predefined regions-of-interest are taken as data instead of the voxel-wise signal. If the regions of interest are defined by a brain parcellation, we call the corresponding analysis *a parcel-based analysis*. The model (1) is fit at the parcel level or at the voxel level in the same way, i.e. using (2). The coefficients $\hat{\beta}_{\text{OLS}}$ are tested for non-zero significance and one p-value per descriptor is computed, yielding a statistical parametric *map*. Such maps show the regions of the brain which the target variables are associated with.

Randomized Parcellation Based Inference

RPBI is a neuroimaging group analysis method recently proposed by Da Mota et al. (Da Mota et al., 2013). It notably introduces spatial context in the fitting of the regression model. It consists in performing several independent parcel-based analyses of the data, using several brain parcellations. The results obtained in the different parcellations are

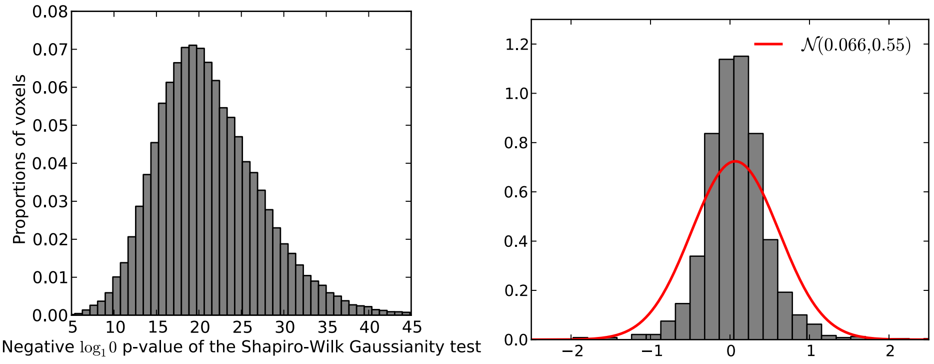


Fig. 1. (Left) Negative log10 p-values of a voxel-wise Shapiro–Wilk Gaussianity test on fMRI contrast images of 500 subjects. This is obtained by computing the p-value of the normality rejection test in each voxel for the sample of subjects. The histogram of the negative log10 p-values across voxels clearly demonstrates that the data are non-Gaussian in all brain voxels, although this assumption is often made in practice. Considering signal averages within parcels does not help, as the non-Gaussianity results from the inter-subject statistical distribution. (Right) Distribution of a contrast in a single voxel across 1364 subjects, compared to the best-fitting Gaussian distribution. This is obtained from the datasets described in The Imagen study section. The distribution is super-Gaussian, with a narrow mode and long tails.

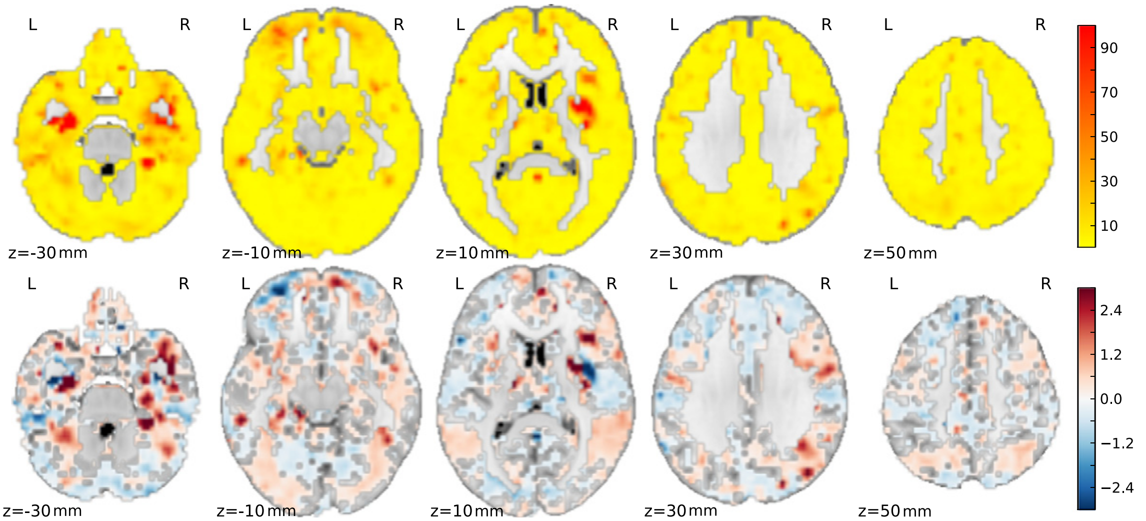


Fig. 2. Kurtosis (top row) and skewness (bottom row) across the fMRI contrast maps of 1500 subjects. This is obtained from the datasets described in The Imagen study section. Overall, all the voxel-based distributions are super-Gaussian, which is that the observed distributions have heavier tails than Gaussian distributions. Note that both kurtosis and skewness are zero for Gaussians.

then pooled together at the voxel level: the score of a voxel is the number of times across analyses that this voxel was part of a parcel having a p-value smaller than a predefined significance threshold (usually P < 0.1, Bonferroni corrected). The resulting *counting statistic* is converted into a p-value using a permutation test. Randomized Parcellation Based Inference has been shown to be more sensitive than other conventionally used methods (Da Mota et al., 2013). In the last section of this paper, we demonstrate that using robust regression in combination with RPBI further improves the sensitivity of the method.

Robust regression

While $\hat{\beta}_{\text{OLS}}$ is the maximum likelihood estimate of β for Gaussian-distributed noise ($\boldsymbol{\epsilon}$), this assumption does not hold for neuroimaging data, as shown in Figs. 1 and 2. Corrupted data and observations that deviate from the population-representative pattern potentially introduce bias in the parameter estimates. We call these unusual observations *outliers*, and the proportion of outliers in a dataset is called the *contamination*. Formally, outliers have large residual values that give them excessive importance in the OLS criterion presented in Eq. 2, resulting in a poor estimation of β.

As outliers are heavily weighted by the square function, Huber (2005) proposed a robust regression criterion (RLM, for *Robust Linear Model*) in which the square function is replaced by a function ρ (Fig. 3) that dampens the influence of the outliers:

$$\hat{\beta}_{\text{RLM}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right). \tag{3}$$

The dampening is non-linear and thus offers an interesting trade-off between statistical efficiency under the Gaussian model and resilience to outliers. σ is the standard deviation of the residuals, and acts as a scaling factor to tune the non-linearity induced by ρ. In practice, σ is not known and has to be estimated while the model is fit, yielding a joint estimation challenge (i.e. one has to estimate $\hat{\beta}$ and $\hat{\sigma}$ at the same time):

$$\hat{\beta}_{\text{RLM}}, \hat{\sigma}_{\text{RLM}} = \underset{(\beta,\sigma)}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right) \right\}. \tag{4}$$

A standard choice for ρ is the following Huber loss function:

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if}\,|x| < c, \\ c|x| - \frac{1}{2}c^2, & \text{if}\,|x| \geq c, \end{cases} \tag{5}$$

with $c = 1.345$ for 95% asymptotic efficiency on the standard Gaussian distribution (Huber, 2005). We use this definition of ρ in the remaining of the article.

Fitting a linear model with criterion (5) is equivalent to downweighting the observations according to their residual value with respect to the *true model*, while the variance of the residuals is robustly estimated. Thus, beyond outlier resistance, a robust regression criterion ensures that the fit does not depend on the data in the tails of the distribution. This is also the case for the p-values that are derived from the associated robust test defined in Significance testing section.

Algorithm

Efficient algorithms to solve problem (4) have been proposed for the *fixed scale* case (O'Leary, 1990; Baryamureeba, 2000) — where the *scale* is defined as the robust residual magnitude. The *Iteratively Reweighted Least Squares* (*IRLS*) (Algorithm 1) is used to solve the problem when the scale is not fixed. One important step to ensure the convergence of the algorithm is the *scale step*, that corresponds to the update of $\hat{\sigma}$. In the applied literature, a robust estimate of the residuals' standard deviation is taken so as to update $\hat{\sigma}$ (e.g. in (Phillips and Eyring, 1983; Chatterjee and Mächler, 1997)), but Huber raises the point that no theoretical proof of the algorithm convergence has been given in that setting and suggests a more complex update that guarantees the algorithm convergence when a convex weighting function is used (Huber, 2005). We use a Python implementation of robust regression available in the statsmodels[2] library, which we optimized for our application (handling of multiple targets and extraction of Huber's regression from the general implementation of robust linear models). Our implementation strictly follows Huber's definition of the scale update step.

[2] http://statsmodels.sourceforge.net.

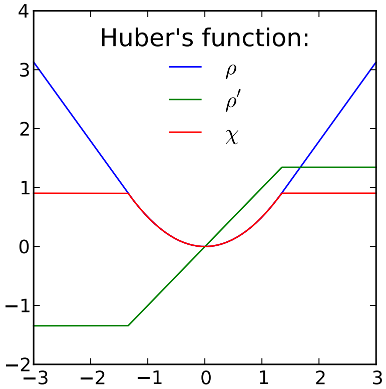


Fig. 3. Huber's ρ function. ρ is a parabola that is continued by a line on its both ends. The square part of the function smoothly downweights the observations as they depart from the regression hyperplane. This prevents the model to be dominated by strong outliers as their influence is bounded (see the flat part of ρ'). χ is introduced for technical reasons, and is defined as $\chi : x \mapsto x\rho'(x) - \rho(x)$.

Algorithm 1. Iteratively reweighted least squares

```
Require: $\boldsymbol{X}, \boldsymbol{y}, \rho$.
Ensure: $\epsilon = 10^{-8}$, $\boldsymbol{W}_{\text{old}} = \infty$, $p = rk(X)$, $h(p)$ a normalization factor.
 1: define function $\chi : x \mapsto x\rho'(x) - \rho(x)$
 2: $\boldsymbol{\beta} \leftarrow (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$
 3: $\sigma^2 \leftarrow \text{Var}[y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j]$
 4: $\boldsymbol{W} \leftarrow \mathbf{1}_n$
 5: while $||\boldsymbol{W}_{\text{old}} - \boldsymbol{W}||_\infty > \epsilon$ do
 6:     $\boldsymbol{W}_{\text{old}} \leftarrow \boldsymbol{W}$
 7:     $\boldsymbol{W} \leftarrow \left( \frac{\rho'\left((y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)/\sigma\right)}{(y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)/\sigma} \right)_{i \in \{1..n\}}$ (reweighting)
 8:     $\boldsymbol{\beta} \leftarrow (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{y}$
 9:     $\sigma^2 \leftarrow \frac{1}{n\,h(p)} \sum_{i=1}^{n} \chi\left(\frac{y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j}{\sigma}\right)\sigma^2$ (scale step)
10: end while
11: $\text{cov}(\hat{\boldsymbol{\beta}}) = K \frac{[1/(n-p)]\sum_{i=1}^{n} \rho''(y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)^2}{(1/n)\sum_{i=1}^{n} \rho''(y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)} \boldsymbol{W}^{-1}$,
12: where $K = 1 + \frac{p}{n} \frac{\text{Var}[\rho''(y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)]}{\left(\frac{1}{n}\sum_{i=1}^{n} \rho''(y_i - \sum_{j=1}^{p} x_{i\,j}\beta_j)\right)^2}$
```

Significance testing

Huber (2005) proposed to adapt the standard *t*- and *F*-tests to robust regression by considering a robust unbiased estimate of $\text{cov}\left(\hat{\beta}\right)$ (Algorithm 1, line 11) instead of the standard estimates. Such an analytic testing procedure is however crucial to us as the IRLS algorithm is prohibitive for use with permutation testing. Another problem related to the use of permutation testing is the presence of covariates in the model, which yields complex procedures to avoid bias (Anderson and Robinson, 2001). We dedicate Simulations section to a validation of this testing procedure as it has, to our knowledge, not been done previously.

Alternative robust regression criteria

Support vector regression

Support Vector Regression (*SVR*) (Drucker et al., 1997) is a regression model based on the *Support Vector Machine* (*SVM*) algorithm (Cortes and Vapnik, 1995). Fitting a SVR is fast because it can be solved in the same fashion as SVMs. The algorithm is routinely used with the ϵ-insensitive loss (Rosasco et al., 2004), in a version also known as the ϵ-*SVR* model. A key feature of this loss is its linear (instead of quadratic) increase for large values of the target variable:

$$\mathcal{H}(x) = \begin{cases} 0, & \text{if} |x| < \epsilon, \\ |x - \epsilon|, & \text{otherwise}. \end{cases} \tag{6}$$

SVR estimates can therefore be considered as robust estimates. To the best of our knowledge, no statistical test on the estimated model coefficients of SVR regression has been derived so far. We therefore resort to permutation tests. Permutation testing is computationally expensive and should be done with an appropriate test statistic. In our experiments, we use the following decision statistic:

$$t_{\text{SVR}} = \frac{\mathbf{c}^\top \hat{\beta}_{\text{SVR}}}{\text{MAD}\left(y - X\hat{\beta}_{\text{SVR}}\right)},$$

where $\mathbf{c}$ is a contrast vector corresponding to the performed test, $\hat{\beta}_{\text{SVR}}$ are the linear model coefficient estimates obtained with the SVR algorithm, and MAD stands for *median absolute deviation*, a robust version of the standard deviation. This statistic is readily comparable to the statistics used in the OLS/LTS and RLM tests.

Least trimmed squares regression

Some alternative regression criteria target a *breakdown point* of 50%, i.e. their goal is to ensure a correct estimation of the regression hyperplane under amounts of contamination going up to 50%. While this robustness property is crucial regarding applications such as outlier detection or analysis of a high-dimensional multimodal cohort, the use of high breakdown point regression criteria should be limited to diagnosis purpose and is inefficient with uncontaminated or weakly contaminated models. Here, we want to investigate whether high-breakdown point methods are worth the computational effort in the context of neuroimaging studies. We consider least trimmed squares (LTS) regression (Rousseeuw, 1984), the state-of-the-art high breakdown point robust regression technique:

$$\hat{\beta}_{\text{LTS}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{h} \left(r^2\right)_{i:n}, \tag{7}$$

where $(r^2)_{i\,:\,n} = ((y_i - \sum x_{ij}\beta_j)^2)_{i\,:\,n}$ is the *i*th ordered squared residual and $\frac{n}{2} \leq h \leq n$ sets up the breakdown point of the corresponding regression estimate. The set of the *h* observations that are considered for computation of the LTS regression criterion is called the LTS *support*. An important drawback of the method is that there is no algorithm for computing LTS regression estimates that converges globally — even with fixed *h*. This is because the loss function is not convex by design. We use the standard procedure described in Rousseeuw and Van Driessen (2006) which is however slower as compared to the algorithms used to fit other types of regression: In our implementation, the time required for the computation of the LTS regression estimate is ten times larger than that of RLM, which is itself between ten and one hundred times greater than that of OLS.

The p-values associated with the statistical test on the estimated model coefficients are obtained through a permutation test, which is another drawback of the method. An alternative procedure to avoid a permutation test and obtain approximated p-values is to fit an OLS regression algorithm on the observations contained in the LTS support.[3] We consider this method in our simulations.

[3] The estimated scale of the model residuals needs to be adjusted to take into account the fact that the fit is performed on a reduced number of observations (Croux and Rousseeuw, 1992), (Pison et al., 2002).

Robust analysis of simulated and real datasets

Simulations

We first conduct an empirical validation of the robust regression testing procedure, and compare it with standard OLS. We use the following model to generate n observations $\{y_1, \ldots, y_n\}$:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{a}_q \circ \boldsymbol{\epsilon} + \alpha\left(\mathbf{1}_n - \mathbf{a}_q\right) \circ \boldsymbol{\epsilon}, \tag{8}$$

where $\mathbf{X}$ is a random $(n \times r)$ design matrix, β is the $(r \times 1)$ vector of the model coefficients, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ models a Gaussian noise, $\boldsymbol{a}_q$ is a n-dimensional vector with coordinates drawn from a Bernoulli distribution $\mathcal{B}(1-q)$, $\alpha > 1$ is a scalar, and $\circ$ is the element-wise multiplication (Hadamard product). q is thus the average contamination in the generated dataset, and α is a parameter that controls how strongly the outliers deviate from the regular model. We set α to 5 so as to obtain potential gross outliers.

Control of the type I error rate

To verify that we can control the rate of type I error under the null hypothesis, we generate data under the null hypothesis by setting a column of β to 0 in the model (8), say column j. For various values of q, we fit a standard and two robust linear models (RLM and LTS) to a dataset generated according to the model described above, respectively yielding matrices of estimated model coefficients $\hat{\beta}_{OLS}$, $\hat{\beta}_{RLM}$ and $\hat{\beta}_{LTS}$. Finally, we test the parameters corresponding to the j-th column of $\hat{\beta}_{OLS}$, $\hat{\beta}_{RLM}$ and $\hat{\beta}_{LTS}$, that are associated with a null true effect. The proportion of null rejection should thus be equal to or less than the nominal test p-value. For instance, provided that the testing procedures are unbiased, it should happen in 5% of the cases that the regression method reports a significant effect at $p < 0.05$ uncorrected. We verify that this holds for any threshold with the use of Q–Q plots (Wilk and Gnanadesikan, 1968) corresponding to predetermined amounts of contamination.

Statistical power (type II error rate)

We show that in the presence of outliers, the statistical power of the robust test associated with RLM is higher than that achieved by an F-test subsequent to an OLS fit. The simulation framework is the same as in the previous experiment, except that we use non-zero coefficients β: we perform tests on a variable with a non-zero effect. We construct Precision–Recall curves for RLM and OLS according to the true/false acceptance/rejection rate of the null hypothesis (i.e. "no correlation exists between the tested variable(s) and the fMRI signal values"). This is done by varying the p-value under which we reject the null. We include LTS in these simulations in order to assess the performance of the testing procedure described in Least trimmed squares regression section. We have however no alternative to permutation testing for SVR, so we do not include the method in the simulation.

Synthetic neuroimaging data

We measure the accuracy improvement yielded by robust regression by first applying it on synthetic neuroimaging data. We use a generative model that makes it possible to compare the analysis results to a ground truth. The parameters of the simulation are set to yield data that model real neuroimaging data well. We simulate fMRI contrast images as volumes of shape 40 × 40 40 voxels. Each contrast image contains a simulated, irregularly shaped activation patch at a given location, jittered by a random translation sampled according the three-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_3)$ (coordinates of the jitter are rounded to the nearest integers). The strength of the activation is set so that the signal to noise ratio (SNR) peaks at 2 in the most associated voxel. The background noise is drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ distribution, Gaussian-smoothed at σ_{noise} isotropic and normalized by its global empirical standard deviation. After superimposing noise and signal images, we optionally smooth at 5 voxels full width at half maximum (FWHM). Voxels with a probability above 0.1 to be active in a large sample test are considered as part of the ground truth. Ten samples of 100 images are then generated and analyzed. Each sample was then contaminated by 15% of outliers, i.e. in each sample, we replace 15 observations by images that contains noise only. The amplitude of the noise is increased by an arbitrary factor equal to 3. All the others parameters are equal to those of the 85% regular observations.

We perform two voxel-intensity based analyses on each subgroup: the first one using standard regression (OLS) and the second one using robust regression (RLM). In both cases, the output statistical maps corresponding to the ten groups are concatenated so as to obtain one single statistical map. We then threshold that map at various levels to obtain a receiver operating characteristic (ROC) curve, using the ground truth described above to identify true/false detections.

Real data

The IMAGEN study

IMAGEN is a large-scale European multicentre study investigating genetic and environmental risk factors in brain development, brain function and the onset of common psychiatric disorders in adolescence (Schumann et al., 2010). It includes four fMRI tasks and a resulting neuroimaging database of 99 different contrast images in more than 2000 subjects, who gave informed signed assent (parents gave informed consent). In the following examples, the faces task (Grosbras and Paus, 2006; Tahmasebi et al., 2012) was used, with the [*angry faces — control*] contrast, i.e. the difference between watching angry faces and non-biological stimuli (concentric circles). We also used the Stop Signal Task (Logan, 1994) (SST), and specifically the [*success — fail*] contrast described further in this paper.

Eight different 3 T 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 using the SPM8 software and its default parameters; functional images were resampled at 3 mm resolution. All images were warped in the MNI152 coordinate space using a study-specific template. Obvious outliers detected using simple rules such as large registration or segmentation errors or very large motion parameters were removed after this step. BOLD time series was recorded using Echo-Planar Imaging, with TR = 2200 ms, TE = 30 ms, flip angle = 75° and voxel size 3.4 mm × 3.4 mm × 3.4 mm. Gaussian smoothing at 5 mm-FWHM was eventually applied. Contrasts

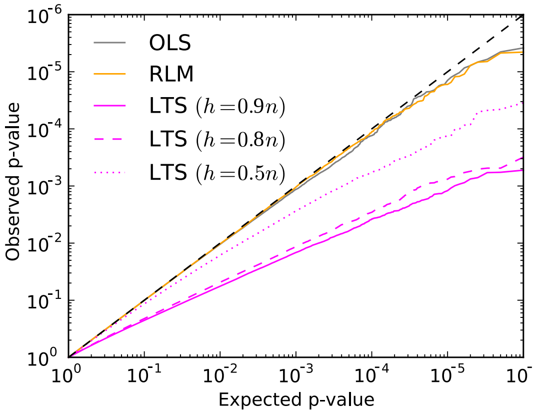


Fig. 4. Q–Q plot showing the proportion of type I errors for OLS, RLM and LTS, estimated on 10^6 independent tests performed under a null hypothesis. SVR is not plotted because permutation testing has to be used to provide an exact test. The experimental design involves 400 observations ($n = 400$), 1 tested variable and 10 variables of no interest. The amount of contamination is 20%. Both axes are presented in logarithmic scale.

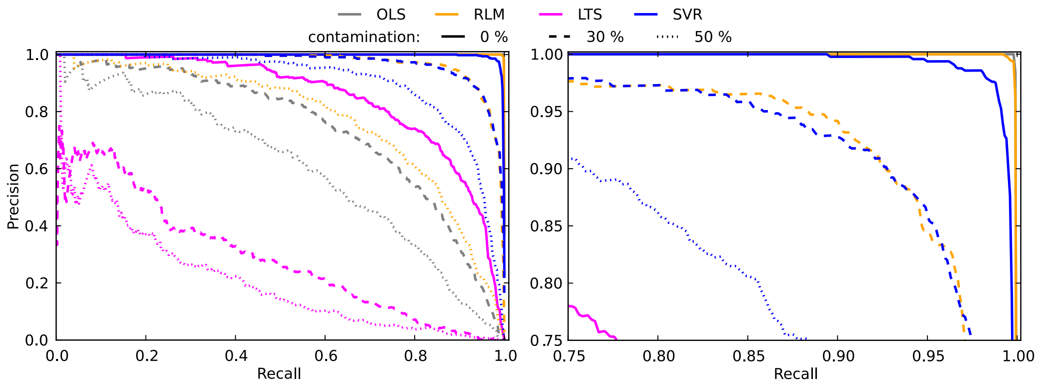


Fig. 5. Precision–recall curve showing the accuracy of OLS, RLM, SVR and least trimmed squares (LTS) regression under various amounts of contamination (0%, 20%, 40%). (a) Full curve; (b) focus on the top-right corner. $n = 400$, $p = 10$. RLM, SVR and their associated testing procedures always achieve a better compromise between type I and type II errors than OLS does in the presence of outliers. LTS appears clearly suboptimal.

were obtained using a standard linear model, based on the convolution of the time course of the experimental conditions with the canonical haemodynamic response function, together with standard high-pass filtering procedure (period = 120 s) and temporally auto-regressive noise model. The estimation of the first-level was carried out using the SPM8 software. T1-weighted MPRAGE anatomical images were acquired with a voxel size of 1.1 mm × 1.1 mm × 1.1 mm, and gray matter probability maps were available for 1986 subjects as outputs of the SPM8 "New Segmentation" algorithm applied to the anatomical images. A mask of the gray matter was built by averaging and thresholding the individual gray matter probability maps. More details about data preprocessing can be found in Thyreau et al. (2012). Genotyping was performed genome-wide using Illumina Quad 610 and 660 chips, yielding approximately 600,000 autosomal SNPs. 477,215 SNPs are common to the two chips and pass *plink* standard parameters (minor allele frequency >0.05, Hardy-Weinberg Equilibrium $p < 0.001$, missing rate per SNP <0.05).

Stability of the detections in a group analysis

We consider 10 random groups of 50 randomly drawn subjects from the Imagen database. We specifically consider the [*angry faces* — *control*] fMRI contrast, on which we perform a group analysis of the average subject activity for each of the ten groups. The analyses are conducted twice: with OLS and RLM. We smoothed the contrast maps at FWHM = 5 voxels as this is known to improve the performance of voxel-level analysis (Worsley et al., 1996). Missing values were replaced by the median value of the corresponding voxels in the other subjects of the same group. We provide a stability study of the strongest detections across the ten groups for OLS and RLM.

Imaging genetics analyses

We apply robust regression to a imaging genetics study examining gene × environment (G × E) interaction effects on fMRI BOLD activity to angry faces (Grosbras and Paus, 2006) in a sub-sample of $n = 392$ subjects for whom genotyping was available at the beginning of this project. Datasets exhibiting severe motion or deformation artifacts as well as those detected using a multivariate outlier procedure covering the whole brain, were removed. All of the 392 available observations are thus considered as correct upon manual (visual) quality check. This experiment checks the common hypothesis that genetic effects on brain function (and behavior) may often only be detected under certain environmental conditions (Caspi et al., 2010). Consequently, compared to main effect models, tests of the G × E interaction term require sensitive inference procedures. As in many imaging genetics studies we employed an unbalanced design, comparing 65 minor allele carriers of a common *Single Nucleotide Polymorphism* (*SNP*) in the oxytocin receptor gene (rs2268494) to 327 major-allele homozygotes. The model covariates were genotype, environmental risk (number of stressful life events, SLE), sex, puberty development, study center and handedness. Our aim was to compare the ability of standard and robust regression to uncover interesting effects at a fixed specificity level, i.e. the sensitivity of both methods. We perform a voxel-wise Bonferroni-corrected analysis, using standard and robust regressions.

Robust regression can straightforwardly be combined with more advanced analysis methods, such as *Randomized Parcellation Based Inference* (*RPBI*) (Da Mota et al., 2013). We demonstrate that such combinations actually yield more detections than the standard, non-robust version of the methods. We apply RPBI twice: the first time with a standard regression algorithm (RPBI_{OLS}), the second time with a robust regression algorithm (RPBI_{RLM}). We construct 100 brain parcellations with 1000 parcels each using Ward's clustering (Ward, 1963) on the contrast images of 300 bootstrapped subjects amongst the available 392. Each parcellation is used to convert the contrast images of the 392 subjects into neuroimaging features by averaging the voxels signal within each parcel.

Behavioral correlates of response inhibition

As another example of the importance of using robust regression, contrast images of 1364 subjects from the Imagen study were regressed against an impulsivity factor variable using Randomized Parcellation Based Inference. Covariates such as handedness, acquisition center, sex and age were included. The task was a stop-signal task and the contrast corresponds to the [*success* — *fail*] contrast (Logan, 1994), where the *stop success* (respectively *stop fail*) condition corresponds to the event where the subject managed (respectively failed) to inhibit its response when asked to do so (i.e. not pressing a button when the first intention was to press it but a stop signal occurred). As above, 300 bootstrapped subjects were used to build 100 different Ward's parcellations of 1000 parcels each. RPBI is then carried out with standard and robust regression.

Results

Simulation

Type I error control

The control of type I error obtained with the testing procedures associated with OLS and RLM is shown in Fig. 4. This shows that the analytical test is exact for thresholds up to $p < 10^{-5}$ uncorrected. Given the number of tests (10^6) performed in the simulation, we cannot assert that the control is accurate for lower thresholds.[4] This result holds when the amount of contamination or the number of observations involved in the simulation are varied (contamination: from 0 to 40%, number of observations: from 50 to 1000, results not shown). We also

[4] Although we went up to 10^7 draws for OLS and RLM (not shown), the results of which confirmed the accuracy of the type I error control.

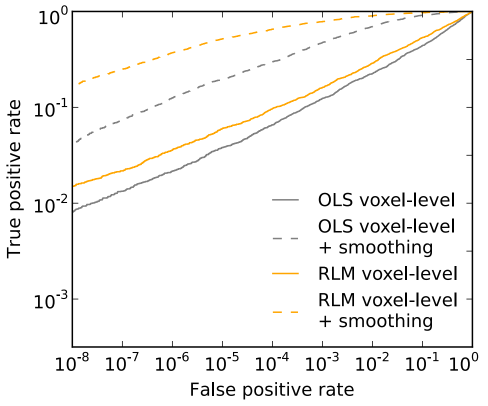


Fig. 6. Synthetic data. Sensitivity/specificity trade-off (log-scale ROC curves) for standard (OLS) and robust (RLM) regression in voxel-level group analysis in the presence of 15% outliers. Robust regression achieves more sensitivity than standard regression, because it accommodates deviations from the model assumptions. This is likely to happen in the context of neuroimaging. The performance of the two algorithms is almost equivalent when no outlier is present (confounded curves, not shown), which confirms the high power of Huber's robust regression at the uncontaminated model.

obtained a strict control when various numbers (between 1 and $n/2$) of covariates were included in the model, and with multivariate tests (i.e. several columns of the design matrix associated with null coefficients were tested for a joint effect). The efficient test designed for the LTS regression method yields very conservative p-values, which is acceptable but results in a power loss. We observe that the difference between expected and observed LTS p-values increases with the amount of contamination (not shown). This emphasizes the need of permutation tests to obtain accurate p-values when real, heavily contaminated data are considered.

Type II error minimization

The Precision–Recall curves presented in Fig. 5 illustrate the ability of the testing procedures associated with OLS, RLM, LTS and SVR to detect a significant non-null effect in the presence of outliers. Outliers potentially mislead OLS while RLM and SVR still have a good accuracy, as defined by the trade-off between correct and incorrect null-hypothesis rejections. Regarding SVR, given (*i*) the computation cost of the method due to permutation testing, and (*ii*) the lack of a standard test statistic, we decide to exclude the SVR method from our experiments in the remaining experiments. LTS appears to have poor recovery properties (conditional to the test statistic that we use). The curves may drop as more variables of no interest are included in the experimental design, but the relative performance of the regression framework is preserved. Conversely, testing several variables at a time increases the performance of the methods.

RLM offers the best compromise between robustness and computation time. Unlike alternative methods, it can be used in practice with a good control of its statistical properties/behavior as the significance level of its statistic is independent of the contamination rate (see Fig. 4).

Synthetic neuroimaging data

In Fig. 6, we show how the improved sensitivity/specificity tradeoff in non-Gaussian but i.i.d. settings carries over to neuroimaging settings: if the data are not contaminated, robust regression performs similarly to OLS regression (not shown, since the curves are virtually identical), as predicted by theory. This simulation suggests that there is a systematic statistical benefit in using robust regression in neuroimaging, because the quality of the data cannot be easily checked, especially in large and complex cohorts.

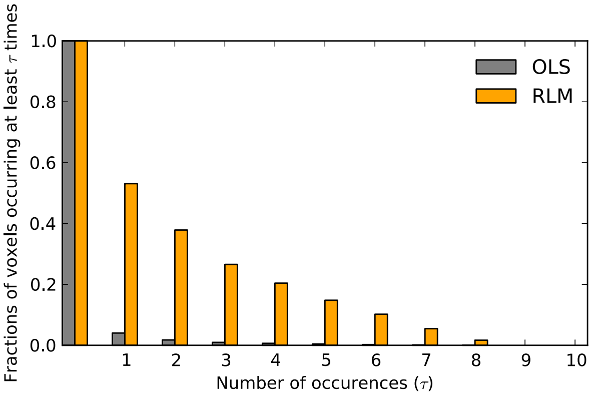


Fig. 7. Real data. Stability of strongest detections in the voxel-level group analysis of 10 random groups of 50 subjects ([*angry faces — control*] fMRI contrast, intercept test). Robust regression shows a greater stability than ordinary least squares regression.

Real data

Stability study

Fig. 7 is the reverse cumulative normalized histogram representing the number of detections across ten different groups. We limited our observation to the 500 most active voxels of each individual analysis. This shows that RLM results are much better replicated than those of OLS.

Imaging genetics study

Table 1 summarizes the results of voxel-level analyses performed with OLS and RLM. A significant G × E effect was observed in only one voxel located in the right ventral striatum with OLS. With RLM, a the significant effect was observed in five voxels of the left and the right ventral striatum. Carriers of the minor-allele showed higher ventral striatum activity than major-allele homozygotes in favorable environments (indicating higher sensitivity to anger signals) but also a significantly greater reduction of ventral striatum activity as a function of stressful life events (indicating greater sensitivity to stressful experiences). This finding supports a priori hypotheses, as the ventral striatum plays a key role in the processing of positive and negative reward signals, including anger expressions and mirror results obtained in a more comprehensive study on the OXTR gene in a larger cohort (Loth et al., 2013). Robust regression increases the number of detections and interestingly uncovers the symmetric activation as well. In Fig. 8, voxel-wise p-values obtained by permutation testing are plotted against p-values obtained with an analytic test (negative $\log_{10}$ p-values are shown). Regarding OLS regression, the deviation that is observed from the identity line is considered to be small in most of neuroimaging applications (Mumford and Nichols, 2009), and researchers often choose to trust the analytical test that has the advantage of being fast as compared to

Table 1
Whole-brain voxel-wise regression analysis results. Only one activation is found in the right ventral striatum when the analysis is performed with OLS. Robust regression is more sensitive as it yields more detections in the same brain area. Notably, a symmetric activation in the left striatum is recovered too.

Method/brain area	MNI coordinates (mm)	p-Value (Bonferroni-corrected)
OLS		
Right striatum	15, 17, 7	0.040
RLM		
Right striatum	15, 17, 10	0.010
	15, 11, 10	0.047
	15, 11, 7	0.042
	12, 17, 10	0.005
Left striatum	−6, 11, 4	0.025

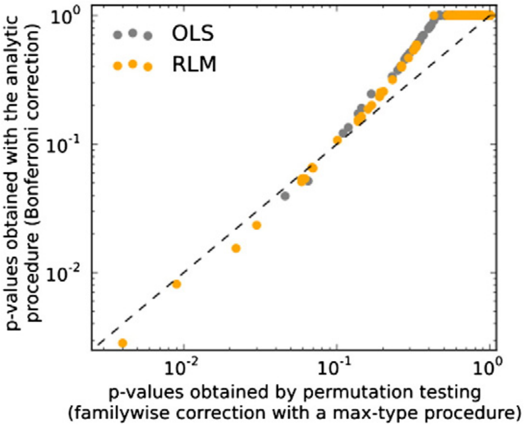


Fig. 8. p-Values obtained by a permutation test versus p-values obtained with an analytic test for OLS and RLM. We observe a deviation from the identity line, which characterizes the difference between the two sets of p-values. However, this difference is considered as small enough and is ignored in most neuroimaging applications. Bonferroni correction was implemented as a multiplication of the uncorrected p-values by the number of tests, yielding p-values potentially greater than 1. Such p-values were given the value 1 in order to preserve the definition of a p-value, as it can be observed in the top right corners of the plots.

permutation testing. The same situation is observable for robust regression as the deviation is very similar. This validates the use of the analytic testing procedure associated with RLM regarding our real data applications.

The results obtained with the RPBI method confirm those of voxel-level analyses: five brain locations were reported as significantly associated with a non-null effect when applying RPBI_{RLM} to the imaging genetics study of this real data application. Only three of them were reported by RPBI_{OLS}, as shown in Fig. 9. The activation in the left amygdala ($z = 7$ mm slice) is larger and more significant according to RPBI_{RLM}.

Neuroimaging study with behavioral features

As shown in Fig. 10, a standard regression framework reports a significant effect of the impulsivity factor within the right thalamus ($p < 0.1$ Bonferroni corrected) while a robust algorithm does not report anything. Further investigation on the data shows that one subject is an outlier and influences OLS regression. As an illustration, we focused on a single parcel that overlaps with the thalamus location. Fig. 11 represents the corresponding data in a scatter plot, and notably one subject having both a very low BOLD signal value in this parcel and a high impulsivity score. We presented OLS and RLM regression lines within the same scatter plot. The shift created by the outlier observation can be observed for each individual parcellation, in the parcel that matches the right thalamus at best. Re-running the experiment without the outlier results in the disappearance of the significant effect observed in the thalamus with OLS.

Discussion

The current study shows that Huber's robust regression (referred to as *RLM* in this paper) yields more sensitive findings than ordinary least squares regression in the context of neuroimaging studies. After ensuring that the testing procedure associated with robust regression comes with a good control of the type I errors, we showed that resistance to outliers yields a better stability and, in turn, increases the number of detected regions. These results are confirmed on real neuroimaging data.

Alternatives to Huber's robust regression

Support vector regression (SVR) and least trimmed squares (LTS) regression are two types of robust regression that can be considered as alternative to RLM. Their main drawback is that they both need to resort to costly permutation tests in order to obtain p-values that measure the significance of the tested effect. While there is no analytical result to compute p-values associated with an SVR fit, we have shown that the approximation that we use for the LTS p-values is strongly conservative (it yields correct type I error control). It is indeed difficult to obtain an accurate estimate of the scale of the model residuals when it is computed from a reduced number of observations (LTS only fits a fixed proportion of observations). Some consistency factors exist under the assumption that no outlier has been included in the LTS support. This assumption does, however, not hold if outliers are considered to be drawn from a distribution that has the same mean (location) but a larger variance than that of the regular observations. Avoiding permutation testing is crucial in some application since one might want to embed robust regression into more complex analysis frameworks, that may require permutation tests. Considering the aforementioned difficulties to

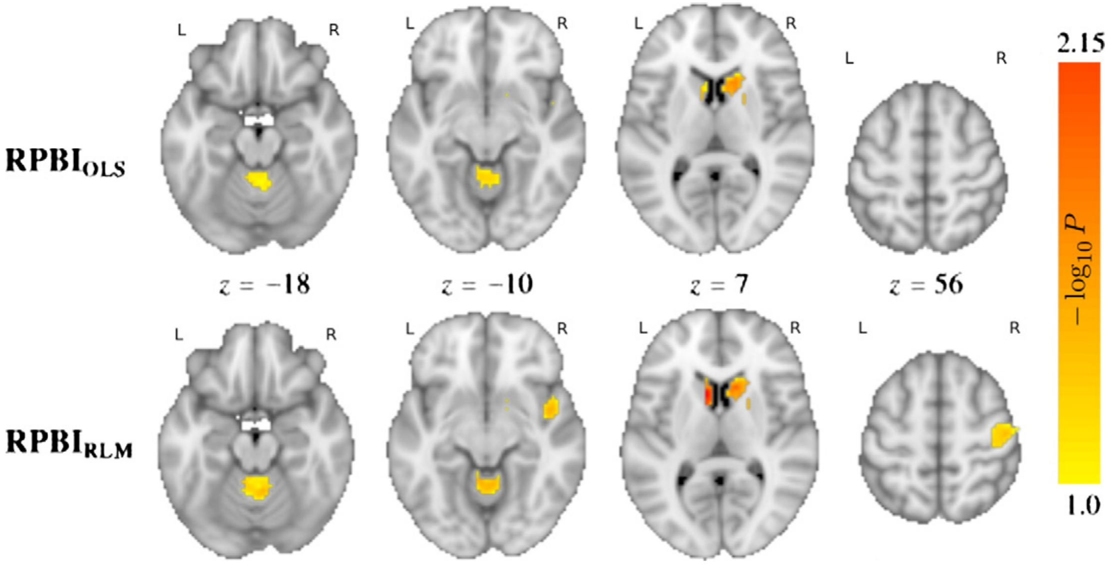


Fig. 9. Voxel-level FWER-corrected p-values maps given by RPBI_{OLS} and RPBI_{RLM} in the imaging genetics study. Five brain regions are associated with a significant non-null effect according to the robust version of RPBI, while only three of them are reported by standard RPBI. The significant associations observed in the left and right ventral striatum (third column, $z = 7$ mm) are particularly relevant to the study, as the ventral striatum plays a key role in the processing of positive and negative reward signals, including anger expressions.

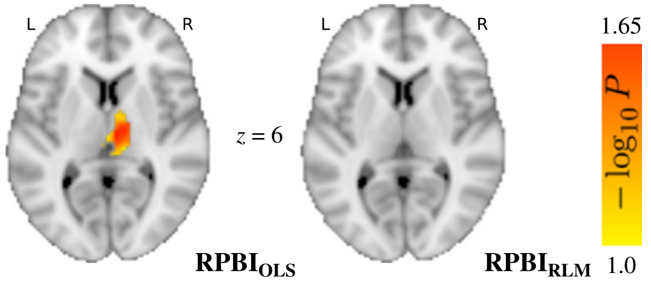


Fig. 10. RPBI analysis with standard regression (OLS) and robust regression (RLM). RPBI$_{OLS}$ (left) reports a significant effect for a group of voxel within the right thalamus while RPBI$_{RLM}$ (right) does not report any significant effect. The difference is explained by the presence of a gross outlier (details given in Fig. 11).

adapt SVR and LTS to practical applications, we have not investigated their use on real data. We note however that permutation testing is mandatory to obtain accurate p-values in many settings (family-wise error control, cluster-size tests, TFCE test); see e.g. Wager et al. (2007). For such cases, SVR might be worth considering.

Computation time

Unlike support vector regression and other alternative robust regression algorithms, Huber's robust regression has the advantage that an analytic procedure exists to test the estimated model coefficients. This reduces the running time of the algorithm in neuroimaging applications, where the ultimate goal is to find significant associations between experimental variables and brain imaging variables. We optimized the implementation of robust regression so that we can perform voxel-level analyses of a cohort of hundreds of subjects in a few minutes on a desktop computer. A robust fit is still 10 to 100 times slower than an OLS fit, which prevents RLM to be routinely used with permutation testing, including the scope of a more complex statistic such as RPBIRLM or robust cluster-size inference. We use a cluster of computers to run the analyses with RPBIRLM. An interesting direction to speed up the computation of the RLM estimate would be to keep the value of the scale constant after a few iterations or to use a conjugate gradient method to estimate both the model parameters and the scale.

Robust regression and outlier detection

Robust statistical tools are mainly used for their outlier-resistance properties. Alternatively, some studies do not use robust tools because the data are previously quality checked and should not contain any outlier. This step is relevant and we recommend to perform it in order to remove gross outliers and increase the homogeneity of the data. Outlier detection algorithms (Fritsch et al., 2012) or least trimmed squares regression (Rousseeuw, 1984) can be useful for an automated diagnosis. However, the poor performance of LTS stresses the fact that a robust fit does not simply boil down to discarding potential outliers. Deviation from normality is more general than the contamination by outlier values and is a prominent concern in all brain image modalities. More subtle deviations from the model assumptions than gross outliers cannot be systematically detected and robust tools still turn out to be useful in standard settings, as our real data experiments demonstrate. Indeed, a larger number of detections is achieved by robust regression in the neuroimaging genetic experiment, while we control the specificity accurately. The experiment with behavioral factors also shows differences between a robust and a non-robust fit, and further investigation revealed the presence of a multivariate outlier that could not be detected with a mere quality check: the impulsivity score of the main outlier is not an extreme value, but its local BOLD value is.

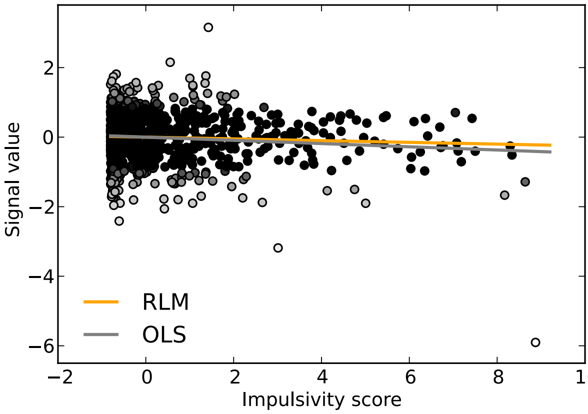


Fig. 11. Relationship between the mean signal within a parcel centered at the thalamus and the impulsivity factor (1364 subjects). Covariates effect has been removed from the two values. Regression lines have been drawn on top of the data for standard (OLS) and robust regression (RLM). The observation weights are represented with shades of gray.

Embedded robust regression

Our experiments demonstrate that robust regression can successfully be combined with state-of-the-art neuroimaging analysis methods for improved accuracy. This approach, albeit more computationally expensive than traditional inference schemes, is promising. We have demonstrated its usefulness in a large neuroimaging genetic cohorts. The number of studies of this kind is growing. Those datasets have a complex statistical structure. Specific statistical procedures are therefore required to address this challenging problem. We focused on Randomized Parcellation Based Inference because it outperforms the others established methods, but robust regression would be embedded in cluster-size inference (Roland et al., 1993; Friston et al., 1993; Poline and Mazoyer, 1993; Woo et al., 2014) or TFCE (Smith and Nichols, 2009; Salimi-Khorshidi et al., 2011) as well.

To conclude, the results presented strongly advocate for the use of robust statistics in large neuroimaging cohorts analyses. An interesting question for future work is the combination of robust regression with mixed-effects model that distinguish between the first and second levels of variance (Roche et al., 2007).

Acknowledgment

This work was supported by Digiteo grants (HiDiNim project No. 2010-42D and ICoGeN project), the ANR grant ANR-10-BLAN-0128, and the Microsoft Inria joint center grant *A-brain*.

This work was supported by the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behavior in normal brain function and psychopathology) (LSHM-CT-2007-037286), the FP7 projects IMAGEMEND and MATRICS and the Innovative Medicine Initiative Project EU-AIMS (115300-2), Medical Research Council Programme Grant “Developmental pathways into adolescent substance abuse” (93558), as well as the Swedish funding agency FORMAS. Further support was provided by the Bundesministerium für Bildung und Forschung (eMED SysAlc; AERIAL; 1EV0711). The authors thank Dr. Danilo Bzdok, Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Germany, for the valuable suggestions regarding the quality of the manuscript. They also thank the *Centre d'Analyse et Traitement des Images* (*CATI*) for giving access to their cluster.

References

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43, 75–88.

Atlas, L.Y., Bolger, N., Lindquist, M.A., Wager, T.D., 2010. Brain mediators of predictive cue effects on perceived pain. J. Neurosci. 30, 12964–12977.

Baryamureeba, V., 2000. Solution of Robust Linear Regression Problems by Preconditioned Conjugate Gradient Methods. Department of Informatics, University of Bergen.

Bennett, C.M., Miller, M., Wolford, G., 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. NeuroImage 47, S125.

Caspi, A., Hariri, A.R., Holmes, A., Uher, R., Moffitt, T.E., 2010. Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. Am. J. Psychiatry 167, 509.

Chatterjee, S., Mächler, M., 1997. Robust regression: a weighted least squares approach. Commun. Stat. Theory Methods 26, 1381–1394.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Croux, C., Rousseeuw, P.J., 1992. A class of high-breakdown scale estimators based on subranges. Commun. Stat. Theory Methods 21, 1935–1951.

Da Mota, B., Fritsch, V., Varoquaux, G., Frouin, V., Poline, J.B., Banaschewski, T., Barker, G.J., Bokde, A.L., Bromberg, U., Conrod, P., Gallinat, J., Garavan, H., Martinot, J.L., Nees, F., Paus, T., Pausova, Z., Rietschel, M., Smolka, M.N., Ströhle, A., the IMAGEN consortium, Thirion, B., 2013. Randomized parcellation based inference. Neuroimage 89, 203–215.

Dodge, Y., 1987. Statistical data analysis based on the L1-norm and related methods. Elsevier Science Inc.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systemspp. 155–161.

Erasmus, L., Hurter, D., Naudé, M., Kritzinger, H., Acho, S., 2004. A short overview of MRI artefacts: review article. SA J. Radiol. 8, 13.

Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., Poline, J.B., 2002. Improved detection sensitivity in functional MRI data using a brain parcelling technique. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002. Springer, pp. 467–474.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1993. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220.

Friston, K.J., Holmes, A.P., Poline, J., Grasby, P., Williams, S., Frackowiak, R.S., Turner, R., 1995. Analysis of fMRI time-series revisited. Neuroimage 2, 45–53.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B., 2012. Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Med. Image Anal. 16, 1359–1370.

Grosbras, M.H., Paus, T., 2006. Brain networks involved in viewing angry hands or faces. Cereb. Cortex 16, 1087–1096.

Hampel, F.R., 1975. Beyond location parameters: robust concepts and methods. Bull. Int. Stat. Inst. 46, 375–382.

Hansen, L.K., Larsen, J., Nielsen, F.Å., Strother, S.C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O.B., 1999. Generalizable patterns in neuroimaging: How many principal components? NeuroImage 9, 534–544.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. Neuroimage 22, 676–687.

Huber, P.J., 2005. Robust Statistics. John Wiley & Sons, Inc., p. 149 (chapter 7).

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27, 685–691.

Kober, H., Mende-Siedlecki, P., Kross, E.F., Weber, J., Mischel, W., Hart, C.L., Ochsner, K.N., 2010. Prefrontal–striatal pathway underlies cognitive regulation of craving. Proc. Natl. Acad. Sci. 107, 14811–14816.

Logan, G.D., 1994. On the ability to inhibit thought and action: a user's guide to the stop signal paradigm. Psychol. Rev. 91, 295–327.

Loth, E., Poline, J.B., Thyreau, B., Jia, T., Tao, C., Lourdusamy, A., Stacey, D., Cattrell, A., Desriviéres, S., Ruggeri, B., Fritsch, V., Banaschewski, T., Barker, G.J., Bokde, A.L., Büchel, C., Carvalho, F.M., Conrod, P.J., Fauth-Buehler, M., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Bruehl, R., Lawrence, C., Mann, K., Martinot, J.L., Nees, F., Paus, T., Pausova, Z., Poustka, L., Rietschel, M., Smolka, M., Struve, M., Feng, J., Schumann, G., the IMAGEN Consortium, 2013. Oxytocin receptor genotype modulates ventral striatal activity to social cues and response to stressful life events. Biol. Psychiatry 76, 367–376.

Lund, T.E., Madsen, K.H., Sidaros, K., Luo, W.L., Nichols, T.E., 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29, 54–66.

Luo, W.L., Nichols, T.E., 2003. Diagnosis and exploration of massively univariate neuroimaging models. Neuroimage 19, 1014–1032.

McRae, K., Hughes, B., Chopra, S., Gabrieli, J.D., Gross, J.J., Ochsner, K.N., 2010. The neural bases of distraction and reappraisal. J. Cogn. Neurosci. 22, 248–262.

Moorhead, T.W.J., Job, D.E., Spencer, M.D., Whalley, H.C., Johnstone, E.C., Lawrie, S.M., 2005. Empirical comparison of maximal voxel and non-isotropic adjusted cluster extent results in a voxel-based morphometry study of comorbid learning disability with schizophrenia. Neuroimage 28, 544–552.

Mumford, J.A., Nichols, T., 2009. Simple group fMRI modeling and inference. Neuroimage 47, 1469–1475.

Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., Guenther, F.H., 2003. Region of interest based analysis of functional imaging data. Neuroimage 19, 1303–1316.

O'Leary, D.P., 1990. Robust regression computation using iteratively reweighted least squares. SIAM J. Matrix Anal. Appl. 11, 466–480.

Ochsner, K.N., Hughes, B., Robertson, E.R., Cooper, J.C., Gabrieli, J.D., 2009. Neural systems supporting the control of affective and cognitive conflicts. J. Cogn. Neurosci. 21, 1841–1854.

Pausova, Z., Paus, T., Abrahamowicz, M., Almerigi, J., Arbour, N., Bernard, M., Gaudet, D., Hanzalek, P., Hamet, P., Evans, A.C., et al., 2007. Genes, maternal smoking, and the offspring brain and body during adolescence: design of the Saguenay Youth Study. Hum. Brain Mapp. 28, 502–518.

Petersson, K.M., Nichols, T.E., Poline, J.B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. Philos. Trans. R. Soc. Lond. B Biol. Sci. 354, 1261–1281.

Phillips, G.R., Eyring, E.M., 1983. Comparison of conventional and robust regression in analysis of chemical data. Anal. Chem. 55, 1134–1138.

Pison, G., Van Aelst, S., Willems, G., 2002. Small sample corrections for LTS and MCD. Metrika 55, 111–123.

Poldrack, R.A., 2007. Region of interest analysis for fMRI. Soc. Cogn. Affect. Neurosci. 2, 67–70.

Poline, J.B., Mazoyer, B.M., 1993. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. J. Cereb. Blood Flow Metab. 13, 425–437.

Roche, A., Mriaux, S., Keller, M., Thirion, B., 2007. Mixed-effect statistics for group analysis in fMRI: a nonparametric maximum likelihood approach. Neuroimage 38, 501–510.

Roland, P.E., Levin, B., Kawashima, R., Åkerman, S., 1993. Three-dimensional analysis of clustered voxels in 15-o-butanol brain activation images. Hum. Brain Mapp. 1, 3–19.

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., Verri, A., 2004. Are loss functions all the same? Neural Comput. 16, 1063–1076.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.

Rousseeuw, P.J., Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Min. Knowl. Disc. 12, 29–45.

Salimi-Khorshidi, G., Smith, S.M., Nichols, T.E., 2011. Adjusting the effect of nonstationarity in cluster-based and TFCE inference. Neuroimage 54, 2006–2019.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P.J., Dalley, J.W., 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.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Ströhle, A., Struve, M., IMAGEN consortium, 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Siegel, A.F., 1982. Robust regression using repeated medians. Biometrika 69, 242–244.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98.

Spetsieris, P.G., Ma, Y., Dhawan, V., Eidelberg, D., 2009. Differential diagnosis of parkinsonian syndromes using pca-based functional imaging features. Neuroimage 45, 1241–1252.

Tahmasebi, A.M., Artiges, E., Banaschewski, T., Barker, G.J., Bruehl, R., Büchel, C., Conrod, P.J., Flor, H., Garavan, H., Gallinat, J., et al., 2012. Creating probabilistic maps of the face network in the adolescent brain: a multicentre functional MRI study. Hum. Brain Mapp. 33, 938–957.

Thirion, B., Faugeras, O., 2003. Dynamical components analysis of fMRI data through kernel PCA. NeuroImage 20, 34–49.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B., 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Hum. Brain Mapp. 27, 678–693.

Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstdt-Klein, S., Paus, T., Artiges, E., Conrod, P.J., Schumann, G., Whelan, R., Poline, J.B., Consortium, I.M.A.G.E.N, 2012. Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. Neuroimage 61, 295–303.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., et al., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62, 2222–2231.

Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J., 2005. Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26, 99.

Wager, T.D., Scott, D.J., Zubieta, J.K., 2007. Placebo effects on human mu-opioid activity during pain. Proc. Natl. Acad. Sci. U. S. A. 104, 11056–11061.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis for the analysis of data. Biometrika 55, 1–17.

Woo, C.W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. Neuroimage 91, 412–419.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C., 1996. Searching scale space for activation in PET images. Hum. Brain Mapp. 4, 74–90.