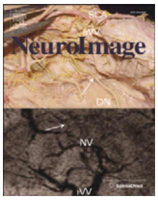
NeuroImage 111 (2015) 431–441

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.