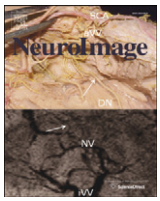
NeuroImage 54 (2011) 2198–2217

Contents lists available at ScienceDirect

NeuroImage

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



Linked independent component analysis for multimodal data fusion

Adrian R. Groves [a,*], Christian F. Beckmann [a,b,c], Steve M. Smith [a], Mark W. Woolrich [a,d]

[a] FMRIB (Oxford University Centre for Functional MRI of the Brain), Dept. Clinical Neurology, University of Oxford, UK
[b] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, The Netherlands
[c] MIRA Institute for Biomedical Technology and Technical Medicine, University of Twente, The Netherlands
[d] OHBA (Oxford University Centre for Human Brain Activity), Dept. Psychiatry, University of Oxford, UK

ARTICLE INFO

Article history:
Received 18 June 2010
Revised 15 September 2010
Accepted 27 September 2010
Available online 14 October 2010

ABSTRACT

In recent years, neuroimaging studies have increasingly been acquiring multiple modalities of data and searching for task- or disease-related changes in each modality separately. A major challenge in analysis is to find systematic approaches for fusing these differing data types together to automatically find patterns of related changes across multiple modalities, when they exist. Independent Component Analysis (ICA) is a popular unsupervised learning method that can be used to find the modes of variation in neuroimaging data across a group of subjects. When multimodal data is acquired for the subjects, ICA is typically performed separately on each modality, leading to incompatible decompositions across modalities. Using a modular Bayesian framework, we develop a novel "Linked ICA" model for simultaneously modelling and discovering common features across multiple modalities, which can potentially have completely different units, signal- and contrast-to-noise ratios, voxel counts, spatial smoothnesses and intensity distributions. Furthermore, this general model can be configured to allow tensor ICA or spatially-concatenated ICA decompositions, or a combination of both at the same time. Linked ICA automatically determines the optimal weighting of each modality, and also can detect single-modality structured components when present. This is a fully probabilistic approach, implemented using Variational Bayes. We evaluate the method on simulated multimodal data sets, as well as on a real data set of Alzheimer's patients and age-matched controls that combines two very different types of structural MRI data: morphological data (grey matter density) and diffusion data (fractional anisotropy, mean diffusivity, and tensor mode).

Introduction

One of the greatest strengths of MR neuroimaging is its flexibility; by using different pulse sequences in a single scanning session, one can acquire information about the subject's tissue volume and morphology (using high-resolution structural scans), functional activity (using BOLD FMRI), white matter integrity (using diffusion-weighted imaging), perfusion (using ASL), and other distinct acquisition types. The result of this is that many recent studies have acquired these multimodal MRI data sets for each subject and analysed them separately to find changes in different aspects of the brain. For example, several recent studies have used structural and diffusion tensor imaging (DTI) to find changes in grey matter density and white matter tracts that are related to schizophrenia (Douaud et al., 2007) or learning (Scholz et al., 2009). Other possible combinations are DTI and task-related FMRI (Watkins et al., 2008) or structural, diffusion, and resting-state FMRI (Filippini et al., 2009).

A major challenge is to find systematic approaches for fusing data across multiple MRI modalities, in order to find any patterns of related change that may be present. We develop a model based on Bayesian ICA to extract linked components from multimodal data, using as inputs the subject-wise contrast images from *modality-specific* analyses. For example, these inputs could be GLM contrasts from FMRI, cortical thickness or VBM maps from structural MRI, and skeletonised tensor measures from diffusion-weighted imaging. ICA is a particularly effective model for finding meaningful, spatially-independent components in an unsupervised setting because it searches for non-Gaussian spatial sources that are likely to represent real structured features in the data. This is because linear mixing processes tend to turn non-Gaussian independent sources into more Gaussian observed signals, so seeking non-Gaussianity is an unsupervised way of isolating the original independent sources.

Standard ICA decompositions treat the input data as a 2D matrix, typically voxels × timepoints or voxels × subjects. Multimodal data does not naturally fit into this form and there are a number of different configurations one could consider for performing combined ICA on multimodal data:

- Separate ICA analysis of each modality reveals the salient features for each modality. Since some of these features are caused by distributed neurological variations they could be visible (to varying degrees) in all modalities, with similar subject-courses.
  Corresponding components can then be matched up using heuristics; however there is no guarantee that components with strongly

* Corresponding author. Fax: +44 1865222717.
E-mail address: adriang@fmrib.ox.ac.uk (A.R. Groves).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.09.073

correlated subject-courses will be extracted, for example a single component in one modality might be explained as a mixture of components in another. When potential matches are found, it can be difficult to determine if they are simply noisy estimates of the same subject-course or whether the underlying subject-courses are different but correlated.
A slightly more sophisticated approach to this is the Parallel ICA method described by Liu et al. (2009) which runs separate ICAs on each modality simultaneously; when correlated components are detected, it adds terms to the cost function to encourage these components to become more correlated in later iterations. This relies on a number of tunable constraints (learning rates and weights) to ensure convergence and balance between modalities. Furthermore, it is still not clear how to interpret paired components where the subject-courses are significantly, but not perfectly, correlated.

- Spatial concatenation has also been used for analysing multimodal data, combining all of the data from each subject into a single dataset with more voxels. This "joint ICA" method has been used before for simultaneously analysing functional maps and grey matter maps (Calhoun et al., 2006), and has been used to extract correlations in structural grey matter/white matter density data (Xu et al., 2009). Since concatenation is a preprocessing step, the ICA model is completely unaware of which voxels belong to which modality.
However, different modalities may have different spatial source histograms. ICA effectively assumes that each component has a single, non-Gaussian histogram as the prior distribution for all voxels in its spatial map.
If this map consists of voxels from several different modalities, the modelled histogram (which is effectively an estimate of the source distribution) may have to compromise. For example, this can occur if one modality has a small area of strong activation (or signal change in the case of structural modalities), while the other has a large region of weak activation. This can cause sub-optimal estimates of intensities in spatial maps.
A related problem is that the contribution each modality makes to the ICA cost function greatly depends on the scaling. One of the difficulties of concatenating multimodal data is that the modalities may have different noise levels and different numbers of voxels. If the scaling is mismatched, unsupervised methods such as PCA and ICA will be dominated by the largest-variance modalities, or those with the most voxels. Typically these concatenation methods also require the same resolution and smoothing for all modalities, rather than using optimized values for each.
There is also an issue of noise covariance, for example due to spatial smoothing; in particular, adding more smoothing to one modality reduces the noise level but leaves the number of voxels unchanged. The proposed method deals with this explicitly using a precalculated correction for the number of effective degrees of freedom (eDOF), which is closely related to the number resolution elements (RESELs) in the image (Worsley et al., 1995).
We also expect that some of the structured signals modelled by ICA will be observable in only one modality, and may be extremely weak or even absent in some of the other modalities. It would therefore be useful for sources to be "switched off" in the models where they are not needed, just as it is important to eliminate unneeded components in the single-modality Bayesian ICA model (Choudrey and Roberts, 2001).

- Tensor ICA stacks the modalities to create a 3D data matrix. This has been used for multi-subject FMRI analysis, with dimensions of voxels × time × subjects (Beckmann and Smith, 2005). In the multimodal scenario this would most likely translate into voxels × subjects × modalities. This is related to the PARAFAC model (see (Nielsen, 2004) for a VB-based implementation) but with the addition of spatial-independence priors. This method assumes that each component has a single spatial map for all modalities, applied to each modality with different weightings. This can be a beneficial feature because it avoids unnecessary duplication of the spatial maps and can allow them to be inferred more accurately when the assumption holds. However this is effectively a strong prior on the nature of the spatial distribution and it may be inappropriate, for example if the number of voxels is different or if the spatial maps in different modalities are not similar.

Using a modular Bayesian framework, we have developed a novel "Linked ICA" general model that allows for either tensor ICA or spatially-concatenated ICA, or a combination of both at the same time. The same subject loading matrix is shared between all of the modalities, so each component consists of a single subject-course and one spatial map in each of the modalities. The subject-weighting matrix automatically balances information from all of the modalities. This novel Linked ICA method will be applied to a data set with four different modalities, acquired from 93 subjects (probable-Alzheimer's patients and age-matched controls). One of these modalities is a grey matter partial volume map ("GM") derived from Voxel-Based Morphometry (VBM) methods (Ashburner and Friston, 2000), and the other three are measures of white matter integrity: Fractional Anisotropy (FA), Mean Diffusivity (MD), and an orthogonal Tensor Mode (MO) described in Ennis and Kindlmann (2006). These last three modalities have been projected onto a two-dimensional white matter surface (the "skeleton") using a Tract-Based Spatial Statistics (TBSS) analysis (Smith et al., 2006).

Theory

Linked ICA model for multimodal data sets

We assume that the data set is from a group of R subjects, each scanned using several different modalities. It should be noted that the proposed method has the potential to be applied in any situation where multiple modalities have been collected across a single shared dimension (subjects, trials, timepoints, etc.). Each of the scans is prepared for analysis using whatever methods are recommended for a linear regression analysis (or a single-modality ICA) of the group data. This produces maps for each modality, which can have different spatial masks and different numbers of voxels. In this model, "modality" is defined as referring to a single contrast image (per subject) that refers to a particular output extracted from the data. Typically, different modalities will have different units, different scalings and different noise levels. In some cases, a single analysis may result in several different contrast images; for example, a diffusion tensor imaging (DTI) analysis can produce maps of FA (fractional anisotropy), MD (mean diffusivity) and MO (tensor mode). These are treated as separate modalities as they contain distinct, complementary, biophysical information.

However, to maintain the benefits of tensor ICA (inferring the same spatial patterns across modalities) as much as possible, similar modalities can be collected into K "modality groups". Modalities in the same modality group must be observations of the same points in space; this means the modalities must be spatially aligned to each other and have the same spatial mask, and should also have similar spatial properties (for example, the same amount of smoothing). A good example of this are multiple diffusion-derived measures projected onto a white matter skeleton using TBSS. The data can then be packed into a set of 3D arrays $\mathbf{Y}^{(k)} \in R^{N_k \times T_k \times R}$, where N_k is the number of voxels in the shared spatial map and $T_k \geq 1$ is the number of modalities in the kth modality group. Each modality group is modelled using a Bayesian tensor ICA model. This general configuration is shown in Fig. 1. Note that the Bayesian ICA differs from standard methods like FastICA (Hyvärinen and Oja, 2000) in that it incorporates dimensionality reduction into the ICA method itself by the use of automatic relevance determination (ARD) priors on

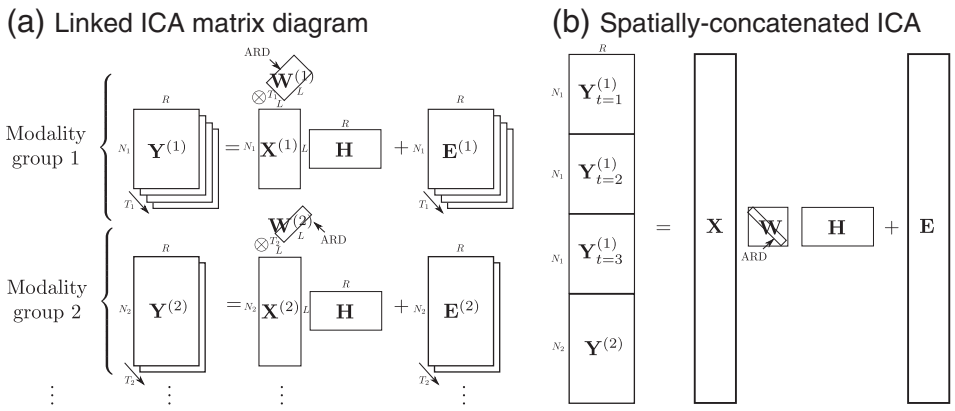


Fig. 1. (a) The main matrices of the Linked ICA that models multimodal data **Y**. Note that the same subject loading matrix **H** is used for all of the modality groups, but otherwise they are *K* separate Tensor ICAs, each with separate data dimensions $N_k \times T_k \times R$ (voxels × modalities × subjects). Each of the modality groups contains one or more modalities stacked together, expressed in terms of spatial maps $\mathbf{X}^{(k)}$, modality weights $\mathbf{W}^{(k)}$, a shared subject-weighting matrix **H**, and additive noise $\mathbf{E}^{(k)}$. (b) The spatially-concatenated ICA configuration, for comparison. This model is almost identical to a standard Bayesian ICA.

components (Choudrey and Roberts, 2001; Bishop, 1999). The model works on the full-dimensionality data directly and has an additive noise model. The Bayesian ICA also models an explicitly parametrized non-Gaussian source model (in this case a Gaussian mixture model) instead of maximizing negentropy (as used in FastICA).

Bayesian tensor ICA model

Within each modality group k the data is modelled as a sum of components using a tensor decomposition. Each component $i = 1 \ldots L$ can be expressed as the tensor product of one spatial map, one subject-course, and one modality-course. These model the data in modality group $k = 1:K$, modality $t = 1:T_k$, subject $r = 1:R$ and voxel $n = 1:N_k$ as

$$\mathbf{Y}^{(k)}_{n,t,r} = \sum_{i=1}^{L} \mathbf{X}^{(k)}_{n,i} \mathbf{W}^{(k)}_{t,i} \mathbf{H}_{i,r} + \mathbf{E}^{(k)}_{n,t,r} \tag{1}$$

where $\mathbf{X}^{(k)}_{n,i}$ are the spatial maps for component i in modality group k, $\mathbf{W}^{(k)}_{t,i}$ are the modality weightings for component i in modality t (of modality group k), and $\mathbf{H}_{i,r}$ are the weights for component i in subject r. For simplicity this model is used even when $T_k = 1$, so that $\mathbf{W}^{(k)}_{:,i}$ is just a scalar. Crucially, the same **H** matrix is shared between all of the modality groups; this forms a link between the different modality groups, which are otherwise modelled completely separately. The ith component has the same subject weightings across modality groups but each group has its own spatial map. Thus the number of repeats R and the maximum number of components L must be the same everywhere, because these dimensions are shared, while N_k and T_k are not. Uncorrelated Gaussian residuals are assumed, with the modality-dependent noise precision (inverse variance) $\lambda^{(k)}_t$:

$$\mathbf{E}^{(k)}_{n,t,r} \sim N(0, 1 / \lambda^{(k)}_t). \tag{2}$$

Note that this assumes the same noise variance for each voxel, while in the original data there may actually be large (orders-of-magnitude) differences in the white noise intensity. To correct for this, we rely on a robust preprocessing method called variance normalization which is widely used for ICA on functional MRI (Beckmann and Smith, 2004).

The sketch of the Linked ICA matrices is shown in Fig. 1, and Fig. 2 shows how these variables fit into the full Linked ICA graphical model; this also includes the hyperparameters explained in the next two sections. Aside from the shared matrix **H**, linked ICA model is identical to performing separate tensor ICA analyses: each modality group k has its own separate source mixture model, as well as having its own noise model and separate ARD priors to drive different patterns of sparsity. Note that r indexes the "repeats" dimension and is the dimension that is shared across modality groups, for example r indexes subjects in the multi-subject application.

Adaptive modality-weighting

The tensor model (Eq. (1)) implies that the same spatial sources $\mathbf{X}^{(k)}_{:,i}$ are used for all of the different maps $t \in 1..T_k$, with weightings given by $\mathbf{W}^{(k)}_{t,i}$. In previous tensor ICA applications (Beckmann and Smith, 2005), this t dimension indexes over repeats of the same scan, such as in a multi-subject FMRI data from a study with identical stimulus timings. In that case it makes sense to assume the same noise level for all timepoints, i.e. use only a scalar $\lambda^{(k)}$. Instead, the tensor model is being used here to consolidate several different contrast maps produced by the analysis of a single MRI scan. Since in this case each t refers to different underlying modalities of data, it needs to model differences in scaling and noise levels, hence the modality-specific $\lambda^{(k)}_t$ is used.

To adapt to different scalings of the *signal* in each modality, an ARD prior (Wipf and Nagarajan, 2008) is used on the modality-courses (**W**). Just like the noise level, the relative scaling of the data in each t needs to be determined independently; thus independent ARDs are placed on each *element* of **W**.

$$\mathrm{P}\left(\mathbf{W}^{(k)}\right) = \prod_{t=1}^{T_k} \prod_{i=1}^{L} N(\mathbf{W}^{(k)}_{t,i} \,|\, 0, (\omega^{(k)}_{t,i})^{-1}) \tag{3}$$

with an approximately scale-free prior on each $\omega^{(k)}_{t,i}$; as $\omega^{(k)}_{t,i} \to \infty$ that will effectively eliminate a source from that timepoint by forcing $\mathbf{W}^{(k)}_{t,i}$ to zero with very high precision. In this approach, the number of sources is not explicitly chosen, but the method automatically determines the number of sources that are needed to optimally describe the data. We start with a full set of sources and allow the model to gradually downweight and eliminate sources that are too weak.

This means that it is now possible to eliminate a source from some modalities while keeping it in others, so it is possible to model effects like single-modality structured noise/artefacts. This means that the subject-course no longer needs an ARD, so it has a simple fixed prior:

$$\mathrm{P}(\mathbf{H}) = \prod_{i=1}^{L} \prod_{r=1}^{R} N(\mathbf{H}_{i,r} \,|\, 0, 1). \tag{4}$$

This dimensionless prior on **H** is analogous to the fixed unit variance priors used in variational PCA (Bishop, 1999). When a source has not been eliminated, the ARD priors on **W** will tend to balance with this fixed-scale prior to keep the rows of **H** close to a (root-mean-

squared) amplitude of one. This means that each column of **H** is a dimensionless vector summarising everything that varies between different subjects' scans (apart from residual noise). This means that **H** models normalized variability over subjects with the overall scale of this variability in the data being modelled elsewhere, in **W**. This modality-independent hidden state provides a probabilistic link between separate ICAs.

There are situations in which one modality in itself consists of several distinct timepoints; for example, multi-subject whole FMRI scans with synchronised stimuli (Beckmann and Smith, 2005) or identical structural scans acquired longitudinally to image individual neurodegeneration. These can easily be modelled in this framework by returning to a single $\omega_i^{(k)}$ and $\lambda^{(k)}$ for all timepoints of that component.

Independent spatial sources

The driving force behind an ICA decomposition is that the data is derived from a number of statistically independent spatial sources; these are the spatial maps ($\mathbf{X}_{:,i}^{(k)}$ for $i=1\ldots L$). By the central limit theorem, linear mixing of independent sources will produce output that is more Gaussian than the sources. An approach for finding this non-Gaussianity is to explicitly fit a non-Gaussian distribution to each source by assuming a particular functional form. This is the approach taken here, using an M-component Gaussian mixture model as proposed for independent factor analysis by Attias (1998).

This models the elements of each spatial source ($\mathbf{X}_{:,i}^{(k)}$) as being drawn from an M-component Gaussian mixture model with means $\mu_{i,\,m}$, precisions $\beta_{i,\,m}$, and component proportions $\pi_{i,\,m}$. This is a good approximate model for a variety of underlying distributions (Choudrey and Roberts, 2001). The Gaussian mixture model prior on the spatial maps can be expressed as

$$\mathrm{P}\left(\mathbf{X}_{n,i}^{(k)}\,|\,\mu,\beta,\pi\right) = \sum_{m=1}^{M} \pi_{i,m}^{(k)}\,\mathrm{N}\left(\mathbf{X}_{n,i}^{(k)}\,|\,\mu_{i,m}^{(k)},\,1/\beta_{i,m}^{(k)}\right) \tag{5}$$

and a hidden mixture membership variable $q_{n,\,i}^{(k)}=m$ indicates which mixture component $\mathbf{X}_{n,\,i}^{(k)}$ was drawn from.

For simplicity, the model presented in this paper uses a fixed $M=3$ mixture components. In practice, using 2 or 3 mixture components seems to extract the non-Gaussian sources from noisy simulated data reasonably well, although a Gamma–Gaussian mixture model may actually be a more appropriate model (Makni et al., 2006; Beckmann and Smith, 2004; Woolrich et al., 2005; Hartvig and Jensen, 2000).

The relatively uninformative priors on μ, β, and π are given in Appendix D. It is worth noting that there is theoretically a scale ambiguity in this model, in that simultaneously rescaling μ, $\beta^{-1/2}$, $\mathbf{W}^{-1}$ and $\omega^{1/2}$ can result in the same model fit. However, in practice, the scaling parameter **W** and its ARD prior ω respond much more quickly to produce overall changes in component weight, while this adaptation occurs far more slowly (if at all) when the GMM is relied upon for component scaling and elimination.

Variational Bayesian inference

We fit this model using Variational Bayes (VB), which is a fast iterative approach for approximate Bayesian inference (Attias, 2000). The full posterior distribution is intractable, so the mean field approximation is used and the posterior distribution is factorized as

$$\begin{aligned} &P(\mathbf{Y}\,|\,\mathbf{H},\beta,\mu,\pi,\mathbf{W},\omega,\lambda,\mathbf{X},\mathbf{q}) \\ &\approx P'(\mathbf{H})P'(\beta)P'(\mu)P'(\pi)P'(\mathbf{W})P'(\omega)P'(\lambda)\prod_{k=1}^{K}\prod_{i=1}^{L}P'(\mathbf{X}^{(k,i)},\mathbf{q}^{(k,i)}). \end{aligned} \tag{6}$$

Notice that this explicitly factorizes the spatial sources **X** over components i. The solution is still analytic without this factorization, but the number of components in the joint mixture model grows as M^L (see Attias, 1998). This is an approximation, but a reasonable one because the component sources are assumed to be statistically independent of one another.

For all details of these updates and the free energy F, see Appendix C. The free energy F was used to validate the VB updates (by ensuring that $\Delta F \geq 0$ after each update) and also to monitor convergence. The analysis software is implemented in MATLAB.

Precision contributions

When a component represents some real underlying variation between subjects, it can fuse information across several modalities. We might expect these variations to show up more clearly in some modalities than in others, this simply being a reflection of possibly marked changes in contrast-to-noise ratio (CNR) of each particular feature. This subsection describes a simple measure for assessing the relative influence of each different modality in defining the subject-course of each component.

The latent space **H** is shared between all modalities, so its posterior $\mathrm{P}'(\mathbf{H})$ combines contributions from all of the modalities (as well as the prior). This is the only time that estimates from across modalities are combined, so it is an appropriate place to look in order to find out which modalities are driving a particular component.

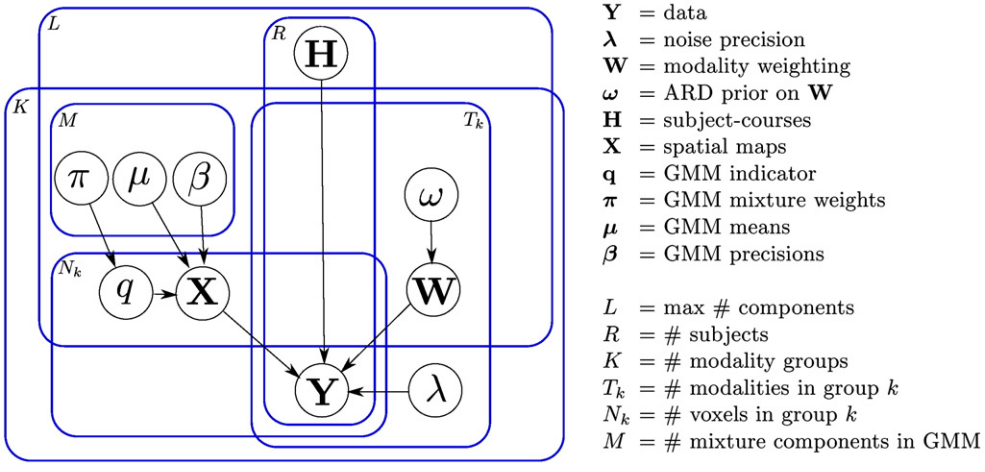


Fig. 2. The graphical model showing the relationships between all parameters and hyperparameters. Plates represent replicated variables or matrix sizes, with dimension in a top corner; for example, **H** is $L\times R$, and **Y** consists of K arrays of dimension $N_k\times T_k\times R$. Fixed hyperpriors have been omitted.

The technical details are explained in Appendix E. Basically, each modality provides its own estimate of $\mathbf{H}$ and the posterior $\langle\mathbf{H}\rangle$ is a precision-weighted average of these. To find the dominant modalities in estimating each source's subject-course, it is informative to look at these precisions.

In the Evaluations and results section the figures will show these "precision contributions" normalized so that the sum of all contributions is 1 for each component i. This makes it easy to see if a component is dominated by one modality or is informed by a combination of several modalities. The prior is also included in this sum, so if a component is eliminated, the dominant precision contribution will be from the prior.

Spatial smoothness correction

Most MRI data used in real analyses will have had some level of spatial smoothing applied as a deliberate preprocessing step. This is primarily because the signals tend to be extended (in the case of BOLD FMRI) or are not perfectly aligned across subjects (in the case of a VBM-style analysis). In these cases, spatially smoothing improves the SNR of the signals of interest. However, we are using an uncorrelated white noise model (Eq. (2)), which means that the model believes it has more independent data points than there are actual degrees of freedom in the data.

There are several approaches available to correct for smoothness. The most direct approach is to use an explicit model of spatially-smooth noise, e.g. using a Gaussian process as the noise model. This can adaptively determine the amount of smoothness in the noise and deconvolve the smoothing kernel. However, this approach is prohibitively slow and can introduce severe numerical problems.

A simpler approach for reducing the impact of this eDOF/voxel mismatch is to decimate the data, reducing the number of voxels while retaining as many independent measurements as possible. There are practical problems with this approach, particularly in the choice of exactly which voxels to remove and the fact that some information will always be lost (and some correlations will always remain).

Instead of decimating the data to remove this spatial correlation structure, we perform a "virtual decimation" by downweighting the effective number of voxels in the VB update equations. This keeps all of the original voxels but downweights any summation over voxels to reflect the effective degrees of freedom (eDOF) ν_k instead of the actual number of voxels N_k.

By keeping this decimation factor fixed, the VB updates and free energy F remain consistent, even permitting model comparison as usual. At present, the smoothness is estimated as a preprocessing step and kept constant across all methods. This approach is fully described in Appendix A and its effect will be demonstrated on simulated data in the Evaluation and results section.

Preprocessing

Each modality's data is de-meaned in the subject dimension. This removes the mean spatial map to emphasize differences between subjects. The mean level of each map (e.g. each subject's total grey matter density) is not removed because this contains important information that distinguishes subjects.

A more serious issue is that the variance can vary enormously from voxel to voxel. This is especially true in FMRI data (where there can be a two order-of-magnitude difference in noise variance between CSF and white matter voxels), and it is also true (to a lesser extent) for the structural modalities used in this paper. The current model assumes the same noise precision $\lambda_t^{(k)}$ for all voxels in a modality. In principle it is possible to estimate per-voxel noise levels but instead we use the well-established empirical method for variance normalization from probabilistic ICA (Beckmann, 2004). This attempts to estimate the per-voxel scaling of the underlying white noise by looking only at the centre of the intensity histograms and ignoring the tails.

As with most ICA implementations, Linked ICA is initialized from a PCA decomposition. For multimodal data, the natural method is to concatenate all of the voxels across modalities k, t (to get a $(\sum_{k=1}^{K} N_k T_k) \times R$ matrix) and then do a PCA decomposition on that. The full details are given in Appendix B.

Component elimination

To provide a large ($\sim 10\times$) speedup in inference, eliminated components (or part-components) are removed from the model completely, avoiding additional inference on these zero-weight spatial maps. These are actually removed from the model as well, removing the free energy cost associated with keeping these parameters around; this cost is partially related to the factorization, and is highly dependent on the priors selected, e.g. the cost is doubled if an "uninformative" $P(\omega) \sim Ga(10^{12}, 10^{-12})$ prior is used instead of a $Ga(10^{6}, 10^{-6})$. This does not affect the estimated posterior in any significant way, but it does greatly change the model comparison, especially with many initial components ($L = 90$). When using this elimination approach, the model comparison results are independent of the number of extra components in the model.

Convergence

Convergence is monitored by evaluating the free energy F. In practice, evaluating F takes longer than a full cycle of updates so it is not done in every iteration. We evaluate F at logarithmically-spaced numbers of iterations, so that it is evaluated when the iteration count is $\lceil(\sqrt{2})^{integer}\rceil$. Convergence is declared when the change in F per iteration drops below 0.1.

Compared approaches

Of the existing data configurations discussed in the introduction, spatially-concatenated ICA (Fig. 1b) represents the only reasonable way to arrange all of the available multimodal data into a single ICA decomposition. This enables all of the data to be used in inferring subject-courses $\mathbf{H}$, but treats all voxels the same regardless of which modality they came from, and also loses the spatial correspondence between the modalities in any modality groups.

All voxels are concatenated across modalities and a single, large spatial map containing all of the modalities is used, with the same noise-normalizing scaling as used for the initial PCA. This is the same configuration of the data used in joint ICA (Calhoun et al., 2006) but we use Bayesian ICA for inference. This "Concatenated ICA" approach is the baseline model against which the Linked ICA models will be evaluated.

In this situation, all of the spatial maps are concatenated voxelwise and the same mixture model is used for the entire concatenated spatial map. Furthermore the tensor model is flattened out so that instead of modalities sharing the same basic pattern with different scalings, the link between corresponding voxels is lost and the spatial map has to be learned separately for each modality, as illustrated in Fig. 1(b). The $\mathbf{W}$ matrix is still present but it can only scale (and eliminate) each component from the entire model. This makes the model identical to the Bayesian ICA model of (Choudrey and Roberts, 2001), aside from the fact that the ARD is on the scaling matrix $diag(\mathbf{W}_{t,\cdot})$ rather than on the rows of $\mathbf{H}$.

For fairness, the correction for spatial smoothness described in Appendix A is also applied to the concatenated data. Because the concatenated modalities may have different amounts of intrinsic smoothness, the weight applied to each voxel will depend on which modality it came from. To our knowledge this correction is not used in existing multimodal methods (Calhoun et al., 2006; Xu et al., 2009) which avoid the issue by using identical spatial smoothing levels even across very different modalities.

Evaluation and results

Simulated multimodal data

This section presents a simulated multimodal data set, which will be analysed using linked tensor ICA and spatially-concatenated ICA to demonstrate the differences between the two approaches in terms of modelling common (multimodal) components and single-modality structured noise components.

A simulated multimodal data set was constructed with four modalities in two modality groups. The first group contains three modalities of 1000 voxels each that share the same spatial patterns with different weightings, and the second group which has a single, 3000-voxel modality. The spatial maps are shown in Fig. 3. For all of the images of simulated data, the spatial maps shown have a consistent colour scale running from $Z = \pm 5$, i.e. five standard deviations of the noise in estimating the spatial map.

There were three common components (labelled C1–C3) that were expressed in each of the modalities, and a structured noise source that was unique to each modality (N1–N4). The subject-courses (not shown, $R = 100$) are white noise, but C1 and C2 were generated with somewhat correlated subject-courses (30%) in order to make crosstalk (cross-contamination) between this pair of sources more likely. This means that the initial PCA will mix the signal components together, so the ICA method must move away from its initialization point and seek non-Gaussian histograms in order to accurately separate components with non-orthogonal subject-courses.

Fig. 3(b) shows the histogram of the true activation levels used in the simulation. These were chosen so that the two modality groups had very different histograms in terms of sparsity, symmetry, voxel count and SNR. Gamma distributions were used because the heavy tails are thought to more accurately reflect the properties of real activation than a Gaussian. The activation distributions shown only account for part of the intensity histogram; the remaining 60% or 90% of voxels are inactive and therefore collect in a large point mass at exactly zero. In the high-noise simulation, the white noise added to the four modalities had standard deviations of 25, 30, 35, and 50 respectively. In the low-noise simulation, the same signal was used but with noise scales of 15, 20, 25, and 40.

The number of components was set to $L = 10$ so that with a true dimensionality of 7 there was work for the dimensionality estimation (via the ARD prior on $\mathbf{W}$) to do. The inference ran until it converged ($\Delta F < 0.1$), which took 200–1000 iterations. This MATLAB code took about 5 min to run each inference on a single core of a 2.4 GHz AMD Opteron 8431 processor.

Results on unsmoothed simulated data

The precision contribution plots are shown in Fig. 4(a,b). In the Linked ICA model at both noise levels, the three shared sources (components 1–3) each split their precision fairly equally between the four modalities. The structured noise sources (4,5,6,7 or 4,5,6,9) are mostly determined by a single modality, and the others are eliminated completely (dominated by the prior precision). In the concatenated model at high noise, only 4 sources are inferred and the rest are eliminated. The last components (4) primarily models the strongest structured noise source.

This occurs in high noise (but not low noise) because these structured noise components are only slightly detectable above the high noise level, so the Concatenated model determined that switching off those additional components (by inferring $\mathbf{W}_i = 0$) provides a more concise explanation of the data. In the Linked ICA model it is possible to switch off each modality's contribution to component i separately (by inferring $\mathbf{W}_{ti}^{(k)} = 0$ for only some modalities k, t), so the complexity penalty for keeping each component is reduced. Model comparison also strongly prefers the correct model (Fig. 4c) over several alternatives. The subject-courses are all extracted slightly more accurately with Linked ICA, but this improvement is relatively small (not shown).

Fig. 5 shows the inferred spatial maps. Linked ICA correctly extracts all seven sources in both the high- and low-noise data, while 3 of the structured noise sources are lost in the high-noise data in the concatenated approach. All of the common-signal spatial maps are recovered well. The signal from C1 is shown contaminating signal C2 in the concatenated model, while linked ICA separates these signals much more cleanly.

The ROC curves show that the Linked ICA method discriminates active from non-active voxels more accurately than the concatenation approach, at both noise levels. This advantage is largest in modality

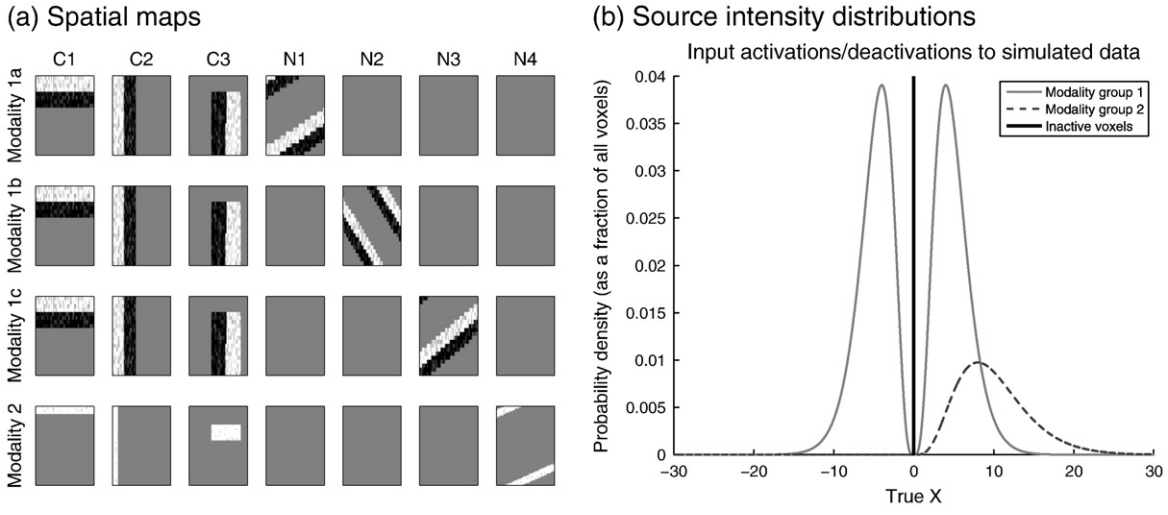


Fig. 3. The simulated multimodal data. (a) There are seven components: three shared sources (C1–C3) that appear in all four modalities, and four structured noise sources (N1–N4) each appearing in a single modality. The first three modalities (1a–1c) have 1000 voxels each (20×50)and are in the same modality group and therefore share the same spatial maps; in this example all of the weights are the same (set to 1) but they have different noise levels. The last modality (labelled modality 2) has 3000 voxels (60×50) and different spatial maps from 1a–c. All of these images are scaled consistently relative to the noise level in the data. (b) The histograms of the two modality groups are very different: group $k = 1$ has 40% of its voxels active (half positive and half negative), while group $k = 2$ has only 10% active (all positive) but the activation is stronger. Note that the remaining 60% or 90% of voxels in both histograms are completely inactive so they are shown by the peak at exactly zero.

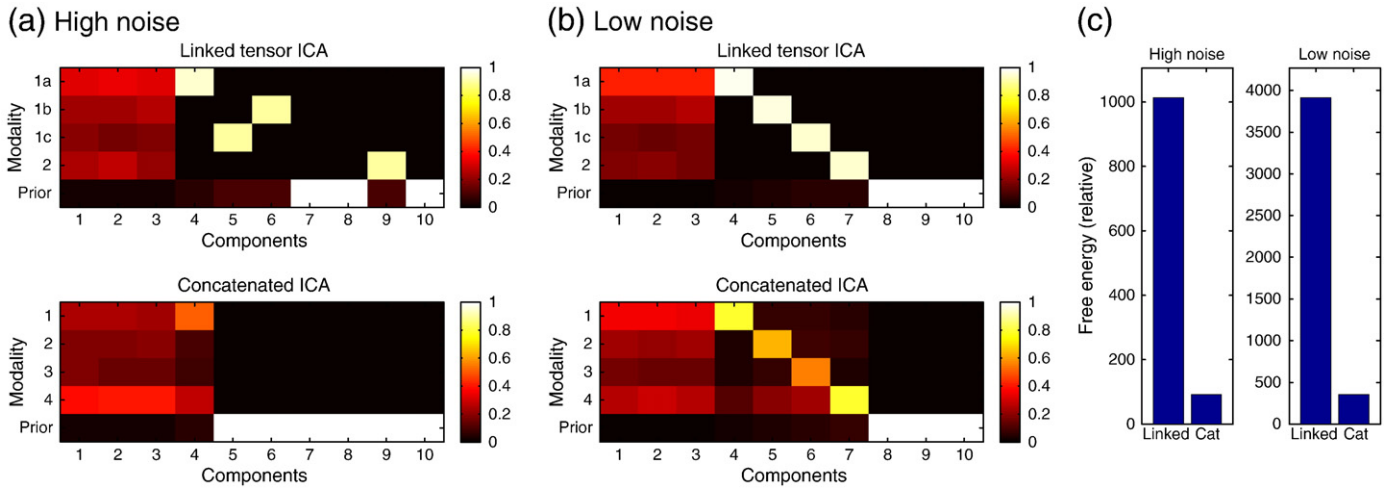


Fig. 4. Inferred precision contributions for the simulated data set, showing which modalities dominate in determining each source's subject-course. Notice that the Linked ICA method consistently identifies the structured noise modalities (components 4–7) while the Concatenated ICA exhibits mixing (most severely in high noise). (c) Model comparison results, showing that the true Linked ICA model is preferred to the Concatenated ICA model in both data sets (more so in low noise).

group 1 because Linked ICA knows that the three modalities share the same spatial map whereas the concatenated approach does not. A small improvement is observed even in the maps for modality 2 where there is no such benefit. This may be partially due to the improvements in subject-course estimates or because it models separate histograms for modality group 1 and modality group 2.

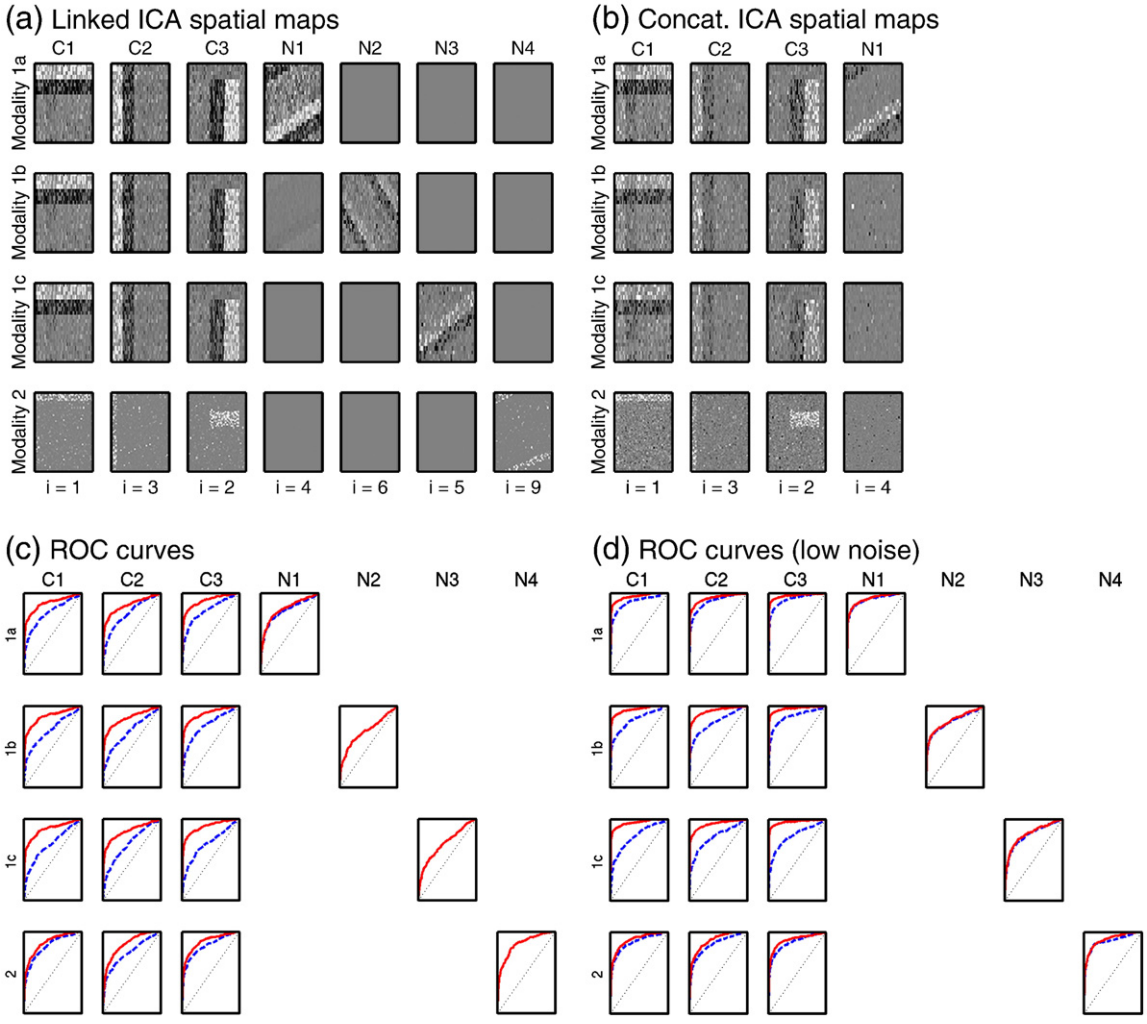


Fig. 5. Inferred spatial maps on simulated data in high noise (a) using linked ICA and (b) using concatenated ICA; the true sources are shown in figure 3(a). The ROC curves showing how well each method's $|\langle \mathbf{H} \rangle|$ discriminates active voxels (both positive and negative) from inactive voxels. The solid red lines show Linked ICA results while the dashed blue lines show the concatenated ICA results, in the (c) high noise and (d) low-noise conditions. The diagonal dotted line indicates chance. Note that there are no dashed blue lines for N2–N4 in high noise because Concatenated ICA did not find components to match those sources.

To assess the sensitivity to initialization, we re-ran these simulations using random initial subject-courses instead of the PCA decomposition. On the low-noise simulation, the results for both Linked ICA and Concat ICA were nearly identical in terms of final subject-courses and spatial maps (although of course the order of the components was different). The modality weights were also very similar, with the final weights varying by ±0.6% on Linked ICA and ± 3.5% on Concatenated ICA.

On the high-noise simulation, some of the weaker components (especially N4) were lost due to the poorer initializations. Note that in these cases we used $L=90$ so that there were more components to eliminate; starting with $L=10$ gave similar results but was more likely to lose components.

We also initialized using the minor PCA subspace, i.e. discarding the strongest 10 PCA components instead of using them as initial components; again, the resulting subject-courses, spatial maps and modality weights were nearly identical. This illustrates that the initial PCA is not a hard dimensionality-reduction step. The variance that is not part of any initial component is not discarded, but is instead initially treated as part of the noise. The signals buried in this noise can be sought out and recovered within the iterative Bayesian ICA framework.

We also assessed the effect of changing the Gaussian mixture model order (from the default value of $M=3$) on the low-noise simulated data. In Concatenated ICA, using the simpler $M=2$ model made some components slightly worse (e.g. reducing correlation with the true spatial map from 0.772 to 0.767) and others slightly better (e.g. from 0.743 to 0.748). This not surprising because Concatenated ICA uses a single average histogram for two different sources; changing the prior on that histogram can bias the estimates towards having slightly heavier or lighter tails, which may fit one source better and one worse. In the other direction, increasing from $M=3$ to $M=4$ had a far smaller effect. In Linked ICA, the differences were all negligible. This shows that M = 3 is a sensible choice and it is used throughout this paper.

Smoothed simulation results

The results in Fig. 4 were based on simulated data with no spatial correlations in the noise. However, real neuroimaging data often has significant spatial smoothing and this section shows the behaviour of the Linked ICAmodel with and without the smoothness correction (described in Spatial smoothness correction section and Appendix A).

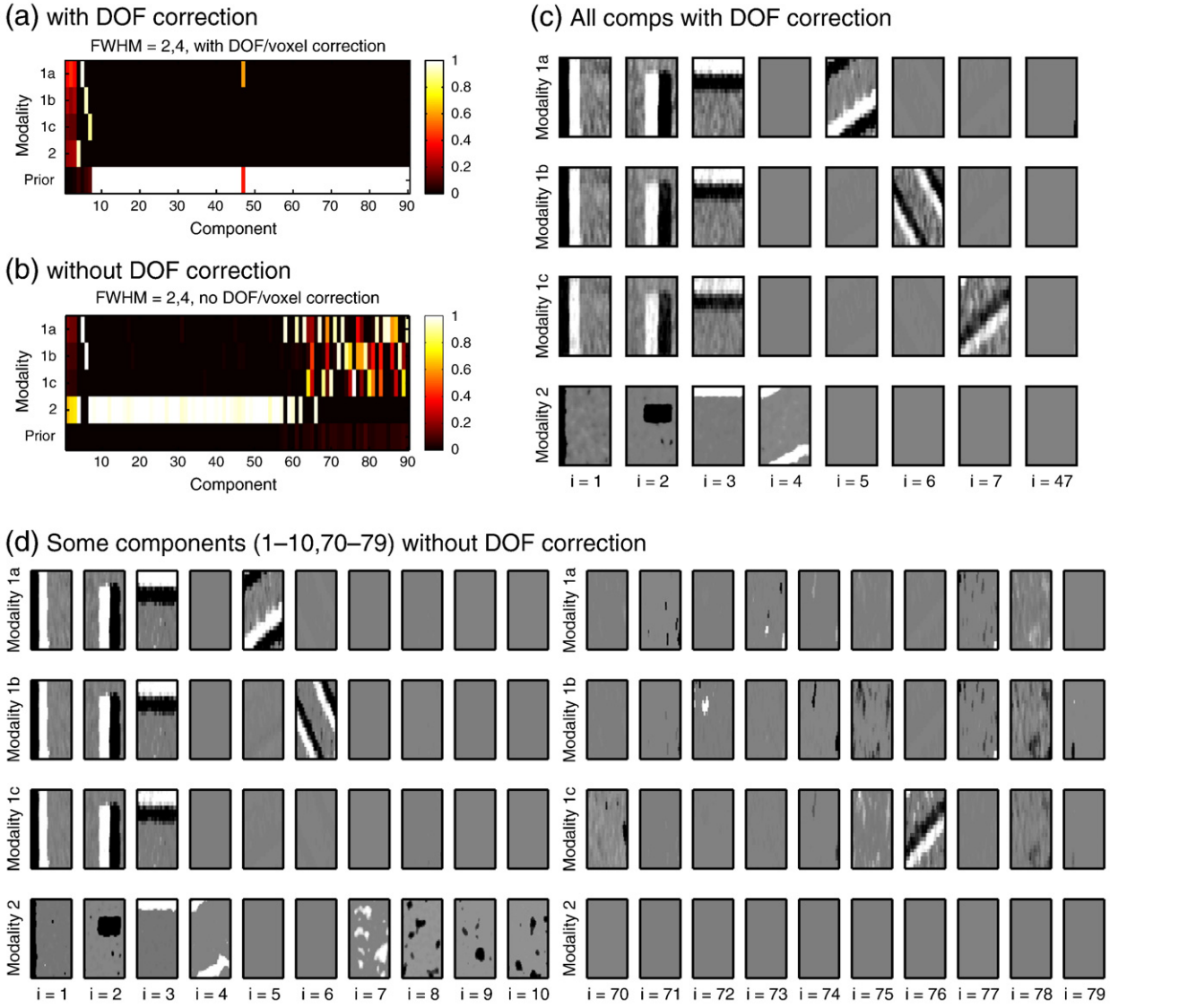


Fig. 6. Analysis of simulated data smoothed with Gaussian kernels with FWHM of 2 voxels and 4 voxels on modality groups 1 and 2 respectively. (a–b) The precision contribution plots with and without the correction for smoothness. (c–d) Spatial maps of the components inferred with and without this correction. Without the correction, many extra components are inferred that model spatially-smooth noise in the data.

The data set from the previous section was re-generated using higher noise standard deviations of [30 40 50 80]. We applied a spherical Gaussian smoothing kernel to each simulated modality, using a FWHM of 2 voxels on modalities 1a–1c, and 4 voxels for modality 2. The exact DOF in each case (as given by Eq. (A.1)) is 0.23 DOF/voxel and 0.058 DOF/voxel respectively. Estimating this from the data yields estimates of 0.18, 0.19, and 0.20 DOF/voxel for modalities 1a, 1b, and 1c respectively; these imply greater smoothness than is really present because they include the very smooth signal in the data, and 1a is more severely affected because it has a higher SNR. Modality 2 is estimated at 0.047 DOF/voxel. When these estimates are computed on the noise only (i.e. the residuals), the estimates are considerably more accurate, 0.25 and 0.061 DOF/voxel.

Clearly it would be more accurate to compute DOF/voxel on the residuals rather than the original data; however, for reasons discussed in Appendix A, the current implementation uses the values derived from the full data set. We also average the estimate of DOF/voxel across modalities 1a–c, because they share spatial maps and therefore ought to have the same true amount of smoothness. Very similar results are obtained if the exact DOF/voxel values are used instead.

The precision contribution plots are shown in Fig. 6, illustrating the necessity of correcting for spatial smoothness. Note how the DOF/voxel weighting enables the method (Fig. 6a) to estimate the dimensionality almost perfectly, even starting from L = 90. Only one extraneous component survives (and the large weighting given to the prior contribution to this component indicates that was close to being eliminated). The inferred components are accurate (Fig. 6c). In contrast, Fig. 6b shows that without the DOF correction all 90 components are kept. A subset of these is shown in Fig. 6d, demonstrating that most of these are modelling spatially-smooth noise patterns.

The common signal and structured noise components are well-recovered by both methods, but in the uncorrected approach note that the common components (C1–C3) dominated by the smoother modality (2), because smoothing reduces the noise level without changing the number of voxels. In the corrected method, the weighting of these components is similar to the unsmoothed case. Also, note that one of the structured noise terms is pushed down to component 76 by the many erroneous noise-smoothness components in modality group 2, which have larger apparent significances.

Separation of correlated subject-courses

All of the configurations discussed in this paper assume that there is a single subject-course for each component, which is identical across modalities. If the subject-courses are different across modalities, even if they are strongly correlated, then they should be split into separate components. The model has to make this hard distinction, so it is possible for two similar components to be combined into one, or for a single component to be split due to noise.

The Linked ICA methods reduce this problem by informing the model where the divisions between the modalities lie, making it easier to split the component apart when the modalities' subject-courses are different. We simulated this by modifying the low-noise simulated data to change the subject-course used to generate modality group 2's data, starting from the subject-courses being identical (correlation = 1) down to them being almost decorrelated (correlation = 0.1). For high correlations (>0.75), both methods combine the components; for low correlations (<0.35), both methods split them. However, Linked ICA splits these subject-courses much earlier as shown in Fig. 7(a).

Furthermore, Fig. 7(b) shows that this earlier splitting results in more accurate recovery of the modality group-2 subject-courses (C1′ to C3′). This indicates that the modalities are being modelled separately, rather than using one component to model a weighted average of the two subject-courses and creating a new "difference" component (e.g. with a subject-course C1′–C1) to soak up the variance caused by the mismatched subject-courses.

Analysis of real structural and diffusion data

To evaluate this method on real multimodal group data, several Linked ICA configurations (Fig. 8) were compared to Concatenated ICA in the task of extracting independent components from a structural and diffusion data set with 47 probable-Alzheimer's patients and 46 age-matched controls. Exploratory techniques can be used to find inter-subject variability and identify whether any of these are correlated with regressors of interest; the Linked ICA approach provides a way to perform this across multiple modalities in a data set. Both grey matter density and white matter integrity have previously been used as biomarkers for neurodegeneration. To assess these, structural and diffusion scans were collected for these

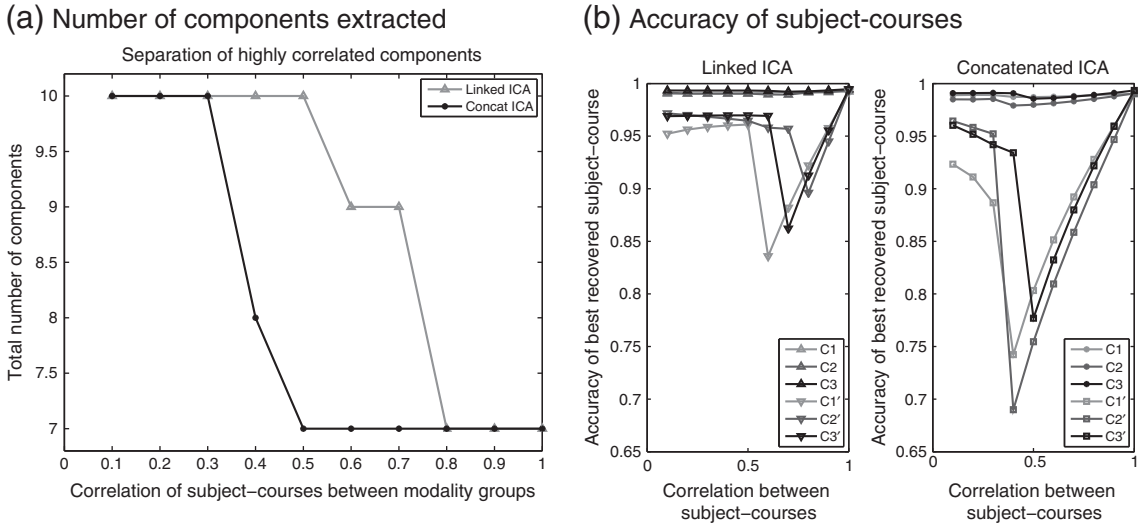


Fig. 7. The effect of progressively decorrelating the subject-courses between the two modality groups: (a) on the number of components extracted by each method, (b) on the accuracy of recovering these subject-courses. C1–C3 denote the subject-courses in modality group 1 of the original three common components, and C1′–C3′ are subject-courses of the similar components in modality group 2. At the right side of each figure (correlation between subject-courses = 1), C1–C3 are identical to C1′–C3′.

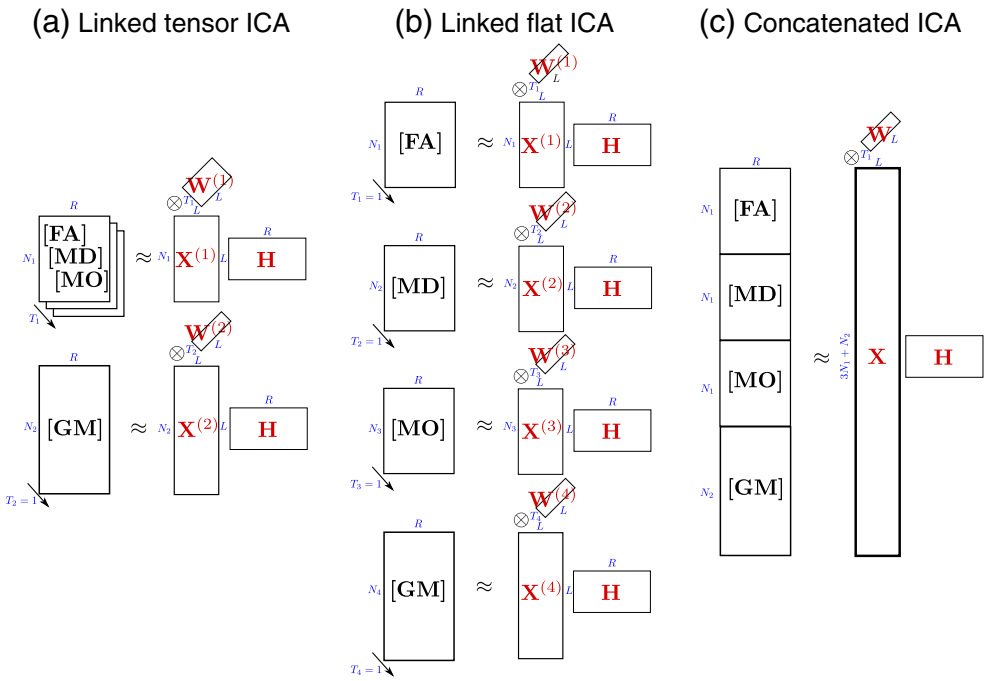


Fig. 8. Matrix diagrams of the ICA configurations evaluated for the real data: (a) Linked ICA with the three DTI modalities configured as a tensor and linked to GM, (b) Linked ICA in flat configuration, with all four modalities linked together by a shared subject-course matrix **H** only, (c) spatially Concatenation ICA, where all modalities share subject-courses **H**, scaling factors **W**, and the GMM source models on **X**.

subjects. The diffusion scans were preprocessed to extract maps of Fractional Anisotropy (FA), Mean Diffusivity (MD) and Tensor Mode (MO). These were projected onto a white matter skeleton using TBSS, which improves registration and makes sure that observed differences are due to white matter tract properties and not just movement or misregistration. The grey matter (GM) partial volume maps were extracted using the FSL-VBM tools (including non-linear registration).

Generally, neurodegeneration results in reduced GM density, decreased FA, and increased MD. Tensor mode MO is another measure derived from the diffusion tensor which is mathematically orthogonal to the other two (FA and MD) and is related to whether diffusion is restricted in a line or in a plane (Ennis and Kindlmann, 2006), and therefore may have significance for assessing degeneration in areas where fibre bundles cross.

The TBSS modalities were resampled to $(2\text{mm})^3$ voxels on the skeleton, yielding $N_1 = 28997$ voxels. The grey matter maps were already smoothed by 7.1mm FWHM to account for anatomical differences, and then subsampled by a factor of two to reduce computation time, yielding $N_2 = 23402$; nearest-neighbour down-sampling the GM to $4\times4\times4$ mm voxels does not discard much information in this case because of the high smoothing.

The degrees of freedom per voxel were estimated to be 0.017 for FA, 0.018 for MD, 0.022 for MO and 0.029 for GM; the first three were averaged yielding $\nu_1/N_1 = 0.019$ and $\nu_2/N_2 = 0.029$.

It is worth noting that these are completely different preprocessing steps, which are not explicitly matched in terms of voxel count, smoothness, or SNR. Instead, these parameters are estimated from the data and the Bayesian model uses this information to automatically weight the modalities appropriately.

Since all the DTI modalities exist in the same space, they can be combined in a single modality group to give a linked tensor model. They can also be linked in a flat formation, with different histograms and spatial maps for each modality. The concatenation approach is used as the baseline. These configurations are presented in Fig. 8.

The following models were evaluated:

- Linked tensor ICA is the Linked ICA model with the three DTI modalities stacked into a tensor, linked to the GM modality; so $K = 2$ modality groups and $T = [3, 1]$ modalities in each.
- Linked flat ICA is the linked ICA model which assumes unrelated spatial maps for each of the four modalities, with $K = 4$ modality groups and a single modality in each group ($T = [1, 1, 1, 1]$).
- Concatenated ICA is the standard spatial-concatenation model. This is different from Linked flat ICA because the same histogram is used for all modalities and there is no sparseness in modalities. Concatenated ICA also assumes the same noise variance for all modalities, but this is a reasonable assumption because the voxelwise noise-variance normalization is designed to leave all voxels (in all modalities) with the same noise level. As far as the model is concerned, there is only one modality ($K = 1$, $T = [1]$).

The VB inference was allowed to run until a fairly stringent $\Delta F < 0.1$ condition was satisfied (around 1000–5000 iterations). The current MATLAB implementation takes around 10–40 h of calculation for $L = 90$.

The precision contributions plots on the resulting fits of each model are shown in Fig. 9. The model comparison results also show that the tensor models are greatly preferred to the flat configuration, and that all Linked models are preferred to the concatenated approach. In Concatenated ICA, most of the components are spread across all of the modalities. In Linked ICA there is much more separation between the modalities: some components appear to explain variability in the white matter only, while some are shared between the white matter and grey matter. This may indicate that there is some variability between subjects in the white matter that is not reflected in grey matter, and vice versa; alternatively, some of these components may be artefacts present in the individual modalities. Both Linked tensor ICA and Linked flat ICA give sparser solutions (by excluding some modalities from some components) and therefore choose to have more components than the Concatenated ICA model. The Linked tensor ICA model has the most restrictive model of each component and therefore uses the most components. In this way, components that have correlated but distinct

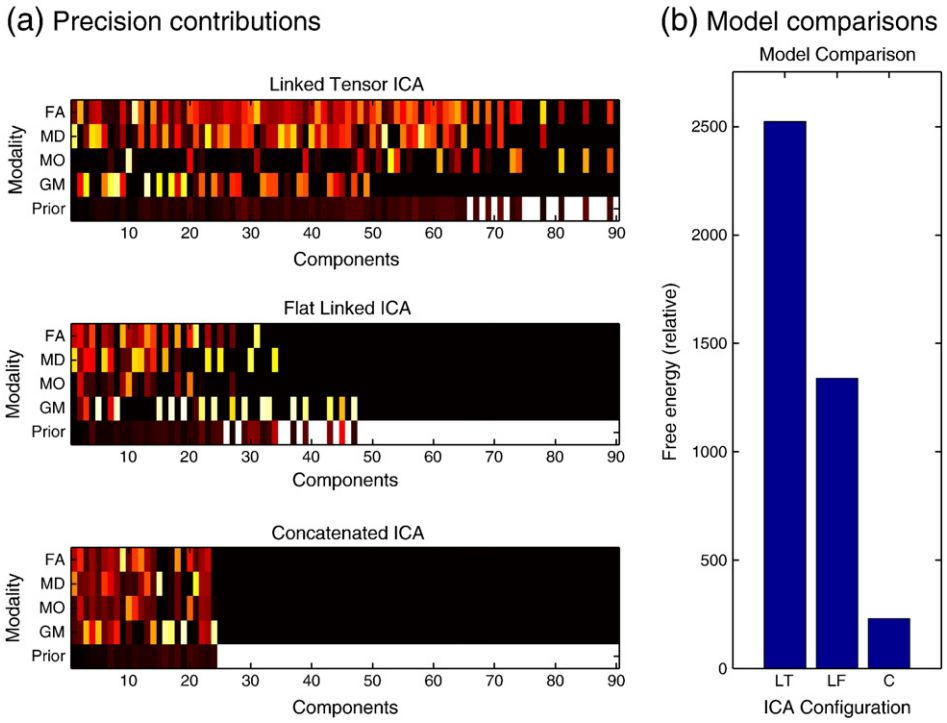


Fig. 9. Left: Precision contribution plots for the Alzheimer's data set; the brightness indicates the relative strength of the modalities' contributions to each component. Linked tensor ICA uses two modality groups: the three DTI modalities (FA, MD, and MO) form one group, while GM is by itself. Linked non-tensor ICA puts each of the four modalities in its own group, yielding 35 components overall (many of them explaining only a single modality). Spatial concatenation yields 24 components, with nearly all of them providing a mixture of signals from across all modalities. Right: Model comparison results for Linked tensor ICA (LT), Linked flat ICA (LF) and Concatenated ICA (C). The strongest evidence is for the LT model, followed by LF.

subject loadings in different modalities are more likely to be split because the Linked (flat and tensor) ICA models are provided with information about where the modality boundaries lie.

Figs. 10 and 11 shows a strong component that is well-preserved across all of these models, and is correlated with age and pathology. In Linked flat ICA (Fig.10a) this component involves all modalities. The widespread increase in MD and decrease in GM are consistent with neural degeneration and brain atrophy (the smaller areas of apparent increase in GM are on the edge of GM, hence are most likely indicative of imperfect alignments between the groups in this fairly high-atrophy dataset). The other DTI patterns are more complicated: in the corpus callosum and forceps major both FA and MO decrease, while both of these measures increase in the internal capsule, corona radiata, and superior longitudinal fasciculus. These areas of increasing FA and MO, for example, where the superior longitudinal fasciculus crosses the descending fibres, are probably due to degeneration of the "weaker" fibre in this crossing-fibre area, as investigated by Douaud et al. (in press). Both of these are also consistent with neurodegeneration, because these signals will decrease in single-fibre direction areas but will increase due to selective degeneration of one tract in areas of crossing fibres. Concatenated ICA (Fig.10b) reveals a similar pattern in this component; however, it is far less extensive in all modalities, which may partially be due to the use of a single, compromise source model for all modalities. Furthermore, MD both increases and decreases in places, which is less physiologically interpretable.

Note that the patterns present in FA and MO are very different from those seen in MD, so this pattern cannot be expressed as a single component in the tensor model. As a result, Linked tensor ICA appears to split this into three components, shown in Fig. 11.

Finally, Fig. 12 shows a component that is extracted very similarly by all three ICA methods. There are a number of components like this in the decomposition, automatically decomposing the white matter into bilateral pairs of tracts that vary across groups of subjects. This component isolates the external/extreme capsule, and in this component the tensor assumption is valid because all methods infer that FA and MD have nearly identical spatial maps (down to a scale factor). The Concatenated ICA does not isolate the tract as strongly and also shows scattered "related" changes in GM that appear to be spurious.

We also re-analysed the real data using a random initialization rather than the PCA decomposition. Free energy results showed that this produced a poorer fit to the data: the final free energy changed by −31 for Concatenated ICA, −110 for the Linked flatICA and −249 for the Linked tensor ICA. For context, a difference of 3 is usually considered strong evidence in favour of the higher-free-energy model over the lower-free-energy model. Looking at this in more detail, the inferred subject-courses and spatial maps can change considerably when the initialization changes. Of course the component orders and signs will also be completely different, so these are paired in a greedy way (the two components with the highest absolute correlation are paired first). The correlations of these resulting pairs are plotted in Fig. 13. In most components, the Concatenated approach is less sensitive to initialization (many correlations near 1), and the Linked ICA methods are considerably more sensitive. However, model comparison provides a good way to choose which initialization produces the best final result, and the size of the drop in free energy is related to how sensitive the results were to initialization. Indeed the model comparisons over different initializations show that the PCA initialization is the better one to use.

Discussion

The Linked ICA method presented in this paper provides a general, flexible way to perform ICA on multimodal data sets that allows components to be sparse in modalities and allows different noise levels and histograms for each modality group. This method also permits part-tensor configurations when the same spatial maps are expected across some of the modalities and allows model comparison.

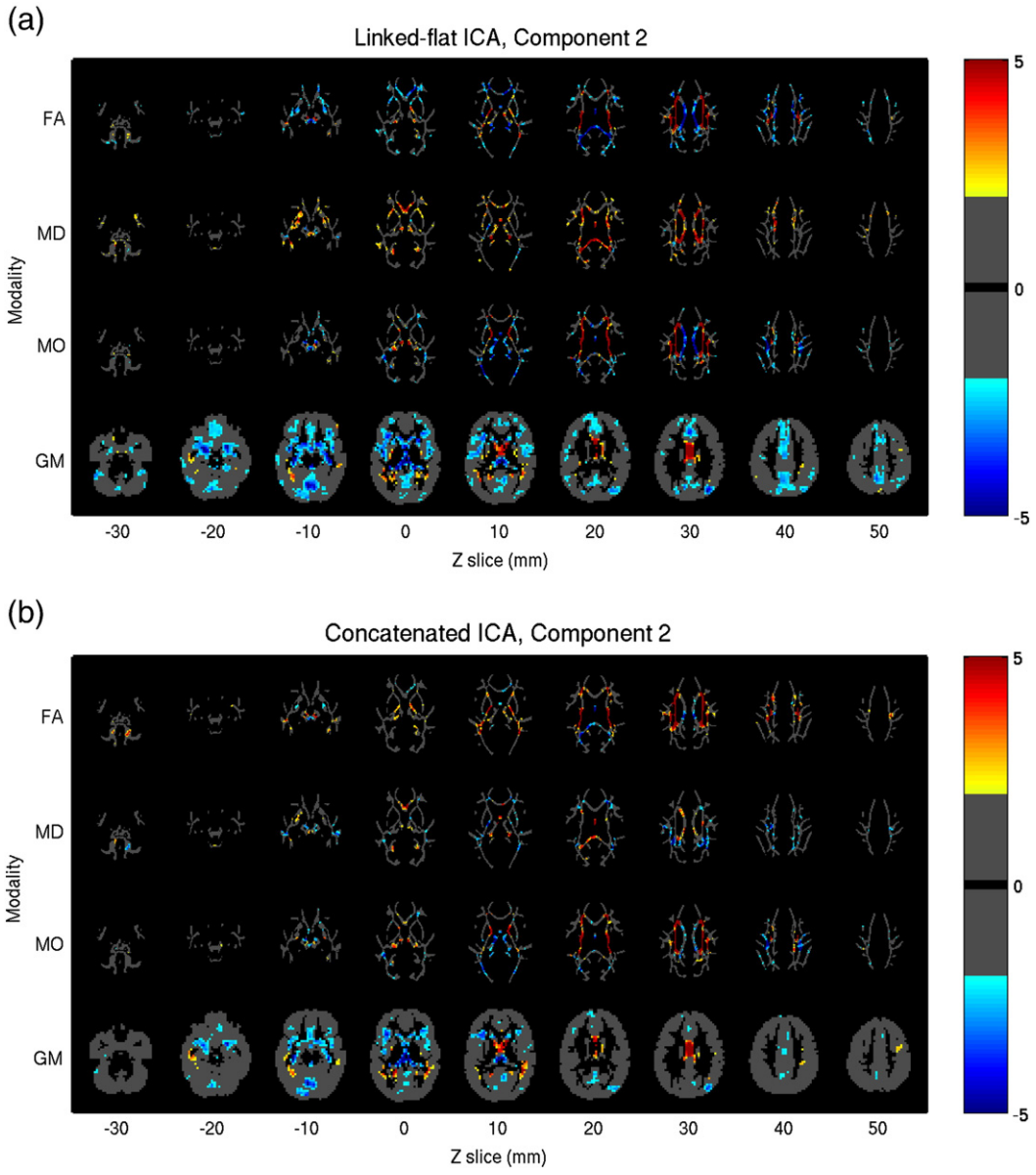


Fig. 10. Spatial maps of a strong component detected by all three models in the real multimodal data analysis: (a) Linked flat ICA and (b) Concatenated ICA are shown here, while Linked tensor ICA is shown in Fig. 11. The components shown are positively correlated with age ($r = 0.49$ for Linked flat ICA and 0.26 Concatenated ICA) and pathology (0.30 and 0.24 respectively). See the text for a full description of these components. Note that these images are thresholded at $|Z|>2$ for display.

Linked tensor ICA performed well in simulated data, and combined information from across modalities more accurately than standard Bayesian ICA with spatial concatenation. In particular, it was better at detecting and isolating single-modality noise, which may help to provide more interpretable components in real data. The linked tensor ICA method estimates the spatial maps more accurately (in terms of ROC performance), benefiting from the partial tensor configuration and from the more accurate histogram estimates found by separating voxels into modality groups.

A simple RESEL-based correction is developed to account for spatial smoothness in the data, and this does not interfere with the VB updates or model comparisons. On smoothed simulated data, we show the importance of correcting for smoothness to allow reasonable dimensionality estimation. The actual smoothness estimation is currently quite basic, being a point estimate derived from the data rather than the residuals. This could be estimated iteratively by examining the residuals, either after a preliminary analysis or as part of the main VB inference. The latter would be at the cost of losing direct model comparability because changing ν_k is effectively changing number of voxels in the data.

One alternative to this virtual decimation approach is to work directly on the unsmoothed data (thus making the unsmoothed noise model more valid) while using adaptive priors to encode the belief that nearby voxels should have similar levels of signal or are likely to have the same mixture-model labels (Woolrich et al., 2005; Hartvig and Jensen, 2000). This introduces additional complexity into the GMM part of the ICA model. Furthermore, one of the goals of this multimodal ICA is to be able to combine modalities that have been previously analysed in a way that is optimal for each modality; spatial smoothness is usually an intrinsic part of optimal preprocessing. Another alternative is to adaptively model the spatial smoothness in the data, much as the temporal smoothness in FMRI is estimated using autoregressive noise models (Woolrich et al., 2004; Roberts and Penny, 2002; Woolrich et al., 2001). Factorization across voxels in $P'(\mathbf{X})$ will probably also be required to achieve a tractable solution, losing some of the benefits of modelling this spatial correlation structure.

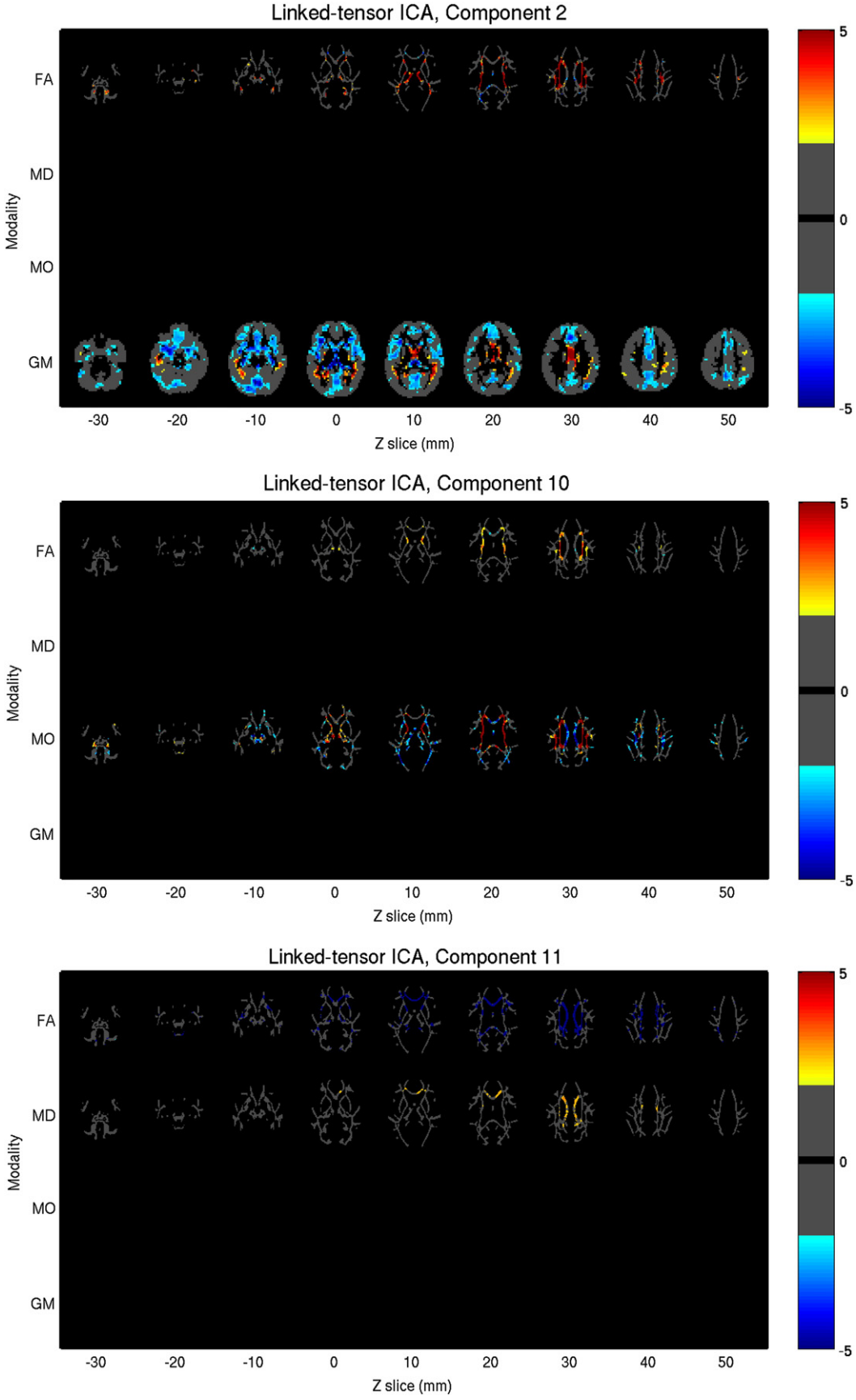


Fig. 11. In the Linked tensor ICA model, the component found in Fig. 10 cannot be expressed as a single component due to the tensor constraint. However, it can be clearly identified across the three components, which have subject-courses that are strongly (0.53–0.78) correlated with each other. These components each correlate positively with age (0.42, 0.51, and 0.58) and pathology (0.29, 0.28, 0.17). In each component, the FA, MD, and MO spatial maps are identical down to a scaling factor; any apparent differences are due to the effect of thresholding $|Z|>2$.

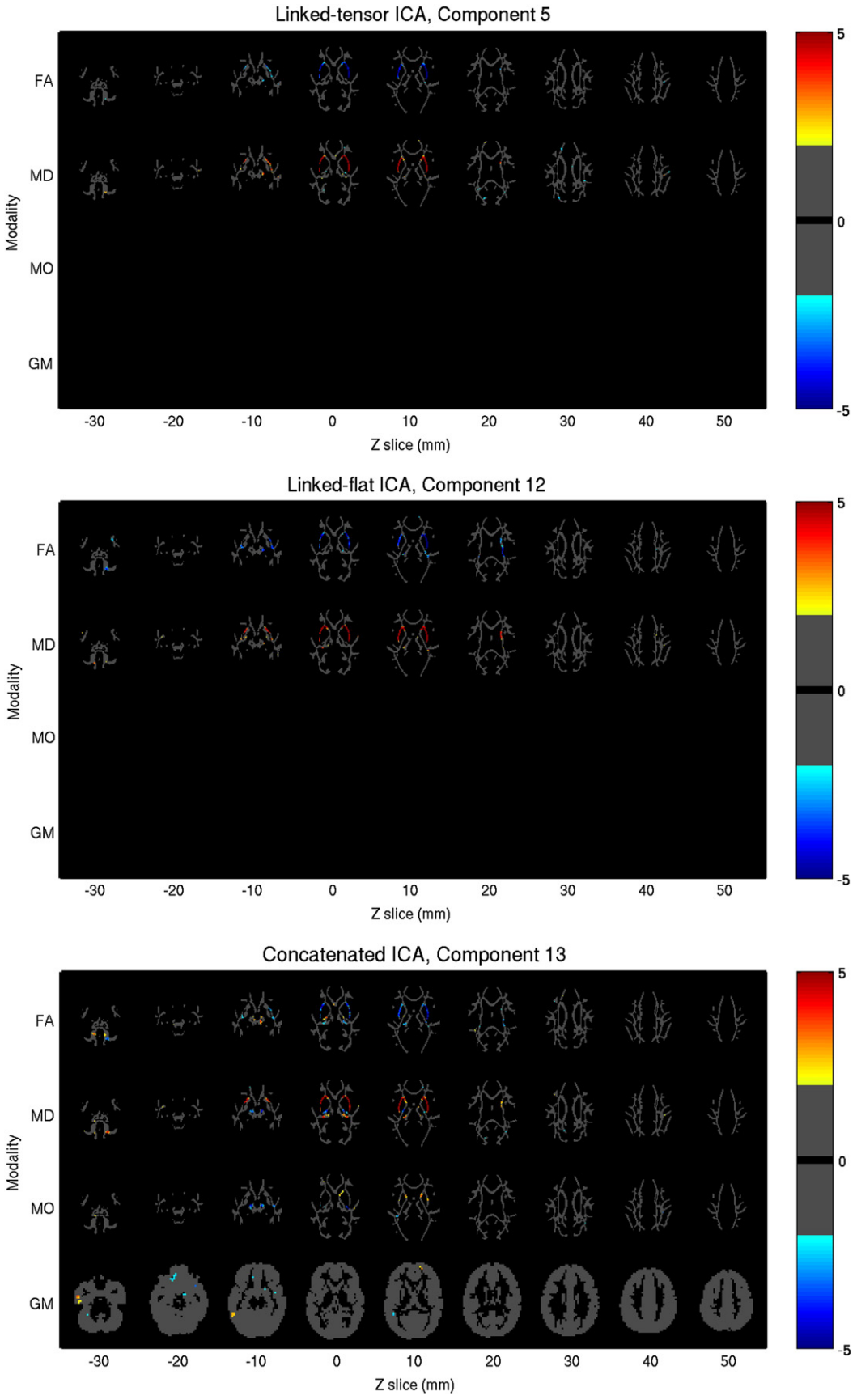


Fig. 12. Components from each of the 3 methods that most cleanly extract the external/extreme capsule. These subject-courses correlate positively with Age (0.23,0.39,0.32) but only very weakly with pathology ($r = 0.08,0.14,0.02$).

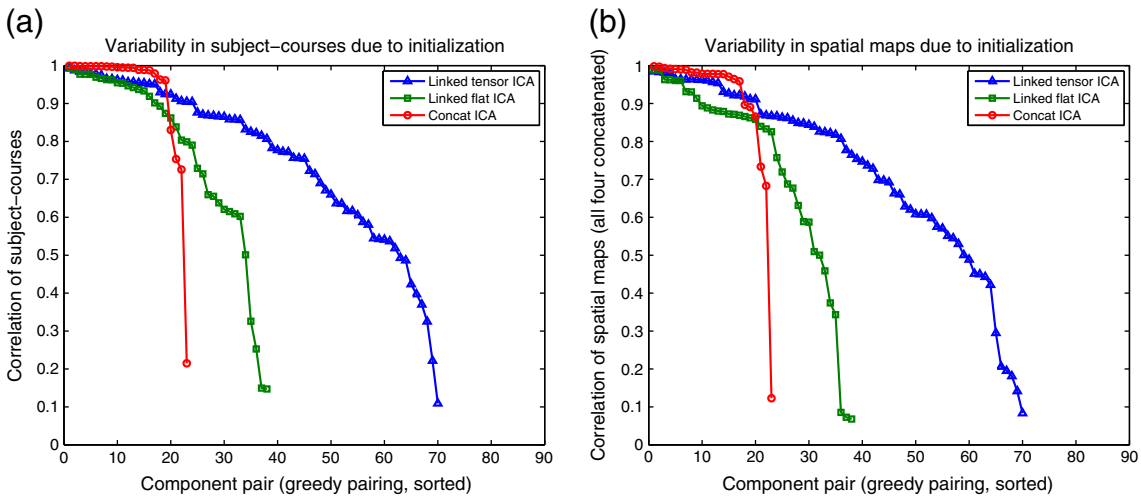


Fig. 13. These figures show the consistency of the inferred subject-courses (a) and spatial maps (b) under the two different initializations (PCA vs. random). Since the component order/sign is irrelevant, components from the two runs are greedily paired, matching up the highest absolute correlations first, and stopping when one run has no components remaining. Ideally, all components should be stable (i.e. all correlations near 1, so each component in one run is paired with a near-identical component in the other run). The proposed linked ICA methods are more sensitive to poor initialization than the Concatenated approach. However, model comparison provides a way to determine which is a better model of the data; in all cases PCA≫random, so only the PCA-initialized results have been given in this paper.

Model comparison showed that the Linked tensor model is preferred for the real multimodal data set because of the ability to re-use spatial maps, since there are some obvious similarities between the patterns observed in FA, MD, and MO. Linked flat ICA was still ranked higher than the Concatenated ICA model on the real data. On the real data set presented, the Flat configuration of Linked ICA produced more interpretable spatial maps than the Tensor configuration by avoiding the assumption of identical spatial patterns between the three DTI modalities. It therefore appears that Linked flat ICA is a more interpretable model for this data. However, this does depend on the question being asked. For example in Fig. 11, the tensor decomposition may be more meaningful in some situations because it attempts to split the DTI changes into several components based on each component having a fixed, proportional relationship between changes in FA, MD, and MO. Further investigation will be required, but this is potentially a meaningful distinction as it may actually be separating fibre regions in a microstructurally-relevant way, such as areas of single-direction vs. crossing-fibre composition. This separation of subtly-different types of change may be valuable in certain applications, for example if it more cleanly separates Alzheimer's-related from age-related differences; this remains a topic for future research.

While model comparison clearly preferred the tensor configuration, this may in part be because the assumption of spatially-similar maps was actually valid for most of the other components. It would be interesting to allow each component to individually decide whether its spatial maps are sufficiently similar to use a tensor configuration; this even more flexible variant would require a modified inference method.

Like any ICA method, Linked ICA relies on certain assumptions about the noise. It is assumed that the noise floor is the same in all voxels and across all subjects; the former should be guaranteed by the variance normalization preprocessing, and the latter effects are likely to be subtle, e.g. noise due to increased head motion in one group. The spatial smoothness of the noise is also assumed to be homogeneous and well-estimated, which may not be true due to the complex nature of noise in VBM or TBSS data. In fact, the voxelwise noise may not be Gaussian at all, and non-parametric statistics (permutation testing) are usually recommended for statistical testing on these modalities.

One important assumption of ICA is that the changes are strictly linear, and the models presented here make the further assumption that this linearity is valid across different modalities. This ignores any saturation effects, for example that FA must by definition be between 0 and 1.

ICA also assumes that the signals are encoded by patterns that vary in intensity, without moving spatially. In a grey matter map, the underlying change is often a volume reduction, which will cause strong effects at the edges that move for significant amounts of atrophy. Although the VBM protocol reduces this effect through the use of non-linear registration, Jacobian-based volume correction and spatial smoothing, it is still likely that the same type of change will show up in somewhat different ways for different levels of atrophy. It would likely be better to use surface-based measures (such as cortical thickness or volume) for ICA, as these might be expected to vary more linearly with the underlying biophysical changes.

There are also limitations in the interpretation of linked components. As demonstrated in Fig. 7, splitting occurs more often as the true subject-courses decorrelate, but this will depend on a number of factors such as noise level. The fact that a component is linked between two modalities does not actually mean that the relationship between the two modalities is significant, but rather it is simply a more concise explanation of the data.

The ICA methods presented here are all susceptible to finding local minima in real data, partially related to the ARD prior which is used to eliminate and part-eliminate components. Although the elimination is data-driven it is also essentially irreversible in practice, and many of these components are eliminated early on when the components are still very different from their final values. Based on the model comparison results, PCA is a much better initialization than random values, but it may still not find the global optimum. Conversely, on the simulated data sets the methods were robust against poor initialization, presumably because there were only seven signals to separate and they were strong independent sources. Improving the initialization and optimization of this method to avoid these local minima is a key direction for future work.

One possible alternative to the ARD-based elimination approach considered in this paper is to use a greedy search method (Friston et al., 2008) to build up the model one component at a time, stopping when the free energy starts to decrease. This could speed up computation by starting with a small number of components and growing, rather than starting with a large number and shrinking. It would also simplify initialization, because only one new component needs to be initialized at a time (either from a PCA decomposition of the residuals, or

randomly). This would allow the strongest components to become settled before increasing the model complexity, which may improve robustness. Unlike the deflation-based approach to fastICA (in which components are extracted sequentially), this would still allow the earlier components to change as new components are extracted and the model becomes more detailed.

There is also the possibility of applying these techniques to non-MRI neuroimaging data, for example by combining FMRI volumes and EEG epochs on a trial-by-trial (rather than subject-by-subject) basis. In MEG and EEG, tensorial decompositions like PARAFAC are a natural way to model the space × time × frequency information in single-subject data (Miwakeichi et al., 2004), and ICA has been used in this context to localise sources.

This framework only requires that the modalities share a single dimension (e.g. subjects or trials), so finding covarying patterns in modalities as different as EEG and FMRI may still be possible.

The major challenge will be in finding appropriate preprocessing methods to keep the data size down while extracting the relevant features.

Because of the Linked ICA's ability to automatically balance information from very different modalities, the same approach can be used to include non-imaging modalities like behavioural regressors or genetic data directly in a multimodal ICA. In practice, some components will be driven by these regressors, while others will model structured noise in the data ((Groves, 2010), Ch. 5).

Since only the single matrix **H** is shared, the generative models of these new modalities are extremely flexible in terms of source models and noise models. In this way, the Linked ICA framework can be extended to bridge the gap between data-driven unsupervised learning and supervised learning of multiple regressors simultaneously.

Acknowledgments

The authors would like to thank Achim Gass and Andreas Monsch for providing the structural and diffusion data, Gwenaëlle Douaud for assistance in interpreting the real data results and Salima Makni for many helpful discussions on Bayesian ICA.

A. Correcting for spatial smoothness

One of the complications of a fully Bayesian inference approach is that it requires a generative model of the data, not just of the signal. A consequence of this is that an inaccurate noise model often leads to biased inferences. Like standard Bayesian ICA, our model assumes uncorrelated white noise (Eq. (2)); however, in the case of neuroimaging data, there is often a very significant amount of spatial smoothness in the noise.

The ICA models are unaware of the spatial structure of the data, i.e. an image is simply a vector of voxels with no attached information about their position relative to one another. When presented with spatially-smoothed data, this spatial structure is learned as a consistent pattern across all images, and as a result many extra components can be inferred in order to model this spatial structure. However, we do not consider this structure to be interesting and wish to remove it, because these extra components obscure and interfere with the components based on the non-Gaussian signals of interest.

There are several approaches available. The most direct approach is to use an explicit model of spatially-smooth noise, e.g. using a Gaussian process as the noise model. This can adaptively determine the amount of smoothness in the noise and has the property of emphasizing sharp edges; this would be of great benefit in detecting sharp edges in the presence of spatially-smooth additive noise. However, this approach introduces a series of practical problems when combined with the ICA model (see discussion).

In neuroimaging, however, most of the smoothness is introduced to the raw signal intentionally. This is primarily because the signals tend to be extended (in the case of BOLD FMRI) or are not perfectly aligned across subjects (in the case of a VBM-style analysis). In these cases, spatially smoothing improves the SNR of the signals of interest. This smoothing of the data tends to reduce the effective degrees of freedom of the image. It is this mismatch between the number of voxels and the number of independent measurements that leads to incorrect estimation in the ICA.

For a given linear $N \times N$ transformation kernel $\mathbf{K}$, the effective degrees of freedom can be calculated exactly as

$$\nu = \mathrm{Tr}\left[\mathbf{K}\mathbf{K}^{\mathrm{T}}\right]^2 / \mathrm{Tr}\left[\mathbf{K}\mathbf{K}^{\mathrm{T}}\mathbf{K}\mathbf{K}^{\mathrm{T}}\right]. \tag{A.1}$$

Thus, for the identity kernel $\mathbf{K} = \mathbf{I}$, $\nu = \frac{N^2}{N} = N$. When $\mathbf{K}$ is a Gaussian smoothing kernel, a simple approximation to the effective DOF is given by (Worsley et al., 1995):

$$\frac{\nu}{N} = \frac{\mathrm{RESELS}}{N}\left(4\log(2)/\pi\right)^{D/2} = \left(\frac{0.9394}{\mathrm{FWHM}_x}\right)\left(\frac{0.9394}{\mathrm{FWHM}_y}\right)\left(\frac{0.9394}{\mathrm{FWHM}_z}\right) \tag{A.2}$$

where the last term is omitted in the case of 2-D data. The smoothing kernel's FWHM is estimated by assessing the correlation between adjacent voxels (in each direction) and

$$\left(\mathrm{FWHM}_x\right)^2 = -2\log(2) / \log(\mathrm{corr}_x) \tag{A.3}$$

where corr_x is the correlation between adjacent voxels in the x-direction.

One approach for reducing the impact of this RESEL/voxel mismatch is to decimate the data, reducing the number of voxels while retaining as many independent measurements as possible. There are practical problems with this approach, particularly in the choice of exactly which voxels to remove and the fact that some information will always be lost (and some correlations will always remain).

Instead, we opt for a virtual decimation approach, in which all N_k voxels in each modality group are retained, but anything that sums over voxels is downweighted by a factor of ν_k / N_k. This is analogous to fixing that only a random fraction of the data points will be kept, but at each stage averaging over all possible choices of decimated voxels. Surprisingly, a variant of the variational free energy F can also be retained in this model. For F to remain valid for model comparison, the ν_k for each modality must remain constant across all models to be compared; the same is also true of real decimation, where changing the decimation fraction would mean changing the number of voxels N_k and thus changing the data.

For the simulations on unsmoothed data (Results on unsmoothed simulated data section), no correction is needed (i.e. $\nu_k = N_k$). For smoothed data (simulated and real), effective degrees of freedom are estimated from the data using the expressions given above. Ideally, the smoothness would be estimated from the model residuals rather than the raw data, because spatially-extended signals will cause smoothness to be overestimated. However, the requirement to keep the F consistent across all approaches makes it impractical to re-estimate this during inference. The Evaluation and results section shows that on simulated data this estimate is reasonable and certainly much better than not correcting at all.

In practice, this adjustment acts much like a regularization weight in a non-Bayesian approach, adjusting the relative cost of accuracy versus complexity. However, this parameter is determined directly from the data before the analysis begins; in fact, if intentional Gaussian smoothing is the dominant form of image smoothness, it could even be obtained simply by inspecting the preprocessing procedures used and applying Eq. (A.1). Thus this is a constant and does not need to be tuned using empirical approaches (such as cross-validation). It is very similar in principle to the correction used in the dimensionality estimation of PICA; however, in that case the output of

the estimator is a hard dimensionality estimate rather than a fixed reweighting constant.

B. Initialization

As with most ICA implementations, Linked ICA is initialized from a PCA decomposition. For multimodal data, the natural method is to concatenate all of the voxels across modalities k, t (to get a $(\sum_{k=1}^{K} N_k T_k) \times R$ matrix) and then do a PCA decomposition on that.

To avoid being overly biased by data scaling, the PCA is done after the data has been de-meaned and variance-normalized. In the current implementation the ν_k/N_k smoothness correction is not taken into account for the PCA, so smoother modalities will tend to be overemphasized in the initial decomposition.

This initial bias is easily undone by the first few VB iterations and is not present in the final ICA components.

Taking the first L components of the decomposition, the loading matrix provides an initial point estimate for the shared subject-course matrix $\mathbf{H}$ and eigenvectors are used to initialize the spatial maps $\mathbf{X}$ as described below. However, this is not a hard dimensionality-reduction step, and the subject-courses are free to take combination of weights in the full R-subject space. Furthermore, we ensure that L is sufficiently large so that there are always extra components (which will be eliminated automatically by the ARD hyperparameters ω).

In the Bayesian ICA model, L is simply the maximum number of components that are allowed. If the number of distinct signals in the data is greater than L, this means that some of them will not be represented and this energy will end up in the white noise term. The upper limit is $L=R-2$, to avoid the degenerate solution where the de-meaned data is perfectly represented by components and there is no residual white noise. The model will automatically eliminate unneeded components using ARD, so the exact number is not particularly important. We chose L close to this upper limit, using $L=90$ for most examples (where $R=93$ or 100). The first examples use the greatly reduced (but still sufficient) $L=10$ to make the figures clearer, but this has little impact on the results.

The initialization of $\mathbf{X}^{(k)}$ depends on whether modality group k is tensor ($T_k>1$) or non-tensor ($T_k=1$). If modality group k is non-tensor then this PCA also yields initial point estimates of $\mathbf{X}^{(k)}$ and the weights $\mathbf{W}^{(k)}$ are initialized to 1. For tensor modality groups, the PCA decomposition yields a $N_k T_k \times 1$ vector. Reversing the original concatenation step reshapes this into one $N_k \times T_k$ matrix per component, but in general this will yield a different spatial map for each modality, while the tensor model demands that each component's spatial map be identical for all modalities (aside from scaling). This $N_k \times T_k$ matrix is approximated by the bilinear matrix $\mathbf{X}^{(k)}_{:,i}\mathbf{W}^{(k)T}_{:,i}$, using another PCA to find the best approximation. This provides a starting point which the Bayesian method can improve upon by looking for independent components.

Once the $\mathbf{X}^{(k)}$ matrices are found, each component's mixture model is initialized by locating means $\mu^{(k)}_{:,i}$ at the 25th, 50th, and 75th percentiles of the spatial map intensities, and setting the standard deviation $\beta^{(k)^{-\frac{1}{2}}}_{:,i}$ equal to half of the spacing between the means.

C. Free energy

In the VB framework, the free energy F is a commonly-used measure for model comparison. It is given by

$$F = \left\langle \log\left(\frac{P(\mathbf{Y}|\Theta)P(\Theta)}{P'(\Theta)}\right)\right\rangle_{P'(\Theta)}, \tag{C.1}$$

where Θ is the set of all model parameters, and the choice of model itself is implicit in each of the $P(\cdot)$ expressions. It can be shown that F is a lower bound on the model evidence $\log P(Y)$, and that they become identical when the factorized posterior matches the real posterior, i.e. $P'(\Theta) = P(\Theta|Y)$. The overall goal of the VB framework is to find the factorized posterior $P'(\Theta)$ that maximizes F for a given model, usually by updating the posterior factors one at a time.

In order to make the free energy consistent with spatial-smoothness adjustments, the $\frac{\nu_k}{N_k}$ correction also needs to be made whenever there is a sum over N_k. This effectively reduces the number of data points in the relevant accuracy and complexity terms of the expression, without actually losing data; this is equivalent to deciding to only use a fixed subset of the data points but not deciding which, and building a single model that fits all possible subsets simultaneously. Since this virtual "decimation factor" remains consistent across all models, model comparison using F is still valid. The (smoothness-corrected) free energy is given by

$$\begin{aligned} F = {} & \sum_k \left[\frac{\nu_k}{N_k}\left\langle \mathrm{P}\left(\mathbf{Y}^{(k)}|\mathbf{X}^{(k)},\mathbf{W}^{(k)},\mathbf{H},\lambda^{(k)}\right)\right\rangle - KL\left(\lambda^{(k)}\right) - KL\left(\omega^{(k)}\right) - KL\left(\mathbf{W}^{(k)}\right)\right] - KL(\mathbf{H}) \\ & + \sum_k \sum_i \left[-KL\left(\mu^{(k,i)}\right) - KL\left(\beta^{(k,i)}\right) - KL\left(\pi^{(k,i)}\right) - \frac{\nu_k}{N_k} KL\left(\mathbf{q}^{(k,i)}\right) - \frac{\nu_k}{N_k} KL\left(\mathbf{X}^{(k,i)}\right)\right] \end{aligned} \tag{C.2}$$

where $KL(\cdot)$ is shorthand for the complexity cost of the specified variable, defined in terms of the K–L divergence of the posterior from the prior

$$KL(z) = KL(P'(z)\,||\,P(z)) = \langle \log P'(z)\rangle_{P'(z)} - \langle \log P(z)\rangle_{P'(z)} \tag{C.3}$$

and the kth likelihood term expands to

$$\begin{aligned} \left\langle \mathrm{P}\left(\mathbf{Y}^{(k)}|\mathbf{X}^{(k)},\mathbf{W}^{(k)},\mathbf{H},\lambda^{(k)}\right)\right\rangle = {} & \sum_{t=1}^{T_k}\left[\frac{N_k R}{2}\left\langle \log\left(\frac{\lambda_t}{2\pi}\right)\right\rangle - \frac{\langle\lambda_t\rangle}{2}\sum_{n=1}^{N_k}\sum_{r=1}^{R}\mathbf{Y}^2_{ntr}\right. \\ & + \langle\lambda_t\rangle \sum_{n=1}^{N_k}\sum_{r=1}^{R}\left(\sum_{i=1}^{L}\left\langle \mathbf{X}^{(k,i)}_n\right\rangle\left\langle \mathbf{W}^{(k)}_{t,i}\right\rangle\langle \mathbf{H}_{ir}\rangle\right)\mathbf{Y}_{ntr} \\ & \left. - \frac{\langle\lambda_t\rangle}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left\langle \mathbf{X}^T\mathbf{X}\right\rangle_{ij}\left\langle \mathbf{W}_{ti}\mathbf{W}_{tj}\right\rangle\left\langle \mathbf{H}\mathbf{H}^T\right\rangle_{ij}\right] \end{aligned} \tag{C.4}$$

and the KL divergence on the Gaussian mixture model expressed as

$$KL\left(\mathbf{X}^{(k,i)}\right) = \sum_{n=1}^{N_k}\sum_{m=1}^{M} P'\left(q_n^{(k,i)}=m\right) KL\left(P'\left(\mathbf{X}_n^{(k,i)}|q_n^{(k,i)}=m\right)\Big\|P\left(\mathbf{X}_n^{(k,i)}|q_n^{(k,i)}=m\right)\right) \tag{C.5}$$

where both the conditional prior and conditional posterior are normal distributions.

D. Priors and VB updates

Shared latent space (subject-course) matrix $\mathbf{H}$

The VB updates are found in a similar way to F in Eq. (C.1), but integrating over all factors of the posterior apart from the one being updated:

$$\log P'(\mathbf{H}) = \left\langle \log\left(\frac{P(\mathbf{Y}|\Theta)P(\Theta)}{P'(\neg\mathbf{H})}\right)\right\rangle_{P'(\neg\mathbf{H})} + const, \tag{D.1}$$

where $\neg\mathbf{H}$ is the set of all random variables in Θ except $\mathbf{H}$. The function corresponding to $\mathrm{P}'(\mathbf{H})$ is found by evaluating the right hand side algebraically and matching all terms of $\mathbf{H}$. Many of the terms do not involve $\mathbf{H}$ so in practice (and with the inclusion of the DOF/voxel correction factor as in Eq. (C.2)) this becomes

$$\log P'(\mathbf{H}) = \log P(\mathbf{H}) + \sum_{k=1}^{K}\frac{\nu_k}{N_k}\left\langle \log P(\mathbf{Y}^{(k)}|\mathbf{H},\mathbf{X}^{(k)},\mathbf{W}^{(k)},\lambda^{(k)})\right\rangle + const. \tag{D.2}$$

It turns out that the posterior on $\mathbf{H}$ is a matrix normal distribution:

$$P'(\mathbf{H}) = MN(\mathbf{H}|\mathbf{M}_{\mathbf{H}}, \Omega_{\mathbf{H}}, \sum_{\mathbf{H}}). \tag{D.3}$$

Since all subjects r have the same noise levels and the same prior, the posterior row covariance $\Omega_{\mathbf{H}} = \mathbf{I}_R$. The $L \times L$ column covariance is given by

$$\left(\sum_{\mathbf{H}}^{-1}\right)_{ij} = \left(\sum_{\mathbf{H},0}^{-1}\right)_{ij} + \sum_{k=1}^{K}\left(\frac{\nu_k}{N_k}\sum_{n=1}^{N_k}\left\langle \mathbf{X}_{ni}^{(k)}\mathbf{X}_{nj}^{(k)}\right\rangle\right)\left(\sum_{t=1}^{T_k}\left\langle \mathbf{W}_{ti}^{(k)}\mathbf{W}_{tj}^{(k)}\right\rangle\lambda_t^{(k)}\right) \tag{D.4}$$

while the posterior mean is

$$\left(\mathbf{M}_{\mathbf{H}}\sum_{\mathbf{H}}^{-1}\right)_{ir} = \left(\mathbf{M}_{\mathbf{H},0}\sum_{\mathbf{H},0}^{-1}\right)_{ir} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\left\langle \lambda_t^{(k)}\right\rangle\left\langle \mathbf{W}_{ti}^{(k)}\right\rangle\frac{\nu_k}{N_k}\sum_{n=1}^{N_k}\mathbf{Y}_{n,r,t}^{(k)}\left\langle \mathbf{X}_{n,i}^{(k)}\right\rangle. \tag{D.5}$$

The priors are simply $N(0,1)$ on each element, i.e. $\sum_{\mathbf{H},0} = \mathbf{I}_L$, $\Omega_{\mathbf{H},0} = \mathbf{I}_R$ and $\mathbf{M}_{\mathbf{H},0} = 0$. For MATLAB implementation, these updates are more efficiently expressed in matrix form:

$$\sum_{\mathbf{H}}^{-1} = \mathbf{I}_L + \sum_{k=1}^{K}\left\langle \mathbf{X}^{(k)T}\mathbf{X}^{(k)}\right\rangle \circ \left\langle \mathbf{W}^{(k)T}\,\mathrm{diag}\left\langle \lambda^{(k)}\right\rangle \mathbf{W}^{(k)}\right\rangle \tag{D.6}$$

$$\mathbf{M}_{\mathbf{H}}\sum_{\mathbf{H}}^{-1} = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\mathbf{Y}_{:,:,t}^{(k)T}\left\langle \mathbf{X}^{(k)}\right\rangle \mathrm{diag}\left(\lambda_t^{(k)}\mathbf{W}_{t,:}^{(k)T}\right) \tag{D.7}$$

where $\circ$ represents the elementwise (Schur) product of two matrices.

Noise precision λ

When separate noise is estimated for each modality (k, t), the posterior distribution of the noise precision is found by calculating

$$\log P'(\lambda) = \log P(\lambda) + \sum_{k=1}^{K}\frac{\nu_k}{N_k}\left\langle \log P\left(\mathbf{Y}^{(k)}|\mathbf{H},\mathbf{X}^{(k)},\mathbf{W}^{(k)},\lambda^{(k)}\right)\right\rangle_{P'(\neg\lambda^{(k)})} + const \tag{D.8}$$

which yields

$$P'(\lambda) = \prod_{k=1}^{K}\prod_{t=1}^{T} P'\left(\lambda_t^{(k)}\right) \tag{D.9}$$

$$P'\left(\lambda_t^{(k)}\right) = Ga\left(\lambda_t^{(k)}|b_{kt},c_{kt}\right) \tag{D.10}$$

$$c_{kt} = c_0 + \nu_k R/2 \tag{D.11}$$

$$b_{kt}^{-1} = b_0^{-1} + \frac{1}{2}\frac{\nu_k}{N_k}\sum_{n=1}^{N_k}\sum_{r=1}^{R}\mathbf{Y}_{ntr}^{(k)2} - \frac{\nu_k}{N_k}\sum_{n=1}^{N_k}\sum_{r=1}^{R}\left(\mathbf{Y}_{ntr}^{(k)}\sum_{i=1}^{L}\left\langle \mathbf{X}_{ni}^{(k)}\right\rangle\left\langle \mathbf{W}_{ti}^{(k)}\right\rangle\langle \mathbf{H}_{ir}\rangle\right) + \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left[\left(\frac{\nu_k}{N_k}\sum_{n=1}^{N_k}\left\langle \mathbf{X}_{ni}^{(k)}\mathbf{X}_{nj}^{(k)}\right\rangle\right)\left\langle \mathbf{W}_{ti}^{(k)}\mathbf{W}_{tj}^{(k)}\right\rangle\left(\sum_{r=1}^{R}\left\langle \mathbf{H}_{ir}\mathbf{H}_{jr}\right\rangle\right)\right]. \tag{D.12}$$

Spatial sources $\mathbf{X}$ *and hidden mixture memberships* $\mathbf{q}$

We explicitly model the hidden mixture labels using the categorical variable $q_n^{(k,i)} = m$, which specifies that voxel n (in modality group (k, t) and in ICA component i) is drawn from the mth mixture component. The notation $\mathbf{q}_{:,n}^{(k,i)}$ is also used as the $M \times 1$ vector of all zeros except for a single one in row $q_n^{(k,i)}$. This is convenient because it means that $\left\langle \mathbf{q}_{:,n}^{(k,i)}\right\rangle$ gives the posterior probability of a voxel being drawn from each mixture component.

By conditioning on $q_n^{(k,i)}$ the GMM prior (Eq. (5)) can be rewritten as

$$\left(\mathbf{X}_{n,i}^{(k)}|q_n^{(k,i)}\right) \sim N\left(\mu_{i,q_n^{(k,i)}}^{(k)}, 1/\beta_{i,q_n^{(k,i)}}^{(k)}\right), \qquad \mathbf{q}_{:,n}^{(k,i)} \sim Cat\left(\pi^{(k,i)}\right) \tag{D.13}$$

where the categorical distribution $Cat(\pi)$ is defined by the probability mass distribution

$$Cat(\mathbf{q}|\pi) = \mathbf{q}^{\mathrm{T}}\pi \tag{D.14}$$

assuming that $\mathbf{q}$ has a single element equal to 1 and the rest of the elements equal 0. In other words, $P(q = m|\pi) = \pi_m$.

The posterior distribution is also a Gaussian mixture model, factorized over components:

$$P'(\mathbf{X}) = \prod_{k=1}^{K}\prod_{i=1}^{L} P'\left(\mathbf{X}^{(k,i)}\right) \tag{D.15}$$

$$P'\left(\mathbf{X}^{(k,i)}\right) = \sum_{m=1}^{M} P'\left(\mathbf{X}^{(k,i)}|q_n^{(k,i)} = m\right)P'\left(q_n^{(k,i)} = m\right). \tag{D.16}$$

Note that the Gaussian mixture model prior naturally factors itself into two parts: the posterior distributions conditional on the label, and the mixture labels. However, these remain tightly connected; in particular, the update $P'(\mathbf{q})$ depends on $\mathbf{Y}$, despite $\mathbf{q}$ and $\mathbf{Y}$ not being directly connected in the graphical model. These are therefore treated as a single monolithic update during VB inference.

The conditional posterior distributions are Gaussian, shown below. A more complete derivation can be found in (Choudrey and Roberts, 2001).

$$P'\left(\mathbf{X}^{(k,i)}|q_n^{(k,i)} = m\right) = N\left(\mathbf{X}_n^{(k,i)}|\mathbf{M}_{\mathbf{X},n,m},\sum_{\mathbf{X},n,m}\right) \tag{D.17}$$

$$\sum_{\mathbf{X},n,m}^{-1} = \left\langle \beta_{i,m}^{(k)}\right\rangle + \left(\left\langle \sum_{r=1}^{R}\mathbf{H}_{ir}^2\right\rangle\sum_t\langle\lambda_t\rangle\left\langle \mathbf{W}_{ti}^2\right\rangle\right) \tag{D.18}$$

$$\mathbf{M}_{\mathbf{X},n,m}\sum_{\mathbf{X},n,m}^{-1} = \left\langle \mu_{i,m}^{(k)}\right\rangle\left\langle \beta_{i,m}^{(k)}\right\rangle + \sum_t\left(\left\langle \lambda_t^{(k)}\right\rangle\left\langle \mathbf{W}_{t,i}\right\rangle\sum_{r=1}^{R}\mathbf{Y}_{n,t,r}\left\langle \mathbf{H}_{i,r}\right\rangle - \left\langle \lambda_t^{(k)}\right\rangle\sum_{j\neq i}\left\langle \mathbf{X}_{nj}\right\rangle\left\langle \mathbf{W}_{ti}\mathbf{W}_{tj}\right\rangle\left\langle \mathbf{H}_{ir}\mathbf{H}_{jr}\right\rangle\right) \tag{D.19}$$

while the mixture labels are distributed as

$$P'\left(q_n^{(k,i)}\right) = Cat\left(q_n^{(k,i)}|\mathbf{Q}/\sum_m\mathbf{Q}_m\right) \tag{D.20}$$

$$\log(\mathbf{Q}_m) = \left\langle \log\pi_{i,m}^{(k)}\right\rangle + \frac{1}{2}\left\langle \log\beta_{i,m}^{(k)}\right\rangle - \frac{1}{2}\left\langle \beta_{i,m}^{(k)}\right\rangle\left\langle \mu_{i,m}^{(k)}2\right\rangle - \frac{1}{2}\left\langle \log\sum_{\mathbf{X},n,m}^{-1}\right\rangle + \frac{1}{2}\left(\mathbf{M}_{\mathbf{X},n,m}\right)^2\sum_{\mathbf{X},n,m}^{-1}. \tag{D.21}$$

Mixture component means μ *and precisions* β

The priors are given by

$$P\left(\mu^{(k,i)}\right) = \prod_{m=1}^{M_{k,i}} N\left(\mu_m^{(k,i)}|u_0,v_0\right) \tag{D.22}$$

$$P\left(\beta^{(k,i)}\right) = \prod_{m=1}^{M_{k,i}} Ga\left(\beta_m^{(k,i)}|b_0,c_0\right) \tag{D.23}$$

with the relatively uninformative priors $u_0=0$, $v_0=10^6$, $b_0=10^3$, $c_0=10^{-6}$, and the Gamma distribution in terms of the Gamma function:

$$Ga(x|b,c) = \frac{x^{(c-1)}e^{-x/b}}{\Gamma(c)b^c}. \tag{D.24}$$

The posterior forms are given by

$$P'\left(\mu_m^{(k,i)}\right) = N(u,v) \tag{D.25}$$

$$v^{-1} = v_0^{-1} + \left\langle \beta_m^{(k,i)} \right\rangle \frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left\langle \mathbf{q}_{m,n}^{(k,i)} \right\rangle \tag{D.26}$$

$$uv^{-1} = u_0 v_0^{-1} + \left\langle \beta_m^{(k,i)} \right\rangle \frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left\langle \mathbf{X}_n^{(k,i)} | \mathbf{q}_{m,n}^{(k,i)} = 1 \right\rangle \left\langle \mathbf{q}_{m,n}^{(k,i)} \right\rangle \tag{D.27}$$

$$P'\left(\beta_m^{(k,i)}\right) = Ga(b,c) \tag{D.28}$$

$$\begin{aligned} b^{-1} = b_0^{-1} + \frac{1}{2}\frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left\langle \mathbf{q}_{m,n}^{(k,i)} \right\rangle \Bigg(& \left\langle (\mathbf{X}_{m,n}^{(k,i)})^2 | \mathbf{q}_{m,n}^{(k,i)} = 1 \right\rangle \\ & -2\left\langle \mathbf{X}_{m,n}^{(k,i)} | \mathbf{q}_{m,n}^{(k,i)} = 1 \right\rangle \left\langle \mu_m^{(k,i)} \right\rangle + \left\langle (\mu_m^{(k,i)})^2 \right\rangle \Bigg) \end{aligned} \tag{D.29}$$

$$c = c_0 + \frac{1}{2}\frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left\langle \mathbf{q}_{m,n}^{(k,i)} \right\rangle. \tag{D.30}$$

Mixture model weights π

The priors are given as follows:

$$P(\pi) = \prod_{k=1}^{K} \prod_{i=1}^{L} Dir\left(\pi^{(k,i)} | \mathcal{L}_0\right) \tag{D.31}$$

where the uniform prior ($\mathcal{L}_0 \in R^M$ is a vector of all ones) was used. The Dirichlet distribution is defined as

$$Dir(Q|\pi) \propto \sum_{m=1}^{M} (Q_m)^{\pi_m - 1}. \tag{D.32}$$

The VB updates are given by:

$$P'(\pi) = \prod_{k=1}^{K} \prod_{i=1}^{L} Dir\left(\pi^{(k,i)} | \mathcal{L}^{(k,i)}\right) \tag{D.33}$$

$$\mathcal{L}^{(k,i)} = \frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left\langle q_n^{(k,i)} \right\rangle + \pi_0. \tag{D.34}$$

Modality weight matrix $\mathbf{W}$ *and ARD prior* ω

The posterior distribution of $\mathbf{W}$ is found by expanding

$$\begin{aligned} \log P'(\mathbf{W}) = & \langle \log P(\mathbf{W}|\omega) \rangle_{P'(\omega)} \\ & + \sum_{k=1}^{K} \frac{\nu_k}{N_k} \left\langle \log P\left(\mathbf{Y}^{(k)} | \mathbf{H}, \mathbf{X}^{(k)}, \mathbf{W}^{(k)}, \lambda^{(k)}\right) \right\rangle_{P'(\neg \mathbf{W}^{(k)})} + const. \end{aligned} \tag{D.35}$$

The posterior on $\mathbf{W}$ naturally factors across modalities:

$$P'(\mathbf{W}) = \prod_{k=1}^{K} \prod_{t=1}^{T_k} P'\left(\mathbf{W}_{t,\cdot}\right) \tag{D.36}$$

each of which is a normal distribution given by

$$P'\left(\mathbf{W}_{t,\cdot}\right) = \mathrm{N}\left(\mathbf{W}_{t,\cdot} | \mathbf{m}, \mathbf{V}\right) \tag{D.37}$$

$$\mathbf{V}^{-1} = \langle \omega_t \rangle \mathbf{I} + \langle \lambda_t \rangle \frac{\nu_k}{N_k} \left\langle \mathbf{X}^T \mathbf{X} \right\rangle \circ \left\langle \mathbf{H}\mathbf{H}^T \right\rangle \tag{D.38}$$

$$(\mathbf{m}\mathbf{V}^{-1})_i = \langle \lambda_t \rangle \sum_{r=1}^{R} \langle \mathbf{H}_{ir} \rangle \frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \mathbf{Y}_{nrt} \langle \mathbf{X}_{ni} \rangle. \tag{D.39}$$

The posterior on the ARD parameter ω naturally factorizes as

$$P'(\omega) = \prod_{k=1}^{K} \prod_{t=1}^{T_k} \prod_{i=1}^{L} P'\left(\omega_{ti}^{(k)}\right) \tag{D.40}$$

with each element of ω distributed as

$$P'\left(\omega_{ti}^{(k)}\right) = Ga\left(\omega_t^{(k)} | b, c\right) \tag{D.41}$$

$$b^{-1} = b_0^{-1} + \left\langle (\mathbf{W}_{ti}^{(k)})^2 \right\rangle \tag{D.42}$$

$$c = c_0 + 1/2. \tag{D.43}$$

E. Precision contributions

The VB update for $P'(\mathbf{H})$ is given by Eqs. (D.4) and (D.5). Ignoring the off-diagonal elements of $\sum_{\mathbf{H}}^{-1}$ (which should be small because the spatial maps are independent), this formulation means that each modality provides its own ideal (likelihood-maximizing) estimate of $\mathbf{H}$ and the posterior $\langle \mathbf{H} \rangle$ is a precision-weighted average of these. Conveniently, this precision is the same for each subject r. To find the dominant modalities in estimating each source's subject-course, it is informative to look at these precisions. The "precision contribution" by modality t in modality group k to each source i is defined by looking at the parts of the sum in Eq. (D.4):

$$pc(k,t,i) = \left\langle \frac{\nu_k}{N_k} \sum_{n=1}^{N_k} \left(\mathbf{X}_{n,i}^{(k)}\right)^2 \right\rangle \left\langle \left(\mathbf{W}_{t,i}^{(k)}\right)^2 \right\rangle \left\langle \lambda_t^{(k)} \right\rangle. \tag{E.1}$$

So overall the precision of $P'(\mathbf{H}_{i,\cdot})$ is given by $1 + \sum_{k=1}^{K} \times \sum_{t=1}^{T_k} pc(k,t,i)$, because the prior makes a constant precision contribution of 1. This provides a fixed scale against which to measure these contributions; if $pc(k,t,i) < 1$ then that modality is considered to have been eliminated from that component.

In the Evaluation and results section the figures will show these precision contributions normalized by the overall precision, so that the sum of all contributions is 1 for each component i. This makes it easy to see if a component is dominated by one modality or is informed by a combination of several modalities.

For comparisons, the precision contributions from individual modalities can also be calculated for the Concatenated ICA results by using only the relevant voxels in the spatial maps:

$$pc(\hat{k},\hat{t},i) = \frac{\nu_{\hat{k}}}{N_{\hat{k}}} \sum_{n \in \mathbf{N}^{(\hat{k},\hat{t})}} \left\langle \mathbf{X}_{n,i}^2 \right\rangle \left\langle \mathbf{W}_i^2 \right\rangle \langle \lambda \rangle \tag{E.2}$$

where $\mathbf{N}^{(\hat{k},\hat{t})}$ is the set of voxels relating to modality $k = \hat{k}, t = \hat{t}$ in the Linked ICA.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry — the methods. Neuroimage 110 (6), 805–821.
Attias, H., 1998. Independent factor analysis. Neural Comput. 11, 803–851.
Attias, H., 2000. A variational Bayesian framework for graphical models. Adv. Neural. Inf. Process. Syst. 120 (1–2), 209–215.

Beckmann, C.F., 2004. Independent component analysis for functional magnetic resonance imaging. D.Phil. in information engineering, Image Analysis Group, FMRIB Centre and Robotics Research Group, University of Oxford, UK.
Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent components analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23, 137–152.
Beckmann, C.F., Smith, S.M., 2005. Tensorial extensions of independent component analysis for multisubject FMRI analysis. Neuroimage 250 (1), 294–311.
Bishop, C.M., 1999. Variational principal components. *Artificial Neural Networks*, 7–10 September 19990 (Conference Publication No. 470), pp. 509–514.
Calhoun, V.D., Adali, T., Giuliani, N.R., Pekar, J.J., Pearlson, G.D., 2006. Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. Hum. Brain Mapp. 270 (1), 47–62 Jan.
Choudrey, R.A., Roberts, S.J., 2001. Flexible Bayesian independent component analysis for blind source separation. Proc. Int. Conf. on Independent Component Analysis.
Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., Matthews, P.M., James, A., 2007. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. Brain 130, 2375–2386.
Douaud, G., Jbabdi, S., Behrens, T., Menke, R., Gass, A., Monsch, A., Rao, A., Whitcher, B., Kindlmann, G., Matthews, P., Smith, S., in press. DTI measures in crossing-fibre areas: Increased diffusion anisotropy reveals early white matter alteration in MCI and mild Alzheimer's disease.
Ennis, D.B., Kindlmann, G., 2006. Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. Magn. Reson. Med. 55, 136–146.
Filippini, N., MacIntosh, B.J., Hough, M.G., Goodwin, G.M., Frisoni, G.B., Smith, S.M., Matthews, P.M., Beckmann, C.F., MacKay, C.E., 2009. Distinct patterns of brain activity in young carriers of the APOE-ε4 allele. PNAS 1060 (17), 7209–7214.
Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. Neuroimage 390 (1), 181–205.
Groves, A.R., 2010. *Bayesian Learning Methods for Modelling Functional MRI*. . D.Phil. in Clinical Neurology, Image Analysis Group, FMRIB Centre, University of Oxford, UK.
Hartvig, N.V., Jensen, J.L., 2000. Spatial mixture modeling of fMRI data. Hum. Brain Mapp. 11, 233–248.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 130 (4–5), 411–430.
Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N.I., Calhoun, V., 2009. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Hum. Brain Mapp. 300 (1), 241–255.
Makni, S., Ciuciu, P., Idier, J., Poline, J.-B., 2006. Bayesian joint detection–estimation of brain activity using MCMC with a gamma–gaussian mixture prior model: IEEE International Conference on Acoustics, Speech and Signal Processing, volume 5.
Miwakeichi, F., Martinez-Montes, E., Valdes-Sosa, P.A., Nishiyama, N., Mizuhara, H., Yamaguchi, Y., 2004. Decomposing EEG data into space–time–frequency components using Parallel Factor Analysis. Neuroimage 22, 1035–1045.
Nielsen, F.B., 2004. *Variational Approach to Factor Analysis and Related Models*. Master of Science in Engineering, Intelligent Signal Processing group, Institute of Informatics and Mathematical Modelling – Technical University of Denmark, Anker Engelundsvej 1, Building 101A, 2800 Kgs. Lyngby, Denmark. .
Roberts, S.J., Penny, W.D., 2002. Variational Bayes for generalized autoregressive models. IEEE Trans. Sig. Proc. 500 (9), 2245–2257 Sept.
Scholz, J., Klein, M.C., Behrens, T.E.J., Johansen-Berg, H., 2009. Training induces changes in white-matter architecture. Nat. Neurosci. 120 (11), 1367–1368.
Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Behrens, T.E.J., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487–1505.
Watkins, K.E., Smith, S.M., Davis, S., Howell, P., 2008. Structural and functional abnormalities of the motor system in developmental stuttering. Brain 131, 50–59.
Wipf, D., Nagarajan, S., 2008. A new view of automatic relevance determination. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA, pp. 1625–1632.
Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14, 1370–1386.
Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M., 2004. Fully Bayesian spatio-temporal modeling of FMRI data. IEEE Trans. Med. Imaging 230 (2), 213–231.
Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Smith, S.M., 2005. Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. IEEE Trans. Med. Imaging 240 (1), 1–11.
Worsley, K.J., Poline, J.B., Vandal, A.C., Friston, K.J., 1995. Tests for distributed, nonfocal brain activation. Neuroimage 20 (3), 183–194.
Xu, L., Pearlson, G., Calhoun, V.D., 2009. Joint source based morphometry identifies linked gray and white matter group differences. Neuroimage 440 (3), 777–789.