Draft version June 1, 2017
Typeset using LaTeX default style in AASTeX61

CHEMICAL TAGGING USING MINIMUM MESSAGE LENGTH. I. GAUSSIAN MIXTURE MODELLING

Andrew R. Casey,[1,2] (some order of:,[1] John Lattanzio,[1] David Dowe,[2] Aldeida Aleti,[3]), and others?

[1] *School of Physics and Astronomy, Monash University, Melbourne, Clayton VIC 3800, Australia*
[2] *Faculty of Information Technology, Monash University, Melbourne, Clayton VIC 3800, Australia*
[3] *Faculty of Information Technology, Monash University, Melbourne, Caulfield East VIC 3145, Australia*

(Received; Revised; Accepted)

Submitted to TBD

ABSTRACT

Chemical tagging seeks to identify co-natal stars by their present-day photospheric abundances, long after their phase space similarity is lost. In principle the detailed chemical abundances of a transformative number of stars could be used to reconstruct the evolution of the Milky Way, providing insight on supernovae yields, galactic dynamics, the initial mass function, as well as the formation and destruction of star clusters.

Some progress has been made towards chemical tagging, but here we argue that even the *extent* of the problem has not been realised. We introduce the Real World problems associated with chemical tagging at scale, and describe how currently considered methods will inevitably fail even under the most unrealistically optimistic conditions. We introduce the minimum message length to the astronomical community as a promising alternative to chemical tagging, one which has the capability to address all ensivaged problems with chemical tagging simultaneously. We perform chemical tagging experiments with infinte Gaussian mixture models, and demonstrate using real and generated data that MML outperforms all other considered penalty techniques.

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

1. INTRODUCTION

Freeman & Bland-Hawthorn (2002) introduced the idea of chemical tagging to identify groups of stars that formed together in the same gas cloud, which are no longer identifiable from their physical proximity or orbital properties. This idea is attractive because the observable chemical abundances of stars remains largely unchanged throughout a star's lifetime, whereas phase space similarity can be quickly washed out through dynamical interactions. Indeed, although most stars are thought to form in clusters (**?**), only the most massive or isolated star clusters are able to survive the star formation process (e.g., Lada 2010), leaving most clusters to become dynamically unbound after about 1 Gyr.

Despite the Milky Way being our best laboratory for understanding galaxy formation, a transformative number of stars with precise abundances will be required to chemically tag the Galaxy. Ground- and space-based surveys are poised to deliver those data in the coming decade. The Apache Point Observatory Galactic Evolution Experiment (APOGEE, e.g., Majewski et al. 2015) and the Gaia-ESO survey (GES; Gilmore et al. 2012; Randich et al. 2013) are already delivering 10^5 stars with 20–40 precise abundances, and the Galactic Archaeology with HERMES (GALAH; De Silva et al. 2015) survey seeks to derive ≈30 precise abundances for $\sim 10^6$ stars. Future ground-based surveys will derive abundances from ≈20 million spectra in each hemisphere (e.g., WEAVE, 4MOST; Dalton et al. 2012; de Jong et al. 2012), and the radial velocity spectrometer on *Gaia* can deliver a few abundances for ≈150 million stars.

Chemical tagging is a simple idea, but with this volume of data it represents a formidable parameter estimation and model selection problem. Given the difficulty of the problem, most chemical tagging studies have focussed on addressing one *aspect* of chemical tagging. For example, even the simplest applications of chemical tagging requires the identification of groups in the data, with stars being assigned to certain groups (a clustering problem), and to somehow chose the number of groups that are best represented by the data (a model selection problem). Previous works include definitions of L1-like metrics (Mitschang et al. 2014), where open clusters are used to define confidence intervals of similarity (Mitschang et al. 2013), and whether similar stars constitute a group or not. Other studies have directly applied standard clustering algorithms, (e.g., DBSCAN, k-means; Blanco-Cuaresma et al. 2015; Hogg et al. 2016) and fixed the number of clusters. Inferring the number of true clusters – one of the principal goals of chemical tagging – was not attempted in these simplified experiments.

Other works have focussed [TODO EXPAND]:

- Dimensionality of chemical abundance space (Ting et al).
- Chemical homogenity of open clusters (Bovy, Da Silva)
- Dopplegangers (Ness)
- A single common latent factor, integrated over cosmic time (Bland-Hawthorn)

The aforementioned chemical tagging experiments have sought to address only individual challenges associated with chemical tagging. These challenges are all related, and therefore addressing them separately will give results that ado not make use of all the information available. Here, as we introduce the problem of chemical tagging, we list the inferences we wish to make from a large number of chemical abundances.

1.1. *What's the goal of chemical tagging? What do we want to infer?*

Let us assume that a set of stars have been observed spectroscopically, and chemical abundances have been derived from those data. We will unrealistically assume that the data are noiseless, and the derived abundances are unbiased, known with infinite precision, and complete (e.g., no missing abundances). Regardless of the number of stars observed or abundances derived, there are a number of inferences we seek to make:

1. The number of star clusters (or star-forming events) where at least one star was observed and chemical abundances were derived.
2. The conditional probabilities (or memberships) of each star belonging to every star cluster or star-forming event. Alternatively, the relative weighting of individual mixtures.
3. The mean abundances of each star cluster.

4. The covariance matrix of abundances of each star cluster. This covariance matrix includes the intrinsic (cluster) variance of individual abundances (e.g., the homogeniety), and allows for correlations between abundances.[1]

5. The *number* of common multivariate latent factors, variables that are not observable but affect all or many of the observations. In the context of chemical tagging, the yields of core-collapse supernovae can be considered a common multivariate latent factor: those yields are not directly observable, but they contribute to much of the data. Here we are interested in the *number* of common multivariate latent factors. For example, the number of sources of enrichment: different classes of supernovae, asymptotic giant branch (AGB) stars, etc.

6. The multivariate factor loads of individual latent factors. For example, if core-collapse supernovae and AGB stars had both contributed to the chemical abundances of all stars in the sample, then the *factor loads* would be the *yield* of the supernovae, or the yield of relative abundances that were produced by the AGB star. Each star will be affected differently by those yields (see item 7).

7. The relative *scoring* of different latent factors on each star. This is a scaling that is applied to the latent yields. Continuing our analogy on yields from core-collapse supernovae and AGB stars, the *factor scores* can be thought of as the relative contributions of the different factor loads, for a single star.

8. The intrinsic *specific variances* of the individual abundances. These variances can be thought of as the intrinsic variability from the different factor loads. The intrisic variances affect all observations, before accounting for any homogeneity in individual clusters.

9. The *number* of stellar age- or parameter-dependent latent factors that affect the data. The surface chemical abundances of stars *do change* over its lifetime, at least in part due to atomic diffusion, veiling, and thermohaling mixing. Understanding and modelling these effects will be critical for chemical tagging stars across all stellar ages and evolutionary states.

10. The factor loads of stellar age- or parameter-dependent latent factors. In this example, a single multivariate factor load may represent the age-dependent effects of atomic diffusion.

11. The factor scores of stellar age- or parameter-dependent latent factors.

The number of things we seek to infer from the available chemical abundances is vast. Indeed, even if we ignore the latent factors then a non-trivial model selection problem remains: What is the number of star clusters or star-forming events? In the full description of chemical tagging we seek to know (1) the number of star clusters, (2) the number of latent factors, and (3) the number of age- or stellar parameter-dependent factors. Going further, if we wanted to improve the accuracy and precision of our solution then we may seek to include joint priors on the memberships of stars being associated with specific clusters, either based on their ages or astrometry. In summary, simplified versions of chemical tagging represent formidable model selection problems with non-convex[2] objective functions.

In this *Article* we will introduce a simple probabilistic approach to chemical tagging, where the data are modelled as a mixture of multivariate Gaussian distributions and the number of mixtures is unknown. In Section 2 we describe our model and outline our objective function. Although our model is simplistic in the sense that we will only address some aspects of chemical tagging, in Section 3 our experiments show that our approach is superior to all other methods considered for chemical tagging. In Section 4 we describe how the principles introduced here can be used to construct a fully consistent probabilistic approach to chemical tagging, and simultaneously address the challenges outlined in Section 1.1.

2. THE MODEL

We make the following explicit assumptions:

- We assume that the data (e.g., D detailed chemical abundances for N stars) can be represented by a mixture of K multivariate Gaussian distributions.

[1] It is likely – or hoped – that *true* correlations between abundances will be accounted for through the latent factors, and any remaining correlations are the result of the measurements.

[2] If a problem is convex then any local solution is mathematically guaranteed to be the global solution. If the problem is non-convex, then a local solution may not be the global solution.

- We assume that there are no missing data values. For example, if a star has been observed, then there is a complete set of D chemical abundances available.
- We will assume that the number of *true* multivariate Gaussian distributions K_{true} is not known. This assumption applies to our experiements involving generated data, and those using real data.
- In addition to not knowing the *number* of multivariate normal distributions, we further assume that the means $\boldsymbol{\mu}$ and covariance matrices $\mathbf{C}$ of each K-th mixture is unknown.
- We assume that the covariance matrix of individual components can be described as 'free' or 'full', insomuch that the data within a cluster can have off-diagonal correlation terms, and those terms are unknown.
- In all experiments we will assume that the data have homoskedastic noise properties.
- The relative weights $\mathbf{w}$ (or mixing proportions) of the K distributions is unknown.
- We assume a relative (or conditional) probability distribution for each star belonging to a given cluster. That is to say that we do not adopt a '*hard selection*' approach to cluster modelling, where stars would be assigned as definitively belonging to one cluster or another. Our approach could be described as '*soft* clustering'. This assumption has practical implications for optimization, and for the total message length.

The probability density function f for a multivariate normal distribution with D dimensions is given by

$$f(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}|}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\intercal \mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \tag{1}$$

where $\boldsymbol{\mu}$ and $\mathbf{C}$ is the multivariate mean and covariance matrix, and $\mathbf{y}$ are the data. For a fixed number of K multivariate Gaussian mixtures, the probability density function can be written as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \, f(\mathbf{y}|\boldsymbol{\theta}_k) \tag{2}$$

where $\boldsymbol{\theta}_k \equiv \{\boldsymbol{\mu}_k, \mathbf{C}_k\}$, $\boldsymbol{\theta} \equiv \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, w_1, \ldots, w_K\}$, and w_k is the relative weight (or the mixing probability) of the k-th mixture,[3] and we require that $\sum_{k=1}^{K} w_k = 1$. Although the true number of K mixtures is unknown (as are the means $\boldsymbol{\mu}$ and covariance matrices $\mathbf{C}$), we will address the selection of K and parameter estimates of $\boldsymbol{\theta}$ in Section **??**.

2.1. *Minimum Message Length*

The Minimum Message Length (MML; Wallace & Boulton 1968) principle is a formal description of Occam's razor that is based on information theory. The MML principle states that the best explanation of the data is the model that describes the data using the smallest amount of information (shortest message). Information is described with the same description as Shannon (1948), who showed that that given an event E with probability $P(E)$, the length of the shortest lossless code $I(E)$ to represent that event requires $I(E) = -\log_2 P(E)$ bits of information[4]. Wallace & Boulton (1968) linked this result with Bayes' theorem,

$$P(H|D) = \frac{P(H) \cdot P(D|H)}{P(D)} \tag{3}$$

which states that the joint probability of the data D and hypothesis H, $P(H, D)$, is given by

$$P(H, D) = P(H) \cdot P(D|H) = P(D) \cdot P(H|D) \tag{4}$$

where $P(H)$ is the prior probability of the hypothesis H, $P(D|H)$ is the likelihood, $P(D)$ is the prior probability of the data D, and $P(H|D)$ is the posterior probability of the hypothesis given the data. Specifically, Wallace & Boulton (1968) derived the relationship between conditional probabilities in terms of optimal message lengths,

[3] An alternative representation would be to include integer variables for each of the N stars, where each variable represents the cluster which that star belongs. These two models are mathematically identical (e.g., see discussion in Foreman-Mackey 2014), but the integer representation is highly dimensional and difficult to optimize.

[4] Unless specified – as it is in this instance – throughout this article we will refer to the information I or entropy in natural units (nats). Recall that 1 bit $= \log 2$ nats.

$$I(H, D) = I(H) + I(D|H) = I(D) + I(H|D) \quad . \tag{5}$$

Therefore the total cost of a message is the sum of the message length of the hypothesis H, which takes $I(H)$ bits, and the data D given the hypothesis H, $I(D|H)$. Considering a simple problem of model selection provides good intuition for this two-part message approach. Consider a simple model H_0 where the first part of the message $I(H_0)$ is short, and which provides a poor fit to the data such that $I(D|H_0)$ is long. An alternative model H_1 is more complex, requiring a longer $I(H_1)$, but the increased model flexibility provides a better fit to the data, making $I(D|H_1)$ short. For both parameter estimation *and* model selection, we want to chose H (of H_0, H_1, etc) and the model parameters that minimizes the *total* message length.

Calculating the length of the message is a non-trivial task, especially if the model is even reasonably complex. This makes the MML principle intractable (or uncomputable) in most cases, and forces us to make approximations when calculating the message length. Using a Taylor expansion, Wallace & Freeman (1987) introduce a generalised scheme to estimate a parameter vector $\boldsymbol{\theta}$ (of any distribution), that minimises the message length expression:

$$I(\boldsymbol{\theta}, \mathbf{y}) = \frac{Q}{2} \log \kappa(Q) - \log \left(\frac{p(\boldsymbol{\theta})}{\sqrt{|\mathcal{F}(\boldsymbol{\theta})|}} \right) - \mathcal{L}\left(\mathbf{y}|\boldsymbol{\theta}\right) + \frac{Q}{2} \quad . \tag{6}$$

Here $p(\boldsymbol{\theta})$ is the prior on the hypothesis, q is the number of free parameters in the model, $\kappa\left(Q\right)$ is a function to approximate the lattice quantisation constant for Q free parameters (e.g., Conway & Sloane 1984), and $|\mathcal{F}\left(\boldsymbol{\theta}\right)|$ is the determinant of the *expected* Fisher information matrix (the second-order partial derivatives of the negative log-likelihood function $-\mathcal{L}(\mathbf{y}|\boldsymbol{\theta})$).

The MML principle requires that the message fully specifies all components necessary for the receiver to be able to reconstruct the message, as well as our prior beliefs of the model. The message must include:

1. The number of Gaussian mixtures, K.
2. Encoding the relative weights $\mathbf{w}$ of the K Gaussian mixtures.
3. Encoding the component parameters $\boldsymbol{\mu}$, $\mathbf{C}$ for all K Gaussian mixtures.
4. Encoding the data, given the model parameters.
5. The lattice quantisation constant $\kappa(Q)$ for the number of model parameters Q.

Some of these components require non-trivial derivations or approximations in order to calculate the length of the optimal loseless encoding. For this reason, we will consider the message length of each component in turn before returning to the full message length specified in Equation 6.

2.1.1. *Encoding the number of Gaussian mixtures, K*

We assume that fewer mixtures are more likely than a larger number of mixtures. Specifically, we assume that $p(K) \propto 2^{-K}$, which implies that the optimal message length required to encode K is given by:

$$I(K) = -\log p(K) = K \log 2 + \text{constant} \quad . \tag{7}$$

2.1.2. *Encoding the relative weights,* $\mathbf{w}$

We assume a uniform prior on the mixing weights $\boldsymbol{w}$, only requiring that $\sum_{k=1}^{K} w_k = 1$. This implies that the weights can be treated as parameters of a multinomial distribution, and the length of their optimally encoded message is given by Boulton & Wallace (1969),

$$I(\mathbf{w}) = \frac{K-1}{2} \log N - \frac{1}{2} \sum_{k=1}^{K} \log w_k - \log \Gamma(K) \quad , \tag{8}$$

where $\Gamma(K)$ is the usual gamma function for positive integers $\Gamma(K) = (K-1)!$. A useful extension of this work might be to impose a Dirchlet prior on the mixing weights $\mathbf{w}$, or a distribution inspired by the stellar mass function and/or an observed selection function.

2.1.3. *Encoding the mixture parameters $\boldsymbol{\mu}$ and $\mathbf{C}$*

In order to properly encode the mixture parameters $\boldsymbol{\mu}$ and $\mathbf{C}$ for all K mixtures, we must encode both our prior belief on those parameters, and the determinant of the expected Fisher information matrix. For the k-th mixture this becomes,

$$I(\boldsymbol{\mu}_k, \mathbf{C}_k) = -\log\left(\frac{p(\boldsymbol{\mu}_k, \mathbf{C}_k)}{\sqrt{|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)|}}\right) = -\log p(\boldsymbol{\mu}_k, \mathbf{C}_k) + \frac{1}{2}\log|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| \tag{9}$$

and for all mixtures:

$$I(\boldsymbol{\mu}, \mathbf{C}) = -\sum_{k=1}^{K}\log p(\boldsymbol{\mu}_k, \mathbf{C}_k) + \frac{1}{2}\sum_{k=1}^{K}\log|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| \tag{10}$$

We introduce our priors on $\boldsymbol{\mu}$ and $\mathbf{C}$ before approximating the expected Fisher information matrix. We adopt a uniform prior of $\mathcal{U}(\boldsymbol{\mu}) = [-\infty, \infty]$ on all multivariate mean abundances $\boldsymbol{\mu}$ (in all D dimensions). We note that while this prior is improper, it becomes proper when it enters the (accepted) posterior distribution, and it has no impact on our inferences. Indeed, we similarly could adopt a (proper and) large uniform prior on all mean abundances of $\mathcal{U}(\boldsymbol{\mu}) = [-12, 12]$ (24 orders of magnitude!) and our inference would be the same.

We adopt a congujate inverted Wishart prior for the covariance matrices in the individual mixtures. The joint prior density for the parameters of a single mixture is given by the limiting form of the normal inverted-Wishart density (e.g., Section 5.2.3 of Schafer 1997):

$$p(\boldsymbol{\mu}_k, \mathbf{C}_k) \propto |\mathbf{C}_k|^{\frac{D+1}{2}} \quad . \tag{11}$$

Computing the expected Fisher information $\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)$ requires the second order partial derivatives of the negative log-likelihood $-\mathcal{L}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C})$. We approximate the determinant of the Fisher information $|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)|$ as the product of $|\mathcal{F}(\boldsymbol{\mu}_k)|$ and $|\mathcal{F}(\mathbf{C}_k)|$ (Oliver et al. 1996; Roberts et al. 1998). We begin by taking second derivative of the log-likelihood function $\mathcal{L}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C})$,

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C}) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\Sigma_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{\mu})^{\intercal}\mathbf{C}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) \tag{12}$$

with respect to $\boldsymbol{\mu}$,

$$\frac{\delta\mathcal{L}}{\delta\boldsymbol{\mu}} = \Sigma_{i=1}^{N}\mathbf{C}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \quad \text{and} \quad \frac{\delta^2\mathcal{L}}{\delta\boldsymbol{\mu}^2} = -N\mathbf{C}^{-1} \quad , \tag{13}$$

and for a D-dimensional vector $\boldsymbol{\mu}$ the Fisher information matrix is:

$$|\mathcal{F}(\boldsymbol{\mu}_k)| = \left|-\frac{\delta^2\mathcal{L}}{\delta\boldsymbol{\mu}^2}\right| = N^D|\mathbf{C}|^{-1} \quad . \tag{14}$$

Computing the Fisher information of the covariance matrix $|\mathcal{F}(\mathbf{C})|$ is harder. Using theory of matrix derivatives based on matrix vectorization (Dwyer 1967), Magnus & Neudecker (1988) derived the following analytical expression for the determinant of the Fisher information for a covariance matrix $|\mathcal{F}(\mathbf{C}_k)|$,

$$|\mathcal{F}(\mathbf{C}_k)| = N^{\frac{D(D+1)}{2}}2^{-D}|\mathbf{C}_k|^{-(D+1)} \tag{15}$$

where $D(D+1)/2$ arises from the number of free parameters in a covariance matrix with off-diagonal terms. Equations 14 and 15 allow us to approximate the determinant of the expected Fisher information matrix $|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)|$ as:

$$\begin{aligned}
|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| &\approx |\mathcal{F}(\boldsymbol{\mu}_k)| \cdot |\mathcal{F}(\mathbf{C}_k)| \\
|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| &\approx N^D|\mathbf{C}_k|^{-1}N^{\frac{D(D+1)}{2}}2^{-D}|\mathbf{C}_k|^{-(D+1)} \\
|\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| &\approx N^{\frac{D(D+3)}{2}}2^{-D}|\mathbf{C}_k|^{-(D+2)} \quad .
\end{aligned} \tag{16}$$

Thus for all K mixtures, substituting Equations 11 and 16 into Equation 10 gives:

$$
\begin{aligned}
I(\boldsymbol{\mu}, \mathbf{C}) &= -\sum_{k=1}^{K} \log p(\boldsymbol{\mu}_k, \mathbf{C}_k) + \frac{1}{2}\sum_{k=1}^{K} \log |\mathcal{F}(\boldsymbol{\mu}_k, \mathbf{C}_k)| \\
I(\boldsymbol{\mu}, \mathbf{C}) &= -\sum_{k=1}^{K} \log \left(|\mathbf{C}_k|^{\frac{D+1}{2}} \right) + \frac{1}{2}\sum_{k=1}^{K} \log \left(N^{\frac{D(D+3)}{2}} 2^{-D} |\mathbf{C}_k|^{-(D+2)} \right) \\
I(\boldsymbol{\mu}, \mathbf{C}) &= -\sum_{k=1}^{K} \frac{D+1}{2} \log |\mathbf{C}_k| + \frac{1}{2}\sum_{k=1}^{K} \left[\frac{D(D+3)}{2} \log N - D \log 2 - (D+2) \log |\mathbf{C}_k| \right] \\
I(\boldsymbol{\mu}, \mathbf{C}) &= -\sum_{k=1}^{K} \frac{D+1}{2} \log |\mathbf{C}_k| + \frac{K}{2} \left[\frac{D\,(D+3)}{2} \log N - D \log 2 \right] - \sum_{k=1}^{K} \frac{D+2}{2} \log |\mathbf{C}_k| \\
I(\boldsymbol{\mu}, \mathbf{C}) &= \frac{1}{2}\sum_{k=1}^{K} \log |\mathbf{C}_k| + \frac{K}{2} \left[\frac{D\,(D+3)}{2} \log N - D \log 2 \right] \quad . \qquad (17)
\end{aligned}
$$

2.1.4. *Encoding the data, given the model parameters*

Each datum can only be stated with a precision which is given by the accuracy of the measurement. If we assume homoskedastic noise properties for the observed data, and the precision that each elemental abundance can be measured is $\mathcal{E}$, then the *probability* of a datum is given by $\Pr(y_n) = \mathcal{E}^D \Pr(y_i|\mathcal{M})$ where $\mathcal{M}$ represents the model and $\Pr(y_i|\mathcal{M})$ is the *probability density*. Note that $\mathcal{E}$ is a constant, and is only used to maintain validity between *probability* and *probability density*. For these experiments we adopt $\mathcal{E} = 0.01$, but note that the value of $\mathcal{E}$ has no effect on our inferences.

The length of the total encoding of a datum given the model is,

$$
I(\mathbf{y}_n) = -\log \Pr(\mathbf{y}_n) = -D \log \mathcal{E} - \log \sum_{k=1}^{K} w_k f_k(\mathbf{y}_n|\boldsymbol{\mu}_k, \mathbf{C}_k) \qquad (18)
$$

and for the entire data:

$$
I(\mathbf{y}|\boldsymbol{\theta}) = -ND \log \mathcal{E} - \sum_{n=1}^{N} \log \sum_{k=1}^{K} w_k f_k(\mathbf{y}_n|\boldsymbol{\mu}_k, \mathbf{C}_k) \quad . \qquad (19)
$$

$$
(20)
$$

2.1.5. *Encoding the lattice constant*

The lattice quantisation constant arises from the approximation to the (so-called) strict MML, where parameters are quantised into intervals in high dimensional space. First we will calculate the total number of free parameters in our model. Recall that we assume our covariance matrices have non-zero off-diagonal terms, requiring $\frac{KD(D+1)}{2}$ parameters for all K covariance matrices (where we implicitly assume $D > 1$). The K D-dimensional vector means require a total of KD parameters. Under the assumption $K > 1$, the K relative weights $\mathbf{w}$ only require $K - 1$ parameters because $\sum_{k=1}^{K} w_k = 1$. Therefore the total number of free parameters Q is given by,

$$
\begin{aligned}
Q &= \frac{KD(D+1)}{2} + KD + K - 1 \\
Q &= K \left[\frac{D\,(D+3)}{2} + 1 \right] - 1 \quad . \qquad (21)
\end{aligned}
$$

We use an approximation for the logarithm of the lattice constant (see Sections 5.2.12 and 3.3.4 of Wallace 2005) where,

$$I(y) \approx -\log \frac{p(\theta')}{\sqrt{F(\theta')}} - \log f(y|\theta') - (Q/2)\log 2\pi + \frac{1}{2}\log(\pi Q) - 1 \tag{22}$$

such that,

$$\log \kappa(Q) = \frac{\log Q\pi}{Q} - \log 2\pi - 1 \tag{23}$$

and the expected error on $\log \kappa(Q)$ is less than 0.1 nit. The Q (lattice) terms in Equation 6 becomes:

$$\begin{aligned} \frac{Q}{2}(\log \kappa(Q) + 1) &= \frac{Q}{2}\left(\frac{\log Q\pi}{Q} - \log 2\pi - 1 + 1\right) \\ \frac{Q}{2}(\log \kappa(Q) + 1) &= \frac{1}{2}(\log Q\pi - Q\log 2\pi) \quad . \end{aligned} \tag{24}$$

2.2. *The Objective Function*

The message lengths of individual components allows us to fully specify an objective function to minimise the total message length, given any number of Gaussian mixtures K and parameter estimates $\boldsymbol{\theta}$:

$$I(\boldsymbol{\theta}, \mathbf{y}) = I(K) + I(\mathbf{w}) + \sum_{k=1}^{K} I(\boldsymbol{\theta}_k) + I(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{2}(\log Q\pi - Q\log 2\pi) \quad . \tag{25}$$

Substituting the expressions outlined in previous subsections (except for Q, as per Equation 21, which we leave unexpanded for brevity), we arrive at our complete objective function:

$$\begin{aligned} I(\boldsymbol{\theta}, \mathbf{y}) = {} & K\log 2 + \frac{K-1}{2}\log N - \frac{1}{2}\sum_{k=1}^{K}\log w_k - \log\Gamma(K) + \frac{1}{2}\sum_{k=1}^{K}\log|\mathbf{C}_k| \\ & + \frac{K}{2}\left[\frac{D(D+3)}{2}\log N - D\log 2\right] - ND\log\mathcal{E} - \sum_{n=1}^{N}\log\sum_{k=1}^{K} w_k f_k(\mathbf{y}_n|\boldsymbol{\mu}_k, \mathbf{C}_k) + \frac{1}{2}(\log Q\pi - Q\log 2\pi) \end{aligned} \tag{26}$$

Our objective function includes the number of mixtures, K, a discrete valued parameter. We seek the number of mixtures K and the parameter estimates $\boldsymbol{\theta}$ that minimise our objective function (Eq. 26).

Here we will outline our optimization procedure to minimise the message length when K is fixed, before describing our search strategy to move between trial values of K.

[TODO] - Expectation maximization, using MML (unbiased) updates

2.3. *The Search Strategy*

3. EXPERIMENTS

Here we describe simple chemical tagging experiments that are so optimistic that they could be considered laughable, but ones which will still fail for even the most simplistic chemical tagging approaches.

Each experiment will edge closer to a representative example of chemical tagging.

3.1. *Experiment 0: Position and Velocity information*

3.2. *Experiment 1: Using clusters, full complement of chemical abundances*

3.3. *Experiment 2: Using all clusters, full complement of chemical abundances, random fractions of field stars*

3.4. *Experiment 3: Using clusters, limited set of chemical abundances*

3.5. *Experiment 4: Fake data to replicate clusters, but use latent factor*

3.6. *Experiment 5: Fake data to replicate clusters, but use multiple latent factors*

4. DISCUSSION

This research has made use of NASA's Astrophysics Data System. This research is supported by an Australian Research Council Discovery Project grant (DP160100637). Source code for this project is available at https://github.com/andycasey/snob. This document was compiled on 2017-06-01 from revision hash d2cd3bc in that repository.

Software: astropy (Astropy Collaboration et al. 2013), numpy, scipy, scikit-learn

REFERENCES

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

Blanco-Cuaresma, S., Soubiran, C., Heiter, U., et al. 2015, A&A, 577, A47

Boulton, D. M. & Wallace, C. S. 1969, Journal of Theoretical Biology, 23, 269-278

Conway, J. H. & Sloane, N. J. A. 1984, Journal on Algebraic and Discrete Methods, 5, 3, 294-305

Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, Proc. SPIE, 8446, 84460P

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, Proc. SPIE, 8446, 84460T

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

Dwyer, P. S. 1967, Journal of the American Statistical Association, 62, 318, 607-625

Foreman-Mackey, D. 2014, Blog Post: Mixture Models., Zenodo, https://doi.org/10.5281/zenodo.15856

Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487

Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25

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

Lada, C. J. 2010, Philosophical Transactions of the Royal Society of London Series A, 368, 713

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2015, arXiv:1509.05420

Magnus, J. R. & Neudecker, H. 1998, Matrix differential calculus with applications in statistics and econometrics. Wiley, New York.

Mitschang, A. W., De Silva, G., Sharma, S., & Zucker, D. B. 2013, MNRAS, 428, 2321

Mitschang, A. W., De Silva, G., Zucker, D. B., et al. 2014, MNRAS, 438, 2753

Oliver, J. J., Baxter, R. A., Wallace, C. S. 1996, Unsupervised Learning using MML, Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, 364-372

Randich, S., Gilmore, G., & Gaia-ESO Consortium 2013, The Messenger, 154, 47

Roberts, S., Husmeier, D., Rezek, I., Penny, W. 1998, Bayesian approaches to Gaussian mixture modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 11, 1133-1142

Shannon, C. E. 1948, Bell System Technical Journal, 27, 3

Schafer, J. L. 1997, Analysis of Incomplete Multivariate Data (CRC Press)

Wallace, C. S. & Boulton, D. M. 1968, The Computer Journal, 11, 2

Wallace, C. S. & Freeman, P. R. 1987, Journal of the Royal Statistical Society. Series B (Methodological), 49, 3, 240-265

Wallace, C. S. 2005, Statistical and inductive inference using minimum message length. Springer-Verlag, NJ, USA