Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 7 November 2005 (MN LaTeX style file v2.2)

A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra

Louisa A. Nolan[1], Markus O. Harva[2], Ata Kabán[3] & Somak Raychaudhury[1]
[1]*School of Physics and Astronomy, University of Birmingham, Birmingham, B15 2TT, UK*
[2]*Laboratory of Computer and Information Science, Helsinki University of Technology, Helsinki, P.O. Box 5400, FI-02015 HUT, Finland*
[3]*School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK*

Submitted for publication in MNRAS

ABSTRACT
Efficient predictive models and data analysis techniques for the analysis of photometric and spectroscopic observations of galaxies are not only desirable, but required, in view of the overwhelming quantities of data becoming available. We present the results of a novel application of Bayesian latent variable modelling techniques, where we have formulated a data-driven algorithm that allows one to explore the stellar populations of a large sample of galaxies from their spectra, without the application of detailed physical models. Our only assumption is that the galaxy spectrum can be expressed as a linear superposition of a small number of independent factors, each a spectrum of a stellar sub-population that cannot be individually observed. A probabilistic latent variable architecture that explicitly encodes this assumption is then formulated, and a rigorous Bayesian methodology is employed for solving the inverse modelling problem from the available data. A powerful aspect of this method is that it formulates a density model of the spectra, based on which we can handle observational errors. Further, we can recover missing data both from the original set of spectra which might have incomplete spectral coverage of each galaxy, or from previously unseen spectra of the same kind.

We apply this method to a sample of 21 ultraviolet-to-optical spectra of well-studied early-type galaxies, for which we also derive detailed physical models of star formation history (i.e. age, metallicity and relative mass fraction of the component stellar populations). We also apply it to synthetic spectra made up of two stellar populations, spanning a large range of parameters. We apply four different data models, starting from a formulation of principal components analysis (PCA), which has been widely used. We explore alternative factor models, relaxing the physically unrealistic assumption of Gaussian factors, as well as constraining the possibility of negative flux values that are allowed in PCA, and show that other models perform equally well or better, while yielding more physically acceptable results. In particular, the more physically motivated assumptions of our rectified factor analysis enable it to perform better than PCA, and to recover physically meaningful results.

We find that our data-driven Bayesian modelling allows us to identify those early-type galaxies that contain a significant stellar population that is $\lesssim 1$ Gyr old. This experiment also concludes that our sample of early-type spectra showed no evidence of more than two major stellar populations differing significantly in age and metallicity. This method will help us to search for such young populations in a large ensemble of spectra of early-type galaxies, without fitting detailed models, and thereby to study the underlying physical processes governing the formation and evolution of early-type galaxies, particularly those leading to the suppression of star formation in dense environments. In particular, this method would be a very useful tool for automatically discovering various interesting sub-classes of galaxies, for example post-starburst, or E+A galaxies.

Key words: galaxies: elliptical and lenticular, cD – galaxies: stellar content – methods: data analysis

1 INTRODUCTION

The determination of the star formation history of early-type galaxies has important implications for the still-controversial issue of their formation and evolution. Are the bulk of their stars assembled at high redshift, followed by predominantly passive evolution, or are they formed from low-redshift disk-disk merging? Or, as it seems likely from the variety of detailed elliptical morphology which is being revealed via projects such as those involving integral field spectrographs such as SAURON, does the class of ellipticals contain sub-classes, and are these dependent on the method of their formation, or on their local environment (Bacon *et al.* 2001; Emsellem *et al.* 2004)? We would like to know how these sub-classes depend on local environment. In particular, we wish to know what mechanisms are responsible for the now well-known morphology-density relationship (*e.g.* Dressler 1980), which shows that the proportion of early-type galaxies increases as the proportion of late-type galaxies decreases in regions with higher galaxy density. Furthermore, the recent interest in E+A galaxies (*e.g.* Tran *et al.* 2003), in which the signature of a recent ($\lesssim 1$ Gyr) starburst is seen, but no significant ongoing star formation, suggests that the relation between the evolution of these galaxies and their constituent stellar populations is not well-understood.

However, until recently, only small ensembles of galaxy spectra, spanning a large useful range, were available and the analysis of a large statistical sample of stellar populations of galaxies was not possible. With the recently-completed 2dF Galaxy Redshift Survey (2dFGRS) and the ongoing Sloan Digital Sky Survey (SDSS) or 6dF Galaxy Survey (6dFGS), a few million galaxy spectra will soon be available. The detailed physical modelling of stellar populations is numerically expensive, so the timely extraction of useful knowledge, such as the characteristics of the star formation history of galaxies, from these data will largely depend on developing appropriate data analysis tools that are able to complement more specialised astrophysical studies, in particular by automating parts of the analyses. Some work has already been carried out using an approach which incorporates assumptions from physical models to a greater or lesser extent (*e.g.* Heavens *et al.* 2000; Kauffmann *et al.* 2003; Fernandes *et al.* 2005; Ocvirk *et al.* 2005a,b; Solorio *et al.* 2005).

With the increase of data availability, significant research has recently been focusing on the use of data-driven methods of astrophysical data analysis. Some of these attempt to analyse galaxy properties from spectra, independently from any physical model. The usefulness of the application of Principal Component Analysis (PCA), a classical multivariate statistical analysis technique, has been exploited and demonstrated in various scenarios (see, *e.g.* Connolly *et al.* 1995; Folkes, Lahav & Maddox 1996; Ronen *et al.* 1999; Madgwick *et al.* 2003a,b). Indeed, PCA is an efficient tool for data compression, and as such, suitable for providing a compressed representation for subsequent physical analysis. In a different line of research, the automated partitioning of spectra based on their morphological shapes (Slonim, Somerville & Lahav 2001) has also been explored with reasonable success.

Here, we explore the application of a similar technique to a different problem, namely that of the data-driven identification of the spectra of distinct stellar sub-populations in the spectrum of a galaxy. Our only assumption is that the galaxy spectrum can be expressed as a linear superposition of a small number of independent factors, each a spectrum of a stellar sub-population that cannot be individually observed. We would like to implement an optimal feature transformation that would reveal interpretable and meaningful structural properties (factors) from a large sample of galaxy spectra. This would allow fast, automated extraction of some key physical properties from potentially very large data sets of the kind the Virtual Observatory would offer.

Preliminary results based on a simple non-orthogonal projection method have indicated that an identification of interpretable component spectra may be achievable in a purely data-driven manner (Kabán, Nolan & Raychaudhury 2005). In this work, we follow a more rigorous methodology, explicitly encoding the above hypothesis into appropriate probabilistic data models. We then employ a Bayesian formalism to solve the inverse modelling problem, as well as to objectively assess to what extent such a hypothesis is supported by the data, in terms of both physical interpretability and the ability for predictive generalisation to previously unseen data of the same type.

The Bayesian framework allows us to automate several functional requirements within a single statistically sound formal framework, where all modelling assumptions may be explicitly encoded. Tasks ranging from inference and prediction to inverse modelling and data explanation, data compression, as well as model comparison are all accomplished in a flexible manner, even in conditions of missing values and measurement uncertainties.

In this study we start from a set of synthetic spectra of stellar populations of known age and metallicity from the single stellar population models of Jimenez *et al.* (2004). We also have assembled a set of observed ultraviolet (UV)-optical spectra of early-type galaxies, to which we have fit two component stellar population models, and thus we have some knowledge of their star formation history (i.e. ages, metallicities and mass fractions of the component stellar populations). We then perform our data-driven Bayesian analysis on both synthetic and observed spectra, and compare the derived parameters with those obtained from detailed astrophysical modelling.

In §2, we describe the set of 21 observed early-type galaxy spectra that we analyse here, and describe the synthetic stellar population spectra. §3 details the Bayesian modelling of the data. Our results are presented in and discussed in §4. In §5 we discuss the significance of this work on characterising E+A galaxies. Our conclusions are outlined in §6. Appendix A gives the detailed mathematical framework required to implement our methodology, specifically for the data model estimation of the variational Bayesian rectified factor analysis — a flexible data model for the factorisation of positive data that we developed.

2 THE DATA

2.1 Synthetic stellar population models

We perform the linear independent transformation analyses on a set of two-component model stellar population spectra. We use the single stellar population models of Jimenez *et al.* (2004), which have a Salpeter initial mass function, across the same wavelength range (2500−8000 Å) as the observed spectra, and create a two-component model flux, $F_{2pop,\lambda}$. This is defined as

$$F_{2pop,\lambda} \propto \left(m_i f_{Z_i,\lambda,t_i} + m_j f_{Z_j,\lambda,t_j}\right) \quad (1)$$

where $F_{2pop,\lambda,t}$ is the model flux per unit wavelength in the bin centred on wavelength λ, which is the sum of the two single-metallicity, single-age model fluxes f_{Z_i,λ,t_i} and f_{Z_j,λ,t_j}, which have metallicities and ages, Z_i, t_i and Z_j, t_j respectively. Here m_i, m_j are the fractional contributions (by stellar mass) of each component to the total model population. We set $m_i + m_j = 1$, where m_i can take values 0.25 or 0.5. Ages assigned for t_i, t_j are 0.03, 1, 3, 7 and 12 Gyr, and the metallicities are 0.004, 0.01, 0.02, 0.03 and 0.05, making a total of 50 two-component model spectra.

2.2 Observed early-type galaxy spectra

We have also compiled a set of twenty-one early-type galaxy spectra from the archives for this study, across the wavelength range 2500−8000 Å, although not all objects have data covering the complete range. The spectra of the galaxies in the sample are shown in Fig. 1. As the galaxy sample was assembled from those nearby early-type galaxies with both UV and optical spectra, from various archives, it does not represent a statistical sample. The sources for the data are listed in Table 1, together with the telescopes with which the optical spectra were observed. The UV spectra were observed using the *International Ultraviolet Explorer* (IUE), except for NGC 3605 and NGC 5018, which were observed with the *Faint Object Spectrograph* (FOS) on the *Hubble Space Telescope* (HST).

Since these spectra are compiled from sources with varying spectral coverage, the resulting data matrix has some missing values. The regions of incomplete wavelength coverage are obvious in Fig. 1. It is one of the strengths of our analysis that these missing data regions can be recovered; this is described in §§3.1.3, 3.4, 3.5. Another of the advantages of the probabilistic framework adopted here is that, in contrast with classic non-probabilistic methods, the observational errors on each data point can be taken into account (§3.1.2).

As the data matrix on which the analysis is performed must have the same wavelength binning for all the galaxies, the spectra are de-redshifted and binned to the same wavelength resolution as the synthetic stellar population spectra ($\sim$20 Å; the observed spectra have wavelength resolution which varies from $\sim$5-18 Å). For each galaxy, the various spectral sections, taken in turn, are normalised to unity in the overlap regions, and spliced, to create a single continuous spectrum for each object. Although the different spectral pieces for any one object were observed separately, and with different telescopes, it should be noted that the apertures used were similar, and therefore each section probes a similar part of each galaxy, and the shape of the continua in the overlap regions agree well.

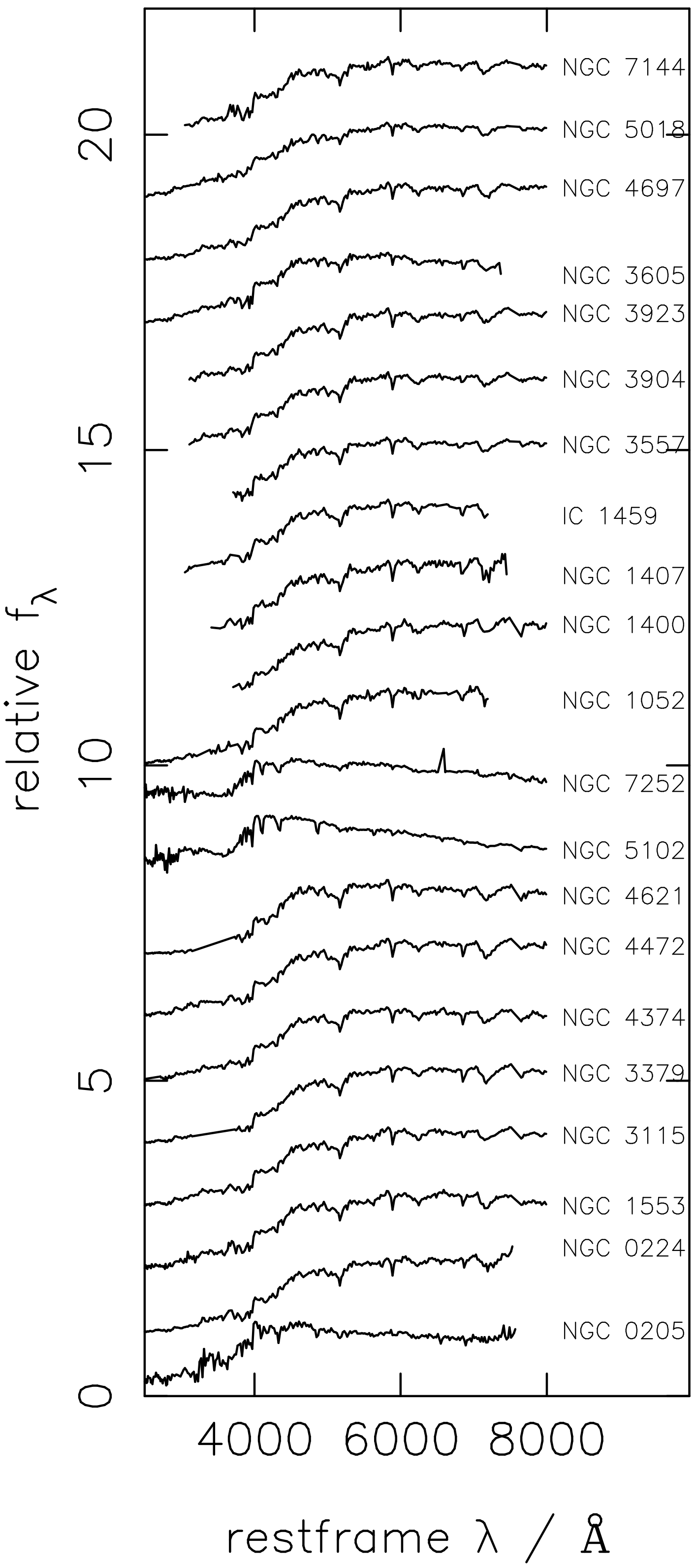

Figure 1. Rest-frame spectra of the 21 galaxies in our sample. See Table 1 and §2 for details. The dotted lines mark some of the absorption features in the spectra which are typically strong in young stellar populations, and the dashed lines mark some of the absorption features which are typically strong in old, metal-rich stellar populations. From left to right, the absorption line species are: MgII (2799 Å), Hε (3970 Å), Hδ (4102 Å), Hγ (4340 Å), Hβ (4861 Å), Mgb (5175 Å), NaD (5893 Å), Hα (6563 Å), TiO (7126 Å). The spectra are arbitrarily shifted along the flux axis for the sake of clarity.

3 DATA MODELLING

Models are simplifications of reality representing relevant aspects while glossing over irrelevant details. Data models are constructed from experimental evidence (data) and available prior knowledge. Once a data model is estimated (which can be done off-line), it offers a tool for making inferences and predictions in a consistent framework.

In Bayesian data modelling, all information is encoded in probability distributions (Bernardo & Smith 1994). The joint probability distribution of the observed data and other unobservable variables of interest (such as factors or component stellar populations of the representation as well as the parameters of the mixing process, etc.) need to be designed. The unobservable variables in the model are often referred to as hidden or latent variables.

3.1 Designing and building an appropriate latent variable architecture

3.1.1 The independent factor representation model of the spectrum

The spectrum of a galaxy or of a stellar sub-population is essentially a vector, over a binned wavelength range, representing the

Table 1. Table of references and optical telescopes for the observed early-type galaxy spectra.

Object	Reference	Optical telescope
NGC 0205	S-BCK	0.9m, Kitt Peak National Observatory
NGC 0224	S-BCK	0.9m, Kitt Peak National Observatory
NGC 1052	S-BCK	0.9m, Kitt Peak National Observatory
NGC 1400	Q	ESO 1.52m, La Silla
NGC 1407	S-BCK	0.9m, Kitt Peak National Observatory
IC 1459	R,Q	ESO 2.2m, 1.52m, La Silla
NGC 1553	R,Q	ESO 2.2m, 1.52m, La Silla
NGC 3115	R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 3379	Q	ESO 1.52m, La Silla
NGC 3557	Q,B	ESO 2.2m, 1.52m, La Silla
NGC 3605	P,J	F.L. Whipple Observatory, 1.5m
NGC 3904	R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 3923	R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 4374	R,Q	ESO 2.2m, 1.52m, La Silla
NGC 4472	R,Q	ESO 2.2m, 1.52m, La Silla
NGC 4621	Q	ESO 1.52m, La Silla
NGC 4697	R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 5018	P,R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 5102	R,Q,B	ESO 2.2m, 1.52m, La Silla
NGC 7144	R Q B	ESO 2.2m, 1.52m, La Silla
NGC 7252	V	1.93m, Haute Provence Observatory

The UV spectra were observed with the IUE, except for NGC 3605 and NGC 5018, which were observed using FOS on the HST.
Sources: J: Jansen *et al.* (2000); P: Ponder *et al.* (1998); S-BCK: Storchi-Bergmann *et al.* (1998); R: Bica & Alloin, unpublished; Q: Bica & Alloin (1987a); B: Bica & Alloin (1987b); V: Bica & Alloin, unpublished, from *ftp://cdsarc.u-strasbg.fr/cats/III/219/*

flux per unit wavelength bin. Consider a set of N spectra of early-type galaxies, each measured at T different wavelength bins. We denote by $\boldsymbol{x}_t \in \mathcal{R}^N$ the vector formed by the flux values at the wavelength bin indexed by t, as measured across the N spectra. The $N \times T$ matrix formed by these vectors may be referred to as $\boldsymbol{X}$ and single elements of this matrix will be denoted as x_{nt}. Similar notational convention will also apply to other variables used.

As stated above, the hypothesis of our model is that each of the N observations can be represented as a superposition of $K < N$ latent underlying component spectra or factors $f(\boldsymbol{s}_t) \in \mathcal{R}^K$, that are not observable by direct measurements but only through an unknown linear mapping $\boldsymbol{A} \in \mathcal{R}^{N\times K}$. Formally, this can be summarised as follows:

$$\boldsymbol{x}_t = \boldsymbol{A} f(\boldsymbol{s}_t) + \boldsymbol{\epsilon}. \tag{2}$$

In (2), the first term of the right is the so-called systematic component and $\boldsymbol{\epsilon}$ is the noise term or the stochastic component of this model, both of which will be further detailed below. The noise term $\boldsymbol{\epsilon}$ is assumed to be a zero-mean independent and identically distributed Gaussian, with variance $e^{-v_{x_n}}$. The diagonal structure of the covariance accounts for the notion that all dependencies that exist in $\boldsymbol{x}_t$ should be explained by the underlying hidden factors.

The K components are assumed to be statistically independent, which is a standard assumption of independent factor models. The role of the function f here is a technically convenient way of imposing certain constraints on the components $f(\boldsymbol{s})$ in order to facilitate the estimation of *interpretable* factors. This will be detailed in subsection 3.1.4.

As in nearly all data modelling applications of factor models, we hope that the feature transformation produced by the mapping $\boldsymbol{A}$ will reveal important structural characteristics of the collection of spectra. It is part of our data modelling design to investigate various ways of facilitating this, as will be detailed in §3.1.4.

3.1.2 Dealing with measurement errors in the data model

Classical non-probabilistic data analysis tools do not offer the flexibility required for taking into consideration the uncertainty that is associated with all real measurements. It is thus an important practical advantage of the probabilistic framework we adopt here that it allows us to take such measurement errors into account as an integral part of the data model. This is achieved simply by making the 'clean' vectors $\boldsymbol{x}_t$ in (2) become hidden variables of the additional error model

$$y_{nt} = x_{nt} + e_{nt}, \tag{3}$$

where e_{nt} is a zero-mean Gaussian noise term with known standard deviations σ_{nt}, for each individual measurement $n = 1 : N, t = 1 : T$.

3.1.3 Dealing with missing values in the data model

Our probabilistic framework also allows us to treat missing values in a principled and straightforward manner under the assumption that they are missing at random (Ghahramani & Jordan 1994; Chan & Lee 2003). Splitting each datum vector into missing and observed parts, i.e. $\boldsymbol{y}_t^T = (\boldsymbol{y}_t^{oT}, \boldsymbol{y}_t^{mT})$, the missing entries simply marginalise out:

$$p(\boldsymbol{y}_t^o) = \int p(\boldsymbol{y}_t) d\boldsymbol{y}_t^m. \tag{4}$$

The superscript 'o' will be omitted hereafter for simplicity, noting that unless otherwise stated, $\boldsymbol{y}_t$ will refer to $\boldsymbol{y}_t^o$ and missing entries are discounted in all the update equations that follow.

Having completed the general structural specification of the proposed probabilistic model, in the following subsection, we will make more specific assumptions about distributional form of the latent spectral representation that emerges from such probabilistic models.

3.1.4 The distributional forms employed in the data model

In order to be able to infer the parameters of the factor model (2), the form of the distributions of the involved variables, which may have a key impact on the subsequent interpretability of the resulting probabilistic model (MacKay 2003), need to be defined. That such assumptions are made explicit is indeed an important feature of the Bayesian framework adopted here. An important advantage of this approach is that of the several alternative forms that could be assumed for these distributional forms (we could call these "hypotheses"), the one best supported by the available data can be determined automatically from the data evidence. More details can be found below, in §3.3.

Before proceeding, we note that such assumptions are present in all automated analysis techniques, although in the case of non-probabilistic methods they are implicit and fixed. In particular, it has been shown that the widely used technique of Principal Component Analysis (PCA) implicitly assumes a standard Gaussian representation and isotropic Gaussian noise (Tipping & Bishop 1999). Although PCA has been useful in a number of ways in astrophysical data analysis (Connolly *et al.* 1995; Folkes, Lahav & Maddox 1996; Ronen *et al.* 1999; Madgwick *et al.* 2003a,b), there is

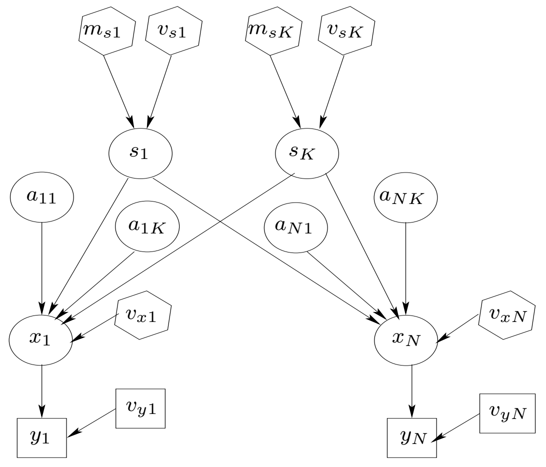

Figure 2. Graphical model of the proposed latent variable architecture (i.e. the data model). Nodes represent random variables in the data model, and arrows indicate causal relationships between variables. Out of the nodes, rectangles refer to observed variables: measured fluxes, y, and known measurement errors, v_y. Circles and hexagons refer to latent (hidden) variables: clean flux, x, along with their variance parameters, v_x, mixing coefficients, a, and components, s, along with their component-wise means, m_s, and variance parameters, v_s). Hexagons indicate the latent variables on the top of this hierarchy, for which further hierarchical priors may be considered.

no guarantee that the assumptions it makes are appropriate in our case. In particular, the assumption of a Gaussian distribution for the spectral elements in our present problem would imply that they are equally likely to have positive and negative flux values. This may make the physical interpretation of the spectra of the individual stellar populations (factors) problematic. Indeed, the physical validity of the isotropic structure of the Gaussian noise imposed by PCA has been questioned by Madgwick *et al.* (2003b).

It is therefore worthwhile to consider alternative factor models to relax such unrealistic constraints, and to introduce more justifiable constraints (such as the positivity of the factors). This is what we do in this section.

In this study we consider four factor models with differing assumptions made regarding the distributional forms employed. These are summarised in Table 3.1.4. The first three of these refer to factor models existing in the literature, and the fourth one has been recently developed by us (Harva & Kabán 2005). All factor models have been implemented in a Bayesian formulation (the so-called Variational Bayesian framework, see details in Section 3.3) in order to ensure a consistent formal framework. The four factor models are described below.

VB-PCA The first row of Table 3.1.4 describes the Variational Bayesian PCA (VB-PCA) data model (Bishop 1999). Due to the widespread use of PCA, we will consider it as the baseline hypothesis. As already mentioned, it assumes that the spectral factors of the representation follow a Gaussian distribution and there is no positivity constraint. This is summarised in the columns $p(s)$ and $f(s)$ respectively of the first row of Table 3.1.4. Further, it assumes that the mixing coefficients of the factors (i.e. the elements of the matrix $\boldsymbol{A}$ of equation (2)) follow a standard Gaussian and finally that the noise term has an isotropic variance structure (i.e. $v_{x_n} = v_x$ is the same for all n).

VB-FA In Table 3.1.4, the Variational Bayesian Factor Analysis (VB-FA) data model (Bishop 1999) relaxes the constraint of the isotropic structure of the Gaussian noise imposed by PCA. Indeed, such a constraint may be physically unjustified (Madgwick *et al.* 2003b). All other assumptions made are in turn identical with those of VB-PCA.

Table 2. Distributional assumptions for the various data models

Model	$p(s)$	$p(a)$	$f(s)$	v_{x_n}
VB-PCA	$\mathcal{N}(s\|m_s, e^{-v_s})$	$\mathcal{N}(a\|0,1)$	s	v_x
VB-FA	$\mathcal{N}(s\|m_s, e^{-v_s})$	$\mathcal{N}(a\|0,1)$	s	v_{x_n}
VB-PFA	$\mathcal{N}^R(s\|0, e^{-v_s})$	$\mathcal{N}^R(a\|0,1)$	s	v_{x_n}
VB-RFA	$\mathcal{N}(s\|m_s, e^{-v_s})$	$\mathcal{N}^R(a\|0,1)$	$\max(0,s)$	v_{x_n}

Data Models: VB-PCA: Variational Bayesian PCA; VB-FA: Variational Bayesian Factor Analysis; VB-PFA: Variational Bayesian Positive Factor Analysis; VB-RFA: Variational Bayesian Rectified Factor Analysis. The columns summarise the assumptions for the distributional form of each variable of Eq. (2) for each of these four data models. Indices have been dropped. The columns $p(s)$ and $p(a)$ indicate the form of the probability density assumed for the elements of $\boldsymbol{s}_t$ and $\boldsymbol{A}$ respectively. The column $f(s)$ specifies the form of the function $f()$ of Eq. (2), and v_{x_n} provides the structural form of the noise variance for the various data models (distinct variances for all data dimensions n except in the case of VB-PCA).

VB-PFA The above two models do not require the individual factors, which are spectra of the stellar sub-populations, to be non-negative, which they need to be in order to be physically acceptable. The remaining two models considered here do so. The third row of Table 3.1.4 describes the Variational Bayesian Positive Factor Analysis (VB-PFA), a data model developed by Miskin (2000), that, in contrast to VB-PCA and VB-FA, imposes the positivity constraint. It does so by assuming that each individual spectrum $p(s)$ (factor) of the representation follows a zero-mode rectified Gaussian distribution (essentially a Gaussian restricted to its positive half and renormalised, defined in Appendix A4). VB-PFA also restricts the mixing coefficients $p(a)$ to be positive, assumed to be standard rectified Gaussians, and maintains the same structural form for the noise variance v_{x_n} as VB-FA.

The positivity constraint just introduced is meant to facilitate the interpretability of the factors as component spectra. Let us observe, however, that in VB-PFA, the $0-$mode of $p(s)$ would mean that the recovered components would have a low flux close to zero occupying most of the spectrum. Although there has been a great deal of interest in factor models that have such sparsely distributed factors for a variety of other purposes (Miskin & Mackay 2001; Lee & Seung 2001; Hinton & Ghahramani 1997; Socci, Lee & Seung 1997; Frey & Hinton 1999), including image and text related applications as well as neurobiological modelling, such a sparsity assumption may be restrictive and often unjustified for the modelling of stellar population spectra. However, as shown in Harva & Kabán (2005), modifying VB-PFA to relax the $0-$mode assumption is technically impossible. Thus we developed a different way of imposing the positivity constraint, which also permits the mode of $p(s)$ to be estimated from the data in a flexible manner. We did not discard the hypothesis of VB-PFA either, since it has been applied to stellar population spectra in the literature (Miskin 2000) and we do not wish to discard any existing hypotheses based on possibly subjective intuitions.

VB-RFA The last row of Table 3.1.4 describes the Variational Bayesian Rectified Factor Analysis (VB-RFA) data model (Harva & Kabán 2005), where the positivity constraint over the factors is ensured through the non-differentiable function $f(s) = \max(0, s)$. With this solution, the factors are now $f(s)$ and they are guaranteed to be positive. Then the variables s become some dummyarguments of $f()$, which are now safely modelled as Gaussians (column $p(s)$ of 4-th row of Table 3.1.4). Therefore, both the means and variances of these Gaussians can now be estimated from the data. The resulting model allows more flexibility regarding the shape of the factors compared with VB-PFA, while ensuring the positivity of the factors. Finally, the mixing coefficients $p(a)$ are assumed to be distributed as standard rectified Gaussians. Note that in this context the zero-mode constraint is appropriate since the sparsity of the mixing matrix $\boldsymbol{A}$ translates to assuming a smooth basis transformation. The technical novelty and computational advantages of the VB-RFA approach are detailed in Harva & Kabán (2005).

As it should now be clear, our scope is neither to apply any off-the-shelf data analysis method in an unjustified manner, nor to rely on subjective intuitions, as both of these approaches may easily lead to biased results. Instead, we have constructed four different hypotheses which we consider equally likely *a priori*. The Bayesian framework allows us to decide which hypothesis is best supported by the data.

3.2 The overall model architecture

The architecture for any of the four data models defined above can be graphically represented as in Fig. 2. This is the joint density of the observational data and all other variables of interest. Nodes denote random variables, rectangles refer to observed variables (i.e. derived from observational data), and circles and hexagons refer to unobservable variables that we are hoping to infer from the data through the formulated probabilistic model. Hexagons denote hidden variables on the top of the hierarchy[1]. Arrows signify generative causal dependency relations.

We now outline the variational Bayesian estimation procedure associated with this architecture.

3.3 Estimation of a data model

Each of the described four data models is estimated using the variational Bayesian methodology described in this section. The specific expressions obtained for the VB-RFA data model are provided in the Appendix. To make the notation concise, we will refer to the latent variables by $\boldsymbol{\theta}$ and to the data by $\boldsymbol{Y}$. Since handling the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ is intractable, we resort to a variational scheme (Jordan *et al.* 1999; Lappalainen 1999; Raiko *et al.* 2005; Attias 2000), where an approximative distribution $q(\boldsymbol{\theta})$ is fitted to the true posterior. This is done by constructing a lower bound of the

[1] In the BayesBlocks based implementation (Valpola *et al.* 2003) adopted here, a maximal flexibility is ensured by further formulating hierarchical priors on all variables m_{s_k}, v_{s_k} and v_{x_n}, as follows: $p(v_{x_n}) \sim \mathcal{N}(m_{v_x}, e^{-v_{v_x}})$, with parameters common to all components of the variable, and these parameters m_{v_x} and v_{v_x} will then have vague priors in the form of a zero mean Gaussian having a large value for variance. The definition is analogous for m_{s_k} and v_{s_k}. These additional variables have been omitted from the diagram on Fig. 2 for simplicity, as they are part of the relatively standard technicalities concerning a full Bayesian approach (Raiko *et al.* 2005) rather than an essential part of the application-specific model design.

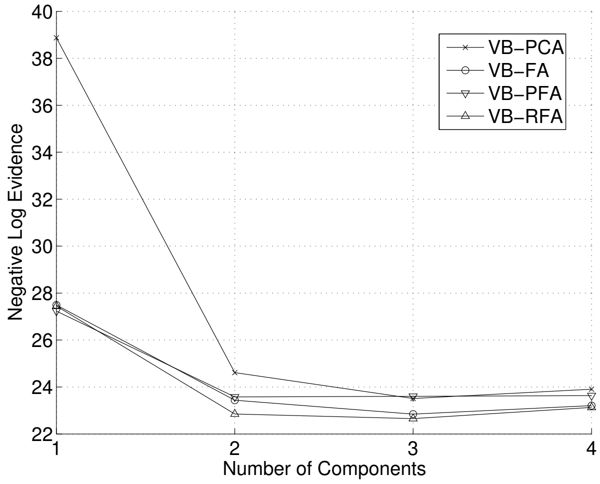

Figure 3. The negative log evidence for the variational Bayesian analyses of the observed spectra, as a function of the number of factors (components), *i.e.* stellar sub-populations. The performance of VB-PCA is clearly inferior. VB-RFA has the highest log evidence, with no significant improvement in increasing K, the number of components, from two to three. This shows that the spectra of early-type galaxies used here can be safely modelled as the sum of the spectra of two distinct major stellar components, and the inclusion of more components is not required by the data.

log evidence, based on Jensen's inequality,

$$\log p(\boldsymbol{Y}) = \log \int d\boldsymbol{\theta}\, p(\boldsymbol{Y}, \boldsymbol{\theta}) = \log \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \frac{p(\boldsymbol{Y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}$$
$$\geq \int d\boldsymbol{\theta}\, q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{Y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} = \langle \log p(\boldsymbol{Y}, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} - \langle \log q(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} \tag{5}$$

where $\langle . \rangle_q$ denotes expectation w.r.t. q.

This inequality holds for any distribution $q(.)$ due to the convexity of the logarithm. A fully-factorial approximation of the posterior (MacKay 2003; Lappalainen 1999; Attias 2000; Bishop 1999; Bishop *et al.* 2002; Raiko *et al.* 2005) will be employed here, which is also known to lead to estimation algorithms that can be implemented efficiently by using local computations only (Lappalainen 1999; Raiko *et al.* 2005; Valpola *et al.* 2003; Bishop *et al.* 2002) The best fully-factorial approximate posterior (referred to as the variational posterior) is computed by functional maximisation of (5), which can be shown (Lappalainen 1999; MacKay 2003; Jordan *et al.* 1999) to be equivalent to minimising the Kullback-Leibler divergence (Cover & Thomas 1991) of the variational posterior from the true one. The fully-factorial variational posterior of the data model in Fig. 2 is simply of the form

$$\begin{aligned} q(\boldsymbol{\theta}) \quad &= \quad \prod_{t=1}^{T} \prod_{n=1}^{N} q(x_{nt}) \times \prod_{n=1}^{N} q(v_{x_n}) \times \prod_{n=1}^{N} \prod_{k=1}^{K} q(a_{nk}) \\ &\times \quad \prod_{k=1}^{K} \prod_{t=1}^{T} q(s_{kt}) \times \prod_{k=1}^{K} q(m_{s_k}) q(v_{s_k}). \end{aligned} \tag{6}$$

Therefore, (5) further decomposes into one plus twice as many terms as there are latent variables in the data model and the optimisation w.r.t. the individual variables is separately performed. The variational posteriors thus obtained, along with the statistics that are required for an implementation[2], are given in Appendix A1 for the VB-RFA model. Some of these posterior computations are non-trivial. Readers interested in more detailed derivations are referred to Harva & Kabán (2005).

The model estimation algorithm then consists of iteratively updating each parameter's posterior statistics in turn, while keeping all other posterior statistics fixed, until a convergence criterion is achieved. For each update, the evidence bound (5) is guaranteed not to decrease (Jordan *et al.* 1999; MacKay 2003), and convergence to a local optimum is guaranteed. It can also be shown that due to the fully-factorial posterior approximation (6), all updates are local, i.e. requiring posterior statistics of the so called Markov blanket only. That is, for updating any of the variable nodes, the posterior statistics of only its children, parents and co-parents are needed. This has been exploited in the efficient implementation realised in Valpola *et al.* (2003), Bishop *et al.* (2002) and Winn & Bishop (2005), and has also been adopted in this work. The scaling of the resulting variational Bayesian estimation algorithm is thus multilinear in the number of factors, observations and wavelength bins of the model. These can also be easily followed from the further details given in Appendix A1.

The evidence bound (5) can be used both for monitoring the convergence of the algorithm and, more importantly, for comparison of different solutions or models. The quantities required for computing this are given in Appendix A2.

3.4 Using a data model to recover missing elements of the data

We have already stressed the advantage of the probabilistic Bayesian framework adopted here in terms of the ability of the resulting data models to predict and recover incomplete data. Recovering missing values can be achieved by computing the posterior of the missing entries given the observed entries:

$$q(\boldsymbol{y}_t^m|\boldsymbol{y}_t^o) = \int d\boldsymbol{\theta} q(\boldsymbol{\theta}) p(\boldsymbol{y}_t^m|\boldsymbol{\theta}). \tag{7}$$

An approximate mean value of this posterior provides the imputation for the missing entry. This is discussed in Appendix A3.

3.5 Using a data model for recovering missing elements from new measurements

Due to the fully generative nature of the adopted framework, once the data model parameters are estimated from a training data set, inferences and predictions can be made for new, previously unseen measurements. This involves performing the inference for those terms of the variational posterior $q(\boldsymbol{\theta})$ that are conditional on the type of new measurements (new galaxy spectra in the same range or various sub-ranges of wavelength for the same galaxies). Using these, as well as the remainder of terms of $q(\boldsymbol{\theta})$ as estimated from the training set, we can then proceed to computing (7). Results will be presented in Section 4.3.

[2] All algorithms investigated here have been implemented as part of the BayesBlocks software (Valpola *et al.* 2003; Raiko *et al.* 2005) and may be made freely available upon publication.

4 RESULTS

One of the strengths of the Bayesian evidence framework employed in this study is that the selection of the best-supported hypothesis can be automatically achieved, in a quantitative manner, based on the data evidence. This is known as model selection.

4.1 Model selection results: how many major stellar populations do we need?

Here we seek a criterion to examine which data model performs the best, as well as to find the dimensionality of the representation space. We would like to have a compact model (small number of factors) both to avoid over-fitting the data, and also for the results to be physically interpretable. At the same time, we would like to keep an adequate number of factors to guarantee a good approximation of the data. Apart from the interpretability issue, the Bayesian evidence framework can be used to automatically select the optimal model and its order from those investigated. Our preferences, motivated by the purpose of the application (e.g. data interpretation, missing value prediction), expressed as prior belief, would then be used to weight the evidence obtained from the data, in order to make a final Bayesian decision on the order of the model.

Fig. 3 shows the negative of the log evidence for the four data models (strictly, the negative log evidence bound), versus the model order, K (i.e. the number of components), for the observed spectra (Fig. 1). It can be seen that the VB-RFA model outperforms the other three models. We also see that increasing the number of factors beyond two does not significantly increase the evidence, which suggests that the bulk of spectra of an early-type galaxy can be reasonably well and compactly described by two factors or components, each interpretable as a stellar sub-population. This is useful for input into more detailed physical models of early-type galaxy spectra. We will also see in §4.2 that the $K = 2$ models produce two factors that are both physically meaningful, while increasing the model order to $K = 3$ does not bring any further interpretable factors.

4.2 Physical interpretation of the results

Here, we compare the results of the data-driven method outlined so far with an entirely independent determination of the star-formation history of our 21 early-type galaxies.

In this more conventional approach, we have fitted two-component detailed physical models, constructed from the single stellar population models of Jimenez *et al.* (2004), to the observed long-baseline spectra of our sample, as shown in Fig. 1. The ages, metallicities and relative mass fractions of the component stellar populations are allowed as free parameters. We construct the two-component spectra as the linear superposition of the two component spectra (1), and allow ages 0.01-14.0 Gyr, and metallicities $Z = 0.01$, 0.2, 0.5, 1.0, 1.5, 2.5 and 5.0 times the solar value ($Z_\odot$). Again, the relative mass fractions have to be jointly constrained, $m_i + m_j = 1$, but m_i is allowed to vary from 0 to 1, in steps of 0.02. A minimum-χ^2 fit was used to determine the best-fit values of m_i, Z_i, t_i, m_j, Z_j and t_j. The whole parameter space was searched, with the best-fitting parameter values corresponding to the point on the age$_i$-age$_j$-Z$_i$-Z$_j$-m_i-m_j grid with the minimum calculated χ^2. Details of the fitting process are given in Nolan (2002) and Nolan *et al.* (2003).

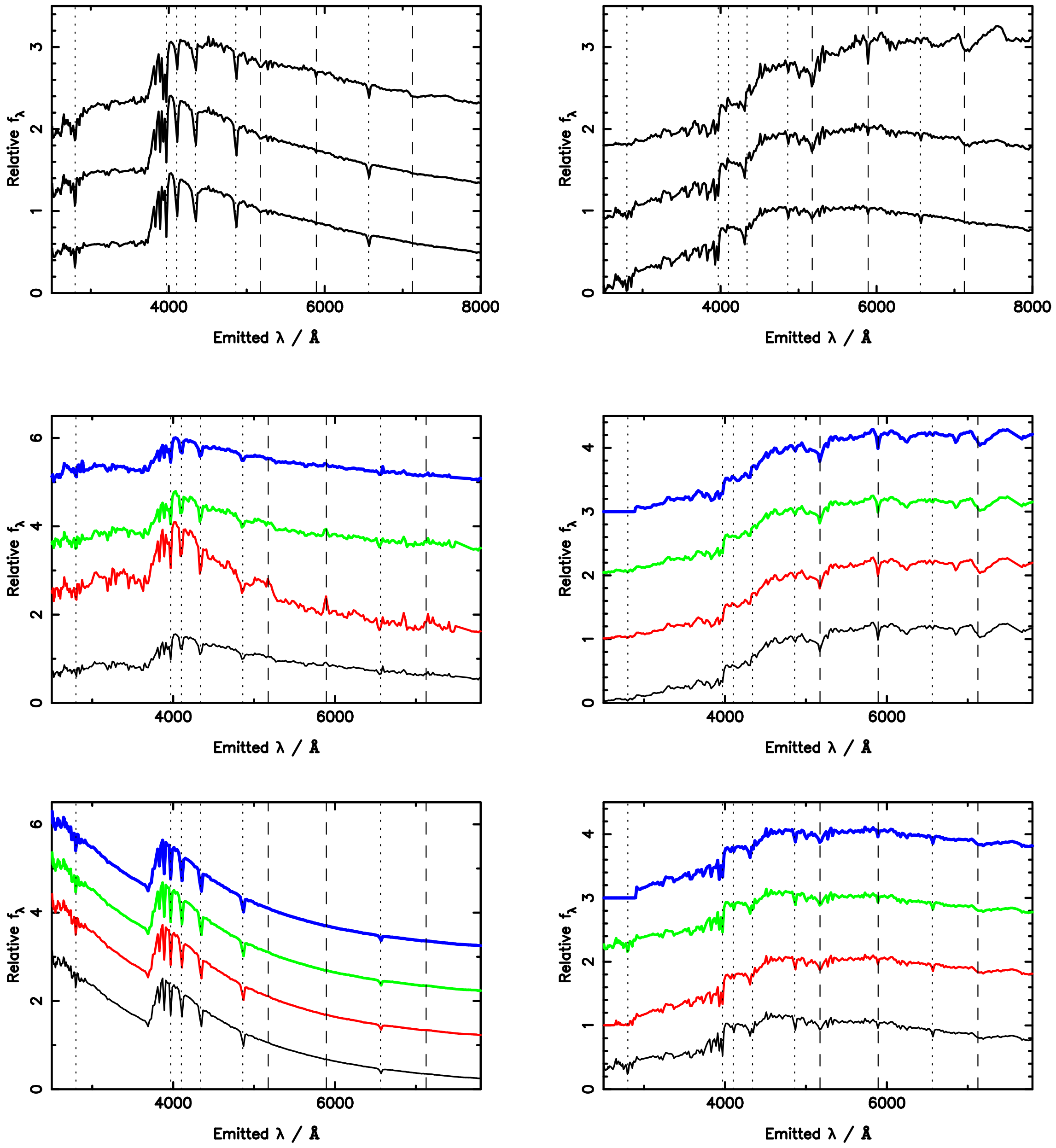

Figure 4. Analysis of the two-component data model. *Top:* Synthetic stellar population spectra (Jimenez *et al.* 2004): *Top Left:* Spectra of a population of age 0.7 Gyr, where metallicity, from bottom to top, is Z =0.2, 1.0 and 2.5 $Z_\odot$; *Top Right:* age = 10 Gyr, same metallicities. The dotted lines mark some of the absorption features in the spectrum which are typically strong in young stellar populations, and the dashed lines mark some of the absorption features which are typically strong in old, metal-rich stellar populations. From left to right, the absorption line species are: MgII (2799 Å), Hε (3970 Å), Hδ (4102 Å), Hγ (4340 Å), Hβ (4861 Å), Mgb (5175 Å), NaD (5893 Å), Hα (6563 Å), TiO (7126 Å). *Middle:* The two components found from the various different linear independent basis transformation analyses, with the number of components $K = 2$, of the observed early-type galaxy spectra. The methods are, from bottom to top: VB-FA, VB-PFA, VB-PCA, VB-RFA. The recovered spectra are convincingly disentangled into one component (middle left) with young stellar population features (MgII, Hε, Hδ, Hγ, Hβ, Hα: dotted lines) and shape, and a second (middle right) with the features (Mgb, NaD, TiO: dashed lines) and shape of an old, high-metallicity stellar population. *Bottom:* The two components found from the same analyses, but from the linear superposition of two synthetic stellar populations. The various curves refer to the same parameters as in the middle panels. The UV spectrum of the 'young' component recovered here rises more steeply than for the same component recovered from the observed spectra, due to the inclusion, in the models, of a very young (0.03 Gyr) synthetic stellar population, which is not seen in our observed spectra.

4.2.1 Analysis of the observed spectra with the two-component data model

In Fig. 4, synthetic spectra (Jimenez *et al.* 2004) of young (0.7 Gyr, left) and old (10 Gyr, right) stellar populations are shown, for metallicities 0.2, 1.0 and 2.5 $Z_\odot$, in each of the upper panels, from top to bottom respectively. Below these are shown the spectra of the individual components, recovered from the four analysis methods described above, for both the observed galaxy spectra (middle panels), and linear superpositions of pairs of synthetic spectra (bottom panels).

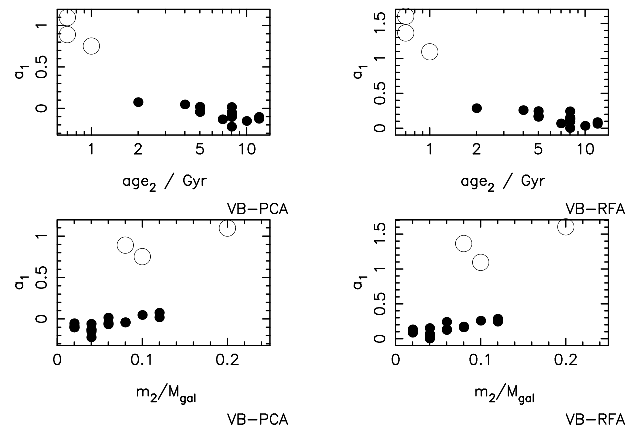

Figure 5. Scatter-plots showing the correlation of *(top)* the age of the younger stellar population and *(bottom)* the mass fraction of the smaller stellar population determined from two-component fitting to the observed spectra with the weight of the first component (a_1) of the VB-PCA (left) and VB-RFA (right) linear basis transformation analyses of the observed spectra, with $K = 2$. A high value of a_1 clearly corresponds to a substantial young ($\lesssim$ 1 Gyr) stellar population. The open circles are for those galaxies with a secondary population of age less than 1 Gyr. These plots also show that the worst-performing (VB-PCA) and best-performing (VB-RFA) algorithms in this case yield almost identical results.

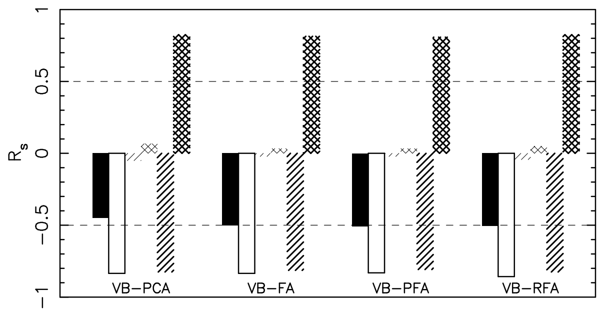

Figure 6. Spearman's rank correlation coefficient (R_S) for the first component (a_1) of the four linear basis transformation analyses with $\mathrm{K} = 2$, of the observed spectra and the best-fitting parameters of the two-component synthetic stellar population fits to the data: solid: age of the older population; outline: age of the younger population; light hatched: metallicity of the older population; light cross-hatched: metallicity of the younger population; dark hatched: mass fraction of the smaller population; dark cross-hatched: mass fraction of the dominant population. $\bmod \mathrm{R}_S > 0.5$ indicates a strong correlation.

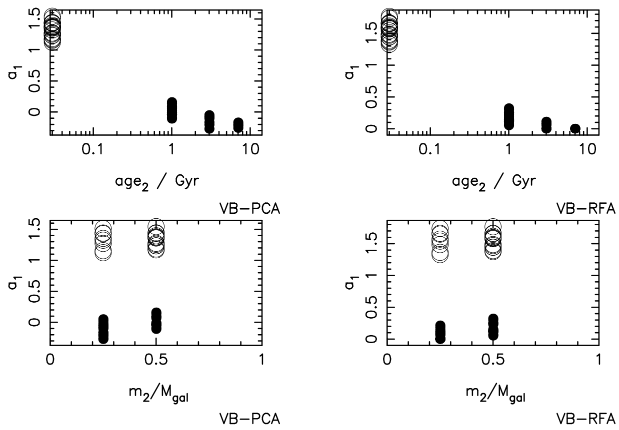

Figure 8. Scatter-plots showing the correlation of *(top)* the age of the younger stellar population and *(bottom)* the mass fraction of the smaller stellar population determined from the synthetic spectra fitting with the weight of the first component of the VB-PCA and VB-RFA linear basis transformation analyses methods (a_1), and $K = 2$. The open circles are for those galaxies with a secondary population of age less than 1 Gyr. A high value of a_1 clearly corresponds to a significant young ($\ll$ 1 Gyr) stellar population.

In general, the similarity between the recovered components, and the corresponding model spectra, is striking, for all four methods. In the younger components displayed on the left hand of Fig. 5, the 4000 Å break, typical of a mature ($\gtrsim$1 Gyr), is not seen, but strong Balmer absorption line features (Hε (3970 Å), Hδ (4102 Å), Hγ (4340 Å), Hβ (4861 Å)) can be clearly identified. These are typical features of the A stars which dominate the emission from young ($<$ 1 Gyr) stellar populations. In the second (older) component, however, the 4000 Å break is apparent, as are other features typical of mature, metal-rich stellar populations, e.g. absorption features corresponding to Mgb (5175 Å), NaD (5893 Å) and TiO (7126 Å). We can rule out either component being representative predominantly of either AGN emission or dust absorption as the characteristic shapes and features of these spectra would be very different from those we recover.

However, the VB-PCA algorithm seems to "find" a spurious sub-component with younger features (middle left panel) that has some features from old, metal-rich stellar populations (notably NaD).

In Fig. 5, we show results for VB-RFA method (right panels), which has the the highest log evidence values, together with the VB-PCA method (left panels), which has the lowest, for comparison. We plot the correlation between the first ('younger') component (a_1) and the age of the younger stellar population (age$_2$, top panels) and the mass fraction of the secondary stellar population component ($m_2/M_{\rm gal}$), as determined from the two-component model-fitting described above (bottom panels). The open points are those galaxies which contain a significantly young stellar population ($\lesssim$1 Gyr).

Fig. 6 plots the results of the Spearman's rank correlation test for the weight of the first component a_1 and the age, metallicity and mass fraction of the two fitted stellar populations for the observed spectra with $K = 2$. A value of the Spearman's rank correlation coefficient $|R_S| > 0.5$ indicates a strong correlation between the parameters. It can be seen that there is a strong correlation between a_1 and the ages of the younger stellar populations, and also between a_1 and the stellar mass fractions ($m_1 M_{\rm gal}$, $m_2/M_{\rm gal}$) of the component stellar populations. The correlation between a_1 and age$_1$ is weak, and there is no correlation between a_1 and the metallicities of the component stellar populations. The trends of the correlations with a_2 follow similar, but opposite, trends to those of a_1. This is not generally true for factor analysis, but arises here as a consequence of the nature of our data set.

The rest-frame spectra of four of the galaxies in our sample, together with the reconstructed data model and the two constituent data model components are presented in Fig. 7. The data model convincingly reproduces the observed spectra.

4.2.2 *Analysis of the synthetic spectra with the two-component data model*

Here we present the results of the analysis of the two-component synthetic spectra, constructed as described in §2.1. They represent

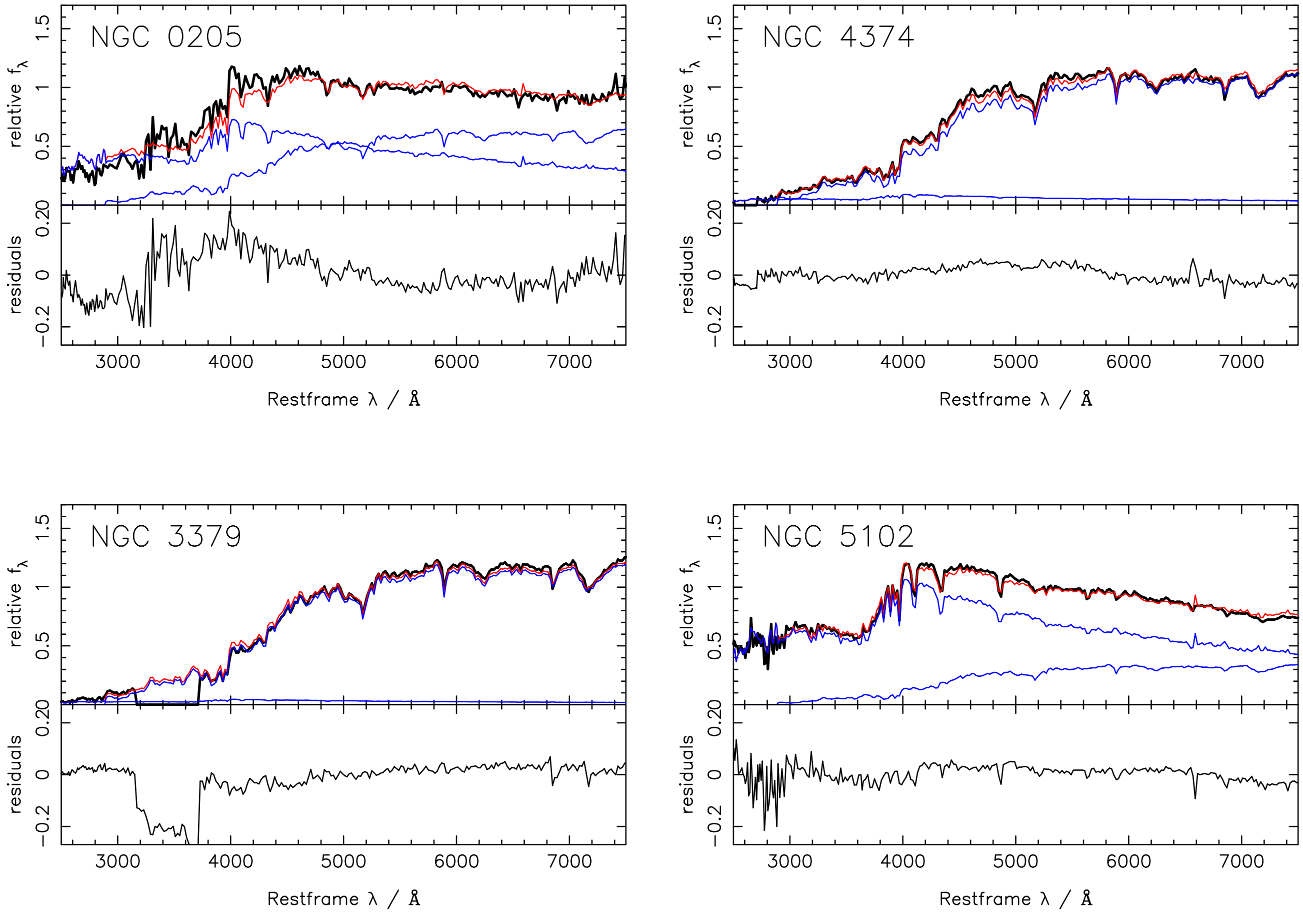

Figure 7. Rest-frame spectra of four of the 21 galaxies in our sample (thick black line), with the reconstructed data model and the two constituent data model components (thin lines) superimposed. The data model convincingly reproduces the observed spectra.

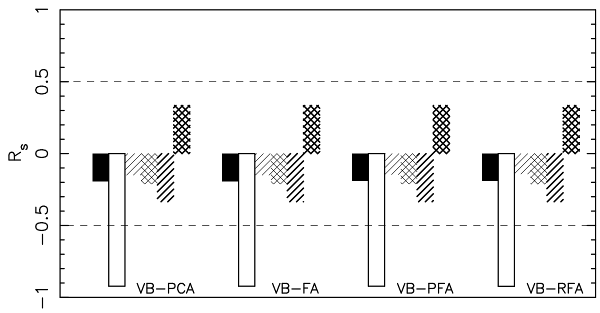

Figure 9. Spearman's rank correlation coefficient (R_S) for the first component (a_1) of the various linear basis transformation analyses with $K = 2$, of the two-component synthetic spectra and the best-fitting parameters of the two-component synthetic stellar population fits to the data: solid: age of the older population; outline: age of the younger population; light hatched: metallicity of the older population; light cross-hatched: metallicity of the younger population; dark hatched: mass fraction of the smaller population; dark cross-hatched: mass fraction of the dominant population. mod $R_S > 0.5$ indicates a strong correlation.

every possible combination of the assigned ages, metallicities and relative stellar mass fractions, and thus allow us to investigate a wider range of known age and metallicity combinations than the observed spectra. It should be noted that not all combinations (for example, those with two components < 1 Gyr old) represent what we might expect to find in the total population of real early-type galaxies, although there could be local regions within these galaxies where, for example, young stars dominate the stellar population. We wanted to cover the complete range of possibilities in order to avoid making any assumptions about the stellar populations of early-types, which, as revealed even in our small sample of twenty-one spectra, can be quite varied.

In this experiment, no noise has been added to the synthetic spectra. This is because the noise from variations in the spectral shape caused by differences in the metallicity is higher than any noise contributions which represent measurement errors.

The bottom panels of Fig. 4 show the spectra of the components recovered using the same four analysis methods as used for the observed spectra. Again, the first component has a shape and features that are similar to those of a young stellar population, and the second component, those of a mature population. The shape of the UV part of the spectrum, together with the featureless nature of the spectrum above ∼5000 Å, are typical of very young stellar populations (∼ a few x 10 Myr). These populations are present in the synthetic data set, but not in significant quantities in the observed data set, hence the differences between the recovered data components.

In Fig. 8, as in Fig. 5, we see that a high value of a_1 (the weight of the 'younger' component from the Bayesian data model) corresponds to the presence of a <1 Gyr stellar population. Fig. 9 plots the Spearman's rank correlation test results for a_1 and the age, metallicity and mass fraction of the synthetic populations. Here, there is an even stronger correlation between a_1 and the age of the younger population than in the case of the observed spectra, and also a strong correlation between a_1 and the age of the older population, but no significant correlation between a_1 and the mass fraction of the secondary population. However, in this exercise with synthetic spectra, we allowed only two values for $m_2/M_{\rm gal}$ (0.25 and 0.5), which does not really allow for a meaningful determination of the correlation coefficient.

These results confirm that the value of a_1 is a clear indicator of the presence of a significant young stellar population, and is particularly unambiguous in discovering components with very young ($<$1 Gyr) stellar populations, as we found in §4.2.1.

4.2.3 *Analysis of the three-component data model*

Fig. 3 shows that increasing the number of component spectra in the Bayesian models from two to three does not significantly increase the evidence for any data model. Indeed, for the positive factor analysis, the evidence decreases as K increases from two to three. These results suggest that the bulk of the stellar content of early-type galaxies can be represented by two components, one which represents contributions from old, metal-rich stellar populations, the other representing contributions from young populations (Fig. 10). It is not easy to find a physical interpretation for the third component. Indeed, the third component is probably a combination of residuals from the sum of the spectra of two populations, combined with correlated components of observational errors.

The results of the analysis performed on the observed spectra are presented in Fig. 11. The components a_1 and a_2 correlate significantly with the age of the younger population (age$_2$), except for the VB-RFA, where the correlation is with a_3 rather than a_1. The relative mass fraction of the two stellar population components (m_1 and m_2) also correlate with a_1 and a_2, and, for the VB-RFA, with a_3 as well.

For the synthetic two-population spectra, age$_2$ correlates strongly with a_1 and a_2, for all the analysis methods, and also with a_3 for the VB-RFA. However, there is no strong correlation ($\mathrm{mod}\ \mathrm{R}_S > 0.5$) for any of the data models with the relative mass fraction of the two populations, although the metallicity of the younger population (Z_2) correlates strongly with a_3 in the case of the VB-PCA and VB-FA methods.

The differences between the correlation coefficients for the observed and synthetic spectra most likely reflect the different distributions of age, metallicity and mass fraction in the two spectral samples. For example, the observed spectra are all dominated by high-metallicity components (2.5 and 5.0 $\mathrm{Z}_\odot$), with small contributions from very low-metallicity populations (0.01 $\mathrm{Z}_\odot$). Only one of the galaxies contains stellar populations with intermediate metallicities. In contrast, the synthesised spectra can take the whole range of combinations of the seven allowed metallicities.

4.2.4 *Interpretation of the eigen-components of PCA*

Although we have managed to identify components with PCA here, it is a well known property of PCA that the solution it provides is rotation-invariant. That is, the eigen-components which have equal variances can be multiplied by an arbitrary orthogonal matrix to produce another equally good solution (in terms of satisfying the same objective function). Therefore PCA is not guaranteed to achieve a successful separation of the underlying independent

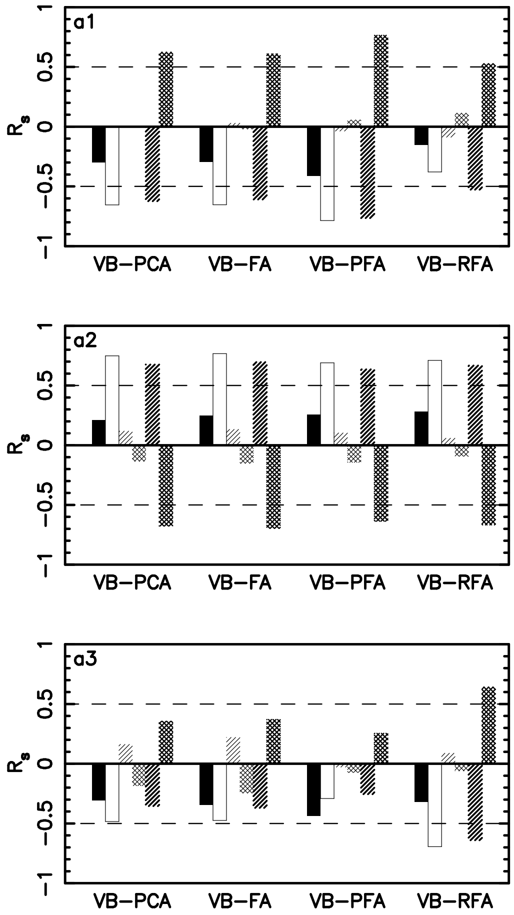

Figure 11. Spearman's rank correlation coefficient (R_S) for the first component (a_1) of the various linear basis transformation analyses with $K = 3$, of the observed spectra and the best-fitting parameters of the two-component synthetic stellar population fits to the data: solid: age of the older population; outline: age of the younger population; light hatched: metallicity of the older population; light cross-hatched: metallicity of the younger population; dark hatched: mass fraction of the smaller population; dark cross-hatched: mass fraction of the dominant population. $\mathrm{mod}\ \mathrm{R}_S > 0.5$ indicates a strong correlation.

source signals that we are looking for. Methods that take into account higher order dependencies need to be used for that purpose (Hyvärinen, Karhunen & Oja 2001).

The only 'lucky' case where PCA can identify components is if the underlying components have suitably different variances. This is true for the work presented here, but there is no reason why this would hold for data sets of galaxy spectra in general. Therefore, even if PCA has been successful here, it comes with no guarantees.

Another way of looking at this problem is through the probabilistic formalism. PCA's implicit assumption (as shown in detail in Tipping & Bishop (1999)) is that the components it will find are Gaussian. Making this assumption explicit, by the probabilistic formalism that we adopted, highlights the lack of plausibility of this assumption: the spectra that we are looking for are not Gaussian, instead they are positive valued only, and have a rather skewed distribution. This renders the use of PCA very problematic in applications that require a clear interpretation of the components.

In turn, by adopting the probabilistic formalism where any assumption in the data model must be made explicit, we are able to specify the distribution that corresponds to components that we are

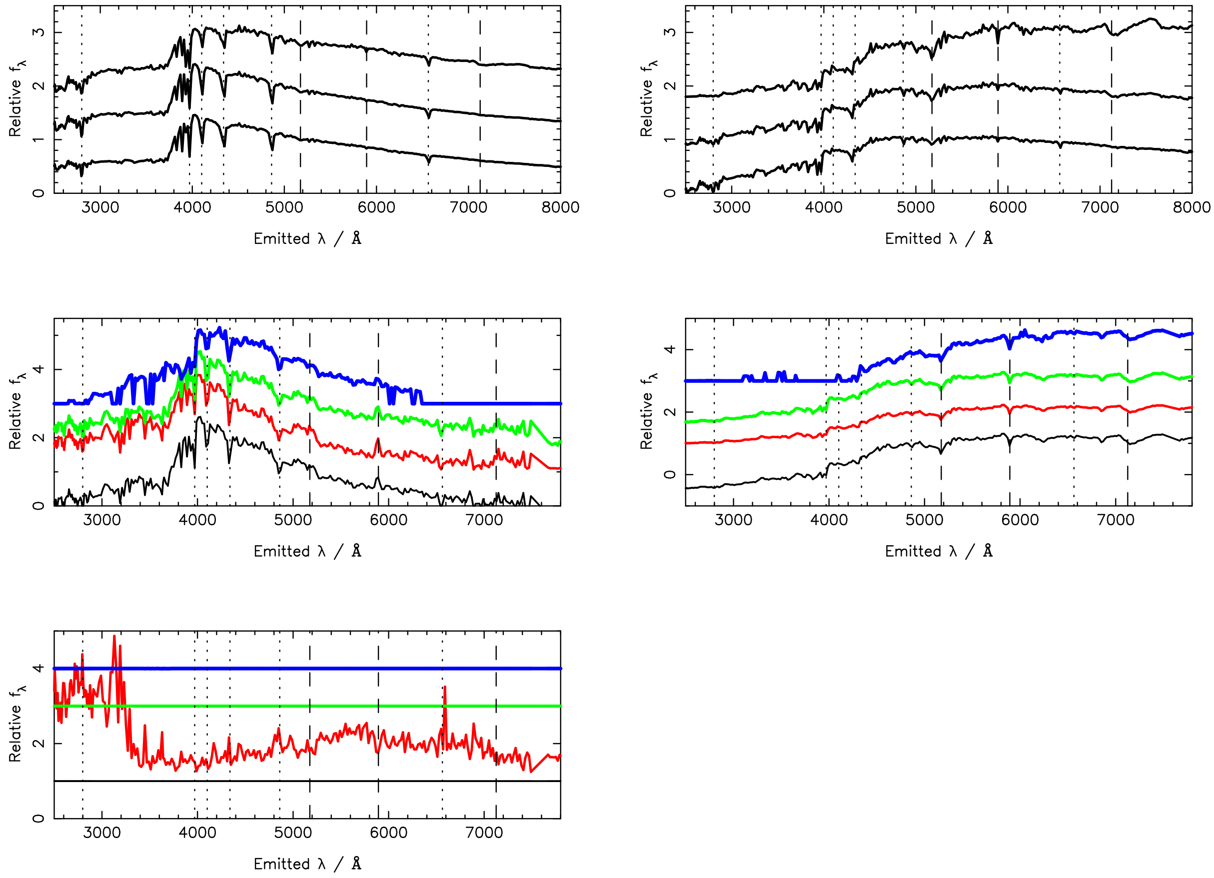

Figure 10. Analysis of the three-component data model. *Top panel:* Synthetic stellar population spectra (Jimenez *et al.* 2004): *top left:* for a stellar population of age 0.7 Gyr, with from bottom to top, metallicity Z = 0.2, 1.0 and 2.5 $Z_\odot$; *top right:* age = 10 Gyr, same metallicities. The dotted lines mark some of the absorption features in the spectrum which are typically strong in young stellar populations, and the dashed lines mark some of the absorption features which are typically strong in old, metal-rich stellar populations. From left to right, the absorption line species are: MgII (2799 Å), Hε (3970 Å), Hδ (4102 Å), Hγ (4340 Å), Hβ (4861 Å), Mgb (5175 Å), NaD (5893 Å), Hα (6563 Å), TiO (7126 Å). *Middle and bottom panels:* The 3 components found from the various different linear independent basis transformation analyses of the model stellar populations, with $K = 3$. The methods are, from bottom to top: VB-FA, VB-PFA, VB-PCA, VB-RFA. The first two recovered spectra are broadly identified with the features of a young and old stellar population in the same way as the $K = 2$ components. The third component (bottom) possibly contains 'residuals', resulting from the fact that the first two components are not a complete detailed reconstruction of the true stellar population components.

looking for. Therefore the chances that we can find what we are looking for, and not something else, are considerably increased. The ultimate conclusion is of course made based on visual inspection of the recovered components, made by human experts.

4.3 The prediction of missing values

As a more objective criterion for the assessment of our data modelling compared to the results obtained by the fitting of detailed spectral models, here we investigate the predictive ability of the modelling approach when faced with missing elements of the spectra.

We begin with demonstrating the recovered values for the missing elements along with their predicted uncertainties (posterior variances), as obtained from VB-RFA on the small real data set, compared with those from the best-fitting physical model. This is shown on the scatter plot in Fig. 12. The correlation coefficient between the posterior mean predictions of the data model and the values imputed from the best-fitting physical model is 0.989, the SNR being 9.3. This is a strong indication that the data-driven analysis and the physical analysis are well matched.

However, in the case of the real data, the true values of the missing entries are unknown. Therefore, in the following section, we employ synthetic data to perform a more thorough and controlled objective evaluation of the predictive capabilities of our probabilistic data model model. Missing entries were simulated, and the true values were used for evaluation only. The synthetic data set was more diverse than the real one, as a variety of parameter combinations that can be simulated, in this process, may not appear in real spectra. Therefore, a wealth of situations were created on which the generalisation abilities of the proposed method were tested. A global and fixed noise level of $\sigma_{nt}^2 = 0.01$ was used in the data models in these experiments [3]. We studied the prediction performance for both two and three component models ($K = 2$, $K = 3$) on new, previously unseen measurements, in the two scenarios as reported below. In this setting, the evidence bound peaked at $K = 3$, as can be seen in Figures 13 and 14.

[3] Here, σ_{nt} can be used to both control the compactness of the representation and to simulate a level of measurement uncertainly similar to that encountered in real data.

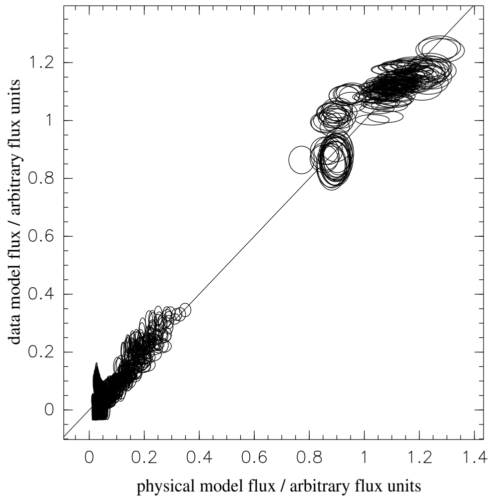

Figure 12. Predictions for the missing values provided by the VB-RFA model versus the physical model. Values of the x-radii of the ellipses represent errors (one standard deviation) obtained from data imputation using synthetic physical models. The y-radius values correspond to one standard deviation of the posteriors of $\boldsymbol{x}$, as obtained using VB-RFA. There are no missing values in the observed spectra in the central region of the wavelength range considered here, which accounts for the gap in the plotted data.

4.3.1 Predictions for previously unseen spectra in the same wavelength range

As the spectra of more and more galaxies become available and are added to public archives, the ability to make predictions for previously unseen spectra, that have a small percentage of missing values within the same overall wavelength range, is an important practical requirement.

It is therefore desirable to be able to infer reliable estimates of missing values, based on a few measurements and the probabilistic representation already created from previously seen spectra. In this scenario, Fig. 13 shows the prediction results obtained by VB-RFA when varying the percentage of missing entries in the test set. The predictions are surprisingly accurate up to a high percentage of missing entries among the previously unseen spectra. Effectively, the high redundancy across the spectra, due to the fact that mature galaxies are very similar to each other, makes this scenario relatively easy to handle.

4.3.2 Predictions for previously unseen wavelength bins

The second scenario investigated here is when the predictions are to be made for missing elements for previously unseen values of wavelength, albeit within the same overall wavelength range.

This is a more difficult task, since the flux varies significantly across different wavelength ranges. We conducted experiments in which the overall wavelength range for each spectrum was split randomly into training sets (regions without missing values in our case, though they need not be so) and test sets, and each model was trained on the training set. Then, to ensure that we retain the missing-at-random (Ghahramani & Jordan 1994) assumption, for randomly picked ranges in the test set, "missing" values were predicted based on the trained model and the prediction errors were averaged over all missing values. This procedure was repeated 10 times with 50 different combinations of training and test sets for each spectrum.

The results are shown in Fig. 14. The percentage of missing values was varied between experiments as is indicated on the x axis. With a complete training set, the predictions are still pretty accurate even when up to 60% of the bins from the test set are missing.

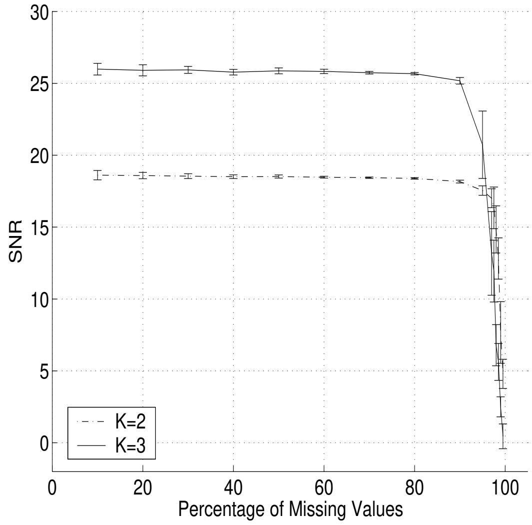

Figure 13. Predictive performance from previously unseen spectra using a VB-RFA model with $K = 2$ (bottom) and $K = 3$ (top), plotted against the percentage of missing entries from the test-set.

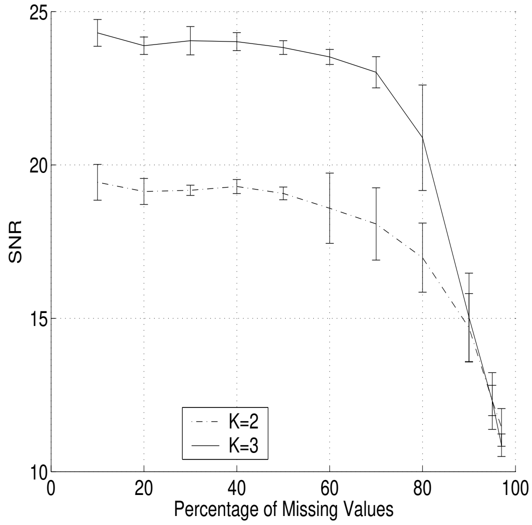

Figure 14. Predictive performance from previously unseen wavelength bins using a VB-RFA model with $K = 2$ (bottom) and $K = 3$ (top). As before, the x-axis indicates the percentage of missing entries from the test-set.

5 FINDING E+A GALAXIES

So far we established generic data-driven procedures for finding evidence of more than one stellar sub-populations, significantly differing in age and metallicity-related characteristic features, in the integrated spectra of galaxies. Here we show an application to a particular data mining project which would benefit from this method, once large sets of galaxy spectra are available for public use.

The determination of the star formation history of early-type galaxies has important implications for the much-debated issue of their formation and evolution. Observational evidence suggests that at least some elliptical galaxies already have the bulk of their stars at high redshift ($z > 2$) (*e.g.* Renzini 1998; Nolan *et al.* 2003), but in hierarchical structure formation models, ellipticals can form at lower redshift from late-type merging (*e.g.* White & Rees 1978; Kauffmann, White, & Guiderdoni 1993). A vital link in understanding how early-type evolution is driven, and under what environmental conditions, is the study of E+A galaxies.

The spectra of these galaxies are characterised by strong hydrogen Balmer absorption-lines together with a lack of significant [OII]3272 Å emission, implying a recent ($\lesssim$1 Gyr) starburst, but no significant on-going star formation. The overall shape of the spectral energy distribution of these galaxies is that expected for an early-type galaxy. They may represent an intermediate stage between disc-dominated rotationally supported systems and spheroid-dominated, pressure-supported systems.

The physical mechanisms responsible for the triggering and subsequent truncation of star-formation in E+A galaxies are currently unclear. They may be the result of disk-disk merging (or galaxy interactions), or cluster-related phenomenon, e.g. ram-pressure stripping, tidal stripping. However, although at intermediate redshift, it has been suggested that E+A-type galaxies are prevalent in clusters (*e.g.* Tran *et al.* 2003), which argues for a cluster-related mechanism, at low-redshift ($z \sim 0.1$), E+As occur in lower-density environments, which is more indicative of merging or interaction processes (Zabludoff *et al.* 1996; Blake *et al.* 2004). Statistical studies of the environments, luminosities and detailed morphologies of these galaxies, across a range of redshifts, are necessary in order to uncover the physical processes governing their star-formation history.

The possible relationship between early-type evolution and E+A galaxies has not as yet been well-studied, as E+A systems are rare (<1 percent of the overall zero-redshift galaxy population, Zabludoff *et al.* (1996)). However, recent and on-going large-scale surveys such as the SDSS and the 2dFGRS offer the opportunity to study these galaxies in a statistically meaningful way, if they can be reliably identified.

In this paper we have developed a rapid, automated method which allows us to identify those early-type galaxies containing a significant young ($<$ 1 Gyr old) population; these galaxies are E+A galaxies. Fig. 15 shows a typical E+A galaxy spectrum, together with its constituent population spectra. Unsurprisingly, given the diagnostic abilities of our components analyses, the component spectra look very similar to the spectra recovered from the components analyses.

Formiggini & Brosch (2004) have analysed the IUE UV spectra of a sample of normal galaxies using PCA. One of their components was associated with a young stellar population, and another with a mature population, consistent with our results. However, they were unable to correlate the principal components with the optical morphology of the galaxies. As we have prior knowledge of the component stellar populations in our galaxy sample, we have been able to make a more quantitative interpretation of our results: if the value of a_1 (the weight of the 'young' component) for a particular early-type galaxy is $>$ 0.5, there is a significant $<$ 1 Gyr-old stellar population present in that galaxy.

Hence, our components analysis is able to rapidly identify post-starburst galaxies, without the uncertainties inherent in the measurement of the equivalent widths of the Balmer absorption and [OII] emission lines. [OII] emission in disk-dominated galaxies may suffer from dust obscuration, and Hγ and Hβ are subject to emission-filling. These, together with the high signal-to-noise requirements, introduce significant uncertainties in the measurement of these lines. In contrast, as our method utilises all the information present in the long-baseline spectra i.e. both the features present and the overall shape of the spectra, it is a more robust diagnostic of recent star-forming activity.

Using the wealth of data available in the SDSS and the 2dFGRS, with signal-to-noise as good or better than the data we analyse here, we may map precisely in what environments star-formation is suppressed (e.g. in groups or clusters), by rapidly and robustly identifying those galaxies hosting only mature (>1 Gyr) stellar populations. This allows us to investigate whether galaxy transformation (from star-forming to passive galaxies) occurs via e.g. ram-pressure stripping, tidal stripping, harassment or suppression of galaxy-galaxy interactions. Although the wavelength ranges of the SDSS and 2dF spectra are shorter than those we have used, we can use a training set of data to identify the linear components (as suggested below), which can then be applied to the SDSS and 2dF data.

With the higher-resolution spectra of these surveys, and higher-resolution model spectra that are available (e.g. (Bruzual & Charlot 2003)), we expect the accuracy of our statistical estimates to increase, as we have more independent samples (wavelength bins) of the mixed components. Whether or not the number of components, K, would increase would depend on the data set used. Both the existence of additional components and the interpretation of any further components can only be explored by carrying out the experiment. However, we would not expect the increase in information that we obtain from higher-resolution spectra to lead to a decrease in K.

We intend to extend our Bayesian modelling technique to identify other classes of galaxies, by using appropriately selected training sets of galaxy spectra and wavelength ranges. For example, starburst galaxies may be identified by the Wolf-Rayet 'bump' around ∼4600–4900 Å, which is a characteristic feature of the emission from the Wolf-Rayet stars in starburst galaxies. By first applying the Bayesian analysis to an appropriate training set of spectra, which include the spectral range of this feature, identification of these galaxies in a large sample should be possible.

6 CONCLUSIONS

Using rigorous Bayesian modelling, we have successfully developed a time-efficient method for the analysis of the spectra of early-type galaxies. At ∼a few seconds per galaxy for this analysis, its speed compares very well with work by other authors (*e.g.* Heavens *et al.* 2000; Fernandes *et al.* 2005). By comparing the results of this analysis with the results of our unique two-component synthetic stellar population fitting to the long-baseline data, and with two-component synthetic spectra, we have shown that the independent components analysis is physically interpretable.

Two linear components are sufficient to describe the bulk of the stellar content of early-type galaxies. One of the components represents a young (<1 Gyr) stellar population, and the other an old, metal-rich population. The relative contribution of these two components allows us to robustly identify which early-type galaxies host significant young stellar populations i.e. which are post-starburst galaxies.

Of the four variational Bayesian methods investigated here, We find the variational Bayesian rectified factor analysis to be the most useful, since it has the greatest flexibility in the shape of the allowed distribution, whilst retaining the astronomically motivated requirement that the flux values in each component spectrum are

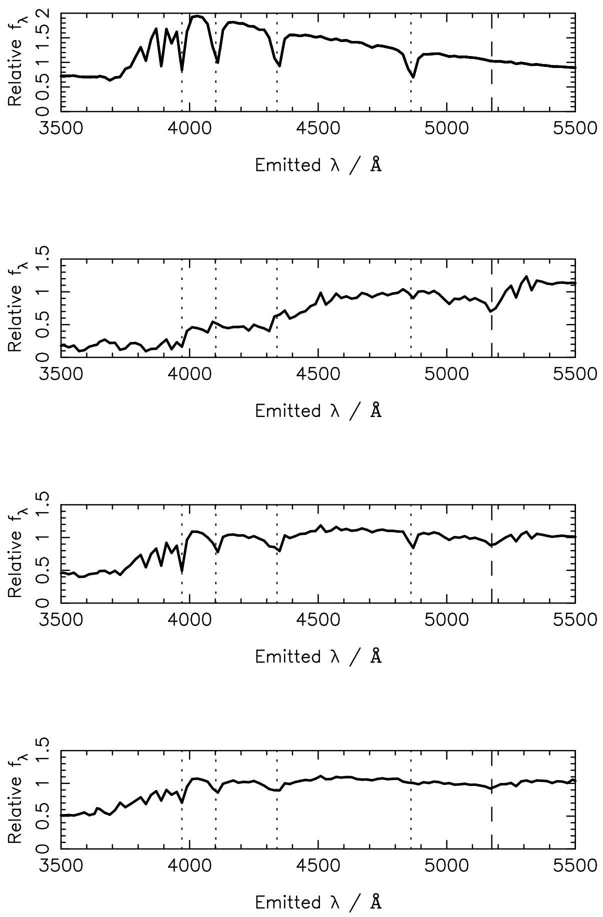

Figure 15. *Top:* the optical spectrum of an A3V star; *second from top:* spectrum of a 12 Gyr stellar population; *third from top:* combined spectrum of the A star plus 12 Gyr stellar population: a typical E+A spectrum. *Bottom:* The observed spectrum of NGC 7252, from our sample, which contains a very young < 1 Gyr stellar population.

non-negative. This method achieved the highest evidence and the highest correlation between the component weights and the population parameters determined from spectral fitting. The method allows the observational errors to be included in the formulation, and has the additional powerful attribute that it can be used to recover missing regions of spectral coverage. The strength of the correlation between the posterior mean predictions of the data model, and the values imputed from the best-fitting physical model, is a strong indication that the data-driven analysis and the physical analysis are well matched.

We intend to apply our analysis to the early-type galaxies, in the 2dF, 6dF and SDSS catalogues, as a useful tool to investigate their evolution with environment and redshift. We have shown that in addition to efficiently finding early-type galaxies with significant young stellar populations, this method would also be an efficient tool to find other distinctive sub-classes, such as E+A galaxies, which is an important missing link in current population scenarios of the formation of early-type galaxies.

In this work, we use model and observed spectra between 2500-8000 Å. However, to be useful as a data mining tool for $z \sim 0$ galaxies in the SDSS, 2dF and 6dF archives, the Bayesian analysis has to work efficiently in the 3700–8000 Å range equally well. Our next step will be to estimate data models that will learn, from the current datasets that include UV spectra, to work on much larger, solely optical datasets.

In future studies, we therefore intend to:

- estimate the data models for the 3700−8000 Å range;
- apply our technique to 2dF, 6dF, SDSS catalogued spectra, to identify E+A galaxies and their environments;
- experiment with identifying starburst galaxies;
- consider the robustness of our data models with respect to uncertainties in the calibration and response of the 2dF/6dF/SDSS fibres, and the effect of random continuum errors, via analyses of simulated data with additional noise factors;
- extend the models to deal with stochastic censoring, i.e. non-ignorable (versus random) missing data.

ACKNOWLEDGEMENTS

Markus Harva was supported by a Paul & Yuanbi Ramsay research award at the School of Computer Science of the University of Birmingham.

REFERENCES

Attias H., 2000, in Advances in Neural Information Processing Systems, vol. 12, pp. 209–215, eds. S. Solla, T.K. Leen and K-L Muller, eds, Cambridge MA, MIT Press
Bacon R., *et al.*, 2001, MNRAS, 326, 23
Bernardo J.M. & Smith A.F., 1994, Bayesian Theory, New York: Wiley
Bica E., Alloin D., 1987a, A&AS, 70, 281
Bica E. & Alloin D., 1987b, A&A, 186, 49
Bishop C.M., 1999, Proc. International Conference on Artificial Neural Networks(ICANN), vol. 1, pp. 509–514
Bishop C.M., Spiegelhalter D. & Winn J., 2002, in Advances in Neural Information Processing Systems, Volume 15, pp. 793−800, eds. S. Becker, S. Thrun, and K. Obermeyer
Blake C., *et al.*, 2004, MNRAS, 355, 713
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Chan K., Lee T.-W., & Sejnowski T.J., 2003, Neural Computation, 15, 1991
Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995, AJ, 110, 1071
Cover T., & Thomas J., 1991, Elements of Information Theory, John Wiley and Sons, New York.
Dressler A., 1980, ApJ, 236, 351
Emsellem E. *et al.*, 2004, MNRAS, 352, 721
Fernandes, R. C., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, MNRAS, 358, 363
Folkes S.R., Lahav O. & Maddox S.J., 1996, MNRAS, 383, 651
Formiggini L. & Brosch N., 2004, MNRAS, 350, 1067
Frey B.J. & Hinton G.E., 1999, Neural Computation, 11(1): 193
Ghahramani Z. & Jordan M., 1994, Technical Report CBCL Paper No. 108, Massachusetts Institute of Technology, Cambridge MA
Harva M. & Kabán A., 2005, Proc. IEEE Int. Conf. Neural Networks, to appear, pp. 185–190
Harva M., 2004, MSc thesis, Helsinki University of Technology
Heavens, A. F., Jimenez, R., & Lahav, O. 2000, MNRAS, 317, 965
Hinton G.E. & Ghahramani Z., 1997, Phil. Trans. Roy. Soc. B. 352, 1177

Hyvärinen A., Karhunen J. & Oja E., 2001, in 'Independent Component Analysis', pub. J. Wiley
Jansen R.A., Fabricant D., Franx M., Caldwell N., 2000, ApJS, 126, 331
Jimenez R., MacDonald J., Dunlop J.S., Padoan P. & Peacock J.A., 2004, MNRAS, 349, 240
Jordan M., Ghahramani Z., Jaakkola T. & Saul L., 1999, Machine Learning, 37, 183
Kabán A., Nolan L.A. & Raychaudhury S., 2005, in proceedings of the SIAM INternational Conference on Data Mining (SIAM DM05), Eds: H. Kargupta, J. Srivastava, C. Kamath and Arnold Goodman, SIAM Press, pp 183-194. (astro-ph/0505059)
Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Kauffmann, G., et al. 2003, MNRAS, 341, 33
Lappalainen H., 1999, in Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99), pp 7–12
Lee D. & Seung S., 2001, Advances in Neural Information Processing Systems, 13,556
MacKay D.J.C., 2003, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, Cambridge, UK
Madgwick D.S., Somerville R., Lahav O. & Ellis R., MNRAS, 2003a, 343, 871
Madgwick D.S., Coil A., Conselice C.J., Cooper M.C., Davis M., Ellis R.S., Faber S.M., Finkbeiner D.P., Gerke B., Guhathakurta P., Kaiser N., Koo D.C., Newman J.A., Phillips A.C., Steidel C.C., Weiner B.J., Willmer N.A. & Yan R., 2003b, ApJ, 599, 997
Miskin J. & Mackay D.J.C., 2001, in Independent Component Analysis: Principles and Practice, pages 209–233, eds. S. Roberts and R. Everson, Cambridge University Press
Miskin J., 2000, PhD Thesis, University of Cambridge
Nolan L.A., 2002, PhD thesis, University of Edinburgh
Nolan L.A., Dunlop J.S., Jimenez R. & Heavens A.F., 2003, MNRAS, 341, 464
Ocvirk, P., Pichon, C., Lancon, A., Thiebaut, A., 2005, accepted for publication in MNRAS, astro-ph/0505209
P. Ocvirk (1), Ocvirk, P., Pichon, C., Lancon, A., Thiebaut, A., 2005, accepted for publication in MNRAS, astro-ph/0507002
Peebles J.E.P., 2002, in A New Era in Cosmology, ASP Conference Proceedings, Vol. 283. eds. N. Metcalfe and T. Shanks.
Ponder J. M., Burstein D., O'Connell R. W., Rose J.A., Frogel J.A., Wu C.-C., Crenshaw D.M., Rieke M.J., Tripicco M., 1998, AJ, 116, 2297
Raiko T., Valpola H., Harva M. & Karhunen J., 2005, Journal of Machine Learning Research, submitted
Renzini A., 1998, in 'When and how do bulges form and evolve?', ed. C.M. Carollo, H.C. Ferguson, R.F.G. Wyse, Cambridge University Press
Ronen S., Aragon-Salamanca A. & Lahav O., 1999, MNRAS, 303, 284
Slonim N., Somerville R., Tishby N. & Lahav O., 2001, MNRAS, 323, 270
Socci N.D., Lee D.D. & Seung H.S, 1997, The Rectified Gaussian Distribution, NIPS'97
Solorio, T., Fuentes, O., Terlevich, R., & Terlevich, E. 2005, MNRAS, 363, 555
Storchi-Bergmann T., Calzetti D., Kinney A. L. 1998, A database of UV−optical spectra of nearby quiescent and active galaxies, (*http://www.stsci.edu/ftp/catalogs/nearby_gal/*).
Tipping M. & Bishop C., 1999, Probabilistic Principal Component Analysis, Journal of the Royal Statistical Society, Series B, 61:3, 611
Tran K. H., Franx M., Illingworth G., Kelson D. D., van Dokkum P., 2003, ApJ, 599, 865
Valpola H., Harva M. & Karhunen J., 2004, Signal Processing, 84(2), 267
Valpola H., Honkela A., Harva M., Ilin A., Raiko T. & Östman T., 2003, Bayes Blocks Software Library http://www.cis.hut.fi/projects/bayes/ software/
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Winn, J., Bishop C. M., 2005, Journal of Machine Learning Research, 6,661-694
Zabludoff A. I., Zaritsky D., Lin H., Tucker D., Hashimoto Y., Shectman S. A., Oemler A., Kirshner R. P., 1996, ApJ, 466, 104

APPENDIX A: IMPLEMENTATION DETAILS OF THE VB-RFA DATA MODEL

A1 Data model estimation

A1.1 Posterior updates for $q(s_{kt})$

The following simplifying notations will be used in this subsection: $s \equiv s_{kt}, m_s \equiv \langle m_{s_k} \rangle, v_s \equiv \langle e^{v_{s_k}} \rangle^{-1}$. Further, the following expressions, involving variables from the Markov blanket (MacKay 2003) of a variable s, will be required: $v_x \equiv \left\{ \sum_{n=1}^{N} \langle e^{v_{x_n}} \rangle \langle a_{nk}^2 \rangle \right\}^{-1}$ and $x \equiv v_x \left\{ \sum_{n=1}^{N} \langle e^{v_{x_n}} \rangle \langle a_{nk} \rangle \left(\langle x_{nt} \rangle - \sum_{j \neq k} \langle a_{nj} \rangle \langle f(s_{jt}) \rangle \right) \right\}$. Using these notations, as well as expressing $f(s) = u(s)s$, where $u(s)$ is the step function, then the outline of the computation of the free form posterior $q(s)$ and its sufficient statistics can be given as follows:

$$q(s) = \frac{1}{Z} \mathcal{N}(x|u(s)s, v_x) \mathcal{N}(s|m_s, v_s) \tag{A1}$$

$$= \frac{1}{Z} w_{u(s)} \mathcal{N}(s|\bar{s}_{u(s)}, \tilde{s}_{u(s)}) \tag{A2}$$

where

$$\tilde{s}_{u(s)} = \{u(s)/v_x + 1/v_s\}^{-1} \tag{A3}$$

$$\bar{s}_{u(s)} = \tilde{s} \{u(s)x/v_x + m_s/v_s\} \tag{A4}$$

$$w_{u(s)} = \mathcal{N}(x|u(s)m_s, u(s)v_s + v_x) \tag{A5}$$

$$Z = \int_{-\infty}^{\infty} ds w_{u(s)} \mathcal{N}(s|\bar{s}_{u(s)}, \tilde{s}_{u(s)})$$

$$= \sum_{u=0}^{1} w_u \operatorname{erfc} \left\{ (-1)^u \bar{s}_u / \sqrt{2\tilde{s}_u} \right\}. \tag{A6}$$

For more detailed derivation, see Harva & Kabán (2005).

Note the non-trivial variational posterior (A2) that resulted from the free-form approximation, a bimodal distribution (A6), which is essentially a mixture of two rectified Gaussians.

The posterior statistics associated with $q(s)$ that are required for updating the posterior statistics of other variables of the model are the first and second order moments, as follows. Denoting $e_u = \operatorname{erfc} \left\{ (-1)^u \bar{s}_u / \sqrt{2\tilde{s}_u} \right\}$ (which is the erfc-term in (A6)), the i-th order moments of s and $f(s)$ can be written as

$$\langle f^i(s) \rangle = \int f^i(s) q(s) ds = M_i(u=1), \tag{A7}$$

where

$$M_i(u) \equiv \frac{w_u}{2Z} e_u \int s^i \mathcal{N}^R(s|(-1)^{1-u} \bar{s}_u, \tilde{s}_u)\, ds,$$

and

$$\langle s^i \rangle = \sum_{u=0}^{1} M_i(u). \tag{A8}$$

The notation $M_i(u)$ introduced in (A7) will be used also in (A11). The remaining integral is the i-th moment of a rectified Gaussian density. The exact expressions for these, are given in §A4 and more details on it can be found e.g. in Harva (2004), along with numerical issues concerning an efficient implementation.

A1.2 Posterior updates for $q(m_{s_k})$

The posteriors $q(m_{s_k}) = \mathcal{N}(m_{s_k}|\bar{m}_{s_k}, \tilde{m}_{s_k})$ are of a Gaussian form, where

$$\tilde{m}_{s_k} = \{\langle e^{v_{m_s}} \rangle + T \langle e^{v_s} \rangle\}^{-1}$$

$$\bar{m}_{s_k} = \tilde{m}_s \left(\langle e^{v_{m_s}} \rangle \langle m_{m_s} \rangle + \langle e^{v_{s_k}} \rangle \sum_{t=1}^{T} \langle s_{kt} \rangle \right)$$

which provide the required posterior expectation $\langle m_{s_k} \rangle = \bar{m}_{s_k}$. Note that if a 'vague prior' was used on m_{s_k}, then the posterior mean estimate $\langle m_{s_k} \rangle$ becomes similar to a maximum likelihood estimate $\sum_t \langle s_{kt} \rangle / T$, as then $\langle m_{m_s} \rangle = m_{m_s} = 0$ and $\langle e^{v_{m_s}} \rangle = e^{v_{m_s}} \approx 0$.

A1.3 Posterior updates for $q(a_{nk})$

The posteriors $q(a_{nk}) = \mathcal{N}^R(a_{nk}|\bar{a}_{nk}, \tilde{a}_{nk})$ are rectified Gaussians, where

$$\tilde{a}_{nk} = \left\{ 1 + \langle e^{v_{x_n}} \rangle \sum_{t=1}^{T} \langle f^2(s_{kt}) \rangle \right\}^{-1}$$
$$\bar{a}_{nk} = \tilde{a}_{nk} \langle e^{v_{x_n}} \rangle \sum_{t=1}^{T} \left(\langle x_{nt} \rangle - \sum_{j \neq k} \langle a_{nj} \rangle \langle f(s_{jt}) \rangle \right) \langle f(s_{kt}) \rangle$$

which provide the required sufficient statistics: $\langle a_{nk}^i \rangle = \int a_{nk}^i \mathcal{N}^R(a_{nk}|\bar{a}_{nk}, \tilde{a}_{nk}) da_{nk}$.

A1.4 Posterior updates for $q(x_{nt})$

The posteriors $q(x_{nt}) = \mathcal{N}(x_{nt}|\bar{x}_{nt}, \tilde{x}_{nt})$ are Gaussian, where

$$\tilde{x}_{nt} = \left\{ \sigma_{y_{nt}}^{-2} b_{nt} + \langle e^{v_{x_n}} \rangle \right\}^{-1}$$
$$\bar{x}_{nt} = \tilde{x}_{nt} \left(\sigma_{y_{nt}}^{-2} b_{nt} y_{nt} + \langle e^{v_{x_n}} \rangle \sum_{k=1}^{K} \langle a_{nk} \rangle \langle f(s_{kt}) \rangle \right)$$

These provide the required posterior expectation $\langle x_{nt} \rangle = \bar{x}_{nt}$. Here, the indicator variable $b_{nt} = 1$ if y_{nt} is observed and zero if it is missing.

A1.5 Posterior updates for variance parameters

Following Raiko *et al.* (2005) and Valpola, Harva & Karhunen (2004), in the exp-parameterisation (see §4.1.1), Gaussian priors are imposed on v_θ, where θ stands for any of the variables x, s or m in Fig. 2. This is then solved by a fixed form Gaussian variational posterior,

$$q(v_\theta) = \mathcal{N}(v_\theta|\bar{v}_\theta, \tilde{v}_\theta), \tag{A9}$$

and this has been employed in the experiments described in this paper.

Here, $\bar{v}_\theta$ and $\tilde{v}_\theta$ are computed by numerical optimisation (Raiko *et al.* 2005; Valpola, Harva & Karhunen 2004). Alternatively, a direct parameterisation may be used with Gamma priors and Gamma posteriors may be sufficient for this model.

A2 The log evidence bound

Computing the log of the evidence bound is useful for monitoring the convergence of the iterative estimation process, and more importantly, for comparing different modelling hypotheses. As mentioned, (5) decomposes into a number of terms.

$$\log p(\boldsymbol{Y}) \geq \sum_{nt} \langle \log p(y_{nt}|x_{nt}) \rangle + \sum_{\theta \in \boldsymbol{\theta}} \left\{ \langle \log p(\theta|\operatorname{pa}(\theta)) \rangle - \langle \log q(\theta) \rangle \right\}. \tag{A10}$$

The p and q terms for Gaussians will be given only once, while those that are different will be detailed separately.

A2.1 p- and q-terms of Gaussian variables

The p and q terms in (A10), in the case of Gaussian variables, can be written as:

$$\langle \log p(y_{nt}|x_{nt}, \sigma_{nt}^2) \rangle = -\frac{1}{2} \left\{ \frac{1}{\sigma_{nt}^2} \left[(y_{nt} - \langle x_{nt} \rangle)^2 + \operatorname{Var}\{x_{nt}\} \right] + \log 2\pi\sigma_{nt}^2 \right\}$$
$$\langle \log p(\theta|m_\theta, e^{-v_\theta}) \rangle = -\frac{1}{2} \left\{ \langle e^{v_\theta} \rangle \left[(\bar{\theta} - \langle m_\theta \rangle)^2 + \tilde{\theta} + \operatorname{Var}\{m_\theta\} \right] - \langle v_\theta \rangle + \log 2\pi \right\}$$
$$-\langle \log q(\theta) \rangle = \frac{1}{2} + \frac{1}{2} \log 2\pi\tilde{\theta},$$

where $\bar{\theta}$ and $\tilde{\theta}$ are the posterior mean and variance of θ. If the exp-parameterisation is used for the variances, then the term $\langle e^{v_\theta} \rangle = e^{\bar{v}_\theta + \tilde{v}_\theta/2}$ — see Valpola, Harva & Karhunen (2004).

A2.2 p- and q-terms of a_{nk}

The p and q terms of a_{nk}, which are the elements of the matrix $\boldsymbol{A}$ in (A1), are given by:

$$\langle \log \mathcal{N}^R(a|0,1) \rangle = -\frac{1}{2} \left\{ \log 2\pi + \langle a \rangle^2 + \operatorname{Var}\{a\} \right\} + \log 2$$
$$-\langle \log q(a) \rangle = \frac{1}{2\tilde{a}} \left\{ \operatorname{Var}\{a\} + (\langle a \rangle - \bar{a})^2 \right\} - \frac{1}{2} \log \frac{2}{\pi\tilde{a}} + \log \operatorname{erfc}(-\bar{a}/\sqrt{2\tilde{a}}),$$

where we have dropped the indices.

A2.3 The q-term of s_{kt}

The p-term of variable s is computed as in the case of other Gaussian variables. The q-term gives the following:

$$-\langle \log q(s)\rangle = \sum_{u=0}^{1} \left\{ \left[\frac{\bar{s}_u^2}{2\tilde{s}} - \log \frac{w_u}{Z\sqrt{2\pi\tilde{s}_u}} \right] M_0(u) - \frac{\bar{s}_u}{\tilde{s}_u} M_1(u) + \frac{1}{2\tilde{s}_u} M_2(u) \right\}. \tag{A11}$$

For detailed derivation see Harva & Kabán (2005).

A3 Missing value prediction

A Gaussian approximation of the predictive distribution can be obtained as follows.

$$q(y_{nt}^m|\boldsymbol{y}_t^o) = \int d\boldsymbol{\theta} p(y_{nt}^m|\boldsymbol{\theta})q(\boldsymbol{\theta}) \tag{A12}$$

$$= \int d\boldsymbol{\theta}_{-x_{nt}} \int dx_{nt} p(y_{nt}|x_{nt})p(x_{nt}|\boldsymbol{\theta}_{-x_{nt}})q(\boldsymbol{\theta}_{-x_{nt}}) \tag{A13}$$

$$= \int d\boldsymbol{\theta}_{-x_{nt}} \mathcal{N}(y_{nt}^m | \sum_k a_{nk} f(s_{kt}), \sigma_{y_{nt}^m}^2 + e^{-v_{x_n}}) q(\boldsymbol{\theta}_{-x_{nt}}) \tag{A14}$$

$$\approx \mathcal{N}(y_{nt}^m | \sum_k \langle a_{nk}\rangle\langle f(s_{kt})\rangle, \sigma_{y_{nt}^m}^2 + \langle e^{v_{x_n}}\rangle^{-1}) \tag{A15}$$

where $\boldsymbol{\theta}_{-x_{nt}}$ denotes hidden variables above x_{nt} in the conditional dependency architecture of the model. The mean of this distribution provides a prediction for the missing entry y_{nt}. Conveniently, for missing entries the mean of (A15) is identical to that of $q(x_{nt})$ (as $b_{nt} = 0$), so that the imputation of missing values takes no additional computation. Further, evaluating (A15) when $\sigma_{y_{nt}}^2 = 0$ (i.e. no measurement of errors) – in order to test the prediction performance – also happens to be identical to evaluating $q(x_{nt})$ from (A15) and §A1.4.

A4 Required properties of $\mathcal{N}^R(\theta|\bar{\theta}, \tilde{\theta})$

The rectified Gaussian distribution along with its first and second order moments, used in §A1.1 and §A1.3, are listed here for completeness.

$$\mathcal{N}^R(\theta|\bar{\theta}, \tilde{\theta}) = \frac{2}{\mathrm{erfc}(\bar{\theta}/\sqrt{2\tilde{\theta}})} u(\theta)\mathcal{N}(\theta|\bar{\theta}, \tilde{\theta}) \tag{A16}$$

$$\langle\theta\rangle = \int \theta\mathcal{N}^R(\theta|\bar{\theta}, \tilde{\theta})d\theta = \bar{\theta} + \sqrt{\frac{2\tilde{\theta}}{\pi}} \frac{1}{\exp\left\{(\bar{\theta}/\sqrt{2\tilde{\theta}})^2\right\} \mathrm{erfc}(-\bar{\theta}/\sqrt{2\tilde{\theta}})}$$

$$\langle\theta^2\rangle = \int \theta^2\mathcal{N}^R(\theta|\bar{\theta}, \tilde{\theta})d\theta = \bar{\theta}^2 + \tilde{\theta} + \sqrt{\frac{2\tilde{\theta}}{\pi}} \frac{\bar{\theta}}{\exp\left\{(\bar{\theta}/\sqrt{2\tilde{\theta}})^2\right\} \mathrm{erfc}(-\bar{\theta}/\sqrt{2\tilde{\theta}})}$$

See Harva (2004) or Miskin (2000) for details.