NeuroImage 155 (2017) 549–564

Contents lists available at ScienceDirect

NeuroImage

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

Inference in the age of big data: Future perspectives on neuroscience

Danilo Bzdok[a,b,c,d,*], B.T. Thomas Yeo[e,f,g,h,i]

[a] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany
[b] JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
[c] IRTG2150 - International Research Training Group, Germany
[d] Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France
[e] Department of Electrical and Computer Engineering, National University of Singapore, 119077 Singapore
[f] Clinical Imaging Research Centre, National University of Singapore, 117599 Singapore
[g] Singapore Institute for Neurotechnology, National University of Singapore, 117456 Singapore
[h] Memory Networks Programme, National University of Singapore, 119077 Singapore
[i] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA

ARTICLE INFO

Keywords:
Systems biology
Epistemology
Hypothesis testing
High-dimensional statistics
Machine learning
Sample complexity

ABSTRACT

Neuroscience is undergoing faster changes than ever before. Over 100 years our field qualitatively described and invasively manipulated single or few organisms to gain anatomical, physiological, and pharmacological insights. In the last 10 years neuroscience spawned quantitative datasets of unprecedented breadth (e.g., microanatomy, synaptic connections, and optogenetic brain-behavior assays) and size (e.g., cognition, brain imaging, and genetics). While growing data availability and information granularity have been amply discussed, we direct attention to a less explored question: *How will the unprecedented data richness shape data analysis practices?* Statistical reasoning is becoming more important to distill neurobiological knowledge from healthy and pathological brain measurements. We argue that large-scale data analysis will use more statistical models that are non-parametric, generative, and mixing frequentist and Bayesian aspects, while supplementing classical hypothesis testing with out-of-sample predictions.

Introduction

During most of neuroscience history, before the emergence of genomics and brain imaging, new insights were "inferred" with little or no reliance on statistics. Qualitative, sometimes anecdotal reports have documented impairments after brain lesion (Harlow, 1848), microscopical inspection of stained tissue (Brodmann, 1909), electrical stimulation during neurosurgery (Penfield and Perot, 1963), targeted pharmacological intervention (Clark et al., 1970), and brain connections using neuron-transportable dyes (Mesulam, 1978). Connectivity analysis by axonal tracing studies in monkeys exemplifies biologically justified "inference" with many discoveries since the 60 s (Köbbert et al., 2000). A colored tracer substance is injected in vivo into source region A, uptaken by local neuronal receptors, and automatically transported in axons to target region B. This observation in *a single monkey* allows *extrapolating* a monosynaptical connection between region A and B to the *entire monkey species* (Mesulam, 2012). Instead, later brain-imaging technology propelled the data-intensive characterization of the mammalian brain and today readily quantifies axonal connections, cytoarchitectonic borders, myeloarchitectonic distributions, neurotransmitter receptors, and oscillatory coupling (Amunts et al., 2013; Frackowiak and Markram, 2015; Kandel et al., 2013; Van Essen et al., 2012). Following many new technologies to generate digitized yet noisy brain data, drawing insight from observations in the brain henceforth required assessment in the statistical arena.

In the quantitative sciences, the invention and application of statistical tools has always been dictated by changing contexts and domain questions (Efron and Hastie, 2016). The present paper will therefore examine how statistical choices are likely to change due to the progressively increasing granularity of digitized brain data. Massive data collection is a game changer in neuroscience (Kandel et al., 2013; Poldrack and Gorgolewski, 2014), and in many other public and private areas (House of Commons, 2016; Jordan et al., 2013; Manyika et al., 2011). There is a growing interest in and pressure for data sharing, open access, and building "big data" repositories (Frackowiak and Markram, 2015; Lichtman et al., 2014; Randlett et al., 2015). For instance, UK Biobank is a longitudinal population study dedicated to the genetic and environmental influence on mental disorders and other medical conditions (Allen et al., 2012; Miller et al., 2016). 500,000 enrolled volunteers undergo an extensive battery of clinical diagnostics

[*] Correspondence to: Department of Psychiatry, Psychotherapy and Psychosomatics, Pauwelsstraße 30, 52074 Aachen, Germany.
E-mail address: danilo.bzdok@rwth-aachen.de (D. Bzdok).

http://dx.doi.org/10.1016/j.neuroimage.2017.04.061
Received 20 August 2016; Received in revised form 25 April 2017; Accepted 25 April 2017
Available online 27 April 2017

from brain scans to bone density with a > 25 year follow-up. In the US, the Precision Medicine Initiative announced in 2015 to profile 1,000,000 individuals (Collins and Varmus, 2015). Targeted analysis of such national and international data collections may soon become the new normal in basic and clinical neuroscience. In this opinion paper, we will inspect the statistical scalability to the data-rich scenario from four different formal perspectives: i) parametric versus non-parametric models, ii) discriminative versus generative models, and iii) frequentist versus Bayesian models, as well as iv) classical hypothesis testing and out-of-sample generalization.

Towards adaptive models

Parametric models seek to capture underlying structure in data, which is representable with a fixed number of model parameters. For instance, many parametric models with Gaussianity assumptions will attempt to fit Gaussian densities regardless of the underlying data distribution. On the other hand, we think of *non-parametric* models as typically making weaker assumptions about the underlying data structure, such that the model complexity is data-driven, the *expressive capacity* does not saturate, the model structure can adapt flexibly, and the prediction can grow more sophisticated (see Box 1 for elaboration). Certain non-parametric models (e.g., Parzen window density estimation) will converge to the true underlying data distribution with sufficient data (although the amount of needed data might be astronomical). With increasing data samples, non-parametric models thus tend to make always-smaller error in capturing underlying structure in data (Devroye et al., 1996; Bickel et al., 2007). Relating these considerations back to the deluge of data from burgeoning neuroscience consortia, "the main concern is underfitting from the choice of an overly simplistic parametric model, rather than overfitting." (Ghahramani, 2015, p. 454). We therefore believe that non-parametric models have the potential to extract arbitrarily complex perceptual units, motor programs, and neural computations directly from healthy and diseased brain measurements.

In our opinion, the expressive capacity of many parametric models to capture cognitive and neurobiological processes is limited and cannot adaptively increase if more input data are provided. For instance, independent component analysis (ICA) is an often-used parametric model that extracts a set of macroscopic networks with coherent neural activity from brain recordings (Calhoun et al., 2001; Beckmann et al., 2009). Applied to human functional magnetic resonance imaging (fMRI) data, ICA reliably yields the default mode network, saliency network, dorsal attention network, and other canonical brain networks (Damoiseaux et al., 2006; Seeley et al., 2007; Smith et al., 2009). Standard ICA is parametric in the sense that the algorithm extracts a user-specified number of spatiotemporal network components, although the "true" number of macroscopic brain networks is unknown or might be ambiguous (Eickhoff et al., 2015). By coupling standard ICA with approximate Bayesian model selection (BMS), Beckmann and Smith (2004) allowed the number of components to flexibly adapt to brain data. The combination of parametric ICA and BMS yields an integrative modeling approach that exhibits the scaling property of non-parametric statistics (Goodfellow et al., 2016, p. 112; Ghahramani, 2015, p. 454): With increasing amount of input data, ICA with BMS adaptively calibrates the *model complexity* by potentially extracting more brain network components, thus enhancing the expressive power of classical ICA.

These advantages are inherent to *non-parametric* models that can potentially extract an always higher number of neural patterns that are *adaptively described by an always higher, theoretically infinite number of model parameters* as the amount of input data increase (Orbanz and Teh, 2011; Ghahramani, 2013). In doing so, we believe non-parametric models can potentially isolate representations of neurobiological phenomena that do not only improve quantitatively (e.g., increased statistical certainty) but also qualitatively (e.g., a much different, more detailed representation). We propose that *non-parametric models are hence more likely to extract neurobiological relationships that exclusively emerge in large brain datasets*. In contrast, parametric models are often more easily interpretable by the investigator, are more stable, and require less data to achieve a satisfactory model fit. Furthermore, parametric statistical tests are often more powerful, assuming the parametric assumptions are correct (cf. Friston, 2012; Eklund et al., 2016). These practical advantages are

Box 1: **Parametric and non-parametric models**

Contrary to common misunderstanding, both *parametric* and *non-parametric* statistical models involve parameters. 'Non-parametric' is typically defined in one of three different flavors (Bishop, 2006; Murphy, 2012; James et al., 2013): The first, perhaps most widespread meaning implies those statistical models that do not make explicit assumptions about a particular *probability distribution* (e.g., Gaussian distribution) from which the data have arisen. As a second and more general definition, non-parametric models do not assume that the *structure of the statistical model* is fixed. The third definition emphasizes that in non-parametric models, *the number of model parameters* increases explicitly or implicitly with the number of available data points (e.g., number of participants in the dataset). In contrast, the number of model parameters is fixed in parametric models and does not vary with sample size (Fig. 1). In its most extreme manifestation, non-parametric models might utilize larger memory than the actual input data themselves. Please note that the non-parametric scaling property of increasing model complexity with accumulating data can be obtained in different ways: i) a statistical model with infinitely many parameters or ii) a nested series of parametric models that can increase the number of parameters as needed (Ghahramani, 2015, page 454; Goodfellow et al., 2016, page 112).

The flexible non-parametric models include random forests (a special kind of decision-tree algorithm), boosting, nearest-neighbor algorithms (where complexity increases with the amount of input data), Gaussian Process methods, kernel support vector machines, kernel principal component analysis (kernel PCA), kernel ICA, kernel canonical correlation analysis, generalized additive models, and hierarchical clustering, as well as many forms of bootstrapping and other resampling procedures. Statistical models based on decision trees often constrain their size, which turns them into parametric models in practice. The more rigid parametric models include Gaussian mixture models, linear support vector machines, PCA, ICA, factor analysis, classical canonical correlation analysis, and k-means clustering, but also modern regression variants using sparsity or shrinkage regularization like Lasso, elastic net, and ridge regression.

Classical statistics has always had a strong preference for low-dimensional parametric models (Efron and Hastie, 2016). It is an advantage of parametric models to express the data compactly in often few model parameters. This increases interpretability, requires fewer data samples, has higher statistical power, and incurs lower computational load. Although the number of parameters in parametric models can be manually increased by the user, only non-parametric models have the inherent ability to automatically scale their *expressive capacity* with increasing data resources. Therefore, as the amount of neuroscience data continues to increase by leaps and bounds, parametric models might underfit the available data, while non-parametric models might discover increasingly complex representations that potentially yield novel neuroscientific insights.