Variational Bayesian Learning of Sparse Representations and Its Application in Functional Neuroimaging

E. Roussos[1], S. Roberts[1], and I. Daubechies[2]

[1] University of Oxford, Dept. of Engineering Science, Oxford, OX1 3PJ, UK
[2] Duke University, Dept. of Mathematics, Durham, NC 27708-0320, USA

Abstract. Recent theoretical and experimental work in imaging neuroscience reveals that activations inferred from functional MRI data have *sparse* structure. We view sparse representation as a problem in Bayesian inference following a machine learning approach and construct a structured generative latent-variable model employing adaptive sparsity-inducing priors. The construction allows for automatic complexity control and regularization as well as denoising. The performance of the proposed algorithm is demonstrated on some representative experiments.

Keywords: Sparse representations, variational Bayesian learning, hierarchical generative models, complexity control, wavelets, fMRI

1 Sparse Representations and Sparse Matrix Factorization

A fundamental operation in data analysis often involves the representation of observations in a 'latent' signal space that best reveals the internal structure of the data. In the linear model for brain activation [1], the spatio-temporal data, $\mathcal{X} = \{X(t,v)\}$, where v is a voxel in a brain volume, $\mathcal{V}$, and $t = 1, ..., T$ are timepoints, is modelled as a linear superposition of different activity patterns: $X(t,v) \approx \sum_{l=1}^{L} S_l(v)\, A_l(t)$, where $A_l(t)$ and $S_l(v)$ represent the dynamics and spatial variation, respectively. Our goal is the decomposition of the data set into *spatio-temporal components*, i.e. pairs $\{(\mathbf{a}_l, \mathbf{s}_l)\}_{l=1}^{L}$ such that the "regressors" $\{\mathbf{a}_l\}_{l=1}^{L}$ capture the 'time courses' and the coefficients $\{\mathbf{s}_l\}_{l=1}^{L}$ capture the 'spatial maps' of the patterns of activation. Unlike model-based approaches, such as the general linear model (GLM), in the 'model-free' case [2], addressed here, *both* factors must be learned from data, without a-priori knowledge of their exact spatial or temporal structure. The above is an *ill-posed* problem, however, without additional constraints. The main tool for exploratory decompositions of neuroimaging data into components currently in use is independent component analysis (ICA). As its name suggests, ICA forces statistical independence in order to derive maximally independent components. However, despite its success, there are both conceptual and empirical issues with respect to using the independence assumption as a prior for brain data analyses [3], [4]. In particular,

there is no physical or physiological reason for the components to correspond to different activity patterns with independent distributions. Recent experimental and theoretical work in imaging neuroscience [4], [5] reveals that activations inferred from functional MRI data are *sparse*. In fact, in Daubechies et al. [4] the key factor for the success of ICA-type decompositions, as applied to fMRI, was identified as the sparsity of the components, rather their mutual independence. Here, we shall exploit this sparseness structure for bilinear decomposition.

In matrix form, the problem of sparsely representing the data set $\mathcal{X}$ becomes a problem of *sparse matrix factorization* of its corresponding $T \times N$ data matrix, $\mathbf{X}$ ("unfolding" each scan at times $t = 1, \ldots, T$ into a row), where $N = |\mathcal{V}|$, the total number of voxels. A classical mathematical formulation of this problem [6] is to setup an optimization functional such as

$$\mathcal{I} = \|\mathbf{X} - \mathbf{AS}\|^2 + \lambda_{\mathbf{S}} \sum_{l=1}^{L} \sum_{n=1}^{N} |s_{l,n}| + \lambda_{\mathbf{A}} \|\mathbf{A}\|^2 \ ,$$

containing an ℓ_1 penalty/prior, enforcing sparse representations, where $\lambda_{\mathbf{S}}$, $\lambda_{\mathbf{A}}$ are regularization parameters. In Li et al. [6], learning of the basis, $\{\mathbf{a}_l\}$, was performed as an external step via the k–means algorithm. More realistic models should include a way for handling *noise* and *uncertainty*, however, and seamlessly fuse information from other parts of the model. In this paper we approach the problem from a Bayesian perspective and propose a fully Bayesian hierarchical model for bilinear decompositions.

2 Bayesian Sparse Decomposition Model

Bayesian inference provides a powerful methodology for machine learning, by providing a principled method for using our domain knowledge and allowing the incorporation of uncertainty in the data and the model into the estimation process, thus preventing 'overfitting'. The outputs of fully Bayesian inference are posterior probability distributions over all variables in the model. We derive a sparse decomposition algorithm viewing bilinear decomposition as a Bayesian generative model. We employ hierarchical source and mixing models, which result in automatic *regularization*. The model also contains an explicit noise model; the benefit of this is that observation noise is prevented from "leaking" into the estimated components, by effectively utilizing an implicit *filter* automatically learned by the algorithm.

We start by forming a representation to an "intermediate" space spanned by a wavelet family of localized time-frequency atoms. The use of wavelets in neuroimaging analysis has become quite widespread in recent years, due to the well-known property of wavelet transforms to form compressed, multiresolution representations of a very broad class of signals. Sparsity with respect to a wavelet dictionary means that most coefficients will be "small" and only a few of them will be significantly different than zero. Furthermore, due to the excellent approximation properties of wavelets, in the standard 'signal plus noise' model,

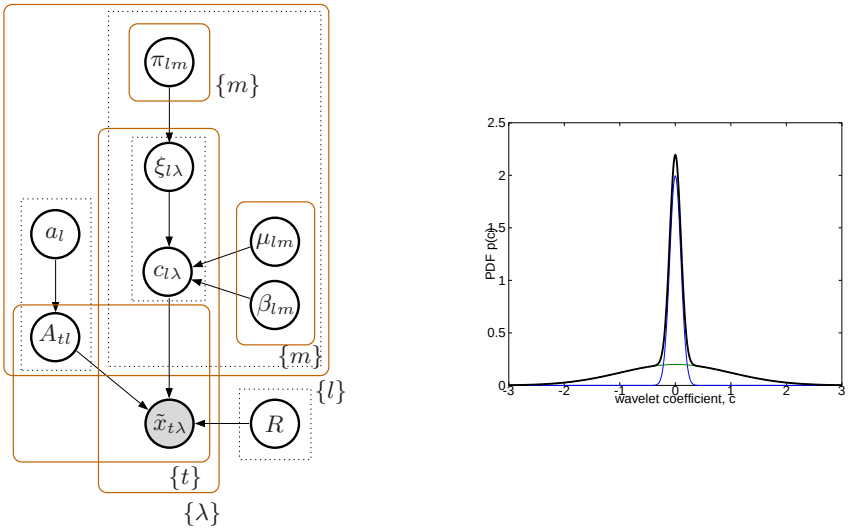

Fig. 1. Left: Variational Bayesian Sparse Representation graphical model. Each module is shown as a dotted box. Repetition over individual indices is denoted by *plates*, shown as brown boxes surrounding the corresponding variables. Instantiated nodes appear shaded. Right: Sparse mixture of Gaussians prior for the coefficients $\{c_{l,\lambda}\}$. Blue/green curves: Gaussian components; thick black curve, mixture density, $p(c_{l,\lambda})$.

decomposing data in an appropriate dictionary will typically result in large coefficients modelling the signal and small ones corresponding to noise. In addition, the wavelet transform largely decorrelates the data. The above properties should be captured by the model.

Following Turkheimer et al. [7], we perform our wavelet analysis in the spatial domain; for examples of the use of wavelets in the temporal dimension see, for example, Bullmore et al. [8]. Using this representation for all signals in the model, we get the noisy observation equation

$$\tilde{\mathbf{X}} = \mathbf{AC} + \tilde{\mathbf{E}} \,, \tag{1}$$

where the matrix $\left[\tilde{\mathbf{x}}_t^{\mathsf{T}}\right]_{t=1}^{T}$ denotes the transformed observations, $\left[\mathbf{c}_l^{\mathsf{T}}\right]_{l=1}^{L}$ the (unknown) coefficients of the wavelet expansion of the latent signals $\{\mathbf{s}_l\}_{l=1}^{L}$, and $\tilde{\mathbf{E}} \sim \mathcal{N}\left(\mathbf{0}_{T\times N}, R^{-1}\mathbf{I}_T\right)$ is a Gaussian noise process.

The probabilistic dependence relationships in our model can be represenented in a directed acyclic graph (DAG) known as a Bayesian network structure, shown in Fig. 1 (left). In this graph, random variables are represented as nodes and structural relationships between variables as directed edges connecting the corresponding nodes. Instantiated nodes appear shaded. The graphical representation offers modularity in modelling and efficient learning, as we can exploit the local (Markovian) structure of the model, captured in the network, as will be shown next. To fully specify the model, the probabilistic specification (priors) for all random variables in the model needs to be given. The learning algorithm then

infers the wavelet coefficients, $\{c_{l,\lambda}\}$, $\forall l, \lambda$, and learns the time-courses, $\{A_{t,l}\}$, $\forall t, l$, the parameters of the sparse prior on the coefficients, and the noise level. These four components of the model are shown as dotted boxes in Fig. 1.

Adaptive sparse prior model for the coefficients. The characteristic shape of the typical empirical histogram of the wavelet coefficients, $\{c_{l,\lambda}\}$, of the spatial maps is highly peaked at zero and heavy tailed. We want to capture this sparsity pattern of the coefficients in probabilistic terms. Our aim is to model a wide variety of sparseness constraints in a tractable (analytic) way and at the same time derive an efficient implementation of our method. In order to achieve this, we use distributions from the conjugate-exponential family of distributions. We enforce sparsity by restricting the general mixture of Gaussians model to be a two-component, zero co-mean mixture over each $\{c_{l,\lambda}\}_{\lambda=1}^{\Lambda}$, for each l. These have respective state variables $\{\xi_{l,\lambda}\}$ and mixing, mean, and precision parameters $\{\pi_{l,m}, \mu_{l,m}, \beta_{l,m}\}$, respectively, forming a parameter vector $\boldsymbol{\theta}_{\mathbf{c}_l}$, indexed by $m = 1, \ldots, M \doteq 2$. The mixture density is then given by

$$p(c_{l,\lambda}|\boldsymbol{\theta}_{\mathbf{c}_l}) = \sum_{m=1}^{M} p(\xi_{l,\lambda} = m|\boldsymbol{\pi}_l) p(c_{l,\lambda}|\xi_{l,\lambda}, \boldsymbol{\mu}_l, \boldsymbol{\beta}_l), \tag{2}$$

where $\mu_{l,m} \doteq 0$, $\forall l, m$, *a-priori*, and $p(\xi_{l,\lambda} = m|\boldsymbol{\pi}_l) = \pi_{l,m}$. The prior hyperparameters of the two components, have zero mean and hyperpriors over the precisions such that one component has a low precision, the other a high precision. These correspond to the two states of the wavelet coefficients, 'large' (carrying signal information) and 'small' (corresponding to "noise"). Figure 1 (right) depicts this scheme. We assign a Gaussian hyperprior on the position parameters $\boldsymbol{\mu}_l$, $p(\boldsymbol{\mu}_l) = \prod_{m=1}^{M} \mathcal{N}(\mu_{l,m}; m_{\mu_{l,0}}, v_{\mu_{l,0}})$, a Gamma on the scale parameters $\boldsymbol{\beta}_l$, $p(\boldsymbol{\beta}_l) = \prod_{m=1}^{M} \mathrm{Ga}(\beta_{l,m}; b_{\beta_{l,0}}, c_{\beta_{l,0}})$, and a Dirichlet on the mixing proportions $\boldsymbol{\pi}_l$, $p(\boldsymbol{\pi}_l) = \prod_{m=1}^{M} \mathrm{Di}(\pi_{l,m}; \alpha_{\pi_{l,0}})$. Note that the sparse MoG (SMoG) model parameters are not fixed in advance, but rather they are automatically learned from the data, adapting to the statistics of the particular spatial maps.

Hierarchical mixing model and Automatic Relevance Determination. The prior over the timecourses, $\{A_{t,l}\}$, is a zero-mean Gaussian with precision hyperparameter a_l over the lth column vector, $\mathbf{a}_l$: $p(\mathbf{a}_l|a_l) = \mathcal{N}\left(\mathbf{a}_l; \mathbf{0}_T, a_l^{-1}\mathbf{I}_T\right)$, $l = 1, \ldots, L$. The prior over each a_l is in turn a Gamma distribution, $p(a_l) = \mathrm{Ga}(a_l; b_{a_l}, c_{a_l})$. This hierarchical prior leads to a sparse marginal distribution for $\{\mathbf{a}_l\}_{l=1}^{L}$ (a Student-t, which can be shown if one integrates out the precision hyperparameter, a_l). By monitoring the evolution of the a_l, the relevance of each time-course may be determined; this is referred to as Automatic Relevance Determination, ARD [9]. This allows us to infer the complexity of the decomposition and obtain a sparse matrix factorization in terms of the time-courses as well, by suppressing irrelevant sources.

2.1 Variational Bayesian Inference

In the section above, we stated a generative model for sparse bilinear decomposition: the observed data are explained by propagating probabilities, initially drawn from the roots of the DAG, via the edges, and using the observation model of Eq. (1), performed in wavelet space. Let us collect all unknowns in the set $\mathcal{U} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\xi}, \mathbf{C}, \boldsymbol{a}, \mathbf{A}, R\}$. Exact Bayesian inference in these types of models is generally intractable because we need, in principle, to obtain the joint posterior of $\mathcal{U}$ given the data $\mathbf{X}$, $p(\mathcal{U}|\mathbf{X})$. Instead, we will use the variational Bayesian (VB) framework [10], for efficient approximate inference in high-dimensional settings, such as fMRI. The idea in VB is to approximate the complicated exact posterior with a simpler approximate one, $Q(\mathcal{U})$, that is closest to $p(\mathcal{U}|\mathbf{X})$ in an appropriate sense, in particular in terms of the Kullback-Leibler (KL) divergence. The optimization functional in this case is the (negative) variational *free energy* of the system:

$$\mathcal{F}(Q, \mathcal{X}) = \big\langle \log p(\mathcal{X}, \mathcal{U}) \big\rangle + \mathcal{H}\big[Q(\mathcal{U})\big] \ , \tag{3}$$

where the average, $\langle \cdot \rangle$, in the first term (negative 'variational energy') is over the variational posterior, $Q(\mathcal{U})$, and the second term is the entropy of $Q(\mathcal{U})$. The negative free energy forms a lower bound to the Bayesian log evidence, i.e. the marginal likelihood of the observations. Maximizing the bound minimizes the "distance" between the variational and the true posterior.

We choose to restrict the variational posterior $Q(\mathcal{U})$ to belong to a class of distributions that are factorized over subsets of the ensemble of variables:

$$Q(\mathcal{U}) = \Big(Q(\mathbf{C})Q(\boldsymbol{\xi})\Big)\Big(Q(\boldsymbol{\pi})Q(\boldsymbol{\mu})Q(\boldsymbol{\beta})\Big)\Big(Q(\mathbf{A})Q(\boldsymbol{a})\Big)Q(R) \ . \tag{4}$$

However here, unlike e.g. [14], we will employ variational posteriors that are *coupled* across latent dimensions for the wavelet coefficients of the spatial maps and the time-courses. Performing functional optimization with respect to the distributions of the unknown variables, we obtain the optimal form for the posterior: for $\mathbf{u} \in \mathcal{U}$,

$$Q(\mathbf{u}) \propto \exp\Bigg(\Big\langle \log p(\mathbf{X}, \mathcal{U}) \Big\rangle_{Q(\mathcal{U} \setminus \{\mathbf{u}\})}\Bigg) \ . \tag{5}$$

This results in an system of coupled equations, which are solved in an iterative manner. Theoretical results show that the algorithm converges to a (local) optimum [10]. Since we have chosen to work in the conjugate-exponential family, the posterior distributions have the same functional form as the priors and the update equations are essentially "moves", in parameter space, of the parameters of the priors due to observing the data.

We next show the update equations for the wavelet coefficients of the spatial maps and the time-courses.

Inferring the wavelet coefficients of the sources, $\mathbf{C}$*, and their states,* $\boldsymbol{\xi}$*.* The variational posterior has a Gaussian functional form, $\mathcal{N}\Big(\mathbf{C}; \hat{\boldsymbol{\mu}}_{L \times \Lambda}, \hat{\boldsymbol{\beta}}_{\Lambda \times L \times L}\Big)$,

with mean and precision posterior parameters for the λth wavelet coefficient vector, $\mathbf{c}_\lambda = (c_{1,\lambda}, \ldots, c_{L,\lambda})$, stored in the λth column of matrix $\mathbf{C}_{L\times\Lambda} = \left[c_{l,\lambda}\right]$, given by:

$$\hat{\boldsymbol{\mu}}_\lambda = \left(\hat{\boldsymbol{\beta}}_\lambda\right)^{-1}\left[\bar{\boldsymbol{\mu}}_\lambda + \left\langle \mathbf{A}^\mathsf{T}\right\rangle \langle \mathbf{R}\rangle\, \tilde{\mathbf{x}}_\lambda\right], \quad \text{and} \tag{6}$$

$$\hat{\boldsymbol{\beta}}_\lambda = \bar{\boldsymbol{\beta}}_\lambda + \left\langle \mathbf{A}^\mathsf{T}\mathbf{R}\mathbf{A}\right\rangle, \tag{7}$$

where $\bar{\boldsymbol{\mu}}_\lambda$ and $\bar{\boldsymbol{\beta}}_\lambda$ are 'messages' sent by the parents of the node $\mathbf{c}_\lambda$ to it and are computed by

$$\bar{\mu}_{l,\lambda} = \sum_{m=1}^{M} \hat{\gamma}_{l\lambda m} \left\langle \beta_{l,m}\right\rangle \left\langle \mu_{l,m}\right\rangle, \quad \text{and} \tag{8}$$

$$\bar{\beta}_{l,\lambda} = \sum_{m=1}^{M} \hat{\gamma}_{l\lambda m} \left\langle \beta_{l,m}\right\rangle \ . \tag{9}$$

The weighting coefficient $\hat{\gamma}_{l\lambda m}$, called 'responsibility', encodes the a-posteriori probability of the mth Gaussian kernel generating the λth wavelet coefficient of the lth spatial map. It is defined as the posterior probability of the state variable $\xi_{l,\lambda}$ (corresponding to the wavelet coefficient $c_{l,\lambda}$) being in the mth state: $\hat{\gamma}_{l\lambda m} \stackrel{\text{def}}{=} Q\left(\xi_{l\lambda} = m\right)$. The rest of the update equations for the mixture of Gaussians model take a standard form and can be found e.g. in [11].

Learning the mixing model parameters, $\left\{\mathbf{A}, \boldsymbol{a}\right\}$. The variational posterior over the matrix of the time-courses, $\mathbf{A}_{T\times L}$, is a product of Gaussians with mean and precision parameters for the tth row of $\mathbf{A}$ given by

$$\hat{\mathbf{a}}_t = \left[\mathbf{0}_{1\times L} + \langle R\rangle \left(\sum_{\lambda=1}^{\Lambda} \left\langle \tilde{x}_{t\lambda}\mathbf{c}_\lambda^\mathsf{T}\right\rangle\right)\right]\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{a}_t}\right)^{-1}, \tag{10}$$

$$\hat{\boldsymbol{\Gamma}}_{\mathbf{a}_t} = \operatorname{diag}\left(\langle \boldsymbol{a}\rangle\right) + \langle R\rangle \left(\sum_{\lambda=1}^{\Lambda} \left\langle \mathbf{c}_\lambda\mathbf{c}_\lambda^\mathsf{T}\right\rangle\right) \ . \tag{11}$$

The precision $\boldsymbol{a} = (a_l)$ is given by a Gamma distribution, $a_l \sim \mathrm{Ga}\left(a_l | b_{a_l}, c_{a_l}\right)$, with variational parameters

$$\hat{b}_{a_l} = \left(\frac{1}{b_{a_0}} + \frac{1}{2}\sum_{t=1}^{T}\left\langle {A_{tl}}^2\right\rangle\right)^{-1}, \qquad \hat{c}_{a_l} = c_{a_0} + \frac{1}{2}T \ , \tag{12}$$

for the lth column of $\mathbf{A}$. We set the prior hyperparameters $b_{a_0} = 1$ and $c_{a_0} = 1$.

Learning the noise model parameter, R. Finally, the noise precision has a Gamma distribution with hyperparameters

$$\hat{b}_R = \left(\frac{1}{b_{R_0}} + \frac{1}{2}\left\langle \operatorname{tr}\left[\left(\tilde{\mathbf{X}} - \mathbf{A}\mathbf{C}\right)^\mathsf{T}\left(\tilde{\mathbf{X}} - \mathbf{A}\mathbf{C}\right)\right]\right\rangle\right)^{-1}, \quad \hat{c}_R = c_{R_0} + \frac{1}{2}T\Lambda \ , \tag{13}$$

where we set $b_{R_0} = 1$ and $c_{R_0} = 1$ a-priori.

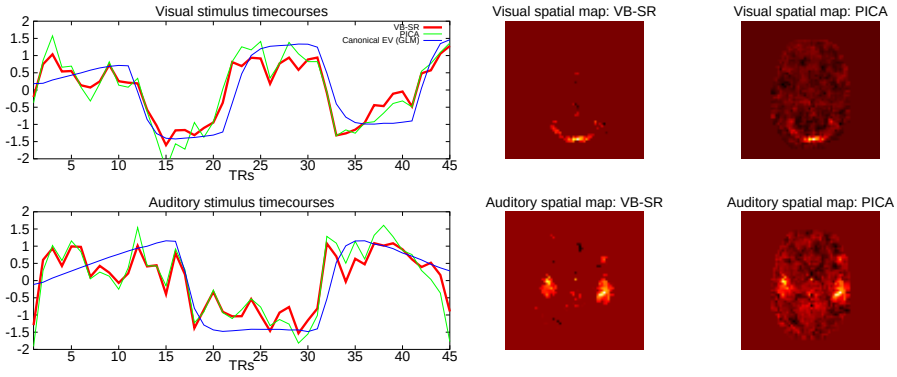

Fig. 2. Time courses and corresponding spatial maps resulting from applying the variational Bayesian sparse decomposition model to a visual-auditory fMRI data set. Red curve: our model; green curve: PICA; blue curve: canonical EVs. Note that the maps are the raw results from the model; no thresholding post-processing was performed.

3 Results

We tested the sparse decomposition model on the well-known audio-visual fMRI data set provided with the FSL FEEDS package [12], used as benchmark. The data set contains 45 time-points and 5 slices of size 64×64 voxels each. It was specifically designed as such in order to make it more difficult to detect the responses. We run our model on the dataset in order to detect 'consistently task related' (CTR) components. We applied the standard preprocessing steps (motion correction, registration, etc.), but no variance normalization or dimensionality reduction. For each of the separated components, we computed the correlation coefficient, r, between the associated timecourse, $\mathbf{a}_l$, and the 'expected timecourses', which were the canonical explanatory variables (EVs) from FEAT. After convergence, the model inferred only $L = 3$ components with $r > 0.3$. The component with the highest value of r was identified as the CTR map. A strong visual and a strong auditory component were extracted by the model; these are shown in Fig. 2. The correlation coefficients were $r_{\text{vis}} = 0.858$ and $r_{\text{aud}} = 0.764$. The corresponding PICA coefficients from Melodic were 0.838 and 0.756, respectively. The result of VB-ICA [13], [14] on the same dataset was 0.780 and 0.676, respectively [15]. It is worth noting that the spatial maps extracted from our model were also much cleaner than both PICA and VB-ICA (not shown), as displayed in Fig. 2. This is due to applying the sparse prior on the maps.

4 Discussion

We have presented a sparse representation model incorporating wavelets and sparsity-inducing adaptive priors under a full Bayesian paradigm. This enables

the estimation of both latent variables and basis functions, in a probabilistic graphical modelling formalism. We employed a variational framework for efficient inference. The preliminary results presented here suggest improved performance compared to other state-of-the-art model-free tools, such as PICA, while potentially allowing for more interpretable activation patterns due to the implicit denoising.

References

1. Worsley K. and Friston K.: Analysis of fMRI time series revisited—again. NeuroImage **2** (1995) 173–181
2. Beckmann C. and Smith S.: Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imag. **23** (2004) 137–152
3. McKeown M., Sejnowski T.: Independent component analysis of fMRI data: examining the assumptions. Hum. Brain Mapp. **6** (1998): 5–6. 368–372
4. Daubechies I., Roussos E., Takerkart S., Benharrosh M., Golden C., D'Ardenne K., Richter W., Cohen J., and Haxby J.: Independent component analysis for brain fMRI does not select for independence. PNAS **106 (26)** (2009) 10415–10412
5. Carroll M., Cecchi G., Rish I., Garg R. and Rao A.: Prediction and interpretation of distributed neural activity with sparse models. NeuroImage **44(1)** (2009) 112–22
6. Li Y., Cichocki A., Amari S-I., Shishkin S., Cao J. and Gu F.: Sparse representation and its applications in blind source separation. In Proceedings of the Annual Conference on Neural Information Processing Systems **17** (2003).
7. Turkheimer, F., Brett, M., Aston, J., Leff, A., Sargent, P., Wise, R., Grasby, P., and Cunningham, V.: Statistical modelling of PET images in wavelet space. Journal of Cerebral Blood Flow and Metabolism **20** (2001) 1610–1618
8. Bullmore, E., Fadili, J., Breakspear, M., Salvador, R., Suckling, J., and Brammer, M.: Wavelets and statistical analysis of functional magnetic resonance images of the human brain. Statistical Methods in Medical Research **12** (2003), 375–399
9. MacKay, D.: Probable Networks and Plausible Predictions - A Review of Practical Bayesian Methods for Supervised Neural Networks. Network: Computation in Neural Systems, **6** (1995), 469–505
10. Attias, H.: A Variational Bayesian Framework for Graphical Models. In Proceedings of Advances in Neural Information Processing Systems **12** (2000)
11. Penny, W, and Roberts, S.: Variational Bayes for 1-dimensional Mixture Models. Techn. Rep. PARG–00–2 (2000), Dept. of Engineering Science, University of Oxford.
12. Smith S., Jenkinson M., Woolrich M., Beckmann C., Behrens T., Johansen-Berg H., Bannister P., De Luca M., Drobnjak I., Flitney D., Niazy R., Saunders J., Vickers J., Zhang Y., De Stefano N., Brady J., and Matthews P.: Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage **23(S1)** (2004) 208–219.
13. Choudrey R., Penny W., and Roberts S.: An Ensemble Learning Approach to Independent Component Analysis. In Proceedings of Neural Networks for Signal Processing (2000)
14. Roussos E, Roberts S, and Daubechies I.: Variational Bayesian Learning for Wavelet Independent Component Analysis. In 25th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering **25** (2005)
15. Groves A.: Bayesian Learning Methods for Modelling Functional MRI. D.Phil. thesis, Department of Clinical Neurology, University of Oxford, 2010