NeuroImage 159 (2017) 79–98

Contents lists available at ScienceDirect

NeuroImage

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

Accelerating permutation testing in voxel-wise analysis through subspace tracking: A new plugin for SnPM

Felipe Gutierrez-Barragan [a,*], Vamsi K. Ithapu [a], Chris Hinrichs [a], Camille Maumet [d], Sterling C. Johnson [c], Thomas E. Nichols [d], Vikas Singh [b,a], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Department of Computer Sciences, University of Wisconsin-Madison, USA
[b] Department of Biostatistics & Med. Informatics, University of Wisconsin-Madison, USA
[c] Department of Medicine, University of Wisconsin-Madison and William S. Middleton Veteran's Hospital, USA
[d] Department of Statistics, The University of Warwick, UK[2]

ARTICLE INFO

Keywords:
Voxel-wise analysis
Hypothesis test
Permutation test
Matrix completion

ABSTRACT

Permutation testing is a non-parametric method for obtaining the max null distribution used to compute corrected *p*-values that provide strong control of false positives. In neuroimaging, however, the computational burden of running such an algorithm can be significant. We find that by viewing the permutation testing procedure as the construction of a very large permutation testing matrix, **T**, one can exploit structural properties derived from the data and the test statistics to reduce the runtime under certain conditions. In particular, we see that **T** is low-rank plus a low-variance residual. This makes **T** a good candidate for low-rank matrix completion, where only a very small number of entries of **T** ($\sim 0.35\%$ of all entries in our experiments) have to be computed to obtain a good estimate. Based on this observation, we present RapidPT, an algorithm that efficiently recovers the max null distribution commonly obtained through regular permutation testing in voxel-wise analysis. We present an extensive validation on a synthetic dataset and four varying sized datasets against two baselines: Statistical NonParametric Mapping (SnPM13) and a standard permutation testing implementation (referred as NaivePT). We find that RapidPT achieves its best runtime performance on medium sized datasets ($50 \leq n \leq 200$), with speedups of 1.5× - 38× (vs. SnPM13) and 20x-1000× (vs. NaivePT). For larger datasets ($n \geq 200$) RapidPT outperforms NaivePT (6× - 200×) on all datasets, and provides large speedups over SnPM13 when more than 10000 permutations (2× - 15×) are needed. The implementation is a standalone toolbox and also integrated within SnPM13, able to leverage multi-core architectures when available.

1. Introduction

Nonparametric voxel-wise analysis, e.g., via permutation tests, are widely used in the brain image analysis literature. Permutation tests are often utilized to control the family-wise error rate (FWER) in voxel-wise hypothesis testing. As opposed to parametric hypothesis testing schemes Friston et al. (1994); Worsley et al. (1992, 1996), nonparametric permutation tests Holmes et al. (1996); Nichols and Holmes (2002) can provide exact control of false positives while making minimal assumptions on the data. Further, despite the additional computational cost, permutation tests have been widely adopted in image analysis Arndt et al. (1996); Halber et al. (1997); Holmes et al. (1996); Nichols and Holmes (2002); Nichols and Hayasaka (2003) via implementations in broadly used software libraries available in the community SnPM (2013); FSL (2012); Winkler et al. (2014).

Running time aspects of permutation testing. Despite the varied advantages of permutation tests, there is a general consensus that the computational cost of performing permutation tests in neuroimaging analysis

* Corresponding author.
E-mail address: fgutierrez3@wisc.edu (F. Gutierrez-Barragan).
[2] http://felipegb94.github.io/RapidPT/.
[1] Data used in preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2017.07.025
Received 2 June 2017; Received in revised form 11 July 2017; Accepted 12 July 2017
Available online 15 July 2017

can often be quite high. As we will describe in more detail shortly, high dimensional imaging datasets essentially mean that *for each permutation*, hundreds of thousands of test statistics need to be computed. Further, as imaging technologies continue to get better (leading to higher resolution imaging data) and the concurrent slowdown in the predicted increase of processor speeds (Moore's law), it is reasonable to assume that the associated runtime will continue to be a problem in the short to medium term. To alleviate these runtime costs, ideas that rely on code optimization and parallel computing have been explored Eklund et al. (2011); Eklund (2012, 2013). These are interesting strategies but any hardware-based approach will be limited by the amount of resources at hand. Clearly, significant gains may be possible if *more efficient schemes* that exploit the underlying structure of the imaging data were available. It seems likely that such algorithms can better exploit the resources (e.g., cloud or compute cluster) one has available as part of a study and may also gain from hardware/code improvements that are being reported in the literature.

Data acquired in many scientific studies, especially imaging and genomic data, are highly structured. Individual genes and/or individual voxels share a great deal of commonality with other genes and voxels. It seems reasonable that such correlation can be exploited towards better (or more efficient) statistical algorithms. For example, in genomics, Cheverud (2001) and Li and Ji (2005) used correlations in the data to estimate the effective number of independent tests in a genome sequence to appropriately threshold the test statistics. Also motivated by bioinformatics problems, Knijnenburg et al. (2009) approached the question of estimating the tail of the distribution of permutation values via an approximation by a generalized Pareto distribution (using fewer permutations). In the context of more general statistical analysis, the authors in Subramanian et al. (2005) proposed Gene Set Enrichment Analysis (GSEA) which exploits the underlying structure among the genes, to analyze gene-sets (e.g., where sets were obtained from biological pathways) instead of individual genes. If the genes within a gene-set have similar expression pattern, one may see improvements in statistical power. This idea of exploiting the "structure" towards efficiency (broadly construed) was more rigorously studied in Efron and Tibshirani (2007) and a nice non-parametric Bayesian perspective was offered in Dahl and Newton (2007). Within neuroimaging, a similar intuition drives Random Field theory based analysis Taylor and Worsley (2008), albeit the objective there is to obtain a less conservative correction, rather than computational efficiency. Recently, motivated by neuroimaging applications and computational issues, Gaonkar and Davatzikos (2013) derived an analytical approximation of statistical significance maps to reduce the computational burden imposed by permutation tests commonly used to identify which brain regions contribute to a Support Vector Machines (SVM) model. In summary, exploiting the structure of the data to obtain alternative efficient solutions is not new, but we find that in the context of permutation testing on *imaging data*, there is a great deal of voxel-to-voxel correlations that if leveraged properly can, in principle, yield interesting new algorithms.

For permutation testing tasks in neuroimaging in particular, several groups have recently investigated ideas to make use of the underlying structure of the data to accelerate the procedure. In a preliminary conference paper (Hinrichs et al. (2013)), we introduced the notion of exploiting correlations in neuroimaging data via the underlying low-rank structure of the permutation testing procedure. A few years later, Winkler et al. (2016) presented the first thorough evaluation of the accuracy and runtime gains of six approaches that leverage the problem structure to accelerate permutation testing for neuroimage analysis. Among these approaches Winkler et al. (2016) presented an algorithm which relied on some of the ideas introduced by Hinrichs et al. (2013) to accelerate permutation testing through low-rank matrix completion (LRMC). Overall, algorithms that exploit the underlying structure of permutation testing in neuroimaging have provided substantial computational speedups.

1.1. Main idea and contributions

The starting point of our formulation is to analyze the entire permutation testing procedure via numerical linear algebra. In particular, the object of interest is the permutation testing matrix, **T**. Each row of **T** corresponds to the voxel-wise statistics, and each column is a specific permutation of the labels of the data. This perspective is not commonly used because a typical permutation test in neuroimaging rarely instantiates or operates on this matrix of statistics. Apart from the fact that **T**, in neuroimaging, contains millions of entries, the reason for not working *directly* with it is because the goal is to derive the maximum null distribution. The central aspect of this work is to *exploit* the structure in **T** – the spatial correlation across different voxel-statistics. Such correlations are not atypical because the statistics are computed from anatomically correlated regions in the brain. Even far apart voxel neighbourhoods are inherently correlated because of the underlying biological structures. This idea drives the proposed novel permutation testing procedure. We describe the contributions of this paper based on the observation that the permutation testing matrix is filled with related entries.

- **Theoretical Guarantees.** The basic premise of this paper is that *permutation testing in high-dimensions (especially, imaging data) is extremely redundant.* We show how we can model **T** as a low-rank plus a low-variance residual. We provide two theorems that support this claim and demonstrate its practical implications. Our first result justifies this modeling assumption and several of the components involved in recovering **T**. The second result shows that the error in the *global* maximum null distribution obtained from the estimate of **T** is quite small.
- **A novel, fast and robust, multiple-hypothesis testing procedure.** Building upon the theoretical development, we propose a fast and accurate algorithm for permutation testing involving high-dimensional imaging data. The algorithm achieves state of the art runtime performance by estimating (or recovering) the statistics in **T** rather than "explicitly" computing them. We refer to the algorithm as *RapidPT*, and we show that compared to existing state-of-the-art libraries for non-parametric testing, the proposed model achieves approximately 20× speed up over existing procedures. We further identify regimes where the speed up is even higher. RapidPT also is able to leverage serial and parallel computing environments seamlessly.
- **A plugin in SnPM (with stand-alone libraries).** Given the importance and the wide usage of permutation testing in neuroimaging (and other studies involving high-dimensional and multimodal data), we introduce a heavily tested implementation of RapidPT integrated as a plugin within the current development version of SnPM — a widely used non-parametric testing toolbox. Users can invoke RapidPT directly from within the SnPM graphical user interface and benefit from SnPM's familiar pre-processing and post-processing capabilities. This final contribution, without a separate installation, brings the performance promised by the theorems to the neuroimaging community. Our documentation Gutierrez-Barragan and Ithapu (2016) gives an overview of how to use RapidPT within SnPM.

Although the present work shares some of the goals and motivation of Winkler et al. (2016) – specifically, utilizing the algebraic structure of **T** – there are substantial technical differences in the present approach, which we outline further below. First, unlike Winkler et al. (2016), we directly study permutation testing for images at a more fundamental level and seek to characterize mathematical properties of relabeling (i.e., permutation) procedures operating on high-dimensional imaging data. This is different from assessing whether the underlying operations of classical statistical testing procedures can be reformulated (based on the correlations) to reduce computational burden as in Winkler et al. (2016). Second, by exploiting celebrated technical results in random matrix