Draft version June 21, 2019
Typeset using LaTeX **twocolumn** style in AASTeX62

A data-driven model of nucleosynthesis with chemical tagging in a lower-dimensional latent space

Andrew R. Casey,[1,2] John C. Lattanzio,[1] Aldeida Aleti,[2] David Dowe,[2] and the GALAH team

[1] *School of Physics & Astronomy, Monash University, Wellington Rd, Clayton 3800, Victoria, Australia*
[2] *Faculty of Information Technology, Monash University, Wellington Rd, Clayton 3800, Victoria, Australia*

(Received 2019; Revised 2019; Accepted 2019)

Submitted to AAS Journals

ABSTRACT

Chemical tagging seeks to identify unique star formation sites from present-day stellar abundances. Previous techniques have treated each abundance dimension as being statistically independent, despite theoretical expectations that many elements can be produced by more than one nucleosynthetic process. In this work we introduce a data-driven model of nucleosynthesis where a set of latent factors (e.g., nucleosynthetic yields) contribute to all stars with different scores, and clustering (e.g., chemical tagging) is modelled by a mixture of multivariate gaussians in a lower-dimensional latent space. We use an exact method to simultaneously estimate the factor scores for each star, the partial assignment of each star to each cluster, and the latent factors common to all stars, even in the presence of missing data entries. We use an information-theoretic Bayesian principle to estimate the number of latent factors and clusters. Using the second *Galah* data release we find that five latent factors are preferred to explain $N = 1{,}072$ stars with 18 chemical abundances. We identify the rapid- and slow-neutron capture processes, as well as latent factors consistent with Fe-peak and α-element production, and another where only K and Zn dominate. When we consider $N \sim 100{,}000$ stars with missing abundances we find another seven factors, as well as 16 components in latent space. Despite these components showing separation in chemistry that is explained through different yield contributions, none show significant structure in their positions or motions. We argue that more data, and joint priors on cluster membership that are constrained by dynamical models, are necessary to realise chemical tagging at a galactic-scale. We release accompanying software that scales well with the available data, allowing for model parameters to be optimised in seconds given a fixed number of latent factors, components, and $\sim 10^7$ abundance measurements.

Keywords: Bayesian statistics (1900), Chemical abundances (224), Galaxy chemical evolution (580)

1. INTRODUCTION

The detailed chemical abundances that are observable in a star's photosphere provide a fossil record that carries with it information about where and when that star formed (Freeman & Bland-Hawthorn 2002). While the photospheric abundances remain largely unchanged throughout a star's lifetime (however see Dotter et al. 2017; Ness et al. 2018a), the dynamical dissipation timescale of open clusters in the Milky Way disc is of order a few gigayears (Portegies Zwart et al. 1998). That makes chemical tagging an attractive approach to identify star formation sites long after those stars are no longer gravitationally bound to each other.

Gravitationally bound star clusters have been useful laboratories for testing the limits and utility of chemical tagging. Although biases arise when only considering star clusters that are still gravitationally bound, the chemical homogeneity of open clusters provides an empirical measure of how similar stars would need to be before they could be tagged as belonging to the same star formation site (Mitschang et al. 2014). However, there are analysis issues in understanding how precisely those chemical abundances can be measured (Bovy 2016), and how chemically similar stars can be that did not form

Corresponding author: Andrew R. Casey
andrew.casey@monash.edu

together (dopplegängers; Ness et al. 2018b). If open clusters were truely chemically homogeneous then under idealistic assumptions our ability to chemically tag the Milky Way would depend primarily on the precision with which we can measure those chemical abundances in stars. Data-driven approaches to modelling stellar spectra are improving upon this precision (Ness et al. 2015; Ness 2018; Ness et al. 2018a; Casey et al. 2016a, 2017; Ho et al. 2017a,b; Leung & Bovy 2018), but more work is needed: astronomers have not yet developed unbiased estimators of chemical abundances that saturate the Cramér-Rao bound (Cramér 1946; Rao 1945).

Chemical tagging experiments require a catalogue of precise chemical abundance measurements for a large number of stars, where those chemical abundances trace different nucleosynthetic pathways. This is the primary goal of the *Galah* survey (De Silva et al. 2015; Martell et al. 2017; Buder et al. 2018), a stellar spectroscopic survey that uses the High Efficiency and Resolution Multi-Element Spectrograph (*HERMES*; Sheinis et al. 2015) on the Australian Astronomical Telescope (AAT). *Galah* will observe up to 10^6 stars in the Milky Way, and measure up to 30 chemical abundances for each star. This includes light odd-Z elements (e.g., Na, K), elements produced through alpha-particle capture (e.g., Mg, Ca, Ti), and elements produced through the slow (e.g., Ba) and rapid neutron-capture process (e.g., Eu). No other spectroscopic survey provides an equivalent set of chemical abundances for a comparable number of stars.

Given these data and the most favourable assumptions in chemical tagging – that star clusters are truely chemically homogenous, that we can measure those abundances with infinite precision, and that those abundances are differentiable between star clusters – then chemical tagging becomes a clustering problem. All clustering techniques applied to chemical tagging thus far have assumed that the data dimensions are independent. That is to say that adding a dimension of say [Ni/H] provides independent information that could not have been predicted from other elemental abundances. Theory and observations agree that this cannot be true. Nucleosynthetic processes produce multiple elements in varying quantities, and the effective dimensionality of stellar abundance datasets has been shown to be lower than the actual number of abundance dimensions (Ting et al. 2012; Price-Jones & Bovy 2018). Any clustering approach that treats each new elemental abundance as an independent axis of information will therefore conclude with biased inferences about the star formation history of our Galaxy.

It is not trivial to confidently estimate the nucleosynthetic yields that have contributed to the chemical abundances of each star. There are qualitative statements that can be made for large numbers of stars, or particular types of stars, but quantifying the precise contribution of different processes to each star is an unsolved problem. For example, the so-called [α/Fe] 'knee' in abundance ratios in the Milky Way can qualitatively be explained by core-collapse supernovae being the predominant nucleosynthetic process in the early Milky Way before Type Ia supernovae made a significant contribution, but efforts to date have not sought to try to explain the detailed abundances of stars as a contribution of yields from different systems (however see West & Heger 2013). This is in part because of the challenging and degenerate nature of the problem as described, and is complicated by the differences in yield predictions that account from prescriptions used in different theoretical models.

New approaches to chemical tagging are clearly needed. Immediate advances would include methods that take the dependence among chemical elements into account within some generative model, or techniques that combine chemical abundances with dynamical constraints to place joint prior probabilities on whether any two stars could have formed from the same star cluster, given some model of the Milky Way.

In this work we focus on the former. Here we present a new approach to chemical tagging that allows us to identify the latent (unobserved) factors that contribute to the chemical abundances of all stars (e.g., nucleosynthetic yields) while simultaneously performing clustering in the latent space. Notwithstanding caveats that we will discuss in detail, this allows us to infer nucleosynthetic yields rather than strictly prescribe them from models. Moreover, the scale of the clustering problem reduces by a significant fraction because the clustering is performed in a lower dimensional latent space instead of the higher dimensional data space. In Section 2 we describe the model and the methods we use to estimate the model parameters. Section 3 describe the experiments performed using generated and real data sets. We discuss the results of these experiments in Section 4, including the caveats with the model as described. We conclude in Section 5.

2. METHODS

Factor analysis is a common statistical approach for describing correlated observations with a lower number of latent variables (e.g., Thompson 2004). Related techniques include principal component analysis (Hotelling 1933) and its variants (Tipping & Bishop 1999), singular value decomposition (Golub & Reinsch 1970), and

other matrix factorization methods. While factor analysis on its own is a useful dimensionality reduction tool to identify latent factors that contribute to the chemical abundances of stars (e.g., Ting et al. 2012; Price-Jones & Bovy 2018), factor analysis cannot describe clustering in the data (or latent) space. Similarly, clustering techniques applied to chemical abundances to date (e.g., Hogg et al. 2016) do not account for the lower effective dimensionality in elemental abundances.

Here we expand on a variant of factor analysis known elsewhere as a mixture of common factor analyzers (Baek et al. 2010), where the data are generated by a set of latent factors that are common to all data, but the scoring (or extent) of those factors is different for each data point, and the data can be modelled as a mixture of multivariate normal distributions in the latent space (factor scores). In this work the data $\mathbf{X}$ is a $N \times D$ matrix where N is the number of stars and D is the number of chemical abundances measured for each star. We assume a generative model for the data

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{LS} + \mathbf{e} \tag{1}$$

where $\mathbf{L}$ is a $J \times D$ matrix of factor loads that is common to all data points, J is the number of latent factors, and the factor scores for the nth data point

$$\mathbf{S}_n \sim \mathcal{N}(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k) \tag{2}$$

are drawn from[1] the kth multivariate normal distribution. The mean vector $\boldsymbol{\mu}$ describes the mean datum in each dimension. The factor scores for all data points $\mathbf{S}$ is then a $N \times J$ matrix, where each data point has a partial association to the components in latent space. We assume $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\mathbf{D}))$ is independent of the latent space, and $\mathbf{D}$ is a vector of variances in each D abundance dimensions. In this model each data point can be represented as being drawn from a mixture of multivariate normal components, except the components are *clustered in the latent space* $\mathbf{S}$ and projected into the data space by the factor loads $\mathbf{L}$.

We assume that the latent space is lower dimensionality than the data space (e.g., $J < D$). Within the context of stellar abundances, the factor loads $\mathbf{L}$ can be thought of as the *mean* yields of nucleosynthetic events (e.g., s-process production from AGB stars averaged over initial mass function and star formation history), and the factor scores are analogous to the relative counts of those nucleosynthetic events. The clustering in factor scores achieves the same as a clustering procedure in data space, except we simultaneously estimate the latent processes that are common to all stars (the so-called factor loads, analogous to nucleosynthetic yields). Within this framework a rare nucleosynthetic event can still be described as a 'factor load' $\mathbf{L}_j$, but its rarity would be represented by associated factor scores being zero for most stars and thus have no contribution to the observed abundances. In practice the factor loads can only be identified up to orthogonality and cannot be expressly interpreted as nucleosynthetic yields because they have limited physical meaning (we discuss this further in Section 4), but this description of typical yields and relative event rates should help build intuition for the model parameters, and provide context within the astrophysical problem it is being applied.

Including latent factors in the model description allows us to account for processes that affect multiple elemental abundances. In this way we are accounting for the fact that the data dimensions are not independent of each other. Another benefit is the scaling with computational cost. If we considered data sets of order $10^{7.5}$ entries (e.g., 30 chemical abundances for 10^6 stars) purely as a clustering problem, then even the most efficient clustering algorithms would incur a significant cumulative computational overhead by searching the parameter space for the number of clusters, and the optimal model parameters given that number of components. However, because the mixture of factor analyzers approach assumes that there is a *lower dimensional latent space* in which the data are clustered, and that clustering is projected into real space by common factor loads, the dimensionality of the clustering problem is reduced from $N \times D$ to $N \times J$. This reduces computational cost through faster execution of each optimization step, and on average fewer optimization steps needed to reach a specified convergence threshold.

From a statistical standpoint, the primary advantage to using a mixture of factor analysers is that we can simultaneously estimate latent factors (e.g., infer nucleosynthetic yields) and perform clustering (e.g., chemical tagging) within a statistically consistent framework. That is to say that we have a generative data-driven model that can quantitatively describe nucleosynthetic yields, and the factor scores can explain the variance in turbulence and gas mixing, or star formation efficiency, and the parameters of this model can be simultaneously estimated in a self-consistent way with a single scalar-justified objective function.

Without loss of generality the density of the mean-subtracted data $\mathbf{X} - \boldsymbol{\mu}$ (which we hereafter will refer to

[1] For clarifying nomenclature across disciplines, the terminology $z \sim \mathcal{N}(0, 1)$ indicates that the z variable *is drawn from* a standard normal distribution.

simply as $\mathbf{Y}$) can be described as

$$f(\mathbf{Y};\mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{Y}; \mathbf{L}\boldsymbol{\xi}_k, \mathbf{L}\boldsymbol{\Omega}_k\mathbf{L}^\intercal + \mathrm{diag}(\mathbf{D})) \quad (3)$$

given J common factor loadings and K components clustered in the latent (factor score) space. Here the parameter vector $\mathbf{\Psi}$ includes $\{\mathbf{L}, \boldsymbol{\pi}, \boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{D}\}$, and $\phi(\mathbf{Y};\boldsymbol{\theta})$ describes the density of a multivariate gaussian distribution, and π_k describes the relative weighting of the kth component in latent space and $\sum \boldsymbol{\pi}^K = 1$. The log likelihood is then given by

$$\log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi}) = \sum_{k=1}^{K} \log f(\mathbf{Y};\mathbf{\Psi}) \quad . \quad (4)$$

The model as described is indeterminate in that there is no unique solution for the factor loads $\mathbf{L}$ and scores $\mathbf{S}$. These quantities can only be determined up until orthogonality in $\mathbf{L}$. However, as we will describe in Section 2.2, with suitable priors on $\mathbf{\Psi}$ one can efficiently estimate the model parameters using the expectation-maximization algorithm (Dempster et al. 1977).

2.1. *Initialisation*

Here we describe how the model parameters are initialised.[2] To initialise the factor loads $\mathbf{L}$ we start by randomly drawing a $D \times D$ matrix from a Haar distribution (Haar 1933), which is uniform on the special orthogonal group SO(n) and therefore guaranteed to return an orthogonal matrix with a determinant of unity (Stewart 1980). We denote the $J \times D$ left-most region of this matrix to be $\mathbf{H}$, and by taking $\mathbf{L}_*$ to be the Cholesky decomposition of $\mathbf{H}^\intercal\mathbf{H}$, we initialise the factor loads as

$$\mathbf{L} = \mathbf{H}\left((\mathbf{L}_*)^{-1}\,\mathbf{I}\right) \quad (5)$$

which ensures that $\mathbf{L}$ is a $J \times D$ matrix of mutually orthogonal vectors. We then initially assign each data point as belonging to one of the K components using the `k-means++` algorithm (Arthur & Vassilvitskii 2007) in the pseudo-latent space $\mathbf{YL}$. Given the initial factor loads and assignments, we then estimate the relative weights $\boldsymbol{\pi}$, the mean factor scores of each component $\boldsymbol{\xi}$, and the covariance matrix of factor scores of each component $\boldsymbol{\Omega}$. Finally, we initialise the specific variance $\mathbf{D}$ in each dimension as the variance in each data dimension. Other initialisation methods for the latent factors include singular value decomposition (Golub & Reinsch 1970) or generating random noise with orthogonal constraints, and random assignment is an alternative method that is available for initialising assignments.

Throughout this work we repeat this initialisation procedure 25 times for every trial of J and K for a given data set. We then run expectation-maximization (Section 2.2) from each initialisation until the log likelihood improves by less than 10^{-5} per step, and we adopt the model with the highest log likelihood as the preferred model given that trial of J, K, and the data. Although this optimisation procedure is not convex, in practice it is normally sufficient to initialise from many points to avoid local minima.

2.2. *Expectation-Maximization*

We use the expectation-maximization algorithm to estimate the model parameters (Dempster et al. 1977). With each expectation step we evaluate the log likelihood given the model parameters $\mathbf{\Psi}$ and we calculate the $N \times K$ responsibility matrix $\boldsymbol{\tau}$ whose entries are the posterior probability that the nth data point is associated to the kth component, given the data $\mathbf{Y}$ and the current estimate of the parameter vector $\mathbf{\Psi}$:

$$\tau_{nk} = \frac{\pi_k \phi(\mathbf{Y}_n; \mathbf{L}\boldsymbol{\xi}_k, \mathbf{L}\boldsymbol{\Omega}_k\mathbf{L}^\intercal + \mathrm{diag}(\mathbf{D}))}{\sum_{g=1}^{G} \pi_g \phi(\mathbf{Y}_n; \mathbf{L}\boldsymbol{\xi}_g, \mathbf{L}\boldsymbol{\Omega}_g\mathbf{L}^\intercal + \mathrm{diag}(\mathbf{D}))} \quad . \quad (6)$$

At the maximization step we update our estimates of the parameters $\mathbf{\Psi}$, conditioned on the data $\mathbf{Y}$ and the responsibility matrix $\boldsymbol{\tau}$. The updated parameters estimates are found by setting the second derivative of the log likelihood (Eq. 4) to zero and solving for the parameter values.[3] In doing so this guarantees that every updated estimate of the model parameters is guaranteed to increase the log likelihood. Although there are no guarantees against converging on local minima, in practice it is sufficient to run expectation-maximization from multiple initialisations (as we do) in order to ensure that the global minima is reached. At the maximization step we first update our estimate of the relative weights $\boldsymbol{\pi}^{(\mathrm{t}+1)}$ given the responsibility matrix $\boldsymbol{\tau}$

$$\pi_k^{(\mathrm{t}+1)} = \frac{1}{N}\sum_{n=1}^{N} \tau_{nk} \quad (7)$$

where the $\mathbf{\Psi}^{(\mathrm{t})}$ superscript refers to the current parameter estimates and $\mathbf{\Psi}^{(\mathrm{t}+1)}$ refers to the updated estimate

[2] This describes the default initialisation approach. Other approaches are available in the accompanying software.

[3] Strictly this introduces a statistical inconsistency in that we should update our parameter estimates by setting the second derivative of our information-theoretic objective function (Eq. 37) to zero instead of the log likelihood, but this inconsistency only becomes serious with small N – precisely the opposite situation of chemical tagging!

for the next iteration. The updated estimates of the mean factor scores $\boldsymbol{\xi}^{(\mathrm{t}+1)}$ for each component are then given by

$$\boldsymbol{\xi}_k^{(\mathrm{t}+1)} = \boldsymbol{\xi}_k^{(\mathrm{t})} + \frac{\mathbf{G}^\intercal(\mathbf{Y}^\intercal - \mathbf{L}^{(\mathrm{t})}\boldsymbol{\xi}_k^{(\mathrm{t})})\boldsymbol{\tau}_k}{N\pi_k^{(\mathrm{t}+1)}} \tag{8}$$

where:

$$\mathbf{W} = (\boldsymbol{\Omega}_k^{(\mathrm{t})})^{-1}\mathbf{I} \tag{9}$$

$$\mathbf{V} = \left(\mathbf{D}^{(\mathrm{t})}\right)^{-1} \tag{10}$$

$$\mathbf{C} = (\mathbf{W} + (\mathbf{L}^{(\mathrm{t})})^\intercal\mathbf{V}\mathbf{L}^{(\mathrm{t})})^{-1}\mathbf{I} \tag{11}$$

$$\mathbf{G} = \left[\mathbf{V} - \mathbf{V}\mathbf{L}^{(\mathrm{t})}\mathbf{C}\left(\mathbf{V}\mathbf{L}^{(\mathrm{t})}\right)^\intercal\right]\mathbf{L}^{(\mathrm{t})}\boldsymbol{\Omega}_k^{(\mathrm{t})} \quad . \tag{12}$$

The covariance matrices of the components of factor scores $\boldsymbol{\Omega}^{(\mathrm{t}+1)}$ are updated next,

$$\boldsymbol{\Omega}_k^{(\mathrm{t}+1)} = \left(\mathbf{I} - \mathbf{G}^\intercal\mathbf{L}^{(\mathrm{t})}\right)\boldsymbol{\Omega}_k^{(\mathrm{t})} + \frac{\mathbf{G}^\intercal\mathbf{Z}\left(\mathbf{Z}\boldsymbol{\tau}_k^\intercal\right)^\intercal\mathbf{G}}{N\pi_k^{(\mathrm{t}+1)}} \tag{13}$$

where

$$\mathbf{Z} = \mathbf{Y}^\intercal - \mathbf{L}^{(\mathrm{t})}\boldsymbol{\xi}_k^{(\mathrm{t})} \quad . \tag{14}$$

After some linear algebra, updated estimates of the common factor loads $\mathbf{L}^{(\mathrm{t}+1)}$ can be found from

$$\mathbf{L}^{(\mathrm{t}+1)} = \mathbf{L}_a\left(\mathbf{L}_b^{-1}\mathbf{I}\right) \tag{15}$$

where:

$$\mathbf{L}_\mathrm{a} = \sum_{k=1}^{K}\left[\boldsymbol{\tau}_k^\intercal\mathbf{Y}\left(\boldsymbol{\xi}_k^{(\mathrm{t})}\right)^\intercal + \mathbf{G}^\intercal\boldsymbol{\tau}_k\mathbf{Z}^\intercal\mathbf{G}\right] \tag{16}$$

$$\mathbf{L}_\mathrm{b} = N\sum_{k=1}^{K}\left[\pi_k^{(\mathrm{t}+1)}\left(\boldsymbol{\Omega}_k^{(\mathrm{t}+1)} + \boldsymbol{\xi}_k^{(\mathrm{t}+1)}\left(\boldsymbol{\xi}_k^{(\mathrm{t}+1)}\right)^\intercal\right)\right] \tag{17}$$

Finally, the updated estimate of the specific variances $\mathbf{D}^{(\mathrm{t}+1)}$ are given by

$$\mathbf{D}^{(\mathrm{t}+1)} = \frac{1}{N}\left[\sum_{k=1}^{K}\boldsymbol{\tau}_k^\intercal\left(\mathbf{Y}\odot\mathbf{Y}\right) - \sum_{j=1}^{J}\left(\mathbf{L}^{(\mathrm{t}+1)}\mathbf{L}_\mathrm{b}\right)\odot\mathbf{L}^{(\mathrm{t}+1)}\right] \tag{18}$$

where $\odot$ denotes is the entry-wise (Hadamard) product. Throughout this work we assume that the data are noiseless and we do not add any observed errors to the constructed covariance matrices.

2.3. *Missing data*

The expectation-maximization procedure as described requires that there be no missing data entries in order to update our estimates of the responsibility matrix $\boldsymbol{\tau}$ and the model parameters $\boldsymbol{\Psi}$. In practice, however, there will often be abundance measurements that are missing for some subset of stars. There are many potential reasons for this, including astrophysical explanations (e.g., an absorption line was not present above the noise), observational limitations (e.g., the signal-to-noise ratio was too low, or contamination by a cosmic ray), or various other reasons that cannot be inferred from the available information.

In this work we will assume that any missing data measurements are missing at random. The missing data points can then be treated as unknown parameters that must be solved for (and updated) at each iteration. Initially we impute zeros for missing data entries in $\mathbf{Y}$, and at each iteration we update these imputed valuee with our estimate of what the missing data values are given the current model parameters. This ensures that the log-likelihood increases with each iteration. Similarly, with each update we inflate our estimates of the specific variances based on the fraction of missing data points in each dimension

$$\mathbf{D}_d^{(\mathrm{t}+1)} = \mathbf{D}_d^{(\mathrm{t}+1)}\left(\frac{N}{N-M}\right) \tag{19}$$

where M is a the number of missing data entries in the dth dimension. In Section 3.2 we show with a toy model that the latent factor loads and scores can be reliably estimated even in the presence of high fractions of missing data (e.g., 40%), conditioned on our assumption that the data are missing at random.

2.4. *Model Selection*

The expectation-maximization algorithm as described requires a specified number of latent factors J and K. In the next Section we describe a toy model using generated data where we will assume that the true number of latent factors and components are not known. We require some heuristic to decide how many latent factors and components are preferred given some data. An increasing number of factors and components will undoubtedly increase the log likelihood of the model given the data, but the log likelihood does not account for the increased model complexity that is afforded by those additional latent factors and components.

One criterion commonly employed for evaluating a class of models is the Bayesian Information Criterion (BIC; Schwarz 1978),

$$\mathrm{BIC} = Q\log N - 2\log\mathcal{L}\left(\mathbf{Y}|\boldsymbol{\Psi}\right) \quad , \tag{20}$$

where Q is the number of parameters in this model:

$$Q = \frac{J}{2}\left[2\left(D-J\right) + K\left(3+J\right)\right] + K + D - 1 \quad . \tag{21}$$

While the BIC does include a penalisation term for the number of parameters (which scales with $\log N$), it does not describe for the increased flexibility that is afforded by the addition of those parameters. For example, adding one parameter to a model will increase the BIC by at most $\log N$, but there are different ways for a single parameter to be introduced. In a fictitious model $y = f(x)$ a parameter b could be added that is a scalar multiple of x, or it could be introduced as x^b. Despite the difference in model complexity, the same penalisation occurs in BIC. Even if the log likelihood were only to improve marginally in both cases, the difference in model complexity is not captured by BIC. In other words, there are situations where we are more interested in balancing the model complexity (or the expected Fisher information and similar properties) with the goodness of fit, instead of penalising the number of parameters.

For these reasons we use the Minimum Message Length (MML; Wallace 2005) principle as a criterion for model selection and evaluation. The classically-described principle of MML is that the best explanation of the data given a model is the one that leads to the shortest so-called two-part message (Wallace 2005), where a *message* takes into account both the complexity of the model and its explanatory power. The complexity of the model is described through the first part of the message, and the second part of the message describes its explanatory power. The *length* of each message part is quantified (or estimated) using information theory, allowing for a fair evaluation between different models of varying complexity or explanatory power. MML has been shown to perform well on a variety of empirical analyses (see, e.g., Viswanathan et al. (1999); Fitzgibbon et al. (2004), with references to further examples in Wallace (2005); Dowe et al. (2007); Dowe (2008, 2011)). Arguments about the statistical consistency (i.e., as the number of data points increases the distributions of the estimates become increasingly concentrated near the true value) of MML are given in Dowe & Wallace (1997); Dowe (2011). The MML principle requires that we explicitly specify our prior beliefs on the model parameters, providing a Bayesian optimisation approach which can be applied across entire classes of models.

The *message* must encode two parts: the model, and the data given the model. The encoding of the message is based on Shannon's information theory (Shannon 1948). The information gained from an event e occurring, where $p(e)$ is the probability of that event, is $I(e) = -\log_2 p(e)$. The information content is largest for improbable outcomes, and smallest for outcomes that we are almost certain about. In other words, an outcome that has a probability close to unity has nearly zero information content because almost nothing new is learned from it, whereas rarer events convey a much higher information content.

In practice calculating the message length can be a non-trivial task, especially for models that are reasonably complex. This makes the strict MML principle intractable (or uncomputable) in many cases and necessitates approximations to the message length. Using a Taylor expansion, a generalised scheme can be calculated to estimate the parameter vector $\mathbf{\Psi}$ that minimises the message length $I(\mathbf{\Psi}, \mathbf{Y})$ (Wallace & Freeman 1987),

$$I(\mathbf{\Psi}, \mathbf{Y}) = \frac{Q}{2} \log \kappa_Q - \log \left(\frac{p(\mathbf{\Psi})}{\sqrt{|\mathcal{F}(\mathbf{\Psi})|}} \right) - \log \mathcal{L}\left(\mathbf{Y}|\mathbf{\Psi}\right) + \frac{Q}{2} \tag{22}$$

where $\log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi})$ is the familiar log likelihood, $p(\mathbf{\Psi})$ is the joint prior density on $\mathbf{\Psi}$, $\mathcal{F}(\mathbf{\Psi})$ is the negative second derivative of the log likelihood, commonly referred to as the expected Fisher information matrix,

$$\mathcal{F}(\mathbf{\Psi}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \mathbf{\Psi}^2} \log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi}) \right] \tag{23}$$

and as before Q is the number of model parameters. Continuous parameters can only be stated to finite precision, which leads to the $\frac{Q}{2} \log \kappa_Q$ term that gives a measure of the volume of the region of uncertainty in which the parameters $\mathbf{\Psi}$ are centred. The $\log \kappa_Q$ term is

$$\log \kappa_Q = -\log 2\pi + \frac{1}{Q} \log Q\pi - \gamma - 1 \tag{24}$$

where γ is Euler's constant.

Like the BIC, the message length is penalised by the number of model parameters through the $\log \kappa_Q$ term. However, the model complexity is also described through the priors and the Fisher information, which describes the curvature of the log likelihood with respect to the model parameters. For these reasons, MML provides a more accurate description of the model complexity (or flexibility) because it naturally includes the curvature of the log likelihood with respect to the model parameters rather than only penalising models based on the *number* of parameters.

We will describe the contributions to the message length in parts. We assume the priors on the number of latent factors J and the number of components K to be $p(J) \propto 2^{-J}$ and $p(K) \propto 2^{-K}$ respectively, such that fewer numbers are preferred. The optimal lossless message to encode each is (Sec. 6.8.2; Knorr-Held 2000),

$$I(J) = -\log p(J) = J \log 2 + \text{constant} \tag{25}$$

$$I(K) = -\log p(K) = K \log 2 + \text{constant} \quad . \tag{26}$$

Only $K-1$ of the relative weights $\boldsymbol{\pi}$ need encoding because $\sum_{k=1}^{K} \pi_k = 1$. We assume a uniform prior on individual weights,

$$p(\boldsymbol{\pi}) = (K-1)! \quad , \tag{27}$$

and the Fisher information is

$$\mathcal{F}(\boldsymbol{\pi}) = \frac{N^{K-1}}{\prod_{k=1}^{K} \pi_k} \quad , \tag{28}$$

which gives the message length of the relative weights $I(\boldsymbol{\pi})$ to be

$$\begin{aligned} I(\boldsymbol{\pi}) &= -\log\left(\frac{p(\boldsymbol{\pi})}{\sqrt{|\mathcal{F}(\boldsymbol{\pi})|}}\right) \\ &= -\log p(\boldsymbol{\pi}) - \frac{1}{2}\log|\mathcal{F}(\boldsymbol{\pi})| \\ &= -\log(K-1)! + \frac{K-1}{2}\log N - \frac{1}{2}\sum_{k=1}^{K}\log\pi_k \\ I(\boldsymbol{\pi}) &= \frac{1}{2}\left((K-1)\log N - \sum_{k=1}^{K}\log\pi_k\right) - \log\Gamma(K) \quad . \end{aligned} \tag{29}$$

We assume uniform priors for the component means in latent space $\boldsymbol{\xi}$, where the bounds are large enough outside the range of observable values such that those priors are proper (integrable) – a necessary condition for the MML principle – and only add constant terms to the message length, which can be ignored. We assume a conjugate inverted Wishart prior for the component covariance matrices $\boldsymbol{\Omega}$ (Section 5.2.3; Knorr-Held 2000),

$$p(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k) \propto |\boldsymbol{\Omega}_k|^{\frac{1}{2}(J+1)} \quad . \tag{30}$$

We approximate the determinate of the Fisher information of a multivariate normal $|\mathcal{F}(\boldsymbol{\xi}, \boldsymbol{\Omega})|$ as $|\mathcal{F}(\boldsymbol{\xi})||\mathcal{F}(\boldsymbol{\Omega})|$ (Oliver et al. 1996; Figueiredo & Jain 2002) where

$$|\mathcal{F}(\boldsymbol{\xi})| = (N\pi_k)^J |\boldsymbol{\Omega}_k|^{-1} \tag{31}$$

$$|\mathcal{F}(\boldsymbol{\Omega})| = (N\pi_k)^{\frac{1}{2}J(J+1)} 2^{-J} |\boldsymbol{\Omega}_k|^{-(N+1)} \tag{32}$$

such that

$$\begin{aligned} I(\boldsymbol{\xi}, \boldsymbol{\Omega}) &= -\sum_{k=1}^{K}\log p(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k) + \frac{1}{2}\sum_{k=1}^{K}\log|\mathcal{F}(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k)| \\ &= \frac{1}{2}\sum_{k=1}^{K}\log\left[(N\pi_k)^{\frac{1}{2}J(J+3)} 2^{-N} |\boldsymbol{\Omega}_k|^{-(N+2)}\right] \\ &\quad \cdots - \sum_{k=1}^{K}\log|\boldsymbol{\Omega}_k|^{\frac{1}{2}(J+1)} \end{aligned} \tag{33}$$

$$\begin{aligned} I(\boldsymbol{\xi}, \boldsymbol{\Omega}) &= \frac{1}{4}J(J+3)\sum_{k=1}^{K}\log N\pi_k - \frac{KD}{2}\log 2 \\ &\quad \cdots - \frac{1}{2}(2J+3)\sum_{k=1}^{K}\log|\boldsymbol{\Omega}_k| \quad . \end{aligned} \tag{34}$$

Previous work on multiple latent factor analysis within the context of MML have addressed the indeterminacy between the factor loads and factor scores by placing a joint prior on the *product* of factor loads and scores (Wallace 1995). Adopting the same prior density in our model is not practical because it would require the priors $p(\boldsymbol{\xi}|\boldsymbol{\tau}, \boldsymbol{\pi})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\tau}, \boldsymbol{\pi})$. That is, we would require a prior density on both the means $\boldsymbol{\xi}$ and covariance matrices $\boldsymbol{\Omega}$ in latent space that requires knowledge about the responsibility matrix $\boldsymbol{\tau}$ and relative weights $\boldsymbol{\pi}$ in order to estimate the effective scores $\mathbf{S}$ for each data point and calculate a joint prior on the product of the factor loads $\mathbf{L}$ and factor scores $\mathbf{S}$. Instead we address this indeterminacy by placing a prior on $\mathbf{L}$ that ensures it is mutually orthogonal. Specifically, we adopt a Wishart distribution with scale matrix $\mathbf{W}$ and D degrees of freedom for the $J \times J$ matrix $\mathbf{M} = \mathbf{L}^\intercal\mathbf{L}$. In other words, $\mathbf{M} \sim W_J(D, \mathbf{W})$ and $\mathbf{W} = \mathrm{Cov}(\mathbf{L}^\intercal)$. This Wishart joint prior density gives highest support for mutually orthogonal vectors,

$$p(\mathbf{L}) = \frac{|\mathbf{L}^\intercal\mathbf{L}|^{\frac{1}{2}(D-J-1)}}{2^{\frac{DJ}{2}}|\mathbf{W}|^{\frac{D}{2}}\Gamma(\frac{D}{2})} \exp\left[-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{L}^\intercal\mathbf{L})\right] \quad . \tag{35}$$

Thus the message length to encode $\mathbf{L}$ is given by

$$\begin{aligned} I(\mathbf{L}) &= -\log\left(\frac{p(\mathbf{L})}{\sqrt{|\mathcal{F}(\mathbf{L})|}}\right) \\ I(\mathbf{L}) &= \frac{1}{2}\mathrm{Tr}(\mathrm{Cov}(\mathbf{L}^\intercal)^{-1}\mathbf{L}^\intercal\mathbf{L}) - \frac{1}{2}(D-J-1)\log|\mathbf{L}^\intercal\mathbf{L}| \\ &\quad \cdots + \frac{1}{2}DJ\log 2 + \frac{1}{2}D\log|\mathrm{Cov}(\mathbf{L})| - \Gamma\left(\frac{D}{2}\right) \end{aligned} \tag{36}$$

Combining equations 25, 26, 29, 34, and 36 with equation 22 leads to the full message length:

$$I(\mathbf{\Psi}, \mathbf{Y}) = -\log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi}) + \frac{1}{4}(J+4)(J-1)\sum_{k=1}^{K}\log \pi_k + \left(K - \frac{1}{2}\right)\log N + \frac{1}{2}D\log|\mathrm{Cov}(\mathbf{L}^\intercal)|$$
$$\cdots - \frac{1}{2}(D-J-1)\log|\mathbf{L}^\intercal\mathbf{L}| + \mathrm{Tr}\left(\mathrm{Cov}(\mathbf{L}^\intercal)^{-1}\mathbf{L}^\intercal\mathbf{L}\right) - \left(J+\frac{3}{2}\right)\sum_{k=1}^{K}\log|\mathbf{\Omega}_k| - \log\Gamma(K) - \Gamma\left(\frac{D}{2}\right)$$
$$\cdots + \frac{Q}{2}\log\kappa_q + \frac{1}{2}\left[J(D+2) + K(2-N)\right]\log 2 \quad . \tag{37}$$

3. EXPERIMENTS

3.1. *A toy model*

Here we introduce a toy model where we use generated data to verify that we recover the true model parameters given some data, and to ensure that the expectation-maximization method is yielding consistent results. We generated a data set with $N = 100{,}000$ data points, each with $D = 15$ dimensions. We adopted a latent dimensional space of $J = 5$ factor loads such that the vector $\mathbf{L}$ has shape $J \times D$, with $K = 20$ clusters in the latent space. We generated the random factor loads in the same way that we initialise the optimisation (Section 2.1). The relative weights $\boldsymbol{\pi}$ are drawn from a multinomial distribution and the means of the clusters in factor scores $\boldsymbol{\xi}$ are drawn from a standard normal distribution. The off-diagonal entries in the covariance matrices in factor scores $\mathbf{\Omega}$ are drawn from a gamma distribution $\mathbf{\Omega}_{k,i,i} \sim \mathbf{\Gamma}(1)$. The variance in each dimension $\mathbf{D}$ are also drawn $\mathbf{D} \sim \mathbf{\Gamma}(1)$. The nth data point (which belongs to the kth cluster) is then generated by drawing $\mathbf{S}_n \sim \mathcal{N}(\boldsymbol{\xi}_k, \mathbf{\Omega}_k)$, projecting by the factor loads $\mathbf{L}$, and adding variance $\mathbf{D}$.

We treat the generated data set as if the number of latent factors and components are not known. Starting with $J = 1$ and $K = 1$, we trialled each permutation of J and K until $J_{\max} = 10$ and $K_{\max} = 40$ (e.g., twice the true values of J_{true} and K_{true}).

We recorded the *negative* log likelihood, the BIC, and the message length[4] for each permutation of J and K. These metrics are shown in Figure 1. Unsurprisingly the negative log likelihood increases with increasing numbers of latent factors J and increasing numbers of components K. The lowest BIC value and message length is found at $J = 5$ and $K = 20$, identical to the true values. It is clear from Figure 1 that a combination of latent factors and clustering in the latent space provides a better description of the (generated) data than a gaussian mixture model without latent factors. Adding components to the model does improves the log likelihood, even with a single latent factor, but the addition of just *one latent factor* improves the log likelihood more so than adding *twenty components*. Not much more can be said for this example because the true data generating process is known, but this toy model does illustrate how clustering in high dimensional data can be better described by latent factors with clustering in the lower dimensional latent space.

Some technical background is warranted before we compare our estimated model parameters to the true values. We previously stated that the latent factors in this model are only identifiable up to an orthogonal rotation. That is to say that if the data were truely generated by latent factors $\mathbf{L}_{\mathrm{true}}$, then our estimates of those latent factors $\mathbf{L}_{\mathrm{est}}$ do not need to be identical to the true values. For example, the ordering of the estimated factors could be different from the true factors, and the ordering of the dimensionality in latent space would then be accordingly different. Since no constraint is placed on the ordering of the factor loads during expectation-maximization, there is no assurance (or requirement) that our factor loads match the true factor loads.

Another possibility is that the estimated factor loads could be flipped in sign relative to the true factor loads, and the scores would similarly be flipped. In both of these situations (reordering or flipped signs) the log likelihood given the data and the estimated factor loads $\mathbf{L}_{\mathrm{est}}$ would be identical to the log likelihood given the data and the true factor loads $\mathbf{L}_{\mathrm{true}}$ despite the difference in ordering and sign. The same can be said for any other scalar metric (e.g., Kullback-Leibler divergence; Kullback & Leibler 1951). These examples serve to illustrate a more general property that the factor loads and factor scores can be orthogonally rotated by *any valid rotation matrix*[5] $\mathbf{R}$. The estimated factor loads $\mathbf{L}_{\mathrm{est}}$ could therefore appear very different from the true values, but they only differ by an orthogonal rotation. We discuss the impact of this limitation on real data in more detail in Section 4.

[4] Omitting constant terms such that negative message lengths are allowed.

[5] Recall that a rotation matrix is valid if $\mathbf{R}\mathbf{R}^\intercal = \mathbf{I}$.

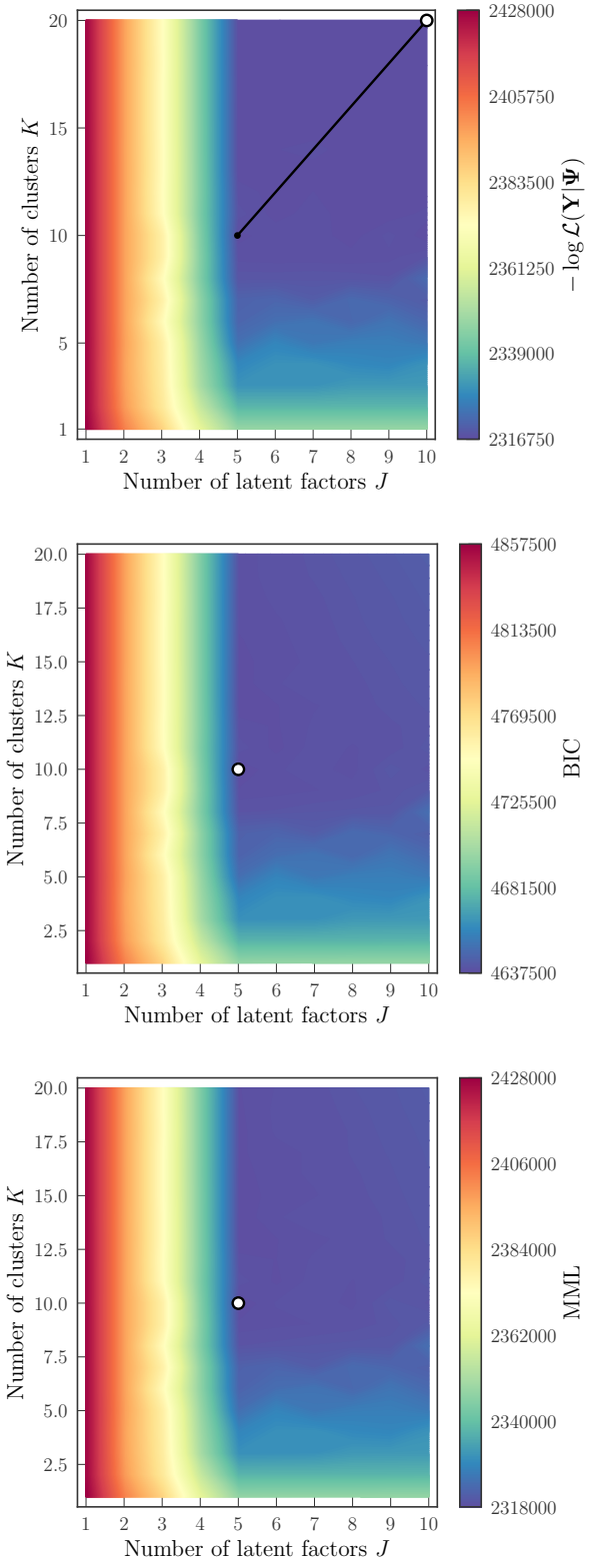

Figure 1. The top panel shows the negative log likelihood $-\log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi})$ evaluated at each combination of latent factors J and number of clusters K using the generated data in our toy model. The middle panel shows the BIC (Eq. 20) for those combinations, and the lower panel shows the message length. The white marker indicates the lowest value in each panel, showing the preferred number of latent factors and components. The black marker indicates the true value.

We took the model with the preferred number of latent factors and components found from a grid search ($K = 20$, $J = 5$; which are also the true values) and applied an orthogonal rotation to the latent space to be as close as possible to the true values. The rotation matrix $\mathbf{R}$ was found by solving for J unknown angle parameters, each of which is used to construct a Givens rotation matrix (Givens 1958), and then we take the product of those Givens matrices to produce a valid rotation matrix $\mathbf{R}$. This process reduces to Euler angle rotation in three or fewer dimensions. This process rotates the latent space ($\mathbf{L}$, $\boldsymbol{\xi}$, $\boldsymbol{\Omega}$), but has no effect on the model's predictive power: the evaluated log likelihood or the Kullback-Leibler divergence (Kullback & Leibler 1951) under the rotated model is indistinguishable from the unrotated model. In Figure 2 we show the estimated factor loads $\mathbf{L}$, factor scores $\mathbf{S}$, and specific variances $\mathbf{D}$ compared to the true values. The agreement is excellent in all model parameters.

3.2. *A toy model with data missing at random*

Here we repeat the toy model used in the previous experiment, but we discard an increasing fraction of the data and evaluate the performance and accuracy of our method in the presence of incomplete data. We considered missing data fractions from 1% to 40%. In each case we treated the model parameters as unknown, assumed the missing data points were missing at random, and initialised the model as per Section 3.1.

In Figure 3 we show the results of this experiment for our worst considered case, where 40% of the data entries are randomly discarded. We find that despite the high fraction of missing entries, our estimates of the model parameters remain unbiased in this example using a toy model. The corrections to our estimates of the specific variances are sufficient, in that the specific variance in each dimension is not systematically underestimated from the true values, despite that 40% of the data entries are missing.

3.3. *The Galah survey*

In this experiment we perform blind chemical tagging using the photospheric abundances released as part of the second *Galah* data release (Buder et al. 2018). This data set includes up to 23 chemical abundances reported for 342,682 stars. In this example we chose to restrict ourselves to stars with a complete set of abundance measurements for a subset of those 23 elements (i.e., no missing data entries). For example, here we will exclude lithium and carbon abundances because the photospheric values will vary throughout a star's life-

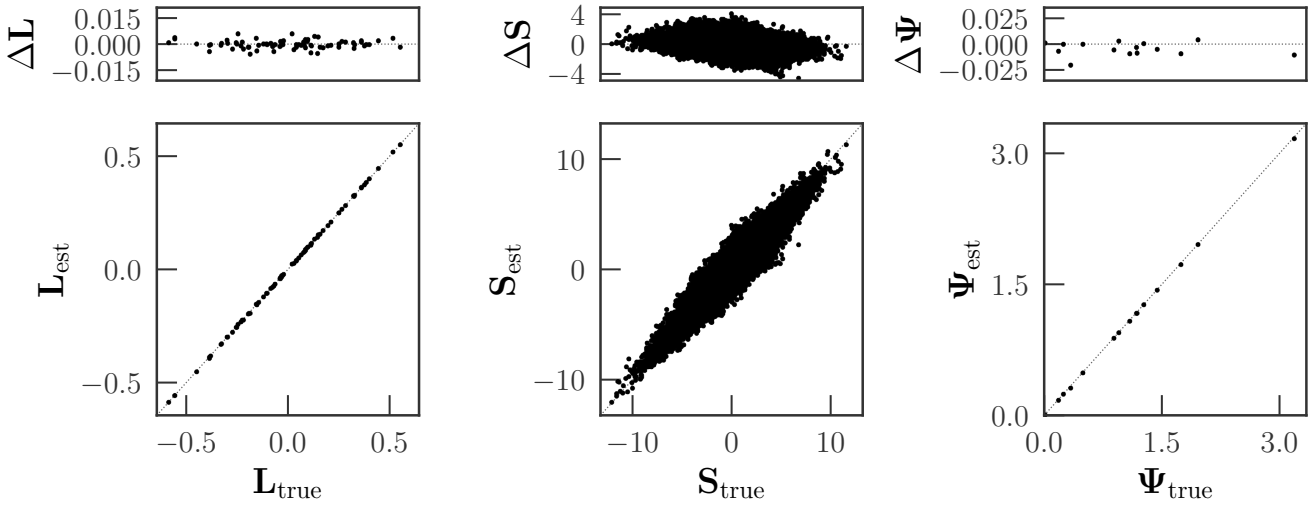

Figure 2. The estimated factor loads **L** (left), factor scores **S** (middle), and specific variances **D** (right) compared to the true data generating values for Experiment 1 (Section 3.1). The agreement is excellent.

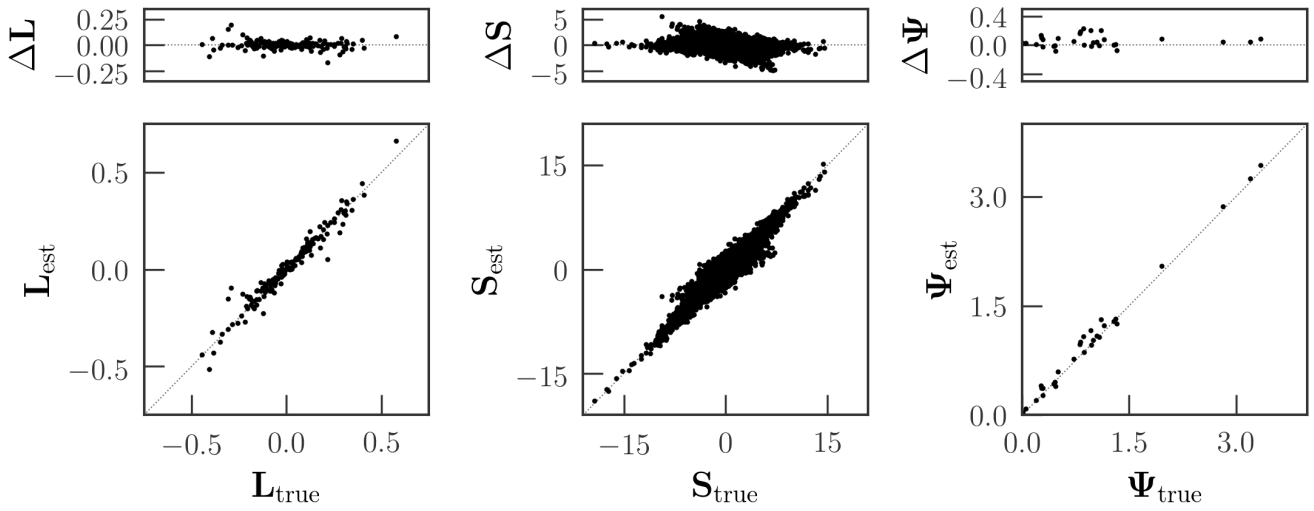

Figure 3. The estimated factor loads **L** (left), factor scores **S** (middle), and specific variances **D** (right) compared to the true data generating values for Experiment 2 (Section 3.2. Here 40% of the data are missing at random. The agreement remains excellent, despite the large fraction of missing data entries.

time (e.g., Casey et al. 2016b, 2019). This is true to a small degree for many elements (e.g., Dotter et al. 2017), but for the purposes of this experiment we assume that all other photospheric abundances remain constant throughout a star's lifetime.

We first selected stars with `flag_cannon = 0` to exclude stars where there is reason to suspect that the stellar parameters (e.g., $T_{\rm eff}$, $\log g$) are unreliable, and as a result the detailed chemical abundances would be untrustworthy. We then took all stars with a signal-to-noise ratio exceeding 40 in the blue arm (`snr_c1 > 40`), and stars with no erroneous flags in all of the following abundances: Na, Al, Si, K, Ca, Sc, Ti, V, Mn, Fe, Ni, Cu, Zn, Y, Ba, La, and Eu. These elements were chosen because they trace multiple nucleosynthetic pathways, and they are more commonly reported in the *Galah* data release, allowing for a larger number of stars with a complete abundance inventory. There are 1,072 stars that met these criteria.

We executed a grid search for the number of latent factors J and the number of components K that were preferred by the data. Starting with $J = 1$ and $K = 1$,

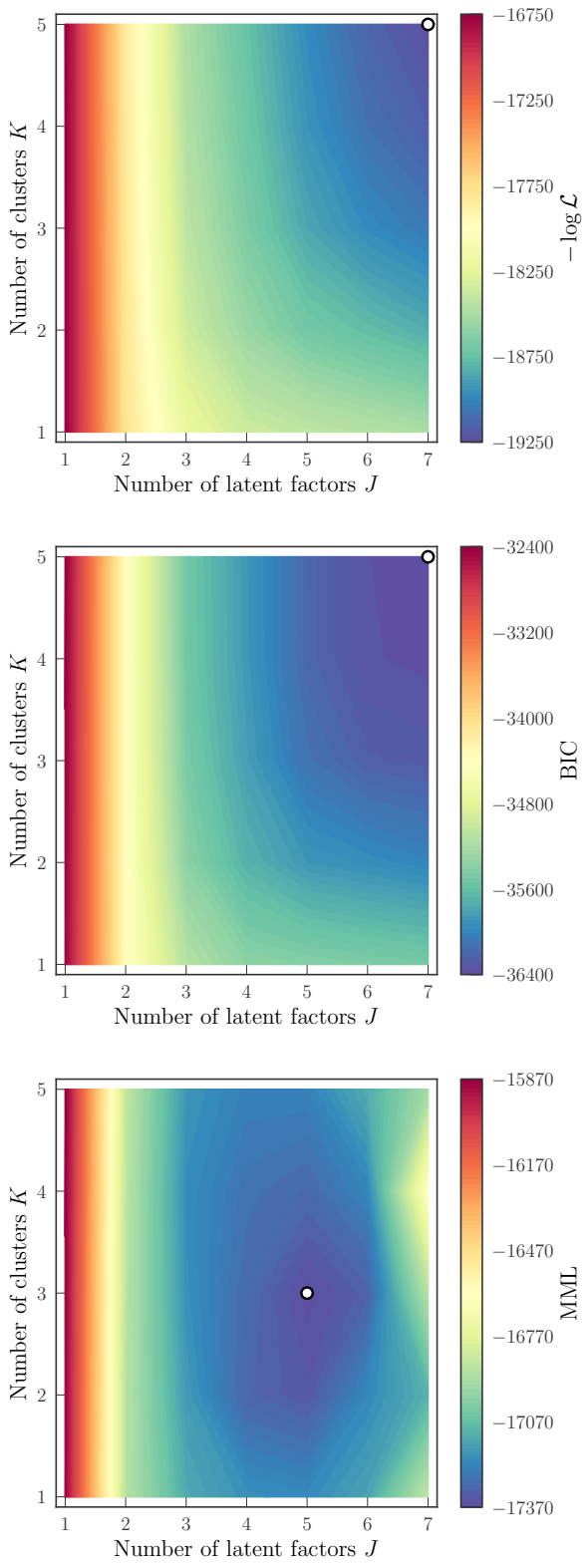

Figure 4. The top panel shows the negative log likelihood $-\log \mathcal{L}(\mathbf{Y}|\mathbf{\Psi})$ evaluated at each combination of latent factors J and number of clusters K using *Galah* data in Experiment 3. The middle panel shows the BIC for those combinations, and the lower panel shows the message length. The white marker indicates the lowest value in each panel, showing the preferred number of latent factors and components.

we trialled each permutation of J and K up until $J = 7$ and $K = 5$. The results of this grid search are shown in Figure 4, where we show the negative log likelihood, the BIC, and message length found for each permutation. The behaviour of the BIC and the message length are very different here, unlike what was observed in our toy model. Here the BIC behaviour appears similar to the negative log likelihood in that the BIC prefers higher components and latent factors than the extent of the grid (e.g., $J > 7$ and $K > 5$). The model with five latent factors and three components ($J = 5$, $K = 3$) is found to have the shortest message length, which we take as our preferred model for these data.

Earlier we described how the latent factors we estimate can only be identified up until an orthogonal rotation. If we want to interpret the latent factors estimated from *Galah* data, then we must specify some target factor loads such that we can identify which factors are most similar to the yields we expect. We specified the following target latent factors where:

- The first factor load should have non-zero entries in Eu and La (e.g., the r-process).
- The second factor load should have non-zero entries in Ba, Y, and La (e.g., the s-process).
- The third factor load should have non-zero entries in Fe-peak elements Sc, V, Mn, Fe, Ni, Cu, and Zn.
- The fourth factor load should have non-zero entries in the α-element tracers Si, Ca, and Ti.
- The fifth factor load should have non-zero entries in the light odd-Z elements Na, Al, and K.

We initially set each non-zero entry in these target factor loads $\mathbf{L}_{\text{target}}$ to $E^{-\frac{1}{2}}$, where E is the number of non-zero entries in that factor load, to ensure that $\mathbf{L}_{\text{target}}$ is mutually orthogonal. We solved for the J unknown angles to produce a valid rotation matrix $\mathbf{R}$ that would make our estimated loads $\mathbf{L}$ as close as possible to the target loads $\mathbf{L}_{\text{target}}$, and then applied that rotation to the model. The target loads and (rotated) estimated loads are shown in Figure 5. Note that the purpose of this procedure is not to 'find' the target loads that we expect, but to provide as little information needed in order to identify and describe all factor loads within an astrophysical context. This procedure still requires that the factors be mutually orthogonal and that they describe the data. For these reasons, we will not always recover the exact target loads we seek: we will only be able to identify factor loads that are closest to the target loads.

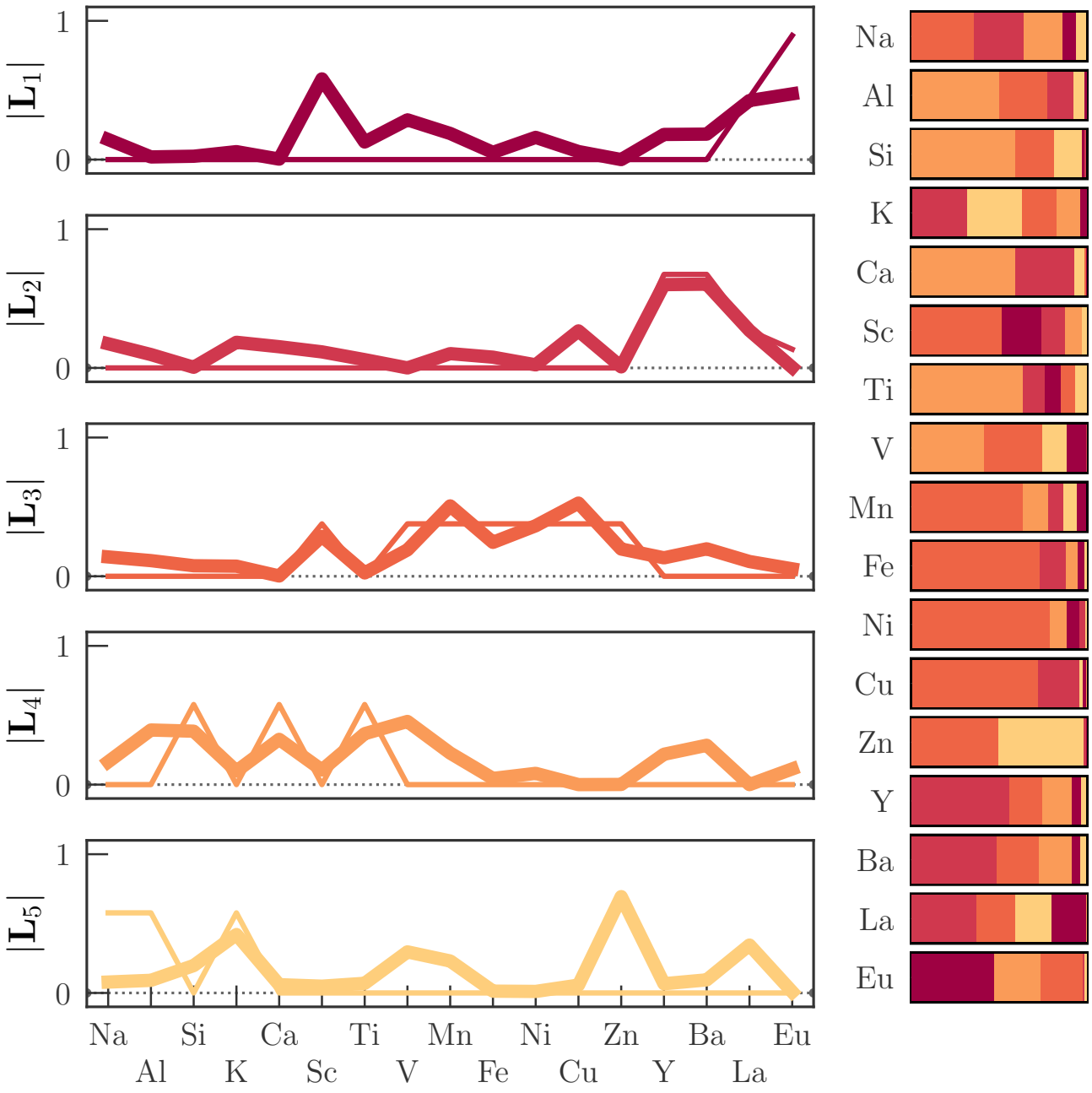

Figure 5. Latent factors inferred from 1,072 stars in *Galah* (Buder et al. 2018, thick lines) with 17 abundance measurements. Left panels show the absolute entries for each factor load, where the thin lines indicate the target latent factors (see Section 3.3). On the right we show the absolute fractional contributions to each element, ordered by the loads that contribute most.

This is demonstrated in Figure 5, where some estimated factor loads match closely to the target load (e.g., $\mathbf{L}_2$ which we identify as the s-process), and some barely match at all (e.g., $\mathbf{L}_5$). Here we show the absolute entry of the factor loads because even if an entry is negative, the corresponding factor scores could also be negative, and their product will contribute to the observed abundances. For this reason the sign does not matter here.

Some of these factor load entries may be non-zero because we require the latent factors to be mutually orthogonal, and not because they truly contribute to the data. To try and disentangle these possibilities, we calculate the fractional contribution that factor load makes to the observed abundances relative to other factor loads. We define the fractional contribution of the

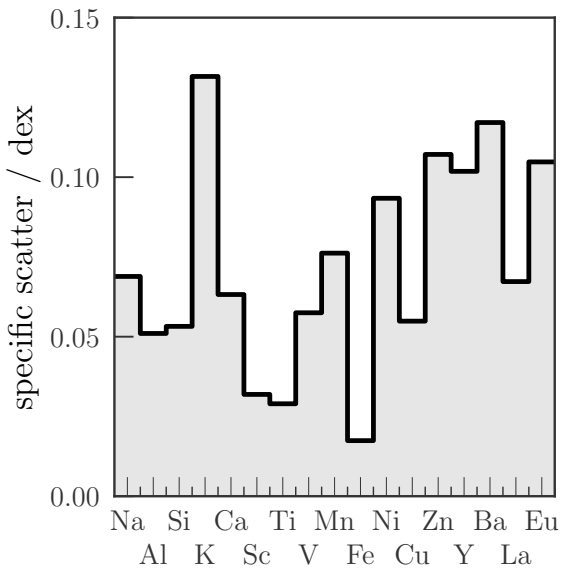

Figure 6. Specific scatter (e.g., $\sqrt{\mathbf{D}}$) remaining in the *Galah* data (Buder et al. 2018) after accounting for the contributions by all latent factors.

jth factor load to the dth data dimension as:

$$\mathbf{C}_{\mathrm{d,j}} = \frac{\sum^{N} |\mathbf{L}_{j,d}\mathbf{S}_{n,j}|}{\sum^{J}\sum^{N} |\mathbf{L}_{j,d}\mathbf{S}_{n,j}|} \quad . \tag{38}$$

The fractional contributions to each element are shown in the right hand side of Figure 5. We identify the first factor $\mathbf{L}_1$ as being most similar to the r-process, and here it is the dominant contributor to Eu, a typical r-process tracer. Surprisingly we also find that this factor load is a reasonable contributor to the Fe-peak element Sc. The specific scatter in Sc is 0.03 dex (Figure 6), suggesting that the Sc abundances are well-described by this latent factor model.

The second latent factor $\mathbf{L}_2$ here is most representative of the slow neutron capture process (s-process), with dominant contributions to Ba, Y, and La. This factor has some support at other elements, notably K. Most Sc is contributed by $\mathbf{L}_3$, our Fe-peak latent factor. In fact $\mathbf{L}_3$ is the primary contributor to nearly all Fe-peak elements, with close to negligible contributions from other factors. The exception here is Zn, where a near-equal contribution comes from $\mathbf{L}_5$. The third latent factor $\mathbf{L}_4$ is reasonably well behaved. It is the dominant contributor to the α-element tracers Si, Ca, and Ti, and surprisingly, Al. It also seems to contribute a non-zero fraction to Eu, with negligible contributions elsewhere. The specific scatter after accounting for these latent factors is smallest for Fe (0.01 dex) and largest for K (0.13 dex; Figure 6). The typical scatter in most elements is about 0.05 dex.

In Figure 7 we show the inferred clustering in latent space, where the separation between components is arguably best seen in the splitting between $\mathbf{S}_5$ with respect to $\mathbf{S}_2$ or $\mathbf{S}_3$. When projected to data space (Figure 8) the third component (dark green) is seen to have relatively higher abundance ratios of [K,Ba,Zn/Fe] at a given [Fe/H], and lower abundance ratios of [V/Fe]. This is consistent with the clustering in latent space.

3.4. *Galah survey data with an increasing number of stars with missing data entries*

Here we extend our experiment in Section 3.3 to progressively include more stars, even though those stars have some abundance measurements missing. Specifically we started with the same subset of 1,072 stars in Section 3.3 and added a random set of stars that met our criteria of `flag_cannon = 0` and `snr_c1 > 40`.

We initially added 100 stars to give a sample of $N =$ 1,172, then repeated the grid search for the number of latent factors and components, and recorded the model with the lowest message length. We then repeated this procedure using 1,000 stars (N = 2,072), again with 10,000 (N = 11,072), and finally using all 99,174 stars that met the criteria of `flag_cannon = 0` and `snr_c1 > 40` to give a total sample size of N = 100,246 stars.

For sample sizes up to N ∼2,000 we found that five latent factors were preferred, and these factors shared common features (Figure **??**). This illustrates that the first[6] set of inferred factor loads inferred from a smaller, complete data set, remain largely unchanged despite the increasing sample size and the increasing number of missing data entries. When the sample size reaches N =11,072 we find another three latent factors are required to best explain the data.

When $N \sim$ 100,000, the preferred number of latent factors rises to twelve. Of note among these factors is $\mathbf{L}_9$, where no 'target load' was prescribed[7], and the non-zero entries mimic what might be expected from a light s-process production. Similarly, $\mathbf{L}_{10}$ has near zero contributions everywhere except among the light elements Al, K, Ca, and Sc. With this data set we find that $K = 16$ components are preferred in latent space.

[6] 'First' has no concept here in terms of factor load ordering, but for the purposes of comparing inferred loads from different data sets we have ordered the loads to be as close to those inferred in Section 3.3.

[7] Although no 'target load' was prescribed here, this statement should be interpreted with caution because the mutual orthogonality constraint exists and the remaining factors do have target loads.

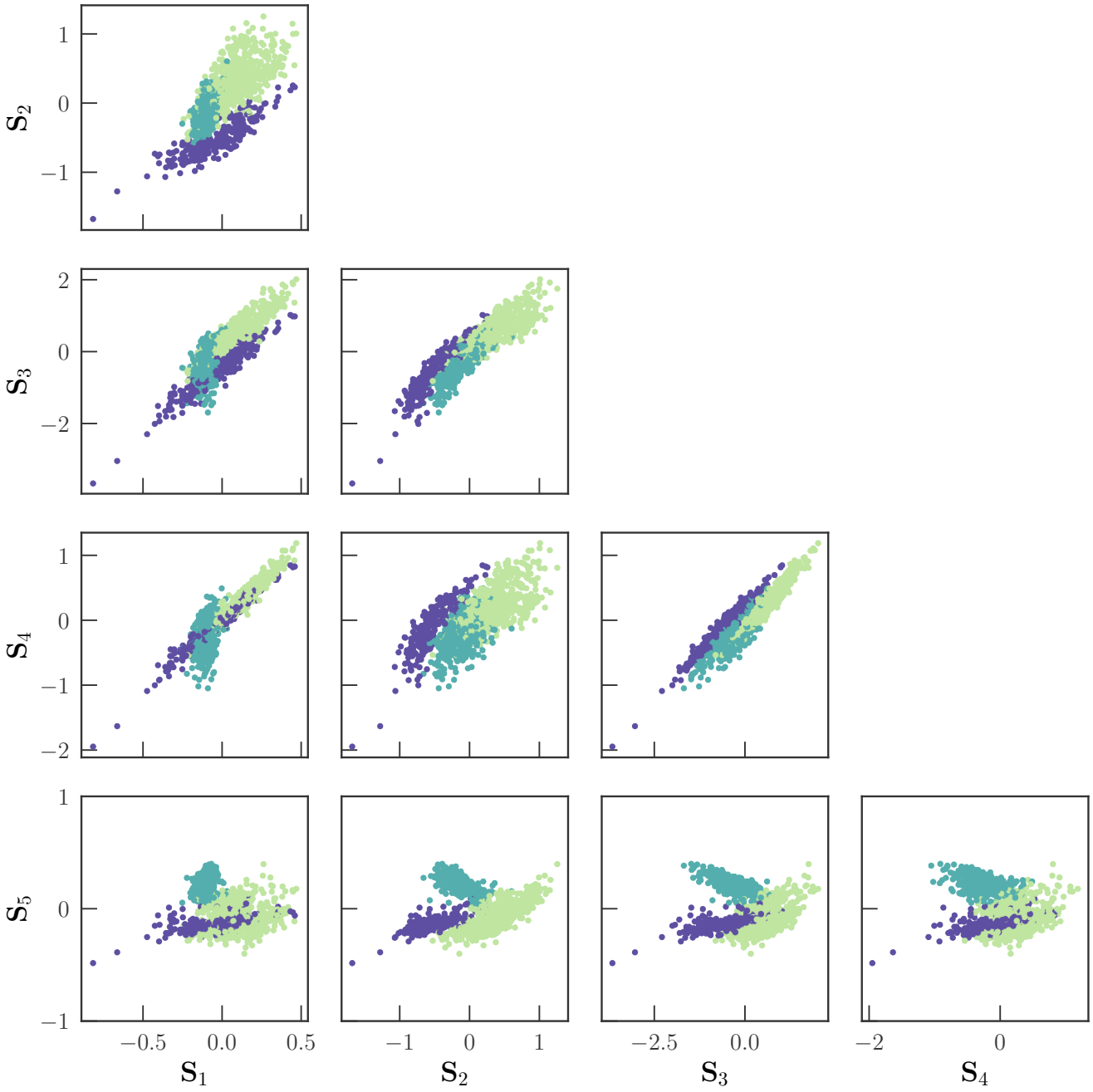

Figure 7. The factor scores **S** estimated in Experiment 3 (Section 3.3 using $N = 1{,}072$ stars in the *Galah* data (Buder et al. 2018) that have 18 abundance measurements. Here each star is coloured by its inferred component.

None of these components appear coherently structured in their positions or motions.

4. DISCUSSION

We have introduced a model to simultaneously account for the lower effective dimensionality of chemical abundance space, and perform clustering in that lower dimensional space. This provides a data-driven model of nucleosynthesis yields and chemical tagging that allows us to simultaneously estimate the latent factors that contribute to all stars, and cluster those stars by their relative contributions from each factor. The results are encouraging in that we find latent factors that are representative of the expected yields from dominant nucleosynthetic channels. However, the model that we describe is very likely *not* the correct model to use to

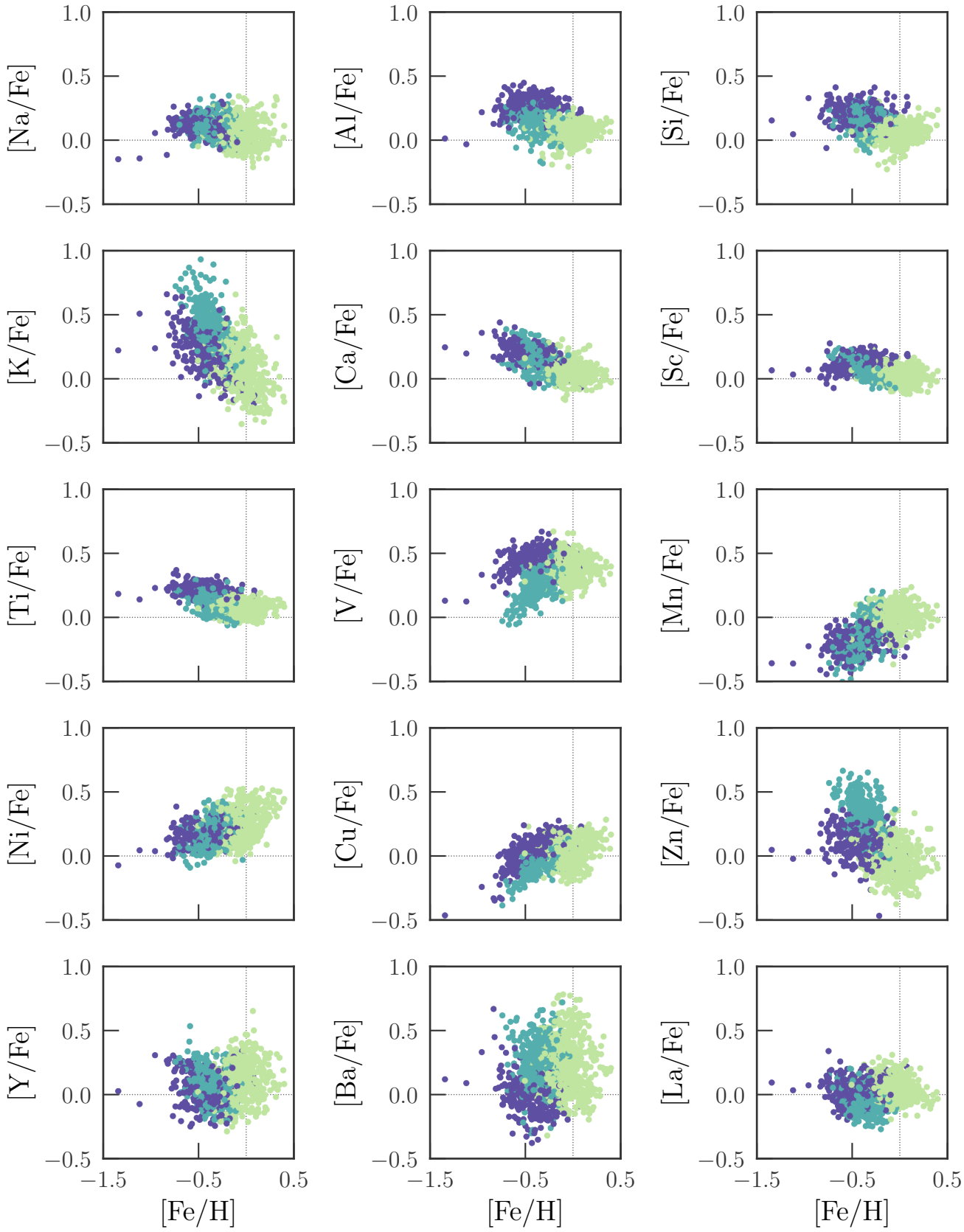

Figure 8. Detailed chemical abundances from the N = 1,072 stars in *Galah* (Buder et al. 2018) that have 18 chemical abundances (Section 3.3). Each star is coloured by its inferred component from the lower-dimensional latent space.

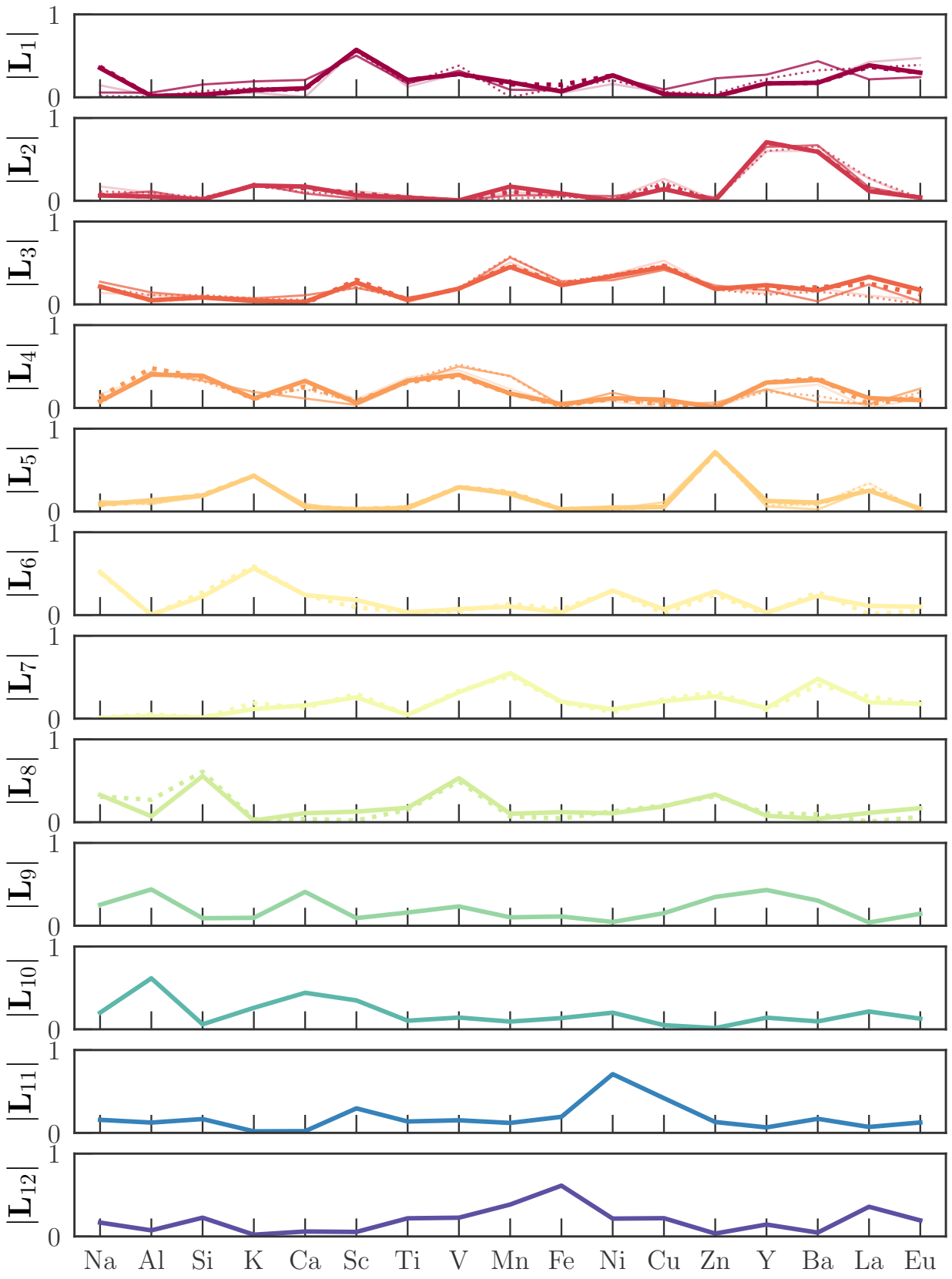

Figure 9. Latent factors found in Section 3.4 using different subsets of *Galah* data (Buder et al. 2018). The thin solid line shows the result from Section 3.3 with $N = 1,072$ stars with 18 abundances and no missing data. Increasing thickness indicates larger samples, up to the solid thick line with $N \sim 100{,}000$ stars, where twelve latent factors are preferred.

represent chemical abundances of stars. Here we discuss the limitations of our model in detail.

We require latent factors to be mutually orthogonal in order to resolve an indeterminacy. This suggests an astrophysical context where the mean nucleosynthetic yields (integrated over all stellar masses and star formation histories) of various nucleosynthetic processes (e.g., *r*-process, *s*-process) are mutually orthogonal to each other. Clearly this assumption is likely to be incorrect: the nuclear physics of one environment where elements are produced will be very different from others, and there is no astrophysical constraint that those yields (or latent factors) should be mutually orthogonal. In principle one could represent the latent factors using a hierarchical data-driven model where the yields contribute as a function of stellar mass, metallicity, and other factors, but in principle to resolve the indeterminacy *in this model* would still require mutual orthogonality on the mean yields. Introducing a constraint on the factor scores that resolves this indeterminacy and allows for more flexible latent factors would be a worthy extension to this work.

The constraint of mutual orthogonality limits the inferences we want to make about stellar nucleosynthetic yields. For example, after accounting for all known sources of potassium production in the Milky Way, galactic chemical evolution models under-predict the level of K in the Milky Way by more than an order of magnitude (Kobayashi et al. 2006). From our inferences using *Galah* data, we find that $\mathbf{L}_1$ – the factor we identify as the s-process – is the dominant contributor to potassium. This latent factor persists even in the presence of missing data, and a sample size two orders of magnitude larger. Does this suggest production of K is linked to the production of much heavier nuclei? If our model could confidently and reliably associate the production of K with other elements or sites then it could help explain the peculiar abundances of stars enhanced in K and depleted in Mg (Mucciarelli et al. 2012; Cohen & Kirby 2012) – a chemical abundance pattern that currently lacks explanation (Iliadis et al. 2016; Kemp et al. 2018). In the (Cohen & Kirby 2012) sample their high [K/Fe] stars also tend to be high in heavier elements, but there are also numerous abundance correlations present. However, is the K contribution that we infer physically realistic, or is it a consequence of requiring that the latent factors are mutually orthogonal? Distinguishing these possibilities is non-trivial, which is in part why caution is warranted when trying to interpret latent factor models. In this situation it is worth commenting that K has the largest specific scatter (Figure 6), suggesting that the contributions of K are perhaps not as well described by the latent factor model as other elements. Still, the specific model scatter is far less than the observed scatter, indicative that the model does have some predictive power.

A similar argument could be made for Sc, where $\mathbf{L}_1$ – a factor load we identify as the r-process – is the secondary contributor. Sc is under-produced in galactic chemical evolution models relative to observations (Kobayashi et al. 2006; Casey & Schlaufman 2015). Based on this work, is the production of Sc linked to the production of heavy nuclei? Unlike K, the specific scatter in Sc is remarkably low: just 0.03 dex, among the best-described elements after Ti and Fe (0.01 dex). This would suggest that the latent factor model is a very good description for the production of Sc, but it does not prove that it is *the* description for the production of Sc.

There are other issues in our model that relate to our assumption of mutual orthogonality. Even if nucleosynthetic yields were truely mutually orthogonal, then the latent factors we infer are only *identifiable* up until an orthogonal basis. As we have seen in our experiments, the ordering and sign of the latent factors is not described *a priori*. This is both a feature and a bug: unrestricted ordering and signs allow for the model parameters to be estimated more efficiently because they can freely rotate as the model parameters are updated, but it does mean that we must 'assign' the latent factors we infer as being described by an astrophysical process (e.g., the first latent factor is r-process). A more general limitation of this is that the latent factors can be multiplied by some arbitrary rotation matrix, leading to latent factor loads that are very different from what was estimated by the model, but still lead to the exact same data (or log likelihood, or Kullback-Liebler divergence, etc). As a consequence, we can only 'identify' latent factors up until this rotation. We have sought to address this by constructing rotation matrices where the entries for each latent factor correspond to our expectations from astrophysical processes (whilst remaining orthogonal), but here we are limited by what astrophysical processes we are *expecting* to find within the constraint of being mutually orthogonal.

This in part constrains our ability to identify new nucleosynthetic processes. For example, let us consider a hypothetical situation where we would only expect there to be four nucleosynthetic processes that predominately contribute to the observed *Galah* abundances, but in practice we found that the data are best explained with five latent factors. We construct a rotation matrix where the first four latent factors describe the nucleosynthetic processes we expect to find. What of the fifth latent factor? We can constrain the possible values

of the fifth latent factor conditioned on the requirement that all factors remain mutually orthogonal, but one can imagine that some (or perhaps many) elements have entries where the fifth latent factor can have near-zero or zero entries. Even if the mean nucleosynthetic yields are mutually orthogonal, there are scenarios that one can imagine where there is a limited amount we can say with confidence about that new nucleosynthetic process.

Notwithstanding these issues, we have shown that a latent factor model which allows for clustering in latent space can adequately describe chemical abundance data. We find five latent factors from a small subset of *Galah* data with complete abundances, and those latent factors can qualitatively be described within the context of astrophysical yields. Those latent factors are recovered in larger samples where the data are incomplete. That did not have to be the case: the mutually orthogonal latent factors could be entirely different from our expectations such that they did not have to match our expectations of nucleosynthetic yields. Indeed, the inferred factors – even after a valid rotation – could have made no astrophysical sense whatsoever. For this reason it is encouraging that there is some interpretability in the latent factors. Indeed, in the elements where we find surprising associations (e.g., Sc and K), these are elements where galactic chemical evolution models are most discrepant from observations, even after accounting for systematic errors in abundance measurements (e.g., violations to the assumption of local thermodynamic equilibrium).

In the subset of *Galah* data with complete abundances we find that three components are preferred. These components can be described as those with (1) low- and (2) high-[α/Fe] abundance ratios, and another (3) primarily differing in K, Ba, Zn, and V abundances at a given [Fe/H] and [α/Fe] abundance ratio. When we include $\sim$100,000 stars with up to 18 abundances, and assume the incomplete abundances are missing at random, we find that 16 components in latent space are preferred to explain the data. By construction these components are structured in their chemical abundances because of the projection from the latent space, and by extension of each component having similar chemistry, each component occupies realistic locations in a Hertzsprung-Russell diagram. When we project these component associations to the data space we find that none of the inferred components are structured or coherent in their positions or motions. However, in this sample of stars there are no gravitationally bound clusters where a reasonable (e.g. $\sim$30) number of stars have been observed. Clearly, more data would help to resolve a higher number of components.

Perhaps it is not so discouraging that none of the inferred components are structured in their positions or motions because there are no gravitationally bound clusters in the data. But there is clearly more that can be done in chemical tagging. Some components we infer have stars with positions and galactic orbits that would imply that they cannot have formed in the same star cluster. In these situations there is likely significant value in including joint probabilities on whether two stars could be associated to the same star formation site based on their dynamic properties. Similarly, although stellar ages are historically difficult to estimate precisely, can this imprecise information help inform weak priors or probabilities of two stars having the same association? There is an incredible amount of dynamical information available from *Gaia*, particularly for stars in the *Galah* survey, and weakly informative priors might be sufficient to help improve the granularity of chemical tagging without being overly constraining on the dynamical and star formation history we seek to infer.

5. CONCLUSIONS

We have introduced a data-driven model of nucleosynthesis by incorporating latent factors that are common to all stars, and allowing for clustering in the lower-dimensional latent space. This approach simultaneously allows us to efficiently tag stars based on their chemical abundances, and to infer the contributions that are common to all stars (e.g., nucleosynthetic yields). Experiments with generated data demonstrate that MML is a useful principle for selecting the appropriate number of latent factors and components. Experiments with *Galah* data reveal latent factors that are qualitatively and quantitatively similar to expected nucleosynthetic yields (e.g., products from the s-process, r-process, et cetera). Interestingly we find that deviations from expected yields occur in elements where observations and galactic chemical evolution models are most discrepant (e.g., K, Sc). While we advise caution in directly interpreting those latent factors as being nucleosynthetic yields, our our model does provide the first data-driven approach to nucleosynthesis and chemical tagging. We advocate that more data, and the inclusion of weakly informative priors – joint probabilities using astrometry and a simplified model of the Milky Way – would help in realising the full potential of chemical tagging.

We acknowledge support from the Australian Research Council through Discovery Project DP160100637. The *Galah* survey is based on observations made at the Australian Astronomical Observatory, under programmes A/2013B/13, A/2014A/25, A/2015A/19,

A/2017A/18. We acknowledge the traditional owners of the land on which the AAT stands, the Gamilaraay people, and pay our respects to elders past and present.

This research has made use of NASA's Astrophysics Data System.

Software: `Astropy` (Astropy Collaboration et al. 2013, 2018), `IPython` (Pérez & Granger 2007), `matplotlib` (Hunter 2007), `numpy` (Walt et al. 2011), `scipy` (Jones et al. 2001–), `Jupyter Notebooks` (Kluyver et al. 2016)

REFERENCES

Arthur, D., & Vassilvitskii, S. 2007, in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. (Society for Industrial and Applied Mathematics Philadelphia, PA, USA), 1027–1035

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Astropy Collaboration, Price-Whelan, A. M., Sipócz, B. M., et al. 2018, AJ, 156, 123

Baek, J., McLachlan, G. J., & Flack, L. K. 2010, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 1298

Bovy, J. 2016, ApJ, 817, 49

Buder, S., Asplund, M., Duong, L., et al. 2018, The Monthly Notices of the Royal Astronomical Society, 478, 4513

Casey, A. R., Hogg, D. W., Ness, M., et al. 2016a, arXiv e-prints, arXiv:1603.03040

Casey, A. R., & Schlaufman, K. C. 2015, ApJ, 809, 110

Casey, A. R., Ruchti, G., Masseron, T., et al. 2016b, MNRAS, 461, 3336

Casey, A. R., Hawkins, K., Hogg, D. W., et al. 2017, ApJ, 840, 59

Casey, A. R., Ho, A. Y. Q., Ness, M., et al. 2019, arXiv e-prints, arXiv:1902.04102

Cohen, J. G., & Kirby, E. N. 2012, ApJ, 760, 86

Cramér, H. 1946, Mathematical Methods of Statistics (Princeton University Press)

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39, 1

Dotter, A., Conroy, C., Cargile, P., & Asplund, M. 2017, ApJ, 840, 99

Dowe, D. L. 2008, Computer Journal, 51, 523 , Christopher Stewart WALLACE (1933-2004) memorial special issue

Dowe, D. L. 2011, in Handbook of the Philosophy of Science - Volume 7: Philosophy of Statistics, ed. P. S. Bandyopadhyay and M. R. Forster (Elsevier), 901–982

Dowe, D. L., Gardner, S., & Oppy, G. 2007, The British Journal for the Philosophy of Science, 58, 709

Dowe, D. L., & Wallace, C. S. 1997, in Proc. Computing Science and Statistics - 28th Symposium on the interface, Vol. 28, 614–618

Figueiredo, M. A. T., & Jain, A. K. 2002, IEEE Transactions on pattern analysis and machine intelligence, 24, 381

Fitzgibbon, L. J., Dowe, D. L., & Vahid, F. 2004, in Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on, IEEE, 439–444

Freeman, K., & Bland-Hawthorn, J. 2002, Annual Review of Astronomy and Astrophysics, 40, 487

Givens, W. 1958, in J. SIAM, 26–50

Golub, G. H., & Reinsch, C. 1970, Numerische mathematik, 14, 403

Haar, A. 1933, The Annals of Mathematics, 34, 147

Ho, A. Y. Q., Rix, H.-W., Ness, M. K., et al. 2017a, ApJ, 841, 40

Ho, A. Y. Q., Ness, M. K., Hogg, D. W., et al. 2017b, ApJ, 836, 5

Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, ApJ, 833, 262

Hotelling, H. 1933, Journal of Educational Psychology, 24, 417

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90

Iliadis, C., Karakas, A. I., Prantzos, N., Lattanzio, J. C., & Doherty, C. L. 2016, ApJ, 818, 98

Jones, E., Oliphant, T., Peterson, P., et al. 2001–, SciPy: Open source scientific tools for Python, ,

Kemp, A. J., Casey, A. R., Miles, M. T., et al. 2018, MNRAS, 480, 1384

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Schmidt, IOS Press, 87 – 90

Knorr-Held, L. 2000, Statistics in Medicine, 19, 1006

Kobayashi, C., Umeda, H., Nomoto, K., Tominaga, N., & Ohkubo, T. 2006, ApJ, 653, 1145

Kullback, S., & Leibler, R. A. 1951, The Annals of Mathematical Statistics, 22, 79

Leung, H. W., & Bovy, J. 2018, MNRAS, 3062
Martell, S. L., Sharma, S., Buder, S., et al. 2017, MNRAS, 465, 3203
Mitschang, A. W., De Silva, G., Zucker, D. B., et al. 2014, MNRAS, 438, 2753
Mucciarelli, A., Bellazzini, M., Ibata, R., et al. 2012, MNRAS, 426, 2889
Ness, M. 2018, Publications of the Astronomical Society of Australia, 35, e003
Ness, M., Hogg, D. W., Rix, H. W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
Ness, M., Rix, H. W., Hogg, D. W., et al. 2018a, ApJ, 853, 198
—. 2018b, ApJ, 853, 198
Oliver, J. J., Baxter, R. A., & Wallace, C. S. 1996, in ICML, 364–372
Pérez, F., & Granger, B. E. 2007, Computing in Science and Engineering, 9, 21
Portegies Zwart, S. F., Hut, P., Makino, J., & McMillan, S. L. W. 1998, A&A, 337, 363
Price-Jones, N., & Bovy, J. 2018, MNRAS, 475, 1410
Rao, C. R. 1945, in Bulletin of the Calcutta Mathematical Society., 81–89
Schwarz, G. 1978, The Annals of Statistics, 6, 461
Shannon, C. E. 1948, Bell System Technical Journal, 27, 379
Sheinis, A., Anguiano, B., Asplund, M., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 035002
Stewart, G. W. 1980, SIAM Journal on Numerical Analysis, 17, 403
Thompson, B. 2004, Exploratory and confirmatory factor analysis: Understanding concepts and applications. (American Psychological Association), doi:10.1037/10694-000
Ting, Y. S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, Monthly Notices of the Royal Astronomical Society, 421, 1231
Tipping, M. E., & Bishop, C. M. 1999, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61, 611
Viswanathan, M., Wallace, C., Dowe, D. L., & Korb, K. 1999, Advanced Topics in Artificial Intelligence, 405
Wallace, C. S. 1995, Technical report, Department of Computer Science, Monash University, 95, 21
Wallace, C. S. 2005, Statistical and inductive inference by minimum message length (Springer Science & Business Media)
Wallace, C. S., & Freeman, P. R. 1987, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 49, 240
Walt, S. v. d., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engg., 13, 22
West, C., & Heger, A. 2013, ApJ, 774, 75