Update
Comment

Modes or models: a critique on independent component analysis for fMRI

Karl J. Friston

This is a commentary on a novel approach to characterizing functional magnetic resonance imaging (fMRI) time-series presented in McKeown *et al.*[1] This characterization uses independent component analysis (ICA) and represents a departure from conventional approaches to fMRI data analysis. The paper is a significant contribution to the literature for a number of reasons: firstly, the interpretation of functional neuroimaging time-series is vitally important given the increasing role of imaging neuroscience in nearly every aspect of systems and cognitive neuroscience[2]. Secondly, ICA has attracted a lot of attention in many fields of research over the past few years, in particular the analysis of electrophysiological (EEG) and neuromagnetic (MEG) time-series[3]. Furthermore, it is closely related to algorithms, adopted in theoretical neuroscience, that emulate the way that the brain extracts information from sensory input in as efficient a way as possible[4].

The aim of ICA is to decompose a multi-channel or imaging time-series into a set of linearly separable 'spatial modes' and their associated time courses or dynamics. By adding together these dynamically changing spatial patterns, one reconstitutes the original observations. In this sense it is similar to principal component or eigenimage analysis[5]. The constraint employed by eigenimage analysis is that both the spatial and temporal profiles are all mutually orthogonal (and uniquely determined such that the first mode accounts for the greatest amount of variance, the second mode for the next greatest and so on). ICA, on the other hand, represents a somewhat more principled decomposition of biological time-series: the analysis renders either the temporal dynamics, or the spatial modes (but not both), not only orthogonal or uncorrelated, but independent. The distinction between 'uncorrelated' and 'independent' relates to correlations between non-linear functions of the variables in question. These are sometimes referred to as 'high-order correlations' (absent if the variables are independent but not necessarily so when they are only uncorrelated). More importantly, ICA does this in a fashion that renders the expression of the components non-Gaussian. In the implementation proposed by McKeown *et al.* these distributions are super-Gaussian or 'sparse'. This simply means that things happen infrequently. Why is a 'sparse', or more generally a non-Gaussian, distribution interesting? The answer to this question is simple and extremely compelling: because measurements of biological systems receive contributions from many sources (e.g. dipoles generated by neuronal activity), the observations usually represent a [roughly linear] mixture of interesting things. By the central limit theorem this mixture conforms to a Gaussian distribution. As mixtures themselves are uninteresting the only interesting things must be non-Gaussian (assuming that Gaussian distributions arise only from mixing). This is the rather beautiful motivation behind ICA. There are many ways of understanding the nature of ICA but this perspective highlights why ICA is so pertinent to biological time-series. In what follows we will look at the particular implementation of ICA in relation to fMRI time-series proposed by McKeown *et al.* and then consider this contribution in the context of extant approaches to fMRI data analysis, and the larger issues it raises in terms of the scientific process in imaging neuroscience.

ICA and fMRI time-series

In application to multi-channel EEG or MEG signals, independent components are generally identified using correlations among channels that are estimated over time. The output comprises a set of non-orthogonal spatial modes whose dynamics are independent and have a sparse distribution. This decomposition can be viewed as an elegant 'un-mixing' of the observed (linearly mixed) time-series to reveal the underlying and independent biological sources. This is very sensible and appeals directly to the conceptual basis of ICA. However, this approach is not that used for fMRI. In fMRI there are many more voxels (i.e. channels) than there are scans (i.e. time points). This is the complement of the situation in EEG and poses a computationally intractable problem if one wanted to apply ICA to the correlations among voxels. The clever trick, adopted by McKeown *et al.*, is literally to transpose the problem and derive independent components based upon correlations among different time points that are evaluated over voxels. The spatial modes that ensue are sparse and independent and express dynamics that are generally correlated. This lends the interpretation of the ensuing modes and their dynamics a very different complexion, relative to ICA analyses of EEG data. There are both pros and cons to the approach of McKeown *et al.* Firstly, the fact that neuronal responses that have distinct causes are likely to be spatially non-overlapping and regionally specific, provides a very nice motivation for identifying spatial modes that are independent and sparse. Secondly, the correlations between temporal dynamics frees one of the constraint that, for example, artifacts due to motion have to be orthogonal to those elicited by experimental design (in conventional analyses any confounding between a signal of interest and an artifact means that one has effectively lost that signal). On the negative side, the very nature of functional integration among brain areas means that large scale neuronal dynamics can share a substantial anatomical infrastructure. ICA would not find it easy to identify these systems[1]. More critically ICA precludes any nonlinear interactions among modes that underpin context-sensitive changes in functional architecture that are the focus of many neuroimaging studies. For example the spatial mode implicated in processing visual motion (that might include the lateral geniculate nucleus and cortical visual areas V1, V2, V5 and V3a) could be substantially modulated by changes in attention to visual motion (such that the contribution of V5 and V3a was increased relative to the other areas) where this modulatory effect might be mediated by an attentional mode involving the prefrontal and posterior

K.J. Friston is at The Wellcome Department of Cognitive Neurology, The National Hospital, Queen Square, London, UK WC1N 3BG.

tel: +44 171 833 7454
fax: +44 171 813 1445
e-mail: k.friston@fil.ion.ucl.ac.uk

 1364-6613/98/$19.00 PII: S1364-6613(98)01227-3

parietal cortices[6]. ICA would not handle this situation very gracefully and would probably fractionate the visual-motion system into two components (that do and do not show attentional modulation). Interestingly non-linearities[7] in the hemodynamic response to underlying neuronal changes at any one point in the brain are not problematic for ICA; the difficulties arise when anatomically distinct neuronal systems interact in a non-linear way[6] (see McKeown *et al.*, p. 184, for a discussion that touches on this).

In short, conventional applications of ICA identify sources with independent and sparse dynamics that may or may not be neuro-anatomically segregated. ICA, as proposed by McKeown *et al.*, identifies sparse, independent spatial modes that may or may not express correlated dynamics. The interpretation of the ensuing modes and their time courses rests upon a proper appreciation of this.

ICA in context

To understand the potential role for ICA, in characterizing fMRI time-series, it is worthwhile considering established techniques. The framework for the analysis of imaging time-series was largely established in positron emission tomography (PET) neuroimaging and has been extended, or recapitulated, for fMRI. In general approaches are either hypothesis-led or data-led (i.e. exploratory). The vast majority of imaging neuroscience depends upon hypothesis-led characterizations that are inferential in nature. These in turn are based upon some form of 'statistical parametric mapping'. Statistical parametric mapping refers to the construction of images using a voxel-specific statistic that tests hypotheses about the dynamics at that voxel. This statistic is usually derived under parametric assumptions using the general linear model (e.g. the T statistic, in testing for a particular compound of parameter estimates using multiple regression, or in the case of just one regressor, the correlation co-efficient)[8]. There have been some developments, such as the use of non-parametric statistics[9] and extensions to the general linear model that allow for serial correlations in fMRI time-series[10], but the overall approach has endured for a decade.

The second class of analyses are data-led and eschew any need to specify a statistical model about which inferences are made. Among these are eigenimage analysis, cluster analysis, multi-dimensional scaling and now ICA. ICA distinguishes itself because it is predicated on assumptions that seem particularly relevant to biological time-series. It should be noted that the conventional 'correlational' analysis[11] used by McKeown *et al.*, as a comparison for ICA, is the simplest form of statistical parametric mapping and uses a statistical model that might be sub-optimal in many circumstances. A typical model would normally include multiple components; for example a series of temporal basis functions for modelling voxel-specific differences in the form of evoked hemodynamic responses (i.e. consistent task-related responses), periodic confounds modelling low frequency drifts and aliased biorhythms and interactions among these terms to accommodate nonlinear responses[7] or time by condition interactions (i.e. transiently task-related responses)[12].

Why is the distinction between hypothesis-led and data-led approaches important? The usefulness of any new technique emerges in the context in which it is used. The facility of being able to apply techniques like ICA to imaging augments the ongoing re-evaluation of inferential approaches in imaging neuroscience. The ubiquity of inferential approaches reflects the (Popperian) scientific process adopted by the imaging neuroscience community (hypothesis generation, refutation and design of ensuing experiments). On the one hand questions have been raised about the usefulness of inferential statistics, especially in relation to fMRI time-series, and there have been calls to move towards alternative characterizations (such as confidence intervals, posterior possibilities, mixture models and nonparametric regression). On the other hand, however, the scientific process based on conventional statistical inference has proved extremely successful. It has allowed for the successive elaboration of increasingly sophisticated experiments, the testing of more refined hypotheses and a comprehensive characterization of functional brain architectures, in terms of both functional specialization and integration. Testing a hypothesis reduces operationally to specifying what one expects to see. In statistical parametric mapping this is realized in terms of the statistical model (i.e. design matrix) that embodies all the explanatory effects that constitute the hypothesis. These explanatory variables are nothing more than assumptions about the spatiotemporal dynamics of fMRI time-series and are central to scientific inference. Data-led techniques like ICA do not appeal to a statistical model and deny the opportunity to test any hypothesis (although they might provide insights that allow better models to be elaborated). Imaging neuroscience is now a maturing discipline wherein some principles of brain organization have already been elucidated. Within these broad principles, for example functional segregation, hundreds of imaging neuroscience programs are following the conventional scientific process to carefully characterize neuronal responses and their relationship to the brain's infrastructure. Working in any imaging laboratory repeatedly exposes one to the importance of being able to test hypotheses, where the results of one experiment inform the next, often generating more questions than resolutions. Furthermore, inferential approaches, based upon statistical models, allow insights from other fields to be adopted easily and powerfully. A clear example of this is the increasing use of multifactorial designs in neuroimaging. These approaches are predicated on advances in psychology (e.g. the additive factors method of Sternberg[13]) made many years before the first PET camera.

In my opinion the hypothesis-based scientific process is serving the imaging community extremely well at the present time. It has arisen not by explicit design (although the constraints of available data acquisition techniques and analysis methods could have played a role) but because it is so efficacious. It enables an internal consistency within scientific programs and a discourse between the imaging and other neuroscience communities. Techniques that seek an entry into the field on the basis of 'minimal assumptions' might not be attractive to those who have treasured assumptions about which they want to make inferences. In relation to the distinction between hypothesis-led and data-led approaches, it is interesting to note that the application of ICA considered here has emerged from a laboratory that is renowned for its insightful use of modelling techniques to emulate the behaviour of neuronal systems. The questions addressed by these approaches pertain to emergent behaviours and the exploration of the parameter space of neuronal models that are a long way away from the statistical models used in neuroimaging. It will be interesting to see if ICA can mediate between these two distinct approaches.

Acknowledgements

I thank Chris Frith for his invaluable comments. K.J.F. is supported by the Wellcome Trust.

References

1 McKeown, M.J. *et al.* (1998) Analysis of fMRI data by blind separation into independent spatial components *Hum. Brain Mapp.* 6, 160–188

2 Friston, K.J. (1997) Imaging cognitive anatomy *Trends Cognit. Sci.* 1, 21–27

3 Makeig, S. *et al.* (1997) Blind separation of auditory event-related responses into independent components *Proc. Natl. Acad. Sci. U. S. A.* 94, 10979–10984

4 Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* 381, 607–609

5 Friston, K.J. *et al.* (1993) Functional connectivity: the principal component analysis of large data sets *J. Cereb. Blood Flow Metab.* 13, 5–14

6 Buchel, C. and Friston, K.J. (1997) Modulation of connectivity in visual pathways by attention: cortical interactions evaluated

with structural equation modelling and fMRI *Cereb. Cortex* 7, 768–778

7 Friston, K.J. *et al.* (1998) Non-linear event-related responses in fMRI *Magn. Reson. Med.* 39, 41–52

8 Friston, K.J. *et al.* (1995) Statistical parametric maps in functional imaging: a general linear approach *Hum. Brain Mapp.* 2, 189–210

9 Kwong, K.K. (1995) Functional magnetic resonance imaging with echo planar imaging *Magn. Reson. Q.* 11, 1–20

10 Worsley, K.J. and Friston, K.J. (1995) Analysis of fMRI time-series revisited – again *NeuroImage* 2, 173–181

11 Bandettini, P.A. *et al.* (1993) Processing strategies for time-course data sets in functional MRI of the human brain *Magn. Reson. Med.* 30, 161–173

12 Büchel, C. *et al.* (1998) Brain systems mediating aversive conditioning: an event-related fMRI study *Neuron* 20, 947–957

13 Sternberg, S. (1969) The discovery of processing stages: extension of Donders method *Acta Psychol.* 30, 276–315

Response from Martin McKeown, Makeig, Brown, Jung, Kindermann, Bell and Sejnowski

As Professor Friston eloquently describes in his commentary[1], current techniques for analysing fMRI data can be dichotomized into either data-driven methods, like independent component analysis (ICA), or hypothesis-driven methods, like the General Linear Model[2]. These two approaches are complementary and mirror the exploratory and confirmatory aspects of scientific investigation. Imaging studies driven by hypotheses derived from cognitive psychology and related disciplines can at best support or refute currently formulated psychological models. Counterintuitive or unanticipated time courses of activation of localized brain areas are less likely to be discovered with such analysis methods. For example, in our paper, we demonstrated transiently task-related (TTR) ICA components whose time courses could not be anticipated before the experiment[3].

The ICA implementation described in our paper fully characterizes the data by separating them into sparse maps, or spatial modes, and associated time courses. Employing an ICA algorithm capable of looking for non-sparse as well as sparse maps found maps that were all sparse[4]. Many of these maps can be identified with known artifacts, such as blood vessel pulsations, head movements and slow drifts. These highly spatially structured signals are not easily modeled by *a priori* estimates as required by hypothesis-driven methods. The other key ICA assumption, that the maps are spatially independent, does not preclude the possibility of spatial overlap, because maximal independence can be achieved with overlap in high-dimensional spaces, especially as positive and negative regions of those maps can cancel.

Friston also rightly points out some of the inherent weaknesses of ICA. In an attempt to find maps that are maximally independent, ICA might tend to fragment broad areas of activation into multiple maps, all having highly correlated time courses. Like all linear models, ICA will be less sensitive to finding non-linear activation relationships between voxels, if these exist. More recent work takes advantage of an important by-product of ICA analysis; that the probability of observing the data conditioned on the ICA model is relatively easy to calculate, in order to estimate how well a linear model fits the data[5]. This report found that some brain regions (e.g. around blood vessels) might fit less well than say, subcortical white matter using an ICA model. Also, purely data-driven techniques like ICA, as Friston says, do not lend themselves to straightforward statistical analyses, potentially limiting their use for functional neuroimaging studies specifically designed to test certain hypotheses, rather than for exploratory analysis.

As the field continues to mature, we foresee the development of hybrid methods that will attempt to take advantage of these two complementary approaches: first employing powerful data-driven techniques to characterize the underlying nature of the signals and noise, then testing hypotheses of interest in the context of this accurate characterization. 'Data-driven' or 'hypothesis-driven' analysis methods will then refer to the extremes of a continuum, to be altered to fit the particular needs of the experimenter.

References

1 Friston, K.J. (1998) Modes or models: a critique on independent component analysis for fMRI *Trends Cognit. Sci.* 2, 373–375

2 Friston, K.J. (1996) Statistical parametric mapping and other analyses of functional imaging data, in *Brain Mapping: The Methods* (Toga, A.W. and Mazziotta, J.C., eds), pp. 363–396, Academic Press

3 McKeown, M.J. *et al.* (1998) Analysis of fMRI data by blind separation into independent spatial components *Hum. Brain Mapp.* 6, 160–188

4 McKeown, M.J. *et al.* (1998) Spatially independent activity patterns in functional magnetic resonance imaging data during the Stroop color-naming task *Proc. Natl. Acad. Sci. U. S. A.* 95, 803–810

5 McKeown, M. and Sejnowski, T. (1998) Independent component analysis of fMRI data: examining the assumptions *Hum. Brain Mapp.* 6, 368–372

Author affiliations

S. Makeig is at the Naval Health Research Center, PO Box 85122, San Diego, CA 92186-5122, and the Department of Neurosciences, School of Medicine, University of California, San Diego, La Jolla, CA 92093, USA.
G.G. Brown and S.S. Kindermann are at the Department of Psychiatry, School of Medicine, University of California, San Diego, La Jolla, CA 92093, USA.
T-P. Jung, A.J. Bell and T.J. Sejnowski are at the Computational Neurobiology Laboratory, Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92186-5800, USA.
T.J. Sejnowski is also at the Department of Biology, University of California, San Diego, La Jolla, CA 92093, USA.

M.J. McKeown is at 227D Bryan Research Building, Box 2900, Duke University Medical Center, Durham, NC 27710, USA.

tel: +1 919 684 0065
fax: +1 919 684 6514
e-mail: martin.mckeown@duke.edu

Affiliations of the remaining authors are at the end of the article.

 PII: S1364-6613(98)01228-5