Some Aspects of Latent Structure Analysis

D.M. Titterington

University of Glasgow, Glasgow, Scotland, UK
mike@stats.gla.ac.uk

Abstract. Latent structure models involve real, potentially observable variables and latent, unobservable variables. The framework includes various particular types of model, such as factor analysis, latent class analysis, latent trait analysis, latent profile models, mixtures of factor analysers, state-space models and others. The simplest scenario, of a single discrete latent variable, includes finite mixture models, hidden Markov chain models and hidden Markov random field models. The paper gives a brief tutorial of the application of maximum likelihood and Bayesian approaches to the estimation of parameters within these models, emphasising especially the fact that computational complexity varies greatly among the different scenarios. In the case of a single discrete latent variable, the issue of assessing its cardinality is discussed. Techniques such as the EM algorithm, Markov chain Monte Carlo methods and variational approximations are mentioned.

1 Latent-variable Fundamentals

We begin by establishing notation. Let y denote observable data for an experimental unit, let z denote missing or otherwise unobservable data and let $x = (y, z)$ denote the corresponding complete data. Then probability functions (densities or mass functions according as the variables are discrete or continuous) are indicated as follows: $f(y, z)$ for y and z jointly, $g(y|z)$ for y conditional on z, and $f(y)$ and $h(z)$ as marginals for y and z respectively. These functions satisfy the relationships

$$\begin{aligned} f(y) &= \int f(y,z)dz \\ &= \int g(y|z)h(z)dz, \end{aligned}$$

where the integration is replaced by summation if z is discrete. Also of interest might be the other conditional probability function, $h(z|y)$, which can be expressed as

$$h(z|y) = f(y,z)/f(y) \propto g(y|z)h(z),$$

which is essentially an expression of Bayes' Theorem.

In latent-structure contexts, z represents latent variables, introduced to create flexible models, rather than 'real' items, although often real, physical interpretations are surmised for the latent variables, in the case of factor analysis, for example. If $y = (y_1, \ldots, y_p)$ is p-dimensional, corresponding to p observable characteristics, then Bartholomew [1] proposes that

$$g(y|z) = \prod_i g_i(y_i|z),$$

which ensures that it is the dependence of the y_i on the latent variables z that accounts fully for the mutual dependence among the y_i.

2 Simple particular cases

There are a number of simple particular cases, depending on the natures of the observed variables y and the latent variables z:

- the case of y continuous and z continuous corresponds to *factor analysis*;
- the case of y continuous and z discrete corresponds to *latent profile analysis* [2] or *cooperative vector quantisation* [3];
- the case of y discrete and z continuous corresponds to *latent trait analysis* [4] or *density networks* [5];
- the case of y discrete and z discrete corresponds to *latent class analysis* [6] or *naive Bayesian networks* [7].

It is noteworthy that, in three of the cases, two nomenclatures are given for the same structure, one from the statistics literature and one more prevalent in the machine learning/computer science literature; this emphasises the fact that these models are of common interest to these two research communities.

A number of general remarks can be made: in all but factor analysis z is usually univariate, although review and development of the case of multivariate z is provided by Dunmur and Titterington [8]; if however z is multivariate then its components are usually assumed to be independent; continuous variables are usually assumed to be Gaussian, especially within z; and discrete variables are usually categorical or binary, rather than numerical or ordinal. Other variations include *mixtures of factor-analysers*, in which y contains continuous variables and z includes some continuous variables and at least one categorical variable [9] [10].

We now present the above four special cases in the form of statistical models.

1. *Factor analysis*:

$$y = Wz + e,$$

where $z \sim N(0, I)$ and $e \sim N(0, \Lambda)$ with z and e independent and Λ diagonal. The dimensionality of z is normally less than p, the dimensionality of y.

2. *Latent profile analysis*:
$$y = WZ + e,$$
where Z is a matrix of indicators for multinomially generated z and e is as in factor analysis.

3. *Latent trait analysis*:
$$g(y_{is} = 1|z) \propto \exp(w_{0is} + w_{is}^{\top} z),$$
where $\{y_{is}\}$ are elements of an indicator vector representing the ith component of y. There are obvious constraints, in that we must have $\sum_s g(y_{is} = 1|z) = 1$ for each i. Also, for identifiability, constraints such as $w_{0i1} = 0, w_{i1} = 0$, for each i, must be imposed. Thus the conditional distributions associated with y given z correspond to linear logistic regression models.

4. *Latent class analysis*:
The conditional probabilities $g(y_{is} = 1|z = u)$ are multinomial probabilities that sum to 1 over s, for each i, and $h(z = u)$ are multinomial probabilities that sum to 1 over u.

3 Issues of Interest

At a general level there are two main issues, listed here in arguably the reverse of the correct order of fundamental importance!

- Estimation of the chosen model, i.e. of relevant parameters, according to some paradigm; we shall describe likelihood-based and Bayesian approaches. By 'parameters' we mean items such as W and Λ in the factor analysis model, and so on.

- Estimation or selection of the model structure itself, and in particular the appropriate level of complexity, in some sense, of z: the more complex the structure of z is, the more 'flexible' is the latent-structure model for y.

Inference will be based on representative data, possibly supplemented with 'prior' information. The data will consist of a number of realisations of the observables:
$$D_y = \{y^{(n)}, n = 1, \ldots, N\}.$$
Thus N represents the number of experimental units in the dataset and often corresponds to 'sample size' in statistical terminology. The fact that the data are 'incomplete', with the latent variables $D_z = \{z^{(n)}, n = 1, \ldots, N\}$ being unobservable, complicates matters, but so also might other aspects of the model

structure, as we shall see.

Suppose that $D_x = \{(y^{(n)}, z^{(n)}), n = 1, \ldots, N\}$ denotes the *complete data*, and that the total set of parameters is

$$\theta = (\phi, \eta),$$

where ϕ and η denote parameters within the models for $y|z$ and z respectively. (In this respect the factor analysis model is rather special, in that $\phi = (W, \Lambda)$ and there is no unknown η.) Then likelihood and Bayesian inference for θ should be based on the observed-data likelihood, defined by the marginal density associated with the observed data but regarded as a function of θ,

$$f(D_y|\theta) = \sum_{D_z} f(D_x|\theta), \tag{1}$$

where here we are assuming that the variables within z are discrete.

Note that Bayesian inference about θ should be based on the posterior density

$$p(\theta|D_y) \propto f(D_y|\theta)\ p(\theta),$$

where $p(\theta)$ is a prior density for θ.

Were D_z not latent but known, then the basis for inference would be the usually much simpler $f(D_x|\theta)$, in the case of likelihood inference, or the corresponding $p(\theta|D_x)$ if the Bayesian approach is being adopted. The marginalisation operation depicted in (1) typically creates quite a complicated object.

4 The Case of a Single Categorical Latent Variable

Suppose that each hidden $z^{(n)}$ is categorical, with K categories, denoted by $\{1, \ldots, K\}$, and that the conditional density of $y^{(n)}$, given that $z^{(n)} = k$, is the *component density* $g_k(y^{(n)}|\phi_k)$. Often Gaussian component densities are used, in which case ϕ_k contains the mean vector and covariance matrix of the kth component density.

As we have just seen, the function of key interest, as the observed-data likelihood, is the joint density for D_y. The complexity of this density will be dictated by the pattern of dependence, marginally, among the $y^{(n)}$'s. We shall assume that the different $y^{(n)}$'s are conditionally independent, given the $z^{(n)}$'s and that $y^{(n)}$ depends only on $z^{(n)}$ and not on any other part of D_z, so that

$$g(D_y|D_z, \phi) = \prod_n g(y^{(n)}|z^{(n)}, \phi);$$

we shall consider different possibilities for the dependence structure among the $z^{(n)}$'s.

1. *Case 1: $z^{(n)}$'s independent.* If the $z^{(n)}$'s are independent then so, marginally, are the $y^{(n)}$'s:

$$\begin{aligned} f(D_y|\theta) &= \sum_{D_z} f(D_y, D_z|\theta) \\ &= \sum_{D_z} g(D_y|D_z, \phi)h(D_z|\eta) \\ &= \sum_{D_z} \{\prod_{n=1}^{N} g(y^{(n)}|z^{(n)}, \phi) \prod_{n=1}^{N} h(z^{(n)}|\eta)\} \\ &= \prod_{n=1}^{N} \{\sum_{k=1}^{K} g_k(y^{(n)}|\phi_k)\eta_k\}, \end{aligned}$$

   where $\eta_k = \text{Prob}(z^{(n)} = k)$. In this case therefore the observed data constitute a sample of size N from a *mixture distribution*, which might otherwise be called a *hidden multinomial*. The mixture density is

$$f(y|\theta) = \sum_{k=1}^{K} g_k(y|\phi_k)\ \eta_k,$$

   and the $\{\eta_k\}$ are called the *mixing weights*.

2. *Case 2: $z^{(n)}$'s following a Markov chain.* In this case the $y^{(n)}$'s correspond to a *hidden Markov (chain) model*, much applied in economics and speech-modelling contexts. (The version with continuous (D_y, D_z) corresponds to state-space dynamic models.) In this case the computation of the observed-data likelihood $f(D_y|\theta)$ is more complicated, essentially because the summation operation cannot be dealt with so simply, but in principle $f(D_y|\theta)$ can be computed by a pass through the data

3. *Case 3: $z^{(n)}$'s following a Markov random field.* In this case the index set is typically two-dimensional, corresponding to a lattice of N grid points. The $y^{(n)}$'s correspond to a noisy/hidden Markov random field model popular in the statistical analysis of pixellated images [11]: here D_z represents the true scene and D_y a noise-corrupted but observable version thereof. This scenario includes simple versions of Boltzmann machines. The computation of the observed-data likelihood $f(D_y|\theta)$ is typically not a practical proposition.

The next two sections deal with inference paradigms, and we shall see that in each case the level of difficulty escalates as our attention moves from Case 1 to Case 2 to Case 3, as has just been mentioned in the context of the calculation of $f(D_y|\theta)$.

5 Maximum likelihood estimation

5.1 The EM Algorithm

The EM algorithm [12] is an iterative algorithm that aims to converge to maximum likelihood estimates in contexts involving incomplete data. From an initial approximation $\theta^{(0)}$, the algorithm generates a sequence $\{\theta^{(r)}\}$ using the following iterative double-step.

E-Step. Evaluate

$$Q(\theta) = E\{\log f(D_x|\theta)|D_y, \theta^{(r)}\}.$$

M-Step. Calculate

$$\theta^{(r+1)} = \arg\max_\theta Q(\theta).$$

Typically, $f(D_y|\theta^{(r+1)}) \geq f(D_y|\theta^{(r)})$, so that the likelihood of interest increases at each stage and, although there are exceptions to the rule, the algorithm converges to at least a local, if not a global, maximum of the likelihood. In summary, the E-Step calculates a (conditional) expectation of the corresponding complete-data loglikelihood function and the M-step maximises that function. The convenience of the algorithm depends on the ease with which the E-Step and M-Step can be carried out, and we discuss this briefly in the context of the three cases identified in the previous section.

So far as the M-Step is concerned, the level of difficulty is the same as that which applies in the corresponding complete-data context. In most mixture problems and most hidden Markov chain contexts this is easy, but in the context of hidden Markov random fields aspects of the M-Step are very difficult, as we shall shortly illustrate. The E-Step for these models amounts to the calculation of expectations of the components of the indicator variables D_z; these expectations are therefore probabilities of the various possible configurations for the latent states, given the observed data. Difficulties in the E-Step are usually caused by intractability of the distribution of $D_x|D_y, \theta^{(r)}$ or equivalently of $D_z|D_y, \theta^{(r)}$. As a result of the independence properties with mixture data, the distribution of $D_z|D_y, \theta^{(r)}$ becomes the product model corresponding to the marginal distributions for $z^{(n)}|y^{(n)}, \theta^{(r)}$, for each n, and the E-Step in this case is straightforward. For the hidden Markov chain case, the dependence among the $z^{(n)}$'s does complicate matters, but the dependence is Markovian and 'one-dimensional', and this leads to the E-Step being computable by a single forwards and then backwards pass through the data [13]. This case is therefore less trivial than for mixtures but is not a serious problem. In the hidden Markov random field case, however, the E-Step is dramatically more difficult; the dependence among the $z^{(n)}$'s is still Markovian, but is 'two-dimensional', and there is no simple analogue of the forwards-backwards algorithm. We illustrate the difficulties in both steps with the simplest example of a hidden Markov random field.

Example. Hidden Ising model.

Suppose the index set for D_z is a two-dimensional lattice and that

$$h(D_z) = h(D_z|\eta) = \{G(\eta)\}^{-1}\exp\{\eta \sum_{s\sim t} z^{(s)}z^{(t)}\},$$

where each $z^{(n)} \in \{-1, +1\}$, so that each hidden variable is binary, η is a scalar parameter, usually positive so as to reflect local spatial association, and the summation is over (s,t) combinations of locations that are immediate vertical or horizontal neighbours of each other on the lattice; this constitues the Ising model of statistical physics. In this model the normalising constant $G(\eta)$ is not computable. This leads to there being no analytical form for the E-step and no easy M-Step for η. For example, as mentioned earlier, the degree of difficulty of the M-Step for η is the same as that of complete-data maximum likelihood, and for the latter one would have to maximise $h(D_z|\eta)$ with respect to η, which is stymied by the complexity of $G(\eta)$.

What can be done if the EM algorithm becomes impracticable? A number of possible approaches exist.

- With the Law of Large Numbers in mind, replace the E-Step by an appropriate sample mean calculated from realisations of the conditional distribution of . However, in the context of the hidden Ising model, simulation from the relevant conditional distribution is not straightforward and itself requires an iterative algorithm of the Markov chain Monte Carlo type.

- Replace the complicated conditional distribution in the E-Step by a deterministic, simpler approximation, e.g. a variational approximation. We give more details of this in the next subsection, concentrating on it because of its prominence in the recent computer science literature.

- Use variational (or other) approximations in the M-Step.

- Other suggestions exist including the consideration of methods other than EM. For example, Younes [14] developed a gradient-based stochastic approximation method and Geyer and Thompson [15] used Monte Carlo methods to approximate the intractable normalisation constant and thereby attack the likelihood function directly. In [16] a number of more ad hoc methods are suggested for the image-analysis context, based on iterative restoration of the true scene, perhaps using Besag's [17] Iterative Conditional Modes algorithm, alternated with parameter estimation with the help of Besag's *pseudolikelihood* [18] for the parameter η. The simplest form of the pseudolikelihood is the product of the full conditional densities for the individual $z^{(n)}$'s, given the rest of D_z. It is much easier to handle than the original $h(D_z|\eta)$ because there is no intractable normalisation constant, and yet the maximiser of the pseudolikelihood is generally a consistent estimator of the true η.

5.2 Variational Approximations

Suppose that $h(D_z)$ is a complicated multivariate distribution, and that $q(D_z)$ is a proposed tractable approximation to $h(D_z)$ with a specified structure. Then one way of defining an optimal q of that structure is to minimise an appropriate measure of distance between q and h, such as the Kullback-Leibler(KL) divergence,

$$\mathrm{KL}(q,h) = \sum_{D_z} q(D_z) \log \{q(D_z)/h(D_z)\}.$$

Often the form of q is determined by the solution of this variational optimisation exercise, as are the values of (variational) hyperparameters that q contains. The simplest model for q would be an independence model, i.e.

$$q(D_z) = \prod_n q_n(z^{(n)}),$$

which leads to *mean field approximations*. Furthermore, variational approximations to the conditional distribution of D_z given D_y lead to lower bounds on the observed-data loglikelihood. To see this note that

$$\begin{aligned} \log f(D_y|\theta) &= \log \{\sum_{D_z} f(D_y, D_z|\theta)\} \\ &\geq \sum_{D_z} q(D_z) \log \{f(D_y, D_z|\theta)/q(D_z)\}, \end{aligned}$$

by Jensen's inequality. Typically, q is chosen to have a structure such that the summation in the lower bound is easily achieved; the choice of a fully factorised q is certainly advantageous in this respect, but it represents a simplifying approximation whose consequences should be investigated. It is straightforward to show that the q, of any prescribed structure, that maximises the above lower bound for $\log f(D_y|\theta)$, minimises the KL divergence between q and the conditional distribution for D_z given D_y and θ. An EM-like algorithm can be evolved in which q and θ are successively updated in the equivalents of the E-Step and M-Step respectively. Convergence of the resulting sequence of iterates for θ, to a local maximum of the lower-bound surface, can be proved, but there is comparatively little theory about the relationship of such a maximum to the maximiser of the observed-data likelihood itself, although some progress is reported in [19]. (If one can show that the maximiser of the lower-bound function tends to the maximiser of the likelihood, asymptotically, then one can claim that the lower-bound maximiser inherits themaximum likelihood estimator's property of being consistent for the true value of θ.)

The practicality of variational approximations relies on the computability of the lower-bound function, and much of the relevant literature is restricted to the case of a fully factorised q_{D_z}. However, more-refined approximations can be developed in some contexts [20] [21]. It would also be of obvious value to obtain

corresponding upper bounds for $\log f(D_y|\theta)$, but they are much harder to come by and a general method for deriving them is as yet elusive.

Wainwright and Jordan [22] take a more general convex analysis approach to defining variational approximations, although operational versions of the method usually amount to optimising a KL divergence.

For an application of these ideas to latent profile analysis see [8], and for a tutorial introduction to variational approximations see [23]. In earlier work, Zhang [24] [25] used mean-field-type approximations within the EM-algorithm in contexts such as image restoration based on underlying Markov random field models.

6 The Bayesian Approach

6.1 Introduction

As already stated, Bayesian inference for the parameters in a model is based on

$$p(\theta|D_y) \propto f(D_y|\theta)\ p(\theta) = \{\sum_{D_z} f(D_x|\theta)\}\ p(\theta),$$

where $p(\theta)$ is a prior density for θ. As with maximum likelihood, Bayesian analysis is often vastly easier if D_z is not missing, in which case we use

$$p(\theta|D_x) \propto f(D_x|\theta)\ p(\theta).$$

In many familiar cases, $f(D_x|\theta)$ corresponds to an exponential family model and then there exists a family of neat closed-form *conjugate* priors for θ; the posterior density $p(\theta|D_x)$ then belongs to the same conjugate family, with hyperparameters that are easily written down; see for example Section 3.3 of [26]. This convenient pattern disappears if there are missing data, in this represented by D_z. Inevitably non-exact methods are then needed, and most common approaches can be categorised either as asymptotically exact but potentially unwieldy simulation methods, usually Markov chain Monte Carlo (MCMC) methods, or as non-exact but less unwieldy deterministic approximations. The latter include Laplace approximations [27] and variational Bayes approximations; we shall concentrate on the latter, again because of its high profile in recent machine-learning literature.

Both these approaches aim to approximate

$$p(\theta, D_z|D_y) \propto f(D_x|\theta)\ p(\theta),$$

the joint posterior density of all unknown items, including the latent variables as well as the parameters. Marginalisation then provides an approximation to $p(\theta|D_y)$.

6.2 The MCMC Simulation Approach

This method aims to generate a set of simulated realisations from $p(\theta, D_z|D_y)$. The resulting set of realisations of θ then form a sample from $p(\theta|D_y)$, and, for example, the posterior mean of θ can be approximated by the empirical average of the realisations of θ.

The now-standard approach to this is to generate a sequence of values of the variables of interest from a Markov chain for which the equilibrium distribution is $p(\theta, D_z|D_y)$. Once the equilibrium state has been reached, a realisation from the required distirubtion has been generated. There are various general recipes for formulating such a Markov chain, one of the simplest being *Gibbs sampling*. This involves recursively sampling from the appropriate set of full conditional distributions of the unknown items. A 'block' version of this for our problem would involve iteratively simulating from $p(\theta|D_z, D_y)$ and $p(D_z|\theta, D_y)$. This is typically easy for mixtures [28] and hidden Markov chains [29] but is problematic with hidden Markov random fields [30]. As with maximum likelihood, the intractability of the normalising constant is the major source of the difficulty.

The development and application of MCMC methods in Bayesian statistics has caught on spectacularly during the quarter-century since the publication of papers such as [31]; see for example [32] [33]. However, important issues remain that are the subject of much current work, including the monitoring of convergence, difficulties with large-scale problems, the invention of new samplers, and the search for perfect samplers for which convergence at a specified stage can be guaranteed. A variety of approximate MCMC approaches are described and compared in [34].

6.3 Variational Bayes Approximations

Suppose that $q(\theta, D_z)$ defines an approximation to $p(\theta, D_z|D_y)$ and suppose we *propose* that q take the factorised form

$$q(\theta, D_z) = q_\theta(\theta) q_{D_z}(D_z).$$

Then the factors are chosen to minimise

$$\mathrm{KL}(q, p) = \int_\theta \sum_{D_z} q \log (q/p),$$

in which p is an abbreviation for $p(\theta, D_z|D_y)$. The resulting $q_\theta(\theta)$ is regarded as the approximation to $p(\theta|D_y)$.

The form of q_θ is often the same as that which would result in the complete-data case: if a prior $p(\theta)$ is chosen that is conjugate for the Bayesian analysis of the complete data, D_x, then q_θ also takes that convenient conjugate form but calculation of the (hyper)parameters within q_θ requires the solution of nonlinear

equations.

Once again, the analysis for mixtures and hidden Markov chains is comparatively 'easy', but the case of hidden Markov random fields is 'hard'; see for example [35] [36] [37] [38]. As in the case of likelihood-based variational approximations, there is not much work on the theoretical properties of the method. However, in a number of scenarios, including Gaussian mixture distributions, Wang and Titterington have shown that the variational posterior mean is consistent, see for example [39], but they have also shown that the variational posterior variances can be unrealistically small, see for example [40].

For a review of variational and other approaches to Bayesian analysis in models of this general type see [41].

7 A Brief Discussion of Model Selection

7.1 Non-Bayesian Approaches

We continue to concentrate on the case of a single categorical latent variable, and especially on mixture models. The model-selection issue of interest will be the determination of an appropriate number K of components to be included in the model. General non-Bayesian approaches include the following:

- selection of a parsimonious model, i.e. the minimum plausible K, by hypothesis-testing;
- optimisation of criteria such as Akaike's AIC [42] and Schwarz's BIC [43].

However, it is well known that standard likelihood-ratio theory for nested models, based on the use of chi-squared distributions for testing hypotheses, breaks down with mixture models. Various theoretical and practical directions have been followed for trying to overcome this; among the latter is McLachlan's [44] use of bootstrap tests for mixtures.

The criteria AIC and BIC impose penalties on the maximised likelihood to penalise highly-parameterised models, so it is plausible that these instruments should select a suitable value of K. This is by no means guaranteed, especially for AIC, but, asymptotically at least, there are results showing that BIC selects the right number of mixture components in at least some cases [45]. It should be pointed out that AIC was developed with scenarios in mind in which, unlike in the case of mixtures, there is the same sort of 'regularity' as is required for standard likelihood-ratio theory.

7.2 Bayesian Approaches

We consider the following three approaches:

– model comparison using Bayes factors;

– use of Bayesian-based selection criteria (DIC);

– generation of a posterior distribution for the cardinality of the latent space.

Approach 1: Bayes factors.

Suppose θ_k represents the parameters present in Model M_k, and that $\{p(M_k)\}$ are a set of prior probabilities over the set of possible models. Then the ratio of posterior probabilities for two competing models M_k and $M_{k'}$ is given by

$$\frac{p(M_k|D_y)}{p(M_{k'}|D_y)} = \frac{f(D_y|M_k)}{f(D_y|M_{k'})} \times \frac{p(M_k)}{p(M_{k'})},$$

where the first ratio on the right-hand side is the *Bayes factor*, and

$$f(D_y|M_k) = \int f(D_y|\theta_k)d\theta_k,$$

in which θ_k are the parameters corresponding to model M_k. A 'first-choice' model would one for which the posterior probability $p(M_k|D_y)$ is maximum. Clearly the calculation of the Bayes factor is complicated in incomplete-data scenarios. For an authoritative account of Bayes factors see [46].

Approach 2: The Deviance Information Criterion

This can be described as a Bayesian version of AIC. First we define 'Deviance' by

$$\Delta(\theta) = -2 \log \{f(D_y|\theta\},$$

and define

$$\begin{aligned} \overline{\Delta(\theta)} &= E_{p(\theta|D_y)}\Delta(\theta) \\ \bar{\theta} &= E_{p(\theta|D_y)}\theta \\ p_\Delta &= \overline{\Delta(\theta)} - \Delta(\bar{\theta}). \end{aligned}$$

Then DIC is defined by

$$\text{DIC} = \overline{\Delta(\theta)} + p_\Delta,$$

clearly to be minimised if model selection is the aim. The method was introduced and implemented in a variety of contexts, mainly involving complete data from generalised linear models, in [47]. A number of somewhat ad hoc adaptations for incomplete-data contexts were proposed and compared, in particular in mixture models, in [48], and C.A. McGrory's Glasgow Thesis [38] will report on the application of variational approximations in the context of DIC, using mixture models, hidden Markov chains and hidden Markov random fields as testbeds.

Approach 3 : Generation of $\{p(k|D_y)\}$.

As remarked when we were discussing Bayes factors, 'exact' computation of $\{p(k|D_y)\}$ is very difficult in incomplete-data contexts, so instead attempts have been made to assess it using MCMC. Two strands have been developed.

1. *Reversible Jump MCMC* [49] . As in 'ordinary' MCMC, a Markov chain is generated that is designed to have, as equilibrium distribution, the posterior distribution of all unknown quantities, *including* k. Therefore, since the value of k can change during the procedure, there is the need to jump (reversibly) between parameter spaces of different dimensions, and this creates special problems. (The reversibility property is required in order to guarantee convergence of the Markov chain Monte Carlo procedure to the desired equilibrium distribution.) Of the problems covered in the present paper, mixtures are dealt with in [50], hidden Markov chains in [51], and some comparatively small-scale spatial problems in [52].
2. *Birth and Death MCMC* [53]. This can be thought of as a 'continuous-time' alternative to reversible jump MCMC. The key feature is the modelling of the mixture components as a marked birth-and-death process, with components being added (birth) of being discarded (death) according to a random process, with 'marks' represented by the parameters of the component distributions.

Cappé et al. [54] showed that the two approaches could be linked and generalised.

8 Acknowledgement

This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the author's views. The paper was prepared for a PASCAL workshop in Bohinj, Slovenia, in February, 2005.

References

1. Bartholomew, D.J.: The foundations of factor analysis. Biometrika **71** (1984) 221–232.
2. Gibson, W.A.: Three multivariate models: factor analysis, latent structure analysis and latent profile analysis. Psychometrika **24** (1959) 229–252.
3. Ghahramani, Z.: Factorial learning and the EM algorithm. In Advances in Neural Information Processing Systems **7** (eds. G. Tesauro, D.S. Touretzky and T.K. Leen). (1996) MIT Press, Cambridge MA.
4. Bartholomew, D.J.: Latent Variable Models and Factor Analysis. (1987) Griffin, London.
5. MacKay, D.J.C.: Bayesian neural networks and density networks. Instr. Meth. Phys. Res. A **354** (1995) 73–80.
6. Hagenaars, J.A.: Categorical Longitudinal Data. (1990) Sage, London.
7. Neal, R.M.: Probabilistic inference using Markov chain Monte Carlo methods. (1993) Tech. Report CRG-TR-93-1, Dept. Comp. Sci., Univ. Toronto.
8. Dunmur, A.P., Titterington, D.M.: Analysis of latent structure models with multidimensional latent variables. In Statistics and Neural Networks: Recent Advances at the Interface (eds. J.W. Kay and D.M. Titterington) (1999) 165–194. Oxford: Oxford University Press.

9. Ghahramani, Z., Beal, M.: Variational inference for Bayesian mixtures of factor analyzers. In Advances in Neural Information Processing, Vol.12 (eds., S.A. Solla, T.K. Leen and K.-R. Müller) (2000) 449–455. MIT Press, Cambridge, MA.
10. Fokoué, E., Titterington, D.M.: Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. Machine Learning **50** (2003) 73–94.
11. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intell. **6** (1984) 721–741.
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statist. Soc. B **39** (1977) 1–38.
13. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77** (1989) 257–285.
14. Younes, L.: Parameter estimation for imperfectly observed Gibbsian fields. Prob. Theory Rel. Fields **82** (1989) 625–645.
15. Geyer, C.J., Thompson, E.A.: Constrained Monte Carlo maximum likelihood for dependent data (with discussion). J.R. Statist. Soc. B **54** (1992) 657–699.
16. Qian, W., Titterington, D.M.: Estimation of parameters in hidden Markov models. Phil. Trans. R. Soc. Lond. A **337** (1991) 407–428.
17. Besag, J.E.: On the statistical analysis of dirty pictures (with discussion). J.R. Statist. Soc. B **48** (1986) 259–302.
18. Besag, J.E.: Statistical analysis of non-lattice data. The Statistician **24** (1975) 179–195.
19. Hall, P., Humphreys, K., Titterington, D.M.: On the adequacy of variational lower bounds for likelihood-based inference in Markovian models with missing values. J. R. Statist. Soc. B **64** (2002) 549–564.
20. Bishop, C.M., Lawrence, N., Jaakkola, T.S., Jordan, M.I.: Approximating posterior distributions in belief networks using mixtures. In Advances in Neural Information Processing Systems, Vol. 10 (eds. M.I. Jordan, M.J. Kearns and S.A. Solla) (1998) 416–422. MIT Press, Cambridge, MA.
21. Humphreys, K., Titterington, D.M.: Improving the mean field approximation in belief networks using Bahadur's reparameterisation of the multivariate binary distribution. Neural Processing Lett. **12** (2000) 183–197.
22. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational approximations. Technical Report 649, (2003) Dept. Statistics, Univ. California, Berkeley.
23. Jordan, M.I., Gharamani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. In Learning in Graphical Models (ed. M. Jordan) (1999) 105–162. MIT Press, Cambridge, MA.
24. Zhang, J.: The Mean Field Theory in EM procedures for Markov random fields. IEEE Trans. Signal Processing **40** (1992) 2570–2583.
25. Zhang, J.: The Mean Field Theory in EM procedures for blind Markov random field image restoration. IEEE Trans. Image Processing **2** (1993) 27–40.
26. Robert, C.P.: The Bayesian Choice, 2nd ed. (2001) Springer.
27. Tierney, L., Kadane, J.B.: Accurate approximations to posterior moments and marginal densities. J. Amer. Statist. Assoc. **81** (1986) 82–86.
28. Diebolt, J., Robert, C.P.: Estimation of finite mixture distributions through Bayesian sampling. J.R. Statist. Soc. B **56** (1994) 363–375.
29. Robert, C.P., Celeux, G., Diebolt, J.: Bayesian estimation of hidden Markov chains: a stochastic implementation. Statist. Prob. Lett. **16** (1993) 77–83.
30. Rydén, T., Titterington, D.M.: Computational Bayesian analysis of hidden Markov models. J. Comp. Graph. Statist. **7** (1998) 194–211.

31. Gelfand, A.E., Smith, A.F.M.: Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. **85** (1990) 398–409.
32. Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.): Markov Chain Monte Carlo in Practice. Chapman and Hall.
33. Doucet, A., de Freitas, N., Gordon, N. (eds.): Sequential Monte Carlo Methods in Practice. Springer.
34. Murray, I., Ghahramani, Z.: Bayesian learning in undirected graphical models: approximate MCMC algorithms. In Proc. 20th Conf. Uncertainty in Artificial Intell. (eds. M. Chickering and J. Halperin) (2004) 577–584. AUAI Press.
35. Corduneanu, A., Bishop, C.M.: Variational Bayesian model selection for mixture distributions. In Proc. 8th Int. Conf. Artific. Intell. Statist. (eds. T. Richardson and T. Jaakkola) (2001) 27–34. Morgan Kaufmann, San Mateo, CA.
36. Ueda, N., Ghahramani, Z.: Bayesian model search for mixture models based on optimizing variational bounds. Neural Networks **15** (2003) 1223–1241.
37. MacKay, D.J.C.: Ensemble learning for hidden Markov models. (1997) Technical Report, Cavendish Lab., Univ. Cambridge.
38. McGrory, C.A.: Ph.D. Dissertation (2005) Dept. Statist., Univ. Glasgow.
39. Wang, B., Titterington, D.M.: Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Analysis **1** (2006) to appear.
40. Wang, B., Titterington, D.M.: Variational Bayes estimation of mixing coefficients. In Proceedings of a Workshop on Statistical Learning. Lecture Notes in Computer Science Vol. ??, (ed. J. Winkler) (2005) pp. ??. Springer.
41. Titterington, D.M.: Bayesian methods for neural networks and related models. Statist. Sci. **19** (2004) 128–139.
42. Akaike, H,: Information theory and an extension of the maximum likelihood principle. In Proc. 2nd Int. Symp. Info. Theory (eds. B.N. Petrov and F. Csaki) (1973) 267–281. Budapest: Akadémiai Kiadó.
43. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6** (1978) 461–466.
44. McLachlan, G.J.: On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. Appl. Statist. **36** (1987) 318–324.
45. Keribin, C.: Consistent estimation of the order of mixture models. Sankhya A **62** (2000) 49–66.
46. Kass, R.E., Raftery, A.: Bayes factors. J. Amer. Statist. Assoc. **90** (1995) 773–795.
47. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of complexity and fit (with discussion). J. R. Statist. Soc. B **64** (2002) 583–639.
48. Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M.: Deviation information criteria for missing data models. (2005) Submitted.
49. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82** (1995) 711–732.
50. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Statist. Soc. B **59** (1997) 731–792.
51. Robert, C.P., Rydén, T., Titterington, D.M.: Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. J. R. Statist. Soc. B **62** (2000) 57–75.
52. Green, P.J., Richardson, S.: Hidden Markov models and disease mapping. J. Amer. Statist. Assoc. **97** (2002) 1055–1070.
53. Stephens, M.: Bayesian analysis of mixtures with an unknown number of components - an alternative to reversible jump methods. Ann. Statist. **28** (2000) 40–74.

54. Cappé, O., Robert, C.P., Rydén, T.: Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo. J. R. Statist. Soc. B **65** (2003) 679–699.