4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

LEARNING HIERARCHICAL DYNAMICS USING INDEPENDENT COMPONENT ANALYSIS

R.A. Choudrey and S.J. Roberts

University of Oxford
Robotics Research
Oxford, U.K.

ABSTRACT

Mixture modelling techniques such as Mixtures of Principal component and Factor analysers [1, 2] are very powerful in representing and segmenting Gaussian clusters in data. Meaningful segmentations may be lost, however, if these self-similar areas are non-Gaussian. For such data, an intuitive model is a Mixture of Independent Component Analysers [3, 4, 5]. Such a model, however, ignores dynamics, both *between* clusters and *within* clusters. The former can be remedied by enforcing a Markov prior over the compnent mixture variables, leading to a Hidden Markov model (HMM) with ICA generators. The latter can be modelled if the source models of these ICA generators are themselves dynamic, for example by utilising HMM sources. HMMs are models for picking up dynamic changes of state in the underlying data generation process, and are therefore useful in capturing high-order temporal information. The proposed method is a piecewise approach to detecting dynamic movement, focussing on abrupt changes in the observation model and/or in the source model, while assuming static statistics in between. The hierarchical approach allows the analysis of signals which have macro- and micro-dynamics, such as stock indices.

1. INTRODUCTION

Virtually all methods for ICA consider observed data to be independently and identically distributed and, as such, any correlations across time are deemed unimportant. Considering the wide spread and successful application of ICA to time series data, dynamics are conspicuous by their absence; considering the extra complexity involved in incorporating a dynamic element, this is understandable. Resolute blindness to temporal information may not be suitable in all cases. In some observations, there may be dynamic changes in the sources, affecting the source pdf statistics, and/or dynamic changes in the mixing process. For example, speech has specific dynamics so using ICA for the blind source separation of mixtures of speakers would benefit greatly if these dynamics could be captured. Similarly, if the mixing of these speakers was non-stationary, modelling this dynamic would improve the separation. More generally, such an ICA model with hierarchical dynamics could model time-series which have both large-scale (macro) and small-scale (micro) dynamic changes more effecively, for example financial time-series. Such a model could also perform filtering (e.g. noise reduction), forcasting, as well as more accurate classification. Unfortunately, an ICA model with full dynamics is horrendously complicated. This paper aims to take a step forward by using Hidden Markov models (HMMs) to detect abrupt changes in macro- and micro-dynamics while assuming static statistics in between. The macro-dynamics will be captured by an HMM switching between components in an ICA mixture model. The micro-dynamics will be modelled by HMM sources in these ICA components. This will be shown to accurate enough to detect macro- and micro-dynamics in stock indices.

Dynamic ICA was first demonstrated by Pearlmutter and Parra's 'contextual' ICA [6] for square, noiseless mixing, where simple temporal information was incorporated into the source densities. This was improved in [7] in which Generalised Auto-Regressive (GAR) sources were used to condition source signals on previous values. A different methodology was proposed by Attias in [8] based on non-stationary, non-instantaneous mixing of stationary sources. This could, for example, be used for blind deconvolution. Learning non-stationary, instantaneous mixing using particle filters has been proposed by Murata *et al.* [9] and Everson and Roberts [10], the latter also making some in-roads into non-stationay sources. Penny and Roberts proposed a simple extension to HMMs, incorporating square, noiseless ICA observation generators with inverse-cosh densities in [11], allowing abrupt changes in the mixing to be captured. This was extended by Penny *et al.* [12] to include ICA GAR sources. Attias extended his IFA framework [13] to include Hidden Markov Model sources in [14].

The Bayesian formulation of models is often desireable when dealing with noisy data and/or small amounts of data as the prior distributions act as natural regularisers. Furthermore, Bayesian learning averages over all possible parameter values giving unbiased estimates, robust predictions and allows the comparison of assumptions underlying different models. Bayesian ICA has been formulated previously [15, 16, 17, 18], as have Bayesian HMMs [19, 20]. Furthermore, Bayesian mixtures of ICAs have been introduced in [4, 5]. The model presented in this paper combines the variational Bayesian ICA mixture model presented in [4] with the variational Bayesian HMM introduced in [19] to produce a variational Bayesian HMM with ICA generators each with HMM sources, able to find nested dynamics in signals.

2. THE MODEL

Each cluster in the data distribution is modelled by an ICA model. Each cluster is described as the linear projection of a set of univariate distributions lying along independent directions, i.e. the independent source distributions of each ICA. To make these source distributions as flexible as possible, each one is modelled by a mixture of Gaussians (MoG). The set of all ICA models comprise an ICA mixture model. Therefore, the overall model is a hierarchy of mixture models. The next section explains mixture models in more detail, followed by a description of the ICA model in section 2.2.

The dynamics of the data as it jumps from one cluster to another (macro-dynamics) are captured by coupling the components in the Mixture of ICAs model (MoICA) using a Markov prior. These state changes are hidden and so transform the MoICA model into a Hidden Markov model with ICA generators. The dynamics of the data *within* each cluster (micro-dynamics) are modelled by coupling the components in each source MoG using another Markov prior. These state changes are also hidden and so transform the MoG model into a Hidden Markov model with Gaussian generators. Section 2.3 discusses HMMs in more detail.

2.1. Mixture models

The probability of generating a data vector $\boldsymbol{x}^t$ from a C-component mixture model given assumptions $\mathcal{M}$ is:

$$p(\boldsymbol{x}^t|\mathcal{M}) = \sum_{c=1}^{C} p(c|\mathcal{M}_0)p(\boldsymbol{x}^t|\mathcal{M}_c, c) \tag{1}$$

A data vector is generated by choosing one of the C components stochastically under $p(c|\mathcal{M}_0)$ and then drawing from $p(\boldsymbol{x}^t|\mathcal{M}_c, c)$. $\mathcal{M} = \{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_C\}$ is the vector of component model assumptions, $\mathcal{M}_c$, and assumptions about the mixture process, $\mathcal{M}_0$. The assumptions represent everything that essentially defines the model - values of fixed parameters, model structure, details of the component switching method, any prior information etc. $p(\boldsymbol{x}^t|\mathcal{M})$ is known as the evidence for model $\mathcal{M}$ and quantifies the likelihood of the observed data under model $\mathcal{M}$.

The variable c indicates which component of the mixture model is chosen to generate a given data vector $\boldsymbol{x}$. If $p(c|\mathcal{M}_0)$ is a vector of probabilities and each component $p(\boldsymbol{x}^t|\mathcal{M}_c, c)$ is a Gaussian, then (1) simply describes a Mixture of Gaussians (MoG). If the MoG is adapted through a maximum likelihood approach then $\mathcal{M}$ represents a list of point estimates for the corresponding parameters. In our mixture model, however, each component has a non-Gaussian density derived from the ICA model presented in the next section, and $\mathcal{M}$ represents assumptions concerning the *distribution* of possible parameter values.

If the data signals contain dynamics, then this temporal information can be exploited by conditioning the probabilistic occupancy of source state at time t, $p(c^t|\mathcal{M}_0)$, on a source state at some previous time, say $t-1$. Such a coupling gives a Hidden Markov model with ICA generators. The ICA model used is discussed in the next section.

2.2. The ICA Model

In common with ICA in the literature, we choose a generative model to work with. The observed variables, $\boldsymbol{x}$, of dimension M, are modelled as a linear combination of statistically independent latent variables, $\boldsymbol{s}_c$, of dimension L_c, with added Gaussian noise

$$\boldsymbol{x} = \boldsymbol{A}_c\boldsymbol{s}_c + \boldsymbol{y}_c + \boldsymbol{n}_c \tag{2}$$

where where $\boldsymbol{A}_c$ is an $M \times L$ mixing[1] matrix, $\boldsymbol{y_c}$ is an M-dimensional *bias* vector, $\boldsymbol{n_c}$ is M-dimensional additive noise and c represents the c^{th} ICA model. In signal processing nomenclature, M is the number of (observed) sensors and L_c is the number of latent (hidden) sources.

Equation (2) acts as a complete description for cluster c in the data density. The bias vector, $\boldsymbol{y_c}$, defines the position of the cluster in the M-dimensional data space, $\boldsymbol{A}_c$ describes its orientation and $\boldsymbol{s}_c$ describes the underlying manifold. The noise, $\boldsymbol{n}_c$, is assumed to be zero-mean Gaussian and isotropic, $p(\boldsymbol{n}_c|0, \Lambda_c, c) = \mathcal{N}(\boldsymbol{n}_c; 0, \lambda_c I)$, where $\lambda_c I$ is the precision. The noise essentially absorbs any (spherical) Gaussianity present in the cluster. The probability of observing data vector $\boldsymbol{x}^t$ under component c is then given by

$$p(\boldsymbol{x}^t|\boldsymbol{W}_c, c) = \left(\frac{\lambda_c}{2\pi}\right)^{\frac{M}{2}} \exp[-E_c^t] \tag{3}$$

where $\boldsymbol{W}_c = \{\boldsymbol{A}_c, \boldsymbol{s}_c^t, \boldsymbol{\lambda}_c, \boldsymbol{y}_c\}$ and where $E_c^t = \frac{\lambda_c}{2}(\boldsymbol{x}^t - \boldsymbol{A}_c\boldsymbol{s}_c^t - \boldsymbol{y}_c)^T(\boldsymbol{x}^t - \boldsymbol{A}_c\boldsymbol{s}_c^t - \boldsymbol{y}_c)$ (^{T} indicates transpose). The distribution over $\boldsymbol{s}_c = \{s_{c,1}, .., s_{c,i}, .., s_{c,L_c}\}$ for data point t can be written as

$$p(\boldsymbol{s}_c^t|\mathcal{M}_{\boldsymbol{s}_c}, c) = \prod_{i=1}^{L_c} p(s_{c,i}^t|\mathcal{M}_{s_{c,i}}, c) \tag{4}$$

where the product runs over the L_c sources of component c and $\mathcal{M}_{\boldsymbol{s}_c}$ is the vector of source model assumptions. Using such a generative approach, one must stipulate the source density.

ICA Source Model

The choice of a flexible and mathematically tractable source model is crucial if a wide variety of source distributions are to be modelled; in particular, the source model should be capable of encompassing distributions with a wide variety of kurtoses and complex, multi-modal distributions.

One such distribution is a factorised mixture of univariate Gaussians with L_c factors (i.e. sources) and m_i components per source

$$\begin{aligned} p(\boldsymbol{s}_c^t|\boldsymbol{\theta}_c, c) &= \prod_{i=1}^{L_c}\sum_{q_i=1}^{m_i} p(q_i^t = q_i|\boldsymbol{\pi}_i, c)p(s_{c,i}^t|\theta_{c,i}, c) \\ &= \prod_{i=1}^{L_c}\sum_{q_i=1}^{m_i} \pi_{i,q_i}\mathcal{N}(s_{c,i}^t; \mu_{i,q_i}, \beta_{i,q_i}) \end{aligned} \tag{5}$$

[1] Please note that the term *mixing* refers to the linear mixing in the ICA model and *mixture* refers to stochastic generating method in a mixture model

where, for brevity, the ICA component subscript c has been dropped from parameters which can be seen to belong to ICA c from context. From now on, all subscripted parameters should be assumed to belong to the c^{th} ICA model, unless otherwise stated. The mean and precision of Gaussian q_i in source i are μ_{i,q_i} and β_{i,q_i} respectively. Equation (5) essentially describes the local features of cluster c - μ_{i,q_i} is the position of feature q_i w.r.t. the cluster centre, β_{i,q_i} is its size, and π_{i,q_i} its 'prominance' w.r.t. other features.

Such a source model is stationary in that there is no temporal coupling across source states. If the source signals contain dynamics, then this temporal information can be exploited by conditioning the probabilistic occupancy of source state at time t, $\boldsymbol{q}_i^t$, on a source state at some previous time, say $t-1$. Such a coupling gives a Hidden Markov model with Gaussian generators.

2.3. The Hidden Markov Model

The HMM presented here is for a 1-dimensional signal, $\boldsymbol{s}$. HMMs for multivariate signals are a straight-forward extension (see [21] for more details). The equivalent model for L statistically independent signals is simply a product of L 1-dimensional HMMs.

In essence, an HMM is a finite-state machine that switches between different probability density functions (pdfs) which represent the observation generators. The state change is a Markov process, whereby the occupancy of state d at time t depends probabilistically *only* on the state at some previous time, $t-\tau$. The model presented below is a first-order Markov process, where $\tau = 1$. Each observation is generated by first moving from state c at time $t-1$ to state d at time t according to transition probability r_{cd}, then stochastically drawing (or picking) from the d^{th} pdf.

The model is defined as follows. Let the model have m states represented by $q = \{1, 2, \ldots, m\}$. Let variable q^t be an m-dimensional vector that indicates which state is chosen at time t. If the HMM is in state d at time t, then q^t has a 1 in the d^{th} entry and zeros everywhere else. Let $\boldsymbol{R}$ represent the matrix of state transition probabilities where the probability of moving from state c to d is given by r_{cd}

$$p(q^t = d | q^{t-1} = c, \boldsymbol{R}) = r_{cd} \tag{6}$$

The probability of the HMM starting in a given state is given by the vector $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_m\}$ such that

$$p(q^1 = c | \boldsymbol{\pi}) = \pi_d \tag{7}$$

Let the observation at time t be denoted s^t. The probability of generating s^t given state $q^t = d$ is given by

$$p(s^t | q^t = d, \theta_d) = p_d(s^t | \theta_d) \tag{8}$$

where θ_d represents the parameters of the d^{th} observation pdf. The parameter set for the m pdfs is $\{\theta_1, \ldots, \theta_m\}$, and for the whole hMM, the total parameter set is $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{R}, [\theta_1, \ldots, \theta_m]\}$.

If there are T data points, then $\boldsymbol{s} = \{s^1, ..., s^T\}$ and $\boldsymbol{q} = \{q^1, ..., q^T\}$. The joint probability of state path $\boldsymbol{q}$ generating observation sequence $\boldsymbol{s}$ given parameter set $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\boldsymbol{s}, \boldsymbol{q} | \boldsymbol{\theta}) &= p(q^1 | \boldsymbol{\pi}) \prod_{t=2}^{T} p(q^t | q^{t-1}, \boldsymbol{R}) \prod_{t=1}^{T} p(s^t | q^t, \theta_{q^t}) \\ &= p(\boldsymbol{q} | \boldsymbol{\theta}) p(\boldsymbol{s} | \boldsymbol{q}, \boldsymbol{\theta}) \end{aligned} \tag{9}$$

where $p(\boldsymbol{q}|\boldsymbol{\theta}) = p(q^1|\boldsymbol{\pi}) \prod_{t=2}^{T} p(q^t|q^{t-1}, \boldsymbol{R})$ and $p(\boldsymbol{s}|\boldsymbol{q}, \boldsymbol{\theta}) = \prod_{t=1}^{T} p(s^t | q^t = d, \theta_d)$. The likelihood that the signals, $\boldsymbol{s}$, were generated by this model using parameters $\boldsymbol{\theta}$ irrespective of state path is

$$p(\boldsymbol{s} | \boldsymbol{\theta}) = \sum_{\{\boldsymbol{q}\}} p(\boldsymbol{q} | \boldsymbol{\theta}) p(\boldsymbol{s} | \boldsymbol{q}, \boldsymbol{\theta}) \tag{10}$$

where $\{\boldsymbol{q}\} = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_{m^T}\}$, the space of all state paths.

An HMM is learnt using the Forward-Backward algorithm [21], an efficient Expectation-Maximisation method. The most likely time-path of the hidden states can be inferred using the Viterbi algorithm [21].

Note the mixture-model form of (10); HMMs are simply mixture models with Markov time dependencies across hidden states $\boldsymbol{q}$. Therefore, any mixture model can be 'made dynamic' by using a Markov prior across the hidden indicator variables. The MoICA model can be made dynamic by placing a Markov prior over the hidden states such that $p(c^t = c)$ becomes $p(c^t = d | c^{t-1} = c)$. These represent the macro-states. Similarly, the ICA source MoGs can be made dynamic by placing a Markov prior over their hidden states. These represent the micro-states.

3. VARIATIONAL BAYESIAN LEARNING

A Bayesian scheme can be used to introduce prior knowledge, control complexity, perform model selection, and integrate out the dependency of a model on its parameters. As ever, a strict Bayesian treatment is computationally expensive and is often intractable. Therefore, a variational Bayesian approximation [22, 23] (also known as ensemble learning) can be used to allow tractability and increase computational efficiency.

Let the weights $\boldsymbol{W}$ be a vector of all hidden variables and unknown parameters. Variational Bayesian learning involves assuming some factored form for the posterior over weights, denoted $p'(\boldsymbol{W})$. The objective function to be maximised is then given by

$$F[\boldsymbol{X}] = \left\langle \log \frac{p(\boldsymbol{X}, \boldsymbol{W})}{p'(\boldsymbol{W})} \right\rangle_{p'(\boldsymbol{W})} \tag{11}$$

The quantity F is called the *negative free-energy* (NFE) for model $\mathcal{M}$ and can be shown [23] to be a strict lower bound to the log evidence, with the difference being the Kullback-Leibler (KL) divergence between the true and approximating posterior. By maximising F, not only do we minimise the KL-divergence between the approximating and true posterior, we also implicitely integrate out the unknowns $\boldsymbol{W}$. By choosing an appropriate form for the approximation $p'(\boldsymbol{W})$, we perform tractable Bayesian learning. As F is a strict lower bound to the model (log) evidence, a wide variety of models and assumptions can be compared and contrasted by calculating F for each model.

The higher F is, the higher the likelihood of the data under that model, and, therefore, the better that model is at 'explaining' the data.

Learning HMMs under a variational Bayesian formultion was derived by MacKay in [19] and was shown to be very similar to the maximum-likelihood Forward-Backward algorithm. The Viterbi path is also similar. Learning mixtures of ICAs under variational Bayes was independently derived by Choudrey and Roberts [4] and Chan *et al* [5]. Combining the two is thus straight forward.

The complete set of weights for our model is given by $\boldsymbol{W} = \{\boldsymbol{c}, \boldsymbol{R}_{\text{mac}}, \boldsymbol{\pi}_{\text{mac}}, [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_C]\}$ where 'mac' refers to the HMM parameters for the dynamic MoICA model and where $\boldsymbol{W}_c$ are the weights for ICA c. These comprise $\boldsymbol{W}_c = \{\boldsymbol{A}_c, \boldsymbol{\Lambda}_c, \boldsymbol{S}_c, \boldsymbol{q}_c, \boldsymbol{\theta}_c\}$ where $\boldsymbol{\theta}_c = \{\boldsymbol{R}_{c,\text{mic}}, \boldsymbol{\pi}_{c,\text{mic}}, \boldsymbol{\mu}_c, \boldsymbol{\beta}_c\}$ are the source HMM parameters for ICA c ('mic' refers to their responsibility for modelling micro-dynamics). The following factorisation is chosen for the approximating posterior

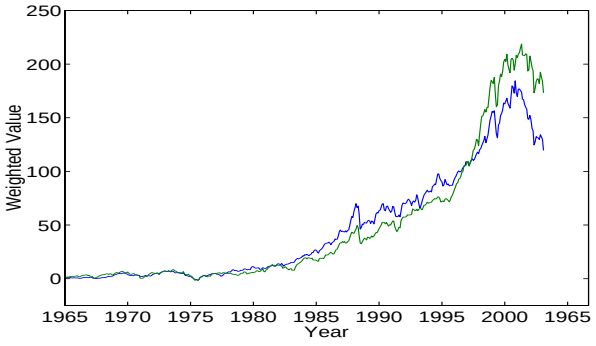
$$p'(\boldsymbol{W}) = p'(\boldsymbol{c})p'(\boldsymbol{R}_{\text{mac}})p'(\boldsymbol{\pi}_{\text{mac}}) \prod_{c=1}^{C} p'(\boldsymbol{W}_c) \qquad (12)$$

where $p'(\boldsymbol{W}_c) = p'(\boldsymbol{A}_c)p'(\boldsymbol{\Lambda}_c)p'(\boldsymbol{S}_c|\boldsymbol{q}_c)p'(\boldsymbol{q}_c)p'(\boldsymbol{\theta}_c)$, $p'(\boldsymbol{\theta}_c) = p'(\boldsymbol{R}_{c,\text{mic}})p'(\boldsymbol{\pi}_{c,\text{mic}})p'(\boldsymbol{\mu}_c)p'(\boldsymbol{\beta}_c)$ The posterior over sources and source states is also chosen to factorise over the L_c sources to reduce the computational load.

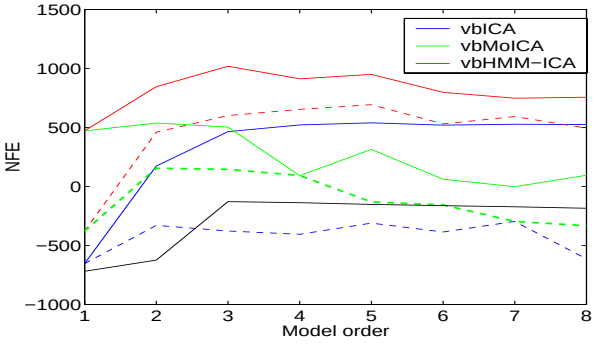
Vague priors are set for the HMM parameters (see [24]) and ICA parameters (see [18]). All the derived posteriors require solving a set of coupled hyper-parameter update equations, which are updated cyclically until convergence. This includes using the forward-backward algorithm [19] twice to learn both sets of HMM parameters. Once trained, the model can be used to reconstruct hidden source signals (to within a scaling and permutation) given a data set by calculating $\langle q_i \rangle$ and $\langle s_{c,i} \rangle$ under their respective posteriors over the whole data-set, and given the (now fixed) model parameter posteriors. The hidden state path for both micro- and macro-HMMs can be inferred using the variational Viterbi algorithm (see [21] and [24]).

4. RESULTS

The dynamic ICA algorithm developed in this paper was used to model the dynamics of the FTSE-100 and Dow Jones Industrial Average-30 (DJIA30) over the last 38 years. Figure 1(a) plots their course from June 1964 to June 2002 at monthly intervals, giving 457 data-vectors in total. These were normalised to unit variance and adjusted to intially start at the weight value of 1 for 6/1964. This data was then fed to a whole gamut of ICA models which are a subset of the model presented here (termed vbHMM^{2}ICA), each trained until the NFE converged to within 0.01%. The models used were vbICA, vbMoICA and vbHMM-ICA each with either MoG or HMM sources, giving 6 flavours in total. The 2-source vbICA models were trained across 1-8 components per source (same number per source), and the vbMoICA and vbHMM models were trained across 1-8 componet ICA generators, each 2-componet ICA with 3 source states each. As a comparison, a vbHMM model with 1-8 Gaussian generators was also trained.



(a) Share indices. Blue: FTSE100 Green: DJIA30

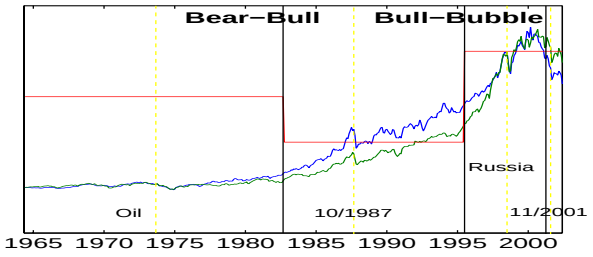


(b) Model selection. Dashed: MoG sources Solid: HMM sources Black: Gaussian vbHMM

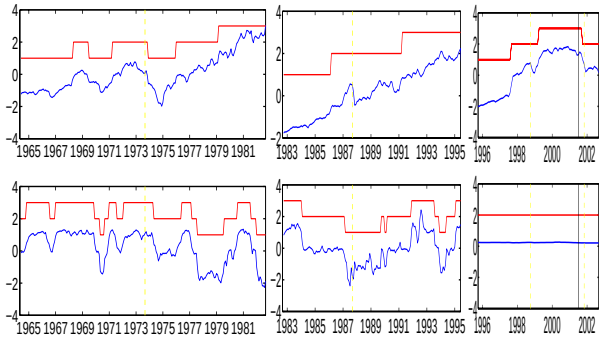
Figure 1: Original data and NFE of various models.

The negative free energy plots for each model is shown in Figure 1(b). By far the most preferred formalism was a vbHMM model with ICA generators, each of which had dynamic HMM source models (vbHMM^{2}ICA). This hierarchical approach scored higher than vbMoICA-HMM, vbICA-HMM and standard vbHMM-ICA, with an HMM with 3 ICA-HMM generators being the most likely. In principle, the NFE can be used to work out the most likely number of micro-states as well, although this search is cumbersome.

Exactly what this model learnt is shown in Figure 2. To give some meaning to these results, important changes and impacting events in the markets are indicated in Figure 2(a). The markets can be divided into 4 regimes, the boundaries of which are marked by solid, black lines. The last sustained, deflationary period of the markets (termed a 'bear market') officially ended on August 12^{th} 1982 [25] and was followed by a sustained, inflationary periord (termed a 'bull market'); this is represented by the first boundary. Shares in the web-browser company 'Netscape' were floated on the 9^{th} of August 1995, almost 13 years to the day later. This is generally regarded [25] as the beginning of what is popularly called the 'internet [speculative] bubble'; this is the second black line. This bubble burst on April 14^{th} 2000, forcing the market - yet again - into a bear market; this is



(a) Regimes and events. Black: Market regimes Yellow: Events: Red: State-path.



(b) ICA2 Recon. (c) ICA1 Recon. (d) ICA3 Recon.

Figure 2: Patterns in share indices.

the final black boundary. Also scribed in dotted yellow are pertinent dates that had marked affects on the markets - the october 1973 raising of oil prices by OPEC, the stock-crash on October 19^{th} 1987 ('Black Monday'), the devaluation of the Russian rouble on the 17^{th} of August 1998, and the September 11^{th} 2001 terrorist attacks on the USA.

Plotted in red on Figure 2(a) is the inferred state-path of the vbHMM^{2}ICA model's macro-states. The 2 changes of state correspond exactly to the dates the markets moved from bear to bull status, and from bull to bubble status. Thus, ICA-HMM generator 2 is responsible for pre-bull market dates, ICA-HMM 1 for the bull market and ICA-HMM 3 for the bubble and subsequent bear market. What vbHMM^{2}ICA has captured are fundamental changes in the markets which manifest themselves as fundamental changes in the *statistics* of the markets. The markets are chaotic, but structured and increase in the long term, so these statistics are non-Gaussian. Although the standard vbHMM-Gaussian model also favoured three states, the associated state-path did not correspond to any recognised changes in the markets' regime. Clearly, the Gaussian generators were not well matched to detecting changes in non-Gaussian statistics.

Plotted below Figure 2(a) are the ICA-HMM source reconstructions together with their state-paths of micro-states. The signals for source 1 represent a 'proto-index' which is the underlying signal common to both the FTSE100 and the DJIA30. The state-paths for these signals essentially track the mean of this signal in a piecewise fashion. On two occasions, the changes of state match with specific stimuli, namely the 1973 oil crisis and the bursting of the internet bubble. Both these led to subsequent bear markets, thus changes in the underlying statistics. The 1987 crash and Russian currency crisis do not register in the state dynamics, both of which had only short-term impacts on an overall bear/bubble market. The effect of 11/9/2001 is harder to resolve as it occurred in an already-falling market, very soon after the bubble burst. The state-change is between the two events, but is most likely to be due to the bubble bursting as this had a far wider and longer impact, leading to the bear market current at the time of writing.

The meaning of the second source is less clear. Whereas the overall vbHMM model and the first source of the constituent ICA-MM models track the changing mean, the second seems to bare little relation to the stock indices. It is tempting to dismiss these signals as non-Gaussian noise, but - for ICA's 1 and 2 at least - there does seem to be some superficial meaning. While source signal 1 tracks the common, underlying proto-index, signal two seems to track the *difference* between the two. This is further supported by entries in the mixing matrices. The appropriate entries for source-to-sensor mapping for source 2 are roughly equal in magnitude for each sensor signal, but with different signs. This would indicate that source 2 quantifies the difference between the two such that

$$\text{FTSE100} \approx A_{11}\text{proto} + \alpha\text{diff} \quad (13)$$

$$\text{DJIA30} \approx A_{21}\text{proto} - \alpha\text{diff} \quad (14)$$

where $\alpha = \frac{1}{2}(|A_{12}| + |A_{22}|)$ for each ICA. This does not seem to be the case for ICA 3, however, where the second signal has been surpressed. The reason for this is unclear.

Note that the final regime change is picked up by a change in ICA-HMM state rather than a change of state in the overall vbHMM, as the previous two were. This is because there are only 15 data vectors after the bursting of the bubble, which is not enough to warrant a vbHMM state of their own - it is cheaper to code them as a state change within an ICA-HMM generator. If there were enough data-vectors, it is hard to tell whether the continuing bear market's statistics are simply a mirror of the bull market's, or are unique to a bear market. In the former case, such a regime would be represented by the same state as the bull market with the ICA source signals coding for the downward trend. In the latter, there would be an extra 4^{th} state solely determined by bear market statistics.

5. DISCUSSION

We have shown how the modelling of hierarchical dynamics using ICA methodology can be used to discover nested dynamics in time series. The temporal information in these series is exploited by utilising hierarchical Hidden Markov models connected using ICA networks. The vbHMM^{2}ICA model was used to find large-scale changes in stock market regimes, while simultaneously capturing more detailed,

short-term changes. Both these levels of dynamics were shown to be interpretable. Bayesian selection between various models was shown to favour vbHMM^{2}ICA as the most representative model.

The vbHMM^{2}ICA model brings together a whole family of methos under one mathematical framework. Factor analysis, mixtures of Factor analysers, ICA with MoG or HMM sources, mixtures of ICA with MoG or HMM sources, HMMs with Gaussian observations, HMMs with ICA observations etc. are all variations of the model presented here as one can choose to 'switch-on' either the macro- and micro-dynamics separately, utilise only 1 ICA component in the ICA mixture, use only a single Gaussian in the ICA sources, or any other permutation. Therefore, vbHMM^{2}ICA can be applied to a vast range of data analysis problems, such as clustering, stationary and non-stationary blind source separation, subspace modelling, speech recognition, image recognition and data coding to name just a few.

6. REFERENCES

[1] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[2] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems*, 2000, vol. 12, pp. 449–455.

[3] T.W. Lee, M.S. Lewicki, and T.J. Sejnowski, "ICA mixture models for unsupervised classification and automatic context switching," in *International Workshop on Independent Component Analysis*, 1999, pp. 209–214.

[4] R. Choudrey and S.J. Roberts, "Variational mixture of Bayesian independent component analysers," *Neural Computation*, vol. 15, no. 1, 2003, To appear.

[5] K. Chan, T-W. Lee, and T. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," in *Proceedings of ICA2001*, 2001.

[6] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," in *ICONIP '96*, 1996, pp. 151–157.

[7] B. Pearlmutter and L. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ICA," in *Advances in Neural Information Processing Systems*, M.C Mozer, M.I Jordan, and T. Petsche, Eds. 1997, vol. 9, pp. 613–619, MIT press.

[8] H. Attias, "Blind source separation and deconvolution: the dynamic component analysis algorithm," *Neural Computation*, vol. 10, pp. 1373–1412, 1998.

[9] N. Murata, S. Ikeda, and A. Ziehe, "Adaptive online learning in changing enviroments," in *Advances in Neural Information Processing Systems*, M.C Mozer, M.I Jordan, and T. Petsche, Eds. 1997, vol. 9, pp. 599–605, MIT press.

[10] R Everson and S.J. Roberts, "Non-stationary Independent Component Analysis," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN'99)*. IEE, 1999.

[11] W.D. Penny and S.J. Roberts, "Hidden Markov models with extended observation densities," Tech. Rep., Imperial College of Science Technology and Medicine, 1998.

[12] W. Penny, S. Roberts, and R. Everson, "Hidden Markov Independent Components Analysis," in *Advances in Independent Component Analysis*, Mark Girolami, Ed. Kluwer Academic Publishers, 2000.

[13] H. Attias, "Independent Factor Analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[14] H. Attias, "ICA and graphical models," in *Independent Component Analysis: Principles and Practice*, S Roberts and R Everson, Eds., chapter 3, pp. 95–112. Cambridge university press, 2001.

[15] K.H. Knuth, "Bayesian source seperation and localisation," in *Proceedings of SPIE'98*, A. Mohammad-Djafari, Ed., 1998, vol. 3459, pp. 147–158.

[16] H Lappalainen, "Ensemble learning for Independent Component Analysis," in *Proceedings of the First International Workshop on Independent Component Analyis*, 1999, pp. 7–12.

[17] J.W. Miskin and D.J.C. MacKay, "Ensemble learning for blind source seperation," in *Independent Component Analysis: Principles and Practice*, S.J. Roberts and R.M. Everson, Eds., chapter 7. Cambridge University Press, 2001.

[18] R.A. Choudrey and S.J. Roberts, "Flexible Bayesian Independent Component Analysis for blind source separation," in *Proceedings of ICA2001*, 2001.

[19] D.J.C. MacKay, "Ensemble learning for hidden Markov models," Tech. Rep., University of Cambridge, 1997.

[20] I. Rezek and S.J. Roberts, "Variational inference for hidden Markov models," *Electronic Letters*, 2001.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 257-286, 1989.

[22] D.J.C. MacKay, "Developments in probabilistic modelling with neural networks - ensemble learning," in *Proceedings of the third Annual Symposium on Neural Networks*, Nijmegen, The Netherlands, 1995, pp. 191–198, Springer.

[23] H. Attias, "Learning parameters and structure of latent variable models by variational Bayes," in *Electronic Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-1999)*, http://www2.sis.pitt.edu/~dsl/UAI/uai.html, 1999, Association for Uncertainty in Artificial Intelligence (AUAI).

[24] R.A. Choudrey and S.J. Roberts, "Bayesian ICA with hidden Markov model sources," Tech. Rep., University of Oxford, 2002, Submitted to ICA2003.

[25] John Cassidy, *Dot.Con*, Penguin, 2002.