Submitted to the Annals of Applied Statistics
arXiv: 0905.2979

EXTREME DECONVOLUTION: INFERRING COMPLETE DISTRIBUTION FUNCTIONS FROM NOISY, HETEROGENEOUS AND INCOMPLETE OBSERVATIONS

By Jo Bovy[*], David W. Hogg[*,†] and Sam T. Roweis

New York University

We generalize the well-known mixtures of Gaussians approach to density estimation and the accompanying Expectation-Maximization technique for finding the maximum likelihood parameters of the mixture to the case where each data point carries an individual d-dimensional uncertainty covariance and has unique missing data properties. This algorithm reconstructs the error-deconvolved or "underlying" distribution function common to all samples, even when the individual data points are samples from different distributions, obtained by convolving the underlying distribution with the unique uncertainty distribution of the data point and projecting out the missing data directions. We show how this basic algorithm can be extended with Bayesian priors on all of the model parameters and a "split-and-merge" procedure designed to avoid local maxima of the likelihood. We apply this technique to a few typical astrophysical applications.

1. Introduction.

Inferring a distribution function given a finite set of samples from this distribution function is a problem of considerable general interest. The literature contains many density estimation techniques, ranging from simply binning the data in a histogram and smoother versions of this collectively known as kernel density estimation, to more sophisticated techniques such as non-parametric (penalized) maximum likelihood fitting (see, e.g., Silverman, 1986, for a review of all of the previous methods) and full Bayesian analyses (e.g., Diebolt and Robert, 1994; Richardson and Green, 1997; Rasmussen, 2000). These techniques perform well in the high signal-to-noise regime, i.e., when one has "good data", however, in scientific applications the data generally come with large and heterogeneous uncertainties and one often only has access to lower dimensional projections of the full

[*] Partially supported by NASA (grant NNX08AJ48G).

[†] During part of the period in which this research was performed, DWH was a research fellow of the Alexander von Humboldt Foundation of Germany at the Max-Planck-Institut für Astronomie, Heidelberg, Germany.

AMS 2000 subject classifications: Primary 62G07; secondary 62H30, 62P35

Keywords and phrases: Bayesian inference, density estimation, Expectation-Maximization, missing data, multivariate estimation, noise

data, i.e., there are often missing data. As a scientist, you are not interested in the observed distribution, what you really want to know is the underlying distribution, i.e., the distribution you would have if the data had vanishingly small uncertainties and no missing data. In this paper we describe a general approach for inferring distribution functions when these complications are present.

A frequently used density estimation technique is to model the distribution function as a sum of Gaussian distributions by optimizing the likelihood of this model given the data (e.g. McLachlan and Basford, 1988). We show that this approach can be generalized in the presence of noisy, heterogeneous, and incomplete data. The likelihood of the model for each data point is given by the model convolved with the (unique) uncertainty distribution of that data point; the objective function is obtained by simply multiplying these individual likelihoods together for the various data points. Optimizing this objective function one obtains a maximum likelihood estimate of the distribution (more specifically, of its parameters).

While optimization of this objective function can, in principle, be performed by a generic optimizer, we develop an Expectation-Maximization (EM) algorithm that optimizes the objective function. This algorithm works in much the same way as the normal EM algorithm for mixture of Gaussians density estimation, except that an additional degree of incompleteness is given by the actual values of the observables, since we only have access to noisy projections of the actual observables; in the expectation step these actual values are estimated based on the noisy and projected measured values and the current estimate of the distribution function. In the limit in which the noise is absent but the data are lower dimensional projections of the quantities of interest, this algorithm reduces to the algorithm described in Ghahramani and Jordan (1994a,b).

We also show how Bayesian priors on all of the parameters of the model can be naturally included in this algorithm as well as how a split-and-merge procedure that heuristically searches parameter space for better approximations to the global maximum can also be incorporated in this approach. These priors and the split-and-merge procedure can be important when applying the EM algorithm developed here in situations with real data where the likelihood surface can have a very complicated structure. We also discuss briefly the practical issues having to do with model selection in the mixture model approach.

Applications to real data sets are discussed in detail in Section 6, both in general terms, i.e., why the approach we put forward here is more appropriate when dealing with noisy, heterogeneous data than the more traditional density estimation techniques, as well as in some concrete examples. These concrete examples show that the technique developed in this paper performs extremely well even when the underlying distribution function has a complicated structure.

The technique we describe below has many applications besides returning a maximum likelihood fit to the error-deconvolved distribution function of a data sample. For instance, when an estimate of the uncertainty in the estimated parameters or distribution function is desired or when a full Bayesian analysis of the mixture model preferred, the outcome of the maximum likelihood technique developed here can be used as a seed for Markov Chain Monte Carlo (MCMC) methods for finite mixture modeling (e.g., Diebolt and Robert, 1994; Richardson and Green, 1997). Another possible application concerns fitting a linear relationship to a data set $\{(x_i, y_i)\}$ when the data has non-negligible uncertainties both in x as well as in y, which can be correlated. This problem can be thought of as fitting the underlying, error-deconvolved distribution of the points (x_i, y_i) with a Gaussian; the linear relationship then corresponds to the direction of the largest eigenvalue of the underlying distribution's covariance matrix. We describe this application in a specific case in section 6.2.

2. Likelihood of a mixture of Gaussian distributions given a set of heterogeneous, noisy samples. Our goal is to fit a model for the distribution of a d-dimensional quantity $\mathbf{v}$ using a set of N observational data points $\mathbf{w}_i$. Therefore, we need to write down the probability of the data under the model for the distribution. The observations are assumed to be noisy projections of the true values $\mathbf{v}_i$

$$\mathbf{w}_i = \mathbf{R}_i \mathbf{v}_i + \text{noise}\,, \tag{1}$$

where the noise is drawn from a Gaussian with zero mean and known covariance tensor $\mathbf{S}_i$. The case in which there is missing data occurs when the projection matrix $\mathbf{R}_i$ is rank-deficient. Alternatively, we can handle the missing data case by describing the missing data as directions of the covariance matrix that have a formally infinite eigenvalue; In practice we use very large eigenvalues in the noise-matrix. When the data has only a small degree of incompleteness, i.e., when each data point has only a small number of unmeasured dimensions, this latter approach is often the best choice, since one often has some idea about the unmeasured values. For example, in the example given below of inferring the velocity distribution of stars near the Sun we know that the stars are moving at velocities that do not exceed the speed of light, which is not very helpful, but also that none of the velocities exceed the local Galactic escape speed, since we can safely assume that all the stars are bound to the Galaxy. However, in situations in which each data point has observations of a dimensionality $\ll d$ using the projections matrices will greatly reduce the computational cost, since, as will become clear below, the most computationally expensive operations all take place in the lower dimensional space of the observations.

We will model the distribution $p(\mathbf{v})$ of the true values $\mathbf{v}$ as a mixture of K Gaussians:

$$p(\mathbf{v}) = \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{v}|\mathbf{m}_j, \mathbf{V}_j)\,, \tag{2}$$

where the amplitudes α_j sum to unity and the function $\mathcal{N}(\mathbf{v}|\mathbf{m}, \mathbf{V})$ is the Gaussian distribution with mean $\mathbf{m}$ and variance tensor $\mathbf{V}$:

$$\mathcal{N}(\mathbf{v}|\mathbf{m}, \mathbf{V}) = (2\pi)^{-d/2} \det(\mathbf{V})^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{v}-\mathbf{m})^\top \mathbf{V}^{-1}(\mathbf{v}-\mathbf{m})\right]\,. \tag{3}$$

For a given observation $\mathbf{w}_i$ the likelihood of the model parameters $\theta = (\alpha_j, \mathbf{m}_j, \mathbf{V}_j)$ given that observation and the noise covariance $\mathbf{S}_i$, which we will write as $p(\mathbf{w}_i|\theta)$, can be written as:

$$\begin{aligned} p(\mathbf{w}_i|\theta) \equiv p(\mathbf{w}_i|\mathbf{R}_i, \mathbf{S}_i, \theta) &= \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{w}_i, \mathbf{v}, j|\theta) \\ &= \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{w}_i|\mathbf{v}) p(\mathbf{v}|j, \theta) p(j|\theta)\,, \end{aligned} \tag{4}$$

where

$$\begin{aligned} p(\mathbf{w}_i|\mathbf{v}) &= \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{v}, \mathbf{S}_i) \\ p(\mathbf{v}|j, \theta) &= \mathcal{N}(\mathbf{v}|\mathbf{m}_j, \mathbf{V}_j) \\ p(j|\theta) &= \alpha_j\,. \end{aligned} \tag{5}$$

This likelihood works out to be itself a mixture of Gaussians[1]

$$p(\mathbf{w}_i|\theta) = \sum_j \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\,\mathbf{m}_j, \mathbf{T}_{ij})\,, \tag{8}$$

where

$$\mathbf{T}_{ij} = \mathbf{R}_i\,\mathbf{V}_j\,\mathbf{R}_i^\top + \mathbf{S}_i\,. \tag{9}$$

[1] Briefly—setting all of the projection matrices equal to the unit matrix in order not to

The free parameters of this model can now be chosen such as to maximize an explicit, justified, scalar objective function ϕ, given here by the logarithm

clutter the algebra—this is shown as follows

$$\begin{aligned}
p(\mathbf{w}_i|\theta) &= \sum_j \alpha_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, \mathcal{N}(\mathbf{w}_i|\mathbf{v},\mathbf{S}_i)\mathcal{N}(\mathbf{v}|\mathbf{m}_j,\mathbf{V}_j) \\
&= \sum_j \alpha_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, (2\pi)^{-d} \det(\mathbf{S}_i)^{-1/2} \det(\mathbf{V}_j)^{-1/2} \\
&\qquad \exp\left[-\frac{1}{2}\left\{(\mathbf{w}_i-\mathbf{v})^\top \mathbf{S}_i^{-1}(\mathbf{w}_i-\mathbf{v}) + (\mathbf{v}-\mathbf{m}_j)^\top \mathbf{V}_j^{-1}(\mathbf{v}-\mathbf{m}_j)\right\}\right] \\
&= \sum_j \alpha_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, (2\pi)^{-d} \det(\mathbf{S}_i)^{-1/2} \det(\mathbf{V}_j)^{-1/2} \\
&\qquad \exp\Bigg[-\frac{1}{2}\{\mathbf{v}^\top(\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1})\mathbf{v} - \mathbf{v}^\top(\mathbf{V}_j^{-1}\mathbf{m}_j + \mathbf{S}_i^{-1}\mathbf{w}_i) \\
&\qquad\quad -(\mathbf{m}_j^\top \mathbf{V}_j^{-1} + \mathbf{w}_i^\top \mathbf{S}_i^{-1})\mathbf{v} + \mathbf{w}_i^\top \mathbf{S}_i^{-1}\mathbf{w}_i + \mathbf{m}_j^\top \mathbf{V}_j^{-1}\mathbf{m}_j\}\Bigg] \,.
\end{aligned}$$

This integral works out to be

$$\begin{aligned}
p(\mathbf{w}_i|\theta) &= \sum_j \alpha_j (2\pi)^{-d/2} \frac{\det(\mathbf{S}_i)^{-1/2}\det(\mathbf{V}_j)^{-1/2}}{\det(\mathbf{S}_i^{-1}+\mathbf{V}_j^{-1})^{1/2}} \\
&\qquad \exp\Bigg[-\frac{1}{2}\{\mathbf{m}_j^\top \mathbf{V}_j^{-1}\left(-\frac{1}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\mathbf{V}_j^{-1}+1\right)\mathbf{m}_j \\
&\qquad\quad +\mathbf{w}_i^\top \mathbf{S}_i^{-1}\left(-\frac{1}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\mathbf{S}_i^{-1}+1\right)\mathbf{w}_i \\
&\qquad\quad -\mathbf{m}_j^\top \mathbf{V}_j^{-1}\frac{1}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\mathbf{S}_i^{-1}\mathbf{w}_i - \mathbf{w}_i\mathbf{S}_i^{-1}\frac{1}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\mathbf{V}_j^{-1}\mathbf{m}_j\}\Bigg] \,,
\end{aligned} \tag{6}$$

which simplifies to

$$\begin{aligned}
p(\mathbf{w}_i|\theta) &= \sum_j \alpha_j (2\pi)^{-d/2} \frac{\det(\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1})^{-1/2}}{\det(\mathbf{S}_i^{-1})^{-1/2}\det(\mathbf{V}_j^{-1})^{-1/2}} \\
&\qquad \exp\left[-\frac{1}{2}(\mathbf{w}_i-\mathbf{m}_j)^\top \mathbf{V}_j^{-1}\frac{1}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\mathbf{S}_i^{-1}(\mathbf{w}_i-\mathbf{m}_j)\right] \\
&= \sum_j \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{m}_j,\mathbf{T}_{ij}) \,,
\end{aligned} \tag{7}$$

where we have defined

$$\mathbf{T}_{ij} = \left(\frac{\mathbf{S}_i^{-1}\mathbf{V}_j^{-1}}{\mathbf{V}_j^{-1}+\mathbf{S}_i^{-1}}\right)^{-1} = \mathbf{V}_j + \mathbf{S}_i \,.$$

(log) likelihood of the model given the data, i.e.,

$$\phi = \sum_i \ln p(\mathbf{w}_i|\theta) = \sum_i \ln \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\,\mathbf{m}_j, \mathbf{T}_{ij})\,. \tag{10}$$

This function can be optimized in several ways, one of which is to calculate the gradients and use a generic optimizer to increase the likelihood until it reaches a maximum. This approach is complicated by parameter constraints (e.g., the amplitudes α_j must all be non-negative and add up to one, the variance tensors must be positive definite and symmetric). In what follows we will describe a different approach: An EM algorithm that iteratively maximizes the likelihood, while naturally respecting the restrictions on the parameters.

3. Fitting Mixtures with heterogeneous, noisy data using an EM algorithm. The problem of finding a maximum likelihood estimate of the parameters of the mixture of Gaussians model by optimizing the total log likelihood given in equation (10) is not a problem with missing data. However, optimization of the total log likelihood is difficult and an analytical solution is not possible for $K > 1$. An analytical solution does exist when $K = 1$ (when the uncertainty covariances $\mathbf{S}_i$ are equal; see below). Therefore, if we knew which Gaussian a specific data point was sampled from, optimization would be simple. The formulation of the Gaussian mixture density estimation as a hidden data problem takes advantage of this fact (Dempster et al., 1977).

When data is actually missing and/or when the data is noisy (uncertainty covariances $\mathbf{S}_i$ not equal to zero), an analytical solution does not exist anymore. As we will show below, in this case formulating the problem as a missing data problem can also lead to a simple, iterative algorithm that leads to a maximum likelihood estimate of the model parameters. First we will briefly recapitulate how the EM algorithm for Gaussian mixtures works by applying it to the basic problem of fitting a set of data points with a mixture of Gaussians, thus we will set all the uncertainty covariances equal to zero. Then we will investigate how the problem can be solved by a similar EM algorithm when we have incomplete and/or noisy data. A short summary of the general properties of the EM methodology is given in Appendix C.

3.1. *The EM algorithm with complete, precise observations.* In the case of complete, precise observations (i.e., $\mathbf{S}_i = 0$, $\mathbf{R}_i = \mathbf{I}$, $\forall\, i$) the log likelihood of the model given the data from equation (10) reduces to the following log likelihood:

$$\phi = \sum_i \ln p(\mathbf{w}_i|\theta) = \sum_i \ln \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{m}_j, \mathbf{V}_j)\,. \tag{11}$$

Formulating this problem as a missing data problem introduces the indicator variables q_{ij} which indicate whether a data point i was sampled from Gaussian j, i.e.,

$$q_{ij} = \begin{cases} 1 & \text{if data point } i \text{ was generated by Gaussian } j \\ 0 & \text{if data point } i \text{ was } \textit{not} \text{ generated by Gaussian } j \end{cases} \tag{12}$$

This variable can take on values between these extreme values, in which case q_{ij} corresponds to the probability that data point i was generated by Gaussian j. In any case, for every data point we have that $\sum_j q_{ij} = 1$.

Using this hidden indicator variable we can write the "full-data" log likelihood as

$$\Phi = \sum_i \sum_{j=1}^{K} q_{ij} \ln \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{m}_j, \mathbf{V}_j)\,. \tag{13}$$

Using Jensen's inequality it is easy to see that optimizing this full-data log likelihood also optimizes the original log likelihood equation (11); we reproduce this proof in our own notation in Appendix C so it can easily be compared to the similar proof in the case of incomplete data below. The two-step EM optimization algorithm that is derived in this proof alternates the following steps:

$$\begin{aligned}
\textbf{E-step:} \quad q_{ij} &\leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{m}_j, \mathbf{V}_j)}{\sum_k \alpha_k \mathcal{N}(\mathbf{w}_i|\mathbf{m}_k, \mathbf{V}_k)} \\
\textbf{M step:} \quad \alpha_j &\leftarrow \frac{1}{N} \sum_i q_{ij} \\
\mathbf{m}_j &\leftarrow \frac{1}{q_j} \sum_i q_{ij}\, \mathbf{w}_i \\
\mathbf{V}_j &\leftarrow \frac{1}{q_j} \sum_i q_{ij} \left[(\mathbf{m}_j - \mathbf{w}_i)\,(\mathbf{m}_j - \mathbf{w}_i)^\top\right]\,,
\end{aligned} \tag{14}$$

where $q_j = \sum_i q_{ij}$.

A fatal flaw of the maximum likelihood technique described here, and we must emphasize that this is a flaw of the objective function and not of the EM algorithm, is that the likelihood is unbounded. Indeed, when one of the means $\mathbf{m}_j$ is set equal to one of the data points, reducing the covariance matrix $\mathbf{V}_j$ of that Gaussian will lead to an unbounded increase in the probability of that data point, i.e., the Gaussian becomes a delta distribution centered on a particular data point and the model therefore has a infinite likelihood. This problem has many different solutions, some of which we explore below. We will describe a solution which assumes a prior for the model covariances, which leads to a regularization of the covariances at every step such that they cannot become zero. However, we will use this technique to deal with the related problem of the model covariances reaching a maximum likelihood at the edge of their domain, i.e., when the covariance becomes zero without making the likelihood infinite. In the next section we will describe a solution to the unbounded likelihood problem that is especially well motivated when dealing with experimental or observational data, i.e., taking into account the measurement uncertainties.

3.2. *The EM algorithm with heterogeneous, noisy data.* The problem we would like to solve has an additional component of incompleteness. The data we are using are noisy, as described by their covariance matrices $\mathbf{S}_i$. This noise can vary between small fluctuations to a complete lack of information for certain components of the underlying quantity $\mathbf{v}$ (as indicated by a formally infinite contribution to the covariance matrix). We can use the EM algorithm to deal with this second kind of incomplete data as well[2].

In the case of full-rank projection matrices, we could try to proceed exactly as we did in the case of complete data with noise covariance matrices $\mathbf{S}_i$ equal to the zero matrix. The E-step remains the same and the part of the M-step that updates the amplitudes will also be the same as before, however, in order to optimize the means and covariances, we would have to solve an equation similar to equation (78), i.e.,

$$\begin{aligned} d\langle\Phi_{\text{red}}\rangle = \sum_i q_{ij}\Big[&\mathbf{T}_{ij}^{-1}d\mathbf{T}_{ij} - \mathbf{T}_{ij}^{-1}d\mathbf{T}_{ij}\mathbf{T}_{ij}^{-1}(\mathbf{w}_i-\mathbf{m}_j)(\mathbf{w}_i-\mathbf{m}_j)^\top - \\ &2\mathbf{T}_{ij}^{-1}(\mathbf{w}_i-\mathbf{m}_j)d\mathbf{m}_j^\top\Big] = 0\ , \end{aligned} \tag{15}$$

in which we simply use $\mathbf{T}_{ij} = \mathbf{V}_j + \mathbf{S}_i$ instead of $\mathbf{V}_j$. This equation cannot be solved analytically to give us update steps for the means and covariances of the Gaussians when the noise covariances $\mathbf{S}_i$ are different for each observation. A reasonable way to deal with this would be to use a generic optimizer

[2] This algorithm was developed independently before (Diebolt & Celeux, 1989, unpublished; Diebolt & Celeux, 1990, unpublished).

to perform the M-step optimization, which would still give a better result than using a generic optimizer for the whole problem since the dimensionality of the problem is greatly reduced, however, we will describe a different procedure which deals with this problem in a similar way to how the EM algorithm dealt with the complete data case, i.e., by introducing the missing data into the likelihood.

Essentially, what we will do is consider the situation of noisy measurements, which may or may not be projections of higher dimensional quantities, as a missing data problem in itself. That is, we will consider the true values $\mathbf{v}_i$ as extra missing data (in addition to the indicator variables q_{ij}). This allows us to write down the "full data" log likelihood as follows

$$\Phi = \sum_i \sum_j q_{ij} \ln \alpha_j \mathcal{N}(\mathbf{v}_i | \mathbf{m}_j, \mathbf{V}_j) \ . \tag{16}$$

We will now show how we can use the EM methodology to find straightforward update steps that maximize the full data likelihood of the model. Then we will prove that these updates also maximize the likelihood of the model given the noisy observations.

The E-step consists as usual of taking the expectation of the full data likelihood with respect to the current model parameters θ. Writing out the full data log likelihood from equation (16) we find

$$\Phi = \sum_i \sum_j q_{ij} \left[\ln \alpha_j - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln \det \mathbf{V}_j - \frac{1}{2} (\mathbf{v}_i - \mathbf{m}_j)^\top \mathbf{V}_j^{-1} (\mathbf{v}_i - \mathbf{m}_j) \right] \ , \tag{17}$$

which shows that in addition to the expectation of the indicator variables q_{ij} for each component we also need the expectation of the $q_{ij}\mathbf{v}_i$ terms and the expectation of the $q_{ij}\mathbf{v}_i\mathbf{v}_i^\top$ terms given the data, the current model estimate and the component j. The expectation of the q_{ij} is again equal to the posterior probability that a data point $\mathbf{w}_i$ was drawn from the component j (see equation 73). The expectation of the $\mathbf{v}_i$ and the $\mathbf{v}_i\mathbf{v}_i^\top$ can be found as follows: Consider the probability distribution of the vector $[\mathbf{v}_i^\top \, \mathbf{w}_i^\top]^\top$ given the model estimate and the component j. From the description of the problem we can see that this vector is distributed normally with mean

$$\mathbf{m}' = \begin{bmatrix} \mathbf{m}_j \\ \mathbf{R}_i \mathbf{m}_j \end{bmatrix} \tag{18}$$

and covariance matrix

$$\mathbf{V}' = \begin{bmatrix} \mathbf{V}_j & \mathbf{V}_j \mathbf{R}_i^\top \\ \mathbf{R}_i \mathbf{V}_j & \mathbf{T}_{ij} \end{bmatrix} \ . \tag{19}$$

The conditional distribution of the $\mathbf{v}_i$ given the data $\mathbf{w}_i$ is normal with mean (see Appendix B)

$$\mathbf{b}_{ij} \equiv \mathbf{m}_j + \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1} (\mathbf{w}_i - \mathbf{R}_i\, \mathbf{m}_j) \tag{20}$$

and covariance matrix

$$\mathbf{B}_{ij} \equiv \mathbf{V}_j - \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1} \mathbf{R}_i\, \mathbf{V}_j\ . \tag{21}$$

Thus we see that the expectation of $\mathbf{v}_i$ given the data $\mathbf{w}_i$, the model estimate, and the component j is given by $\mathbf{b}_{ij}$, whereas the expectation of the $\mathbf{v}_i \mathbf{v}_i^\top$ given the same is given by $\mathbf{B}_{ij} + \mathbf{b}_{ij}\mathbf{b}_{ij}^\top$.

Given this the expectation of the full data log likelihood is given by

$$\begin{aligned}
\langle \Phi \rangle &= \sum_{i,j} q_{ij} \Bigg[\ln \alpha_j - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \mathrm{Trace}\big[\ln \mathbf{V}_j + (\mathbf{B}_{ij} + \mathbf{b}_{ij}\mathbf{b}_{ij}^\top \\
&\qquad\qquad -\mathbf{b}_{ij}\mathbf{m}_j^\top - \mathbf{m}_j \mathbf{b}_{ij}^\top + \mathbf{m}_j \mathbf{m}_j^\top) \mathbf{V}_j^{-1} \big] \Bigg] \qquad (22) \\
&= \sum_{i,j} q_{ij} \Bigg[\ln \alpha_j - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \mathrm{Trace}\big[\ln \mathbf{V}_j \\
&\qquad\qquad +(\mathbf{B}_{ij} + (\mathbf{m}_j - \mathbf{b}_{ij})(\mathbf{m}_j - \mathbf{b}_{ij})^\top) \mathbf{V}_j^{-1} \big] \Bigg]\ . \qquad (23)
\end{aligned}$$

The update step for the amplitudes α_j is given as before by equation (77). Dropping the $\ln(2\pi)$ term the differential of the reduced expectation of the full data log likelihood $\langle \Phi_{\mathrm{red}} \rangle$ is given by

$$\begin{aligned}
\mathrm{d}\langle \Phi_{\mathrm{red}} \rangle = \sum_{i,j} q_{ij} \mathrm{Trace} \Bigg[& \mathbf{V}_j^{-1} d\mathbf{V}_j - \mathbf{V}_j^{-1} d\mathbf{V}_j \mathbf{V}_j^{-1} \left[(\mathbf{b}_{ij} - \mathbf{m}_j)(\mathbf{b}_{ij} - \mathbf{m}_j)^\top + \mathbf{B}_{ij} \right] \\
& - 2\mathbf{V}_j^{-1} (\mathbf{b}_{ij} - \mathbf{m}_j) d\mathbf{m}_j^\top \Bigg]\ .
\end{aligned} \tag{24}$$

The complete EM algorithm given incomplete, noisy observations is then given by

$$
\begin{aligned}
\textbf{E-step:}\quad q_{ij} &\leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\,\mathbf{m}_j, \mathbf{T}_{ij})}{\sum_k \alpha_k \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\,\mathbf{m}_k, \mathbf{T}_{ik})} \\
\mathbf{b}_{ij} &\leftarrow \mathbf{m}_j + \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1}(\mathbf{w}_i - \mathbf{R}_i\,\mathbf{m}_j) \\
\mathbf{B}_{ij} &\leftarrow \mathbf{V}_j - \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1} \mathbf{R}_i \mathbf{V}_j \\
\textbf{M step:}\quad \alpha_j &\leftarrow \frac{1}{N}\sum_i q_{ij} \\
\mathbf{m}_j &\leftarrow \frac{1}{q_j}\sum_i q_{ij}\,\mathbf{b}_{ij} \\
\mathbf{V}_j &\leftarrow \frac{1}{q_j}\sum_i q_{ij}\left[(\mathbf{m}_j - \mathbf{b}_{ij})\,(\mathbf{m}_j - \mathbf{b}_{ij})^\top + \mathbf{B}_{ij}\right]\,, \qquad (25)
\end{aligned}
$$

where, as before, $q_j = \sum_i q_{ij}$.

We can prove that this procedure for maximizing the full data likelihood also maximizes the log likelihood of the data $\mathbf{w}_i$ given the model. We use Jensen's inequality in the continuous case, i.e.,

$$
f\left(\int \mathrm{d}\mathbf{v}\, q(\mathbf{v})\, p(\mathbf{v})\right) \geq \int \mathrm{d}\mathbf{v}\, q(\mathbf{v})\, f(p(\mathbf{v}))\,, \qquad (26)
$$

for a concave function f and a non-negative integrable function q, where we have assumed that q is normalized, i.e., q is a probability distribution. For each observation $\mathbf{w}$ we can then introduce a function $q(\mathbf{v}, j)$ such that

$$
\begin{aligned}
\ln p(\mathbf{w}|\theta) &= \ln \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{w}, \mathbf{v}, j|\theta) \\
&\geq \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, q(\mathbf{v}, j)\, \ln \frac{p(\mathbf{w}, \mathbf{v}, j|\theta)}{q(\mathbf{v}, j)} = F(\mathbf{w}|q, \theta) \\
\ln p(\mathbf{w}|\theta) &\geq F(\mathbf{w}|q, \theta) = \langle \ln p(\mathbf{w}, \mathbf{v}, j|\theta)\rangle_q + \mathcal{H}(q)\,, \qquad (27)
\end{aligned}
$$

where θ, as before, represents the set of model parameters, and $\mathcal{H}$ is the entropy of the distribution $q(\mathbf{v}, j)$. This inequality becomes an equality when we take

$$
q(\mathbf{v}, j) = p(\mathbf{v}, j|\mathbf{w}, \theta)\,, \qquad (28)
$$

since then

$$\begin{aligned}(29)\quad \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}, j|\mathbf{w}, \theta)\, \ln \frac{p(\mathbf{w}, \mathbf{v}, j|\theta)}{p(\mathbf{v}, j|\mathbf{w}, \theta)} &= \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}, j|\mathbf{w}, \theta)\, \ln \frac{p(\mathbf{v}, j|\mathbf{w}, \theta) p(\mathbf{w}|\theta)}{p(\mathbf{v}, j|\mathbf{w}, \theta)} \\ &= \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}, j|\mathbf{w}, \theta)\, \ln p(\mathbf{w}|\theta) \\ &= \ln p(\mathbf{w}|\theta) \sum_j \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}, j|\mathbf{w}, \theta) \\ &= \ln p(\mathbf{w}|\theta)\,.\end{aligned}$$

The above holds for each data point, and we can write

$$(30)\qquad p(\mathbf{v}, j|\mathbf{w}_i, \theta) = p(\mathbf{v}|\mathbf{w}_i, \theta, j)\, p(j|\mathbf{w}_i, \theta)\,.$$

The last factor reduces to calculating the posterior probabilities $q_{ij} = p(j|\mathbf{w}_i, \theta)$ and we can write the F function as (we drop the entropy term here, since it plays no role in the optimization, as it does not depend on the model parameters)

$$\begin{aligned} F &= \sum_{i,j} q_{ij} \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j)\, \ln p(\mathbf{w}_i, \mathbf{v}, j|\theta) \\ &= \sum_{i,j} q_{ij} \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j)\, \ln p(j|\theta) p(\mathbf{w}_i, \mathbf{v}|\theta, j) \\ &= \sum_{i,j} q_{ij} \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j)\, \ln p(j|\theta) p(\mathbf{w}_i|\mathbf{v}) p(\mathbf{v}|\theta, j) \\ &= \sum_{i,j} q_{ij} \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j)\, [\ln \alpha_j + \ln p(\mathbf{w}_i|\mathbf{v}) + \ln p(\mathbf{v}|\theta, j)] \\ &= \sum_{i,j} q_{ij} \Bigg[\ln \alpha_j + \int_{\mathbf{v}} \mathrm{d}\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j) \\ &\qquad \times \left[-\frac{1}{2} \ln\det \mathbf{V}_j - \frac{1}{2} (\mathbf{v} - \mathbf{m}_j)^\top \mathbf{V}_j^{-1} (\mathbf{v} - \mathbf{m}_j) \right] \Bigg]\,, \end{aligned}$$

where we dropped the $\ln p(\mathbf{w}_i|\mathbf{v})$ terms since they don't depend on the model parameters directly. This shows that this reduces exactly to the procedure described above, i.e., to taking the expectation of the $\mathbf{v}_i$ and $\mathbf{v}_i\mathbf{v}_i^\top$ terms with respect to the distribution of the $\mathbf{v}_i$ given the data $\mathbf{w}_i$, the current parameter estimate, and the component j. We conclude that the E-step as described above ensures that the expectation of the full data log likelihood becomes equal to the log likelihood of the model given the observed data. Optimizing this log likelihood in the M-step then also increases the log likelihood of the model given the observations. Therefore the EM algorithm we described will increase the likelihood of the model in every iteration, and the algorithm will approach local maxima of the likelihood. Convergence is identified, as usual, as extremely small incremental improvement in the log likelihood per iteration.

Some care must be taken to implement these equations in a numerically stable way. In particular, care should be taken to avoid underflow when computing the ratio of a small probability over the sum of other small probabilities (see Appendix A). Notice that we don't have to explicitly enforce constraints on parameters, e.g., keeping covariances symmetric and positive definite, since this is taken care of by the updates. For example, the update equation for $\mathbf{V}_j$ is guaranteed by its form to produce a symmetric positive-semidefinite matrix.

In some cases we might want to keep some of the model parameters fixed during the EM steps, and for the means and the covariances this can simply be achieved by not updating them during the M step. However, if we want to keep certain of the amplitudes fixed we have to be more careful, as we have to satisfy the constraint that the amplitudes add up to one at all times. Therefore, if j' indexes the free amplitudes and j'' the set of amplitudes we want to keep fixed, the Lagrange multiplier term we have to add to the functional F is $\lambda\left(\sum_{j'}\alpha_{j'} - 1 + \sum_{j''}\alpha_{j''}\right)$, which leads to the following update equations for the amplitudes

$$\alpha_{j'} \leftarrow \frac{q_{j'}}{\sum_{j'} q_{j'}}\left[1 - \sum_{j''}\alpha_{j''}\right] \ . \tag{31}$$

4. Extensions to the basic algorithm. Singularities and local maxima are two problems, that can severely limit the generalization capabilities of the computed density estimates for inferring the densities of unknown data points. These are commonly encountered when using the EM algorithm to iteratively compute the maximum likelihood estimates of Gaussian mixtures. Singularities arise when the covariance in equation (14) or equation (25) becomes singular, i.e., when a Gaussian in the fit becomes very peaked or elongated. A naive way of dealing with this problem is by artificially enforcing a lower bound on the variances of the Gaussians in the mixture, but this usually results in very low inference capabilities of the resulting mixture estimate. We will present a regularization scheme based on defining a Bayesian prior distribution on the parameter space below.

The problem of local maxima is a natural consequence of the EM algorithm as presented above: as the EM procedure ensures a monotonic increase in log likelihood of the model, the algorithm will exit when reaching a local maximum, since it cannot reach the true maximum without passing through lower likelihood regions. A solution to this problem therefore has to discontinuously change the model parameters, thereby quite possibly ending up in a region of parameter space with a lower likelihood than the first estimate. As a result the log likelihood is no longer guaranteed to increase monotonically, however, decreases in log likelihood are rare and concentrated at these discontinuous jumps. A well-motivated algorithm based on moving Gaussians from overpopulated regions to underpopulated regions of parameter-space is described below. However, instead of this deterministic approach, a stochastic EM procedure could also be used to explore the likelihood surface (Broniatowski et al., 1983; Celeux and Diebolt, 1985, 1986).

We also briefly discuss methods to set the remaining free parameters, the number of Gaussians K and the hyperparameters introduced in the Bayesian regularization described below.

4.1. *Bayesian regularization.* A general Bayesian regularization scheme consists of putting prior distributions on the Gaussian mixtures parameters space $\theta = (\alpha_j, \mathbf{m}_j, \mathbf{V}_j)$ (Ormoneit and Tresp, 1996). A conjugate prior distribution of a single multivariate normal distribution is the product of normal density $\mathcal{N}(\mathbf{m}_j|\hat{\mathbf{m}}, \eta^{-1}\mathbf{V}_j)$ and a Wishart density $\mathcal{W}(\mathbf{V}_j^{-1}|\omega, \mathbf{W})$ (Gelman et al., 2000), where

$$\mathcal{W}(\mathbf{V}_j^{-1}|\omega, \mathbf{W}) = c(\omega, \mathbf{W})|\mathbf{V}_j^{-1}|^{\omega-(d+1)/2} \exp\left[-\mathrm{Trace}\left[\mathbf{W}\mathbf{V}_j^{-1}\right]\right] , \tag{32}$$

with $c(\omega, \mathbf{W})$ a normalization constant. The proper prior distribution on the amplitudes α_j is a Dirichlet density $\mathcal{D}(\alpha|\gamma)$, given by

$$\mathcal{D}(\alpha|\gamma) = b \prod_j \alpha_j^{\gamma_j - 1} , \tag{33}$$

where b is a normalizing factor, $\alpha_j \geq 0$, and $\sum_j \alpha_j = 1$ (as is the case for the amplitudes in the density mixture). The log likelihood we want to maximize then becomes

$$\begin{aligned}\sum_i \ln p(\mathbf{w}_i|\theta) \leftarrow & \sum_i \ln p(\mathbf{w}_i|\theta) + \ln \mathcal{D}(\alpha|\gamma) \\ & + \sum_j \left[\ln \mathcal{N}(\mathbf{m}_j|\hat{\mathbf{m}}, \eta^{-1}\mathbf{V}_j) + \ln \mathcal{W}(\mathbf{V}_j^{-1}|\omega, \mathbf{W})\right] .\end{aligned} \tag{34}$$

We can again use the EM algorithm to find a local maximum to this function. The E-step is the same E-step as before (eq. (25)). In the M step the functional we have to maximize is the same functional as in the unregularized case (given in eq. (22)) with added terms:

(35)
$$\langle \Phi \rangle \leftarrow \langle \Phi \rangle + \mathcal{H}(q_i) + \ln \mathcal{D}(\alpha|\gamma) + \sum_j \left[\ln \mathcal{N}(\mathbf{m}_j|\hat{\mathbf{m}}, \eta^{-1}\mathbf{V}_j) + \ln \mathcal{W}(\mathbf{V}_j^{-1}|\omega, \mathbf{W}) \right] .$$

of which we can take derivatives with respect to the model parameters again, which leads to the following update equations

(36)
$$\begin{aligned} \alpha_j &\leftarrow \frac{\sum_i q_{ij} + \gamma_j - 1}{N + \sum_k \gamma_k - K} \\ \mathbf{m}_j &\leftarrow \frac{\sum_i q_{ij}\, \mathbf{b}_{ij} + \eta \hat{\mathbf{m}}}{q_j + \eta} \\ \mathbf{V}_j &\leftarrow \\ &\frac{\sum_i q_{ij} \left[(\mathbf{m}_j - \mathbf{b}_{ij}) \, (\mathbf{m}_j - \mathbf{b}_{ij})^\top + \mathbf{B}_{ij} \right] + \eta (\mathbf{m}_j - \hat{\mathbf{m}})(\mathbf{m}_j - \hat{\mathbf{m}})^\top + 2\mathbf{W}}{q_j + 1 + 2(\omega - (d+1)/2)} . \end{aligned}$$

These update equations have many free hyperparameters without well-motivated values. All of these could be used in principle to regularize the model parameters. For example, the γ_j could be used to keep the amplitudes α_j from becoming very small. When one does not want to set these hyperparameters by hand, their optimal values can all be obtained by the model selection techniques described below. However, in what follows we will focus on the regularization of the covariances. It is easy to see from the update equation for the variances $\mathbf{V}_j$ that the updated variances are bound from below, since the numerator is greater or equal to $2\mathbf{W}$ and the denominator has an upper limit. Therefore we can focus on the matrix $\mathbf{W}$ for the purpose of the regularization by setting the other parameters to the uninformative values:

$$\gamma_j = 1 \quad \forall j \,, \quad \omega = (d+1)/2 \,, \quad \eta = 0 \,. \tag{37}$$

We can further reduce the complexity to one free parameter by setting $\mathbf{W} = w/2\mathbf{I}$, where $\mathbf{I}$ is the d-dimensional unit matrix. The update equations in the

M step then reduce to

$$\begin{aligned}
\textbf{M step:}\quad \alpha_j &\leftarrow \frac{1}{N}\sum_i q_{ij}\\
\mathbf{m}_j &\leftarrow \frac{1}{q_j}\sum_i q_{ij}\,\mathbf{b}_{ij}\\
\mathbf{V}_j &\leftarrow \frac{1}{q_j+1}\left[\sum_i q_{ij}\left[(\mathbf{m}_j-\mathbf{b}_{ij})\,(\mathbf{m}_j-\mathbf{b}_{ij})^\top+\mathbf{B}_{ij}\right]+w\mathbf{I}\right]. \qquad (38)
\end{aligned}$$

The best value to use for the parameter w is not known a priori but can be determined, e.g., by jackknife leave-one-out cross validation.

4.2. *Split and merge algorithm for avoiding local maxima.* The split and merge algorithm starts from the basic EM algorithm, with or without the Bayesian regularization of the variances, and jumps into action after the EM algorithm has reached a maximum, which more often than not will only be a local maximum. At this point three of the Gaussians in the mixture are singled out, based on criteria detailed below, and two of these Gaussians are merged while the third Gaussian is split into two Gaussians (Ueda et al., 1998). Let us denote the indices of the three selected Gaussians as j_1, j_2, and j_3, where the former two are to be merged while j_3 will be split. An alternative, but similar, approach to local maxima avoidance is given by the birth and death moves in reversible jump MCMC (Richardson and Green, 1997) or variational (Ghahramani and Beal, 2000; Beal, 2003) approaches to mixture modeling. These moves do not conserve the number of mixture components and are therefore less suited for our fixed-K approach to mixture modeling.

The Gaussians corresponding to the indices j_1 and j_2 will be merged as follows: the model parameters of the merged Gaussian j_1' are

$$\begin{aligned}
\alpha_{j_1'} &= \alpha_{j_1}+\alpha_{j_2}\\
\theta_{j_1'} &= \frac{\theta_{j_1}q_{j_1}+\theta_{j_2}q_{j_2}}{q_{j_1}+q_{j_2}}\,, \qquad (39)
\end{aligned}$$

where θ_j stands for $\mathbf{m}_j$ and $\mathbf{V}_j$. Thus, the mean and the variance of the new Gaussian is a weighted average of the means and variances of the two merging Gaussians.

The Gaussian corresponding to j_3 is split as follows:

$$\begin{aligned}
\alpha_{j_2'} &= \alpha_{j_3'} = \alpha_{j_3}/2\\
\mathbf{V}_{j_2'} &= \mathbf{V}_{j_3'} = \det(\mathbf{V}_{j_3})^{1/d}\,\mathbf{I}\,. \qquad (40)
\end{aligned}$$

Thus, the Gaussian j_3 is split into equally contributing Gaussians with each new Gaussian having a covariance matrix that has the same volume as $\mathbf{V}_{j_3}$. The means $\mathbf{m}_{j_2'}$ and $\mathbf{m}_{j_3'}$ can be initialized by adding a random perturbation vector ϵ_{j_m} to $\mathbf{m}_{j_3}$, e.g.,

$$\mathbf{m}_{j_m'} = \mathbf{m}_{j_3} + \epsilon_{j_m}\,, \tag{41}$$

where $\|\epsilon_{j_m}\|^2 \ll \det(\mathbf{V}_{j_3})^{1/d}$ and $m = 1, 2$.

After this split and merge initialization the parameters of the three affected Gaussians need to be re-optimized in a model in which the parameters of the unaffected Gaussians are held fixed. This can be done by using the M step in equation (38) for the parameters of the three affected Gaussians, while keeping the parameters of the other Gaussians fixed, including the amplitudes, i.e., we need to use the update equation (31) for the amplitudes. This ensures that the sum of the amplitudes of the three affected Gaussians remains fixed. This procedure is called the *partial EM procedure*. After convergence this is then followed by the full EM algorithm on the resulting model parameters. Finally the resulting parameters are accepted if the total log likelihood of this model is greater than the log likelihood before the split and merge step. If the likelihood doesn't increase the same split and merge procedure is performed on the next triplet of split and merge candidates.

The question that remains to be answered is how to choose the 2 Gaussians that should be merged and the Gaussian that should be split. In general there are $K(K-1)(K-2)/2$ possible triplets like this which quickly reaches a large number when the number of Gaussians K gets larger. In order to rank these triplets one can define a *merge criterion* and a *split criterion*.

The merge criterion is constructed based on the observation that if many data points have equal posterior probabilities for two Gaussians, these Gaussians are good candidates to be merged. Therefore one can define the merge criterion:

$$J_{\text{merge}}(j, k|\theta) = \mathbf{P}_j(\theta)^\top \mathbf{P}_k(\theta)\,, \tag{42}$$

where $\mathbf{P}_j(\theta) = (q_{i1}, \ldots, q_{iN})^\top$ is the N-dimensional vector of posterior probabilities for the jth Gaussian. Pairs of Gaussians with larger J_{merge} are good candidates for a merger.

We can define a split criterion based on the Kullback-Leibler distance between the local data density around the lth Gaussian, which can be written in the case of complete data as $p_l(\mathbf{w}) = 1/q_l \sum_i q_{il}\delta(\mathbf{w} - \mathbf{w}_i)$, and the lth Gaussian density specified by the current model estimates $\mathbf{m}_l$ and $\mathbf{V}_l$. The Kullback-Leibler divergence between two distributions $p(x)$ and $q(x)$ is given by (MacKay, 2003):

$$D_{\mathrm{KL}}(P||Q) = \int \mathrm{d}x\, p(x) \ln \frac{p(x)}{q(x)} \,. \tag{43}$$

Since the local data density is only non-zero at a finite number of values, we can write this as

$$J_{\mathrm{split}}(l|\theta) = \frac{1}{q_l} \sum_i q_{il} \left[\ln \left(\frac{q_{il}}{q_l} \right) - \ln \mathcal{N}(\mathbf{w}_i|\mathbf{m}_l, \mathbf{V}_l) \right] \,. \tag{44}$$

The larger the distance between the local density and the Gaussian representing it, the larger J_{split} and the better candidate this Gaussian is to be split.

When dealing with incomplete data determining the local data density is more problematic. One possible way to estimate how well a particular Gaussian describes the local data density is to calculate the Kullback-Leibler divergence between the model Gaussian under consideration and each individual data point perpendicular to the unobserved directions for that data point. Thus, we can write

$$J_{\mathrm{split}}(l|\theta) = \frac{1}{q_l} \sum_i q_{il} \left[\ln \left(\frac{q_{il}}{q_l} \right) - \ln \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_l, \mathbf{R}_i\mathbf{V}_l\mathbf{R}_i^{\top}) \right] \,. \tag{45}$$

Candidates for merging and splitting are then ranked as follows: first the merge criterion $J_{\mathrm{merge}}(j,k|\theta)$ is calculated for all pairs j,k and the pairs are ranked by decreasing $J_{\mathrm{merge}}(j,k|\theta)$. For each pair in this ranking the remaining Gaussians are then ranked by decreasing $J_{\mathrm{split}}(l|\theta)$.

4.3. *Setting the remaining free parameters.* No real world application of Gaussian mixture density estimation is complete without a well-specified methodology for setting the number of Gaussian components K and any hyperparameters introduced in the Bayesian regularization described above, the covariance regularization parameter w in our basic scheme. This covariance regularization parameter basically sets the square of the smallest scale of the distribution function on which we can reliable infer small scale features. Therefore, this scale could be set by hand to the smallest scale we believe we have access to based on the properties of the data set.

In order to get the best results the parameters K and w should be set by some objective procedure. As mentioned above, leave-one-out cross validation (Stone, 1974) could be used to set the regularization parameter w, and the number of Gaussians could be set by this procedure as well. The basic idea of leave-one-out cross validation is that one creates N new data sets by sequentially taking one data point out of the original sample. For each of these N new samples we optimize the scalar objective function and then record the logarithm of the likelihood of the data point that was left out under this new optimized parameter set. Summing up all of these cross validation log likelihoods then gives a scalar which can be compared for different models and the best model is the one for which the total cross-validation log likelihood is the largest. This procedure, while simple, can be quite computationally intensive and is therefore often not feasible. The procedure can be simplified by (1) leaving out more than one data point at a time, i.e., creating N' new samples by leaving out $1/N'^{\text{th}}$ of the full sample, optimizing the scalar objective using the new sample and recording the total cross validation likelihood of the data points left out at that step; (2) restricting the optimization, e.g., by starting from the optimized parameters for the full sample and only allowing certain parameters to vary in the individual optimization of the N' new samples. For example, one can choose to only allow the amplitudes of the optimized parameters found for the full sample to change during the cross validation optmizations.

Other techniques include methods based on Bayesian model selection (Roberts et al., 1998) as well as approaches based on minimum coding inference (Wallace and Boulton, 1968; Oliver et al., 1996; Rissanen, 1978; Schwartz, 1978), although these methods have difficulty dealing with significant overlap between components (such as the overlap we see in the example in Figure 2), but there are methods to deal with these situations (Baxter, 1995). Alternatively, when a separate, external data set is available, we can use this as a test data set to validate the obtained distribution function. All of these methods are explored in an accompanying paper on the velocity distribution of stars in the Solar neighborhood from measurements from the *Hipparcos* satellite (see below; Bovy et al., 2009).

A rather different approach to the model selection problem is to avoid it altogether. That is, by introducing priors over the hyperparameters and including them as part of the model it is often possible to infer, or fully marginalize over, them simultaneously with the parameters of the components of the mixture. Such approaches include reversible jump MCMC methods (Richardson and Green, 1997), mixtures consisting of an infinite number of components based on the Dirichlet process (Rasmussen, 2000), or approximate, variational algorithms (Ghahramani and Beal, 2000; Beal, 2003). Extending these approaches to deal with noisy, heterogeneous, and incomplete data is beyond the scope of this paper, but it is clear that this extension is, in principle, straightforward: the MCMC methods mentioned above can include the true values of the observations $\mathbf{v}_i$ and these can be

Gibbs sampled given the current model and the observed values $\mathbf{w}_i$ in an MCMC sweep from the Gaussian with mean given in equation (20) and variance given in equation (21).

5. Overview of the full algorithm.

To summarize the full algorithm we briefly list all the steps involved:

1. Run the EM algorithm as specified in equations (25) and (38). Store the resulting model parameters θ^* and the corresponding model log likelihood ϕ^*.
2. Compute the merge criterion $J_{\text{merge}}(j,k|\theta^*)$ for all pairs j,k and the split criterion $J_{\text{split}}(l|\theta^*)$ for all l. Sort the split and merge candidates based on these criteria as detailed above.
3. For the first triplet (j,k,l) in this sorted list set the initial parameters of the merged Gaussian using equation (39) and the parameters of the two Gaussian resulting from splitting the third Gaussian using equations (40)-(41). Then run the partial EM procedure on the parameters of the three affected Gaussians, i.e., run EM while keeping the parameters of the unaffected Gaussians fixed, and follow this up by running the full EM procedure on all the Gaussians. If after convergence the new log likelihood ϕ is greater than ϕ^*, accept the new parameter values $\theta^* \leftarrow \theta$ and return to step two. If $\phi < \phi^*$ return to the beginning of this step and use the next triplet (j,k,l) in the list.
4. Halt this procedure when none of the split and merge candidates improve the log likelihood or, if this list is too long, if none of the first C lead to an improvement.

Deciding when to stop going down the split-and-merge hierarchy will be dictated in any individual application of this technique by computational constraints. This is an essential feature of any search-based approach to finding global maxima of (likelihood) functions.

6. Applications to real data.

6.1. *General considerations.*

Inferring the distribution function of an observable given only a finite, noisy set of measurements of that distribution function and the related problem of finding clusters and/or overdensities in the distribution is a problem of significant general interest in many areas of science and of astronomy in particular (e.g., McLachlan and Basford, 1988; Rabiner and Biing-Hwang, 1993; Dehnen, 1998; Helmi et al., 1999; Skuljan et al., 1999; Hogg et al., 2005). The description you are interested in as a scientist is *not* the observed distribution, what you really want is the description of the distribution that you would have if you had good data, i.e., data with vanishingly small uncertainties and with all of the dimensions measured. In the low signal-to-noise regime, e.g., large data sets at the bleeding edge in astronomy, the data never have these two properties such that the underlying, true distribution cannot be found without taking the noise properties of the data into account. If you want know the underlying distribution, in order to compare your model with the data, you need to convolve the model with the data uncertainties, not deconvolve the data. When the given set of data has heterogeneous noise properties, that is, when the uncertatinty convolution is different for each data point, each data point is a sample of a different distribution, i.e., the distribution obtained from convolving the true, underlying distribution with the noise of that particular observation. Incomplete data poses a similar problem when the part of the data that is missing is different for different data points. None of the current density estimation techniques confront all of these issues (e.g., McLachlan and Basford, 1988; Silverman, 1986), although techniques that properly account for incomplete data have been developed (Ghahramani and Jordan, 1994a,b).

A first approximation to the problem of modeling the distribution of a set of real-valued data points that is often used is to fit a single (multivariate) normal distribution to the distribution of the data points, e.g., when applying principal component analysis. In general this is hardly ever a good approximation to the full distribution, e.g., when a distribution shows two distinct maxima or significant asymmetry this approximation is poorly suited to the problem at hand. However, fitting a distribution that is the sum of a large enough number of normal distributions provides a good approximation to any reasonably well-behaved underlying distribution. For example, a multi-modal distribution will be described by a sum of normal distributions with varying means, while a single-peak distribution with non-Gaussian tails could be fit by a set of Gaussians with similar means but varying amplitudes and covariance matrices. Therefore, the technique described in the previous sections is perfectly suited for applications with real data sets.

6.2. *A simple application: fitting a line to data with non-trivial uncertainties.* One of the simplest applications of the algorithm described above consists of fitting a linear relationship to a set of data points with, possibly correlated, uncertainties in both the independent and the dependent variable. Many different methods have been devised to fit a straight line to a data set, most of them least-squares procedures (e.g., Isobe et al., 1990; Feigelson and Babu, 1992). When measurement uncertainties are present in both variables a "double weighted" regression can be performed which minimizes errors both in the dependent variables as well as in the independent variable (York, 1966; McIntyre et al., 1966). This procedure can be generalized to the case of correlated uncertainties (York, 1969). In the case of uncorrelated uncertainties, one finds the parameters of the straight line $y = a\,x + b$ which minimize the quantity

$$S = \sum_i \left[(x_i - \hat{x}_i)^2/\sigma_x^2 + (y_i - \hat{y}_i)^2/\sigma_y^2 \right] , \tag{46}$$

where the $(\hat{x}_i, \hat{y}_i)$ are the observed (x_i, y_i) adjusted along a line of slope σ_y^2/σ_x^2 to lie on the straight line. This in fact assumes that the uncertainties are highly correlated, even though by assumption they are not, and does not allow for the full exploration of each pairs' covariance ellipse. In addition to this, it is difficult to implement numerically. A similar but better approach to this problem is given by the total least squares modeling technique (e.g., Golub and Van Loan, 1996), where the objective is to minimize the squared orthogonal distance between the data points and the fitted straight line.

As mentioned in the introduction, the density estimation technique developed in this paper can be applied to this problem as well. Fitting a single Gaussian distribution to the data will pick out the direction of the largest variance, which one can identify with the direction of the linear relation between two variables. Since the technique described above takes care of the heterogeneous uncertainty properties of the data, this technique can be applied straightforwardly to this problem. This is the simplest application of the density estimation technique as it only concerns a single Gaussian component in the mixture, and none of the extensions to the algorithm have to be applied to this problem.

As any method that deals with uncertainties in the measurements of the "independent" variable, this method, by necessity, fits the joint density $p(x, y)$ as opposed to the conditional density $p(y|x)$. The method we proposed in the previous paragraph does this by assuming that the marginal density $p(x)$ is Gaussian. This need not be the case and the method could fail if the distribution of the independent variable is far from Gaussian. This problematic behavior could be resolved by fitting for the marginal distribution $p(x)$ simultaneously with obtaining the best linear fit. This density estimation problem could again be handled by the mixture of Gaussian components approach developed above (this approach is then basically identical to the approach proposed in Kelly, 2007).

In detail, in the Gaussian $p(x)$ approximation, one fits a single Gaussian distribution to the observed data, with no restrictions on the mean or variance of that Gaussian. After convergence, one identifies the direction corresponding to the largest eigenvalue as the direction of the linear relation. The line in this direction going through the mean of the best-fit Gaussian is then the desired linear fit to the data.

This technique is similar to the procedures for fitting a line to data described at the beginning of the section in that it minimizes a quantity similar to the quantity S in equation (46), but it uses the full uncertainty covariance matrices for each datapoint and has a simple numerical implementation given by the EM procedure described in this paper. The objective function given in equation (10) is, as shown in section 2, justifiable under the assumption of Gaussian distributed errors and a underlying Gaussian distribution. The fact that the model space is more general than the the model of a straight line that we want to fit to the data—the straight line model is in fact part of the model space as an infinitely long, infinitely narrow Gaussian distribution—can hardly be thought of as a disadvantage, as it can show that the assumption of a linear relation is suspect through the aspect ratio of the best fit Gaussian.

This procedure was used in Hogg et al. (2005) to fit a line to a set of points with correlated uncertainties in the abscissa and ordinate (Fig 2 in that paper). Here we fit the Tully–Fisher relation in several bands from Hubble Space Telescope (*HST*) data as an example of this procedure. The Tully–Fisher relation is a power-law type relation between the luminosity and the velocity width of spiral galaxies. This relation can be used to measure distances in the Universe because it relates a velocity, which can be measured relatively easily by Doppler measurements, to the intrinsic luminosity, which is hard to observe, but depends on the easily observed apparent brightness of the galaxy by the distance squared (Tully and Fisher, 1977). The relation between luminosity L and velocity width W is of the form

$$L \propto W^{\gamma}\,, \tag{47}$$

where the exponent γ depends on the wavelength at which the luminosity is measured but generally lies somewhere between three and four.

Calibrating the Tully–Fisher relation by measuring accurate Cepheid distances to nearby galaxies was one of the key projects of the Hubble Space Telescope and as an illustration we fit the Tully–Fisher relation in five different bandpasses here from the *HST* data (Table 2 of Sakai et al., 2000). Uncertainties in both the velocity widths (measured at 20 percent of the peak HI flux) of all the galaxies as well as in the absolute magnitudes are large such that our procedure naturally applies (the fits in Sakai et al. 2000 are bivariate fits minimizing errors in both variables). The resulting linear relation is shown in Fig 1 for the different bandpasses. Using a leave-one-out jackknife procedure to establish the uncertainties on the slopes and intercepts, we find the following relations

$$B^c_T = -(8.04 \pm 0.63)\,(\log_{10} W^c_{20} - 2.5) - (19.77 \pm 0.11) \tag{48}$$

$$V^c_T = -(8.86 \pm 0.57)\,(\log_{10} W^c_{20} - 2.5) - (20.33 \pm 0.08) \tag{49}$$

$$R^c_T = -(8.79 \pm 0.50)\,(\log_{10} W^c_{20} - 2.5) - (20.64 \pm 0.07) \tag{50}$$

$$I^c_T = -(9.26 \pm 0.57)\,(\log_{10} W^c_{20} - 2.5) - (21.12 \pm 0.07) \tag{51}$$

$$H^c_{-0.5} = -(11.03 \pm 0.75)\,(\log_{10} W^c_{20} - 2.5) - (21.73 \pm 0.07)\,. \tag{52}$$

These relations agree within the uncertainties with the relations found in Sakai et al. (2000).

6.3. *The velocity distribution from* Hipparcos *data.* We have applied the technique developed in this paper to the problem of inferring the velocity distribution of stars in the Solar neighborhood from transverse angular data from the *Hipparcos* satellite and we present in this section some results of this study to demonstrate the performance of the algorithm on a real data set. A more detailed and complete account of this study is presented elsewhere (Bovy et al., 2009).

The astrometric ESA space mission *Hipparcos*, which collected data over a 3.2 year period around 1990, provided for the first time an all-sky catalogue of absolute parallaxes and proper motions, with typical uncertainties in these quantities on the order of mas (ESA, 1997). From this catalogue of $\sim 100,000$ stars kinematically unbiased samples of stars with accurate positions and velocities can be extracted (Dehnen and Binney, 1998). Since astrometric measurements of velocities basically just compare the position of a star at different times to calculate velocities, the only components of a star's velocity that can be measured astrometrically are the tangential components. The line-of-sight velocities of the stars in the *Hipparcos* sample were therefore not obtained during the *Hipparcos* mission. Since the proper motions that are measured by differencing sky positions are angular velocities, the distance to a star is necessary in order to convert the proper motions to space velocities.

Distances in astronomy are notoriously hard to measure precisely, and at the accuracy level of the *Hipparcos* mission distances can only be reliably obtained for stars near the Sun (out to ~ 100 pc). In addition to this, since distances are measured as inverse distances (parallaxes) only distances that are measured relatively precise will have approximately Gaussian uncertainties associated with them. Balancing the size of the sample with the accuracy of the distance measurement leaves us with distance uncertainties that are typically $\sim$10-percent, such that the tangential velocities that are obtained from the proper motions and the distances have low signal-to-noise.

Of course, if we want to describe the distribution of the velocities of the stars in this sample, we need to express the velocities in a common reference frame, which for kinematical studies of stars around the Sun is generally chosen to be the Galactic coordinate system, in which the x-axis points towards the Galactic center, the y-axis points in the direction of Galactic rotation, and the z-axis points towards the North Galactic Pole (Blaauw et al., 1960; Binney and Merrifield, 1998). The measured tangential velocities are then projections of the three-dimensional velocity of a star with respect to the Sun in the two-dimensional plane perpendicular to the line-of-sight to the star. Therefore, this projection is different for each individual star.

The discussion in the previous paragraphs shows that this sample of stars exhibits all of the properties (incomplete data and noisy measurements, different from observation to observation) for which a procedure such as the one developed in this paper is necessary.

We have studied the velocity distribution of a sample of main sequence stars selected to have accurate distance measurements (parallax uncertainties $\sigma_\pi/\pi < 0.1$) and to be kinematically unbiased (in that the sample of stars faithfully represents the kinematics of similar stars). In detail, we use the sample of $\sim 10,000$ stars from Dehnen and Binney (1998), but we use the new reduction of the *Hipparcos* raw data, which has improved the accuracy of the astrometric quantities (van Leeuwen, 2007a,b). A particular reconstruction of the underlying velocity distribution of the stars is shown in Figure 2, in which 10 Gaussians are used and the regularization parameter w is set to 4 km^2 s^{-2}. We choose this value for w since we believe that the smallest scale of the features we can see is a few km s^{-1}, because of the scale of the differences in rotational velocities around the Galactic center of the stars in the $\sim$100 pc around the Sun (which is rotating around the Galactic center at ~ 200 km s^{-1} at ~ 8 kpc). What is shown are two-dimensional projections of the three-dimensional distribution.

The recovered distribution compares favorably with other reconstructions of the velocity distribution of stars in the Solar neighborhood, based on the same sample of stars (using a maximum penalized likelihood density estimation technique, Dehnen, 1998), as well as with those based on other samples of stars for which three-dimensional velocities are available (Skuljan et al., 1999; Nordström et al., 2004; Famaey et al., 2005; Antoja et al., 2008). All of the features in the recovered distribution function are real and correspond to known features; this includes the feature at $v_y \approx -100$ km s^{-1}, which is known as the Arcturus moving group. Therefore, we conclude that the method developed in this paper performs very well on this complicated data set.

The convergence of the algorithm is shown in Figure 3. Only split-and-merge steps that improved the likelihood are shown in this plot, therefore, the actual number of iterations is much higher than the number given on the x-axis. It is clear that all of the split-and-merge steps only slightly improve the initial estimate from the first EM procedure, but since what is shown is the likelihood per data point, the improvement of the total likelihood is more significant.

7. Implementation and code availability. The algorithm presented in this paper was implemented in the C programming language, depending only on the standard C library and the GNU Scientific Library[3]. The code is available at http://code.google.com/p/extreme-deconvolution/; Instructions for its installation and use are given there. The code can be compiled into a shared object library, which can then be incorporated into other projects or accessed through an IDL[4] wrapper function supplied with the C code.

The code can do everything described above. The convergence of the algorithm is slow, which is mostly due to the large number of split-and-merge steps that can be taken by the algorithm (the split-and-merge aspect of the algorithm, however, can easily be turned off or restricted by setting the parameter specifying the number of steps to go down the split-and-merge hierarchy).

8. Conclusions and Future Work. We have generalized the mixture of Gaussians approach to density estimation such that it can be applied to noisy, heterogeneous, and incomplete data. The objective function is obtained by integrating over the unknown true values of the quantities for which we only have noisy and/or incomplete observations. In order to optimize the objective function resulting from this marginalization we have derived an EM algorithm that monotonically increases the model likelihood; this EM algorithm, in which the E step involves finding the expected value of the first and second moments of the true values of the observables given the current model and the noisy observations, reduces to the basic EM algorithm for Gaussian mixture modeling in the limit of noiseless data. We have shown that the model can incorporate Bayesian priors on all of the model parameters without losing any of its analytical attractiveness and that the algorithm can accomodate the split-and-merge algorithm to deal

[3] http://www.gnu.org/software/gsl/

[4] http://www.ittvis.com/ProductServices/IDL.aspx

with the presence of local maxima, which this EM algorithm, as many other EM algorithms, suffers from.

The work presented here can be extended to be incorporated in various more non-parametric approaches to density modeling, e.g., in mixture models with an infinite number of components based on the Dirichlet Process (e.g., Rasmussen, 2000). In this way current advances in non-parametric modeling can be applied to the low signal-to-noise sciences where the situation of complete and noise-free data is more often than not an untenable and unattainable approximation.

APPENDIX A: LOGSUM

The calculation of the posterior likelihoods q_{ij} in the E step (eq. [25]) can be tricky since it can be the ratio of a small probability over the sum of small probabilities, which can lead to underflow and a significantly different result for the q_{ij}s. For instance, imagine a situation in which only one of the terms is slightly larger than the smallest positive number allowed by the compiler and all the other terms are slightly smaller than this number. The result of this would be that the q_{ij} corresponding to the Gaussian with the largest probability will be set to 1 and all the other q_{ij}s will be set to zero, while in reality the distribution is much more uniform than this.

We will deal with this problem in two steps: (a) we will calculate not the numerator of q_{ij} but its logarithm, which will allow us to keep very small probabilities (since in general the smallest negative number is much smaller than the logarithm of the smallest positive number set by the compiler); (b) we will design a function that given the logarithms of all the q_{ij}s returns the logarithm of their sum. Then we can normalize the numerators by subtracting the logarithm of their sum from their logarithms. This function performs a kind of 'logsum'. Cleverly designing this function will allow us to obtain the sum of a set of numbers, each of which is smaller than the smallest positive number. One approach to this problem that uses the same basic identity given in equation (53) below is the log-sum-exp formula of Press et al. (2007). This formula ensures that at least the largest q_{ij} will not underflow; however, it does not make use of the full dynamic range available and can therefore lead to unnecessary underflow of smaller terms. The method we develop below makes use of the full available dynamic range and leads to the smallest number of ignored terms; it is basically impossible to design a general method that performs better than the method described below.

We will denote the numerators of the q_{ij} by q'_{ij} in what follows. The simplest way of summing a set of numbers given as logarithms would be to exponentiate each number and add it to the sum. However, $\exp(\ln q'_{ij})$ might be below the smallest positive floating point threshold set by the compiler for each term. To avoid this underflow problem, we want to add a constant c to each of the logarithms such that when exponentiated they are all above the threshold, after which we can sum them and take the logarithm of the sum. Finally we can subtract this constant c again from the final answer. In short, we use the identity

$$\ln \sum_j \exp(\ln q'_{ij}) = \left[\ln \sum_j \exp(\ln q'_{ij} + c)\right] - c\,. \tag{53}$$

To find what this constant c should be we will define DBL_MIN to be the smallest positive floating point number available, and DBL_MAX to be the largest floating point number. Avoiding underflow is then achieved when each term is larger than DBL_MIN, or, when

$$\exp\left[\min_j(\ln q'_{ij}) + c\right] > \text{DBL\_MIN}\,, \tag{54}$$

which means

$$c > \ln(\text{DBL\_MIN}) - \min_j(\ln q'_{ij})\,. \tag{55}$$

However, this can be quite large and will therefore sometimes lead to *overflow* for the larger terms in the sum, which is arguably a greater problem than the original underflow. Overflow will be avoided when the sum is smaller than DBL_MAX, which we can ensure by demanding

$$\sum_j \exp(\ln q'_{ij} + c) < K \exp\left[\max_j(q'_{ij}) + c\right] < \text{DBL\_MAX}\,, \tag{56}$$

where K is the number of terms in the sum. This means that we should have

$$c < \ln(\text{DBL\_MAX}) - \max_j(\ln q'_{ij}) - \ln K\,. \tag{57}$$

The second bound obtained here is the most stringent because overflow is worse than underflow. This is simply because the underflow will only ignore a very small term, whereas overflow will lead to an infinite answer, which will affect any following calculation. Moreover, as we are adding probabilities, i.e., positive numbers, overflow of the kind described in the previous paragraph implies that the underflow is probably irrelevant (and if it is not, the situation is completely hopeless). Therefore, we should choose c as

$$c = \min(\ln(\text{DBL\_MIN}) - \min_j(\ln q'_{ij}), \ln(\text{DBL\_MAX}) - \max_j(\ln q'_{ij}) - \ln N)\,. \tag{58}$$

APPENDIX B: MARGINALIZATION AND CONDITIONING OF MULTIVARIATE GAUSSIAN DISTRIBUTIONS

Suppose we are given a Gaussian distribution for a vector $\mathbf{v}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{V}$ which can be partitioned into a d_1-dimensional and a d_2-dimensional component $(d_1 + d_2 = d)$ $(\mathbf{v}_1, \mathbf{v}_2)^\top$, $(\mathbf{m}_1, \mathbf{m}_2)^\top$, and

$$\begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} ,$$

respectively, where $\mathbf{V}_{12}^\top = \mathbf{V}_{21}$. We can now ask what the distribution of $\mathbf{v}_2$ marginalizing over $\mathbf{v}_1$ is, and what the distribution of $\mathbf{v}_1$ given $\mathbf{v}_2$ is? To answer this question we will rewrite the joint distribution function in such a way as to make clear the following identity

$$p(\mathbf{v}) = p(\mathbf{v}_1, \mathbf{v}_2) = p(\mathbf{v}_1|\mathbf{v}_2)p(\mathbf{v}_2) \ , \tag{59}$$

where $p(\mathbf{v}) \equiv \mathcal{N}(\mathbf{v}|\mathbf{m}, \mathbf{V})$.

First we note that we can block-diagonalize the covariance matrix as follows:

$$\begin{aligned} &\begin{bmatrix} \mathbf{I}_{d_1} & -\mathbf{V}_{12}\mathbf{V}_{22}^{-1} \\ 0 & \mathbf{I}_{d_2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{d_1} & 0 \\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & \mathbf{I}_{d_2} \end{bmatrix} \\ &\qquad = \begin{bmatrix} \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & 0 \\ 0 & \mathbf{V}_{22} \end{bmatrix} . \end{aligned} \tag{60}$$

Using this we can write the argument of the exponential in the Gaussian distribution as follows

$$\begin{aligned} &-\frac{1}{2}(\mathbf{v}-\mathbf{m})^\top \mathbf{V}^{-1}(\mathbf{v}-\mathbf{m}) = \\ &\qquad -\frac{1}{2} \begin{bmatrix} \mathbf{v}_1 - \mathbf{m}_1 \\ \mathbf{v}_2 - \mathbf{m}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{v}_1 - \mathbf{m}_1 \\ \mathbf{v}_2 - \mathbf{m}_2 \end{bmatrix} , \end{aligned}$$

which becomes using the identity in equation (60)

$$\begin{aligned} -\frac{1}{2} \quad &\begin{bmatrix} \mathbf{v}_1 - \mathbf{m}_1 \\ \mathbf{v}_2 - \mathbf{m}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_{d_1} & 0 \\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & \mathbf{I}_{d_2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & 0 \\ 0 & \mathbf{V}_{22} \end{bmatrix}^{-1} \\ &\times \begin{bmatrix} \mathbf{I}_{d_1} & -\mathbf{V}_{12}\mathbf{V}_{22} \\ 0 & \mathbf{I}_{d_2} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 - \mathbf{m}_1 \\ \mathbf{v}_2 - \mathbf{m}_2 \end{bmatrix} , \end{aligned} \tag{61}$$

which can be simplified to give the following

$$\begin{aligned} -\frac{1}{2} & \left[\begin{array}{c} \mathbf{v}_1 - \mathbf{m}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2) \\ \mathbf{v}_2 - \mathbf{m}_2 \end{array}\right]^\top \left[\begin{array}{cc} \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & 0 \\ 0 & \mathbf{V}_{22} \end{array}\right]^{-1} \\ & \times \left[\begin{array}{c} \mathbf{v}_1 - \mathbf{m}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2) \\ \mathbf{v}_2 - \mathbf{m}_2 \end{array}\right] . \end{aligned} \tag{62}$$

Introducing $\mathbf{m}_{1|2} \equiv \mathbf{m}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2)$ and $\mathbf{V}_{1|2} \equiv \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$ this becomes

$$\begin{aligned} & -\frac{1}{2}\left[\begin{array}{c} \mathbf{v}_1 - \mathbf{m}_{1|2} \\ \mathbf{v}_2 - \mathbf{m}_2 \end{array}\right]^\top \left[\begin{array}{cc} \mathbf{V}_{1|2} & 0 \\ 0 & \mathbf{V}_{22} \end{array}\right]^{-1} \left[\begin{array}{c} \mathbf{v}_1 - \mathbf{m}_{1|2} \\ \mathbf{v}_2 - \mathbf{m}_2 \end{array}\right] = \\ & -\frac{1}{2}(\mathbf{v}_1 - \mathbf{m}_{1|2})^\top \mathbf{V}_{1|2}^{-1}(\mathbf{v}_1 - \mathbf{m}_{1|2}) - \frac{1}{2}(\mathbf{v}_2 - \mathbf{m}_2)^\top \mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2) . \end{aligned} \tag{63}$$

Using the identity in equation (60) it can also be shown that

$$\det \mathbf{V} = \det \mathbf{V}_{1|2} \det \mathbf{V}_{22} , \tag{64}$$

since the determinants of the left and right multiplying matrices on the left hand side of equation (60) are equal to one (for block triangular matrices the determinant is equal to the product of the determinants of the diagonal blocks, these are both unit matrices). Therefore we can write

$$\begin{aligned} \mathcal{N}(\mathbf{v}|\mathbf{m},\mathbf{V}) &= (2\pi)^{-d/2}\det(\mathbf{V})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{v}-\mathbf{m})^\top\mathbf{V}^{-1}(\mathbf{v}-\mathbf{m})\right) \\ &= (2\pi)^{-(d_1+d_2)/2}\det(\mathbf{V}_{1|2})^{-1/2}\det(\mathbf{V}_{22})^{-1/2} \\ &\qquad \times\exp\Bigg(-\frac{1}{2}(\mathbf{v}_1 - \mathbf{m}_{1|2})^\top \mathbf{V}_{1|2}^{-1}(\mathbf{v}_1 - \mathbf{m}_{1|2}) \\ &\qquad\qquad -\frac{1}{2}(\mathbf{v}_2 - \mathbf{m}_2)^\top \mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2)\Bigg) \\ &= \mathcal{N}(\mathbf{v}_1|\mathbf{m}_{1|2},\mathbf{V}_{1|2})\,\mathcal{N}(\mathbf{v}_2|\mathbf{m}_2,\mathbf{V}_{22}) . \end{aligned} \tag{65}$$

We can now show that this factorization corresponds to the factorization given in equation (59) by marginalizing over $\mathbf{v}_1$, i.e.,

$$\begin{aligned} p(\mathbf{v}_2) &= \int \mathrm{d}\mathbf{v}_1\,\mathcal{N}(\mathbf{v}|\mathbf{m},\mathbf{V}) \\ &= \int \mathrm{d}\mathbf{v}_1\,\mathcal{N}(\mathbf{v}_1|\mathbf{m}_{1|2},\mathbf{V}_{1|2})\mathcal{N}(\mathbf{v}_2|\mathbf{m}_2,\mathbf{V}_{22}) \\ &= \mathcal{N}(\mathbf{v}_2|\mathbf{m}_2,\mathbf{V}_{22}) . \end{aligned} \tag{66}$$

Therefore, we can identify $\mathcal{N}(\mathbf{v}_1|\mathbf{m}_{1|2}, \mathbf{V}_{1|2})$ with $p(\mathbf{v}_1|\mathbf{v}_2)$.

To summarize we can write

$$p(\mathbf{v}_1|\mathbf{v}_2) = \mathcal{N}(\mathbf{v}_1|\mathbf{m}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{v}_2 - \mathbf{m}_2), \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}) \quad \text{(conditioning)} \tag{67}$$

$$p(\mathbf{v}_2) = \mathcal{N}(\mathbf{v}_2|\mathbf{m}_2, \mathbf{V}_{22}) \quad \text{(marginalization).} \tag{68}$$

APPENDIX C: THE EXPECTATION-MAXIMIZATION ALGORITHM AND EXTENSIONS

C.1. Generalities. The EM algorithm is a very general algorithm designed to deal with missing data in maximum likelihood estimates and was first written down in full generality by Dempster et al. (1977), although versions of it had already been around longer (e.g., Wolfe, 1970; Day, 1969). The name Expectation-Maximization captures its main ingredients: the EM algorithm seeks to iteratively maximize the full-data likelihood (i.e., given data + missing data) but since some of the data is missing, it takes the expectation of this full-data likelihood given the data and the previous model estimate; the maximization step then maximizes this expectation of the likelihood with respect to the model parameters. The EM algorithm properly should not be called an "algorithm" since it does not specify the actual sequence of steps actually required to carry out a single E- or M-step, however, this generality has attributed to much of its success and it is now applied to a large variety of problems and numerous extensions and different interpretations of the original algorithm have been proposed (McLachlan and Krishnan, 1997).

The original EM paper established its general properties and applied the algorithm to some specific examples, among which the case of finite mixtures. Among the EM algorithm's properties are that it converges monotonically to a local maximum of the likelihood function (Dempster et al., 1977; Wu, 1983), a property that is shared by the so-called generalized EM algorithms (GEM) in which the M-step merely increases the likelihood instead of maximizing it. The convergence properties of the EM algorithm are generally very good, i.e., given an initial estimate far away from a local maximum it will converge rather quickly to a region of parameter space close to a local maximum and eventually reach the local maximum, as opposed to simple, generic optimizers such as the Newton-Raphson method which need a good initial estimate to find a local maximum and which does not converge monotonically. However, the convergence of the EM algorithm can be very slow, i.e., it is approximately linear, often many orders of magnitude slower than quadratic methods. Numerous methods have been proposed to accelerate the convergence, e.g., using the Jacobian to find improved estimates of the model parameters (Louis, 1982); employing a generalized conjugate gradient approach (Jamshidian and Jennrich, 1993); considering a hybrid approach that switches to a quadratically converging optimization method close to the local maximum (e.g., Redner and Walker, 1984; Jones and McLachlan, 1992; Atkinson, 1992; Aitkin and Aitkin, 1996) (these hybrid approaches make use of the fact that most of the overall change in likelihood is achieved during the first few iterations (Redner and Walker, 1984)); or by extending the EM algorithm through the use of conditional maximization in the M-step (Meng and Rubin, 1993; Liu and Rubin, 1994). These methods are generally less stable and much more complicated than the standard EM algorithm (Lange, 1995). EM is often easy to implement, especially when the E- and M- step are given by closed expressions, but a significant drawback is that it does not give an estimate of the covariance matrix of the MLEs. However, the EM algorithm can be extended to give such estimates, either by considering the information matrix (Louis, 1982; Meng and Rubin, 1989, 1991) or by a bootstrap approach (Efron, 1979; Efron and Tibshirai, 1993).

In a nutshell we can describe the EM algorithm as follows: Suppose we have a likelihood function Φ for a set of parameters θ, dependent on the data $\mathbf{x}$ and the missing data $\mathbf{z}$. The likelihood Φ can then be maximized iteratively in two steps (Dempster et al., 1977): In the $k+1-$th E-step we calculate the expectation of the likelihood given the data and the current parameter estimate, that is, the function

$$Q(\theta|\theta_k) = \langle \Phi|\mathbf{x}, \theta_k \rangle \ . \tag{69}$$

This function is then maximized in the M-step, leading to the new parameter estimate

$$\theta_{k+1} = \mathrm{argmax}_\theta \, Q(\theta|\theta_k) \ . \tag{70}$$

For many problems that on the surface are not missing data problems it can nevertheless be useful to phrase them in such a way that they can be handled by the EM-algorithm, since often the maximization in the M-step is much simpler than the original maximization. For example, below we will show that by formulating the Gaussian mixture density estimation problem as a missing data problem, a completely analytical solution for the E- and M-steps can be found. Even when an analytical solution does not exist for the M-step, that maximization can (1) be lower dimensional than the original maximization; (2) be handled by a GEM algorithm (Dempster et al., 1977) in which a single step of a regular optimizer can be sufficient for overall convergence if it increases the likelihood; or (3) be reduced to several conditional M-steps for which analytical solutions exist or that are of a lower dimensionality still (Meng and Rubin, 1993; Liu and Rubin, 1994).

C.2. The EM algorithm for Gaussian mixtures with complete data. In the case of complete, precise observations (i.e., $\mathbf{S}_i = 0$, $\mathbf{R}_i = \mathbf{I}$, $\forall\ i$) the likelihood is given in equation (11) and one introduces the indicator variables q_{ij} as in equation (12) to obtain the "full-data" likelihood of equation (13). One can use Jensen's inequality to show that optimizing this full-data log likelihood also optimizes the original log likelihood equation (11). Jensen's inequality for a concave function f, numbers x_j, and weights q_j can be stated as

$$f\left(\frac{\sum_j q_j x_{ij}}{\sum_j q_j}\right) \geq \frac{\sum_j q_j f(x_{ij})}{\sum_j q_j} \ . \tag{71}$$

The logarithm is a concave function and, defining $p_{ij} \equiv \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{m}_j, \mathbf{V}_j)$, if we choose $x_{ij} = p_{ij}/q_{ij}$ and weights q_{ij}, the indicator variables defined above, we find that for each data point i (since $\sum_j q_{ij} = 1$)

$$\begin{aligned} \ln \sum_{j=1}^{K} p_{ij} &= \ln \sum_{j=1}^{K} q_{ij} \frac{p_{ij}}{q_{ij}} \\ &\geq \sum_{j=1}^{K} q_{ij} \ln \frac{p_{ij}}{q_{ij}} \equiv F(q, \theta, \mathbf{w}_i) \\ &\geq \sum_{j=1}^{K} q_{ij} \ln p_{ij} - \sum_{j=1}^{K} q_{ij} \ln q_{ij} \ , \end{aligned} \tag{72}$$

with equality when we set

$$q_{ij} = \frac{p_{ij}}{\sum_j p_{ij}} \ , \tag{73}$$

since then

$$\begin{aligned} \sum_{j=1}^{K} q_{ij} \ln \frac{p_{ij}}{q_{ij}} &= \sum_{j=1}^{K} \frac{p_{ij}}{\sum_k p_{ik}} \ln \left(p_{ij} \frac{\sum_k p_{ik}}{p_{ij}} \right) \\ &= \sum_{j=1}^{K} \frac{p_{ij}}{\sum_k p_{ik}} \ln \sum_k p_{ik} \\ &= \left(\ln \sum_k p_{ik} \right) \sum_{j=1}^{K} \frac{p_{ij}}{\sum_k p_{ik}} \\ &= \ln \sum_j p_{ij} \ . \end{aligned} \tag{74}$$

This proves that the EM algorithm applied to the full-data log likelihood in equation (13) leads to a maximum likelihood estimate of the model parameters for the original likelihood, since the calculation of the E-step, i.e., taking the expectation of the full-data log likelihood, essentially reduces to calculating the expectation of the indicator variables q_{ij} given the data and the current model estimate, which is exactly setting the q_{ij} equal to the posterior probability of the data point i belonging to Gaussian j, as given in equation (73). The EM algorithm obtains this maximum likelihood estimate by monotonically increasing the likelihood: optimizing the F function from equation (72), which is a lower bound on the likelihood that matches the likelihood after each E step, increases the likelihood after each M step. This view of the EM algorithm for a mixture of Gaussians (Hathaway, 1986) can also be used to argue for different E- and/or M-steps in order to speed up convergence (Neal and Hinton, 1998). For example, one can choose particular data points to target in the E-step or specific model parameters in the M-step.

In the M step we maximize $\langle\Phi\rangle$ with respect to the model parameters θ. The constraint on the amplitudes (i.e., that they add up to one) can be implemented by adding a Lagrange multiplier λ and an extra term $\lambda\left(\sum_j \alpha_j - 1\right)$ to $\langle\Phi\rangle$. Taking the derivative of this with respect to α_j then leads to

$$\begin{aligned}
\frac{\partial\langle\Phi\rangle}{\partial\alpha_j} = 0 \quad &\Leftrightarrow \quad \frac{\partial}{\partial\alpha_j}\left(\sum_{i,j} q_{ij}\ln\alpha_j + \lambda\left(\sum_j \alpha_j - 1\right)\right)\\
&\Leftrightarrow \quad \frac{\sum_i q_{ij}}{\alpha_j} + \lambda = 0\\
&\Leftrightarrow \quad \alpha_j = -\frac{1}{\lambda}\sum_i q_{ij}\ . \qquad (75)
\end{aligned}$$

The requirement that the α_j add up to one leads to

$$\begin{aligned}
\sum_j \alpha_j = 1 \quad &\Leftrightarrow \quad \sum_j -\frac{1}{\lambda}\sum_i q_{ij} = 1\\
&\Leftrightarrow \quad -\frac{1}{\lambda}\sum_i\sum_j q_{ij} = 1\\
&\Leftrightarrow \quad -\frac{1}{\lambda}\sum_i 1 = 1\\
&\Leftrightarrow \quad \lambda = -N\ , \qquad (76)
\end{aligned}$$

where we have used the fact that $\sum_j q_{ij} = 1$ and N is the number of data points. Therefore the optimal value of α_j is

$$\alpha_j = \frac{1}{N}\sum_i q_{ij} \ . \tag{77}$$

The rest of the optimization reduces to optimizing the reduced log likelihood

$$\langle \Phi_{\text{red}} \rangle = \sum_i \sum_j q_{ij} \left[\ln\det \mathbf{V}_j + (\mathbf{w}_i - \mathbf{m}_j)^\top \mathbf{V}_j^{-1} (\mathbf{w}_i - \mathbf{m}_j) \right] \ ,$$

which we can simplify by using (1) that $\ln\det \mathbf{V}_j = \text{Trace}\ln \mathbf{V}_j$, (2) that for any number a we can write that $a = \text{Trace}(a)$, and (3) the cyclical property of the trace:

$$\begin{aligned} d\langle \Phi_{\text{red}} \rangle = \sum_i q_{ij} \text{Trace}\Big[& \mathbf{V}_j^{-1} d\mathbf{V}_j - \mathbf{V}_j^{-1} d\mathbf{V}_j \mathbf{V}_j^{-1} (\mathbf{w}_i - \mathbf{m}_j)(\mathbf{w}_i - \mathbf{m}_j)^\top \\ & - 2\mathbf{V}_j^{-1}(\mathbf{w}_i - \mathbf{m}_j) d\mathbf{m}_j^\top \Big] \ , \end{aligned} \tag{78}$$

where we have used the fact that $d\mathbf{V}_j^{-1} = -\mathbf{V}_j^{-1} d\mathbf{V}_j \mathbf{V}_j^{-1}$ (which one can derive by differentiating $\mathbf{V}_j \mathbf{V}_j^{-1} = 1$). This is equal to zero when

$$\begin{aligned} \mathbf{m}_j &= \frac{\sum_i q_{ij} \mathbf{w}_i}{\sum_i q_{ij}} \\ \mathbf{V}_j &= \frac{\sum_i q_{ij} (\mathbf{w}_i - \mathbf{m}_j)(\mathbf{w}_i - \mathbf{m}_j)^\top}{\sum_i q_{ij}} \ . \end{aligned} \tag{79}$$

To summarize, the likelihood can be optimized by alternating the following E- and M-steps:

$$\begin{aligned} \textbf{E-step:} \quad q_{ij} &\leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i | \mathbf{m}_j, \mathbf{V}_j)}{\sum_k \alpha_k \mathcal{N}(\mathbf{w}_i | \mathbf{m}_k, \mathbf{V}_k)} \\ \textbf{M step:} \quad \alpha_j &\leftarrow \frac{1}{N} \sum_i q_{ij} \\ \mathbf{m}_j &\leftarrow \frac{1}{q_j} \sum_i q_{ij} \, \mathbf{w}_i \\ \mathbf{V}_j &\leftarrow \frac{1}{q_j} \sum_i q_{ij} \left[(\mathbf{m}_j - \mathbf{w}_i)(\mathbf{m}_j - \mathbf{w}_i)^\top \right] \ , \end{aligned} \tag{80}$$

where $q_j = \sum_i q_{ij}$.

ACKNOWLEDGEMENTS

It is a pleasure to thank Frédéric Arenou and Phil Marshall for comments and assistance.

REFERENCES

M. Aitkin and I. Aitkin. A Hybrid EM/Gauss-Newton Algorithm for Maximum Likelihood in Mixture Distributions. *Statistics and Computing*, 6:127, 1996.

T. Antoja, F. Figueras, D. Fernández, and J. Torra. Origin and Evolution of Moving Groups. I. Characterization in the Observational Kinematic-age-metallicity Space. A&A, 490:135, 2008.

S. E. Atkinson. The Performance of Standard and Hybrid EM Algorithms for ML Estimates of the Normal Mixture Model with Censoring. *Journal of Statistical Computation and Simulation*, 44:105, 1992.

R. A. Baxter. Finding Overlapping Distributions with MML. Technical Report Tech. report 244, Dept. of Computer Science, Monash University, Clayton, Vic. 3168, Australia, 1995.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

J. Binney and M. Merrifield. *Galactic Astronomy.* Princeton University Press, 1998.

A. Blaauw, C. S. Gum, J. L. Pawsey, and G. Westerhout. The New I. A. U. System of Galactic Coordinates (1958 Revision). MNRAS, 121:123, 1960.

J. Bovy, D. W. Hogg, and S. T. Roweis. The Velocity Distribution of Nearby stars from Hipparcos Data I. The Significance of the Moving Groups. ApJ, 700:1794, 2009.

M. Broniatowski, G. Celeux, and J. Diebolt. Reconaissance de Densités par un Algorithme d'Apprentissage Probabiliste. In *Data Analysis and Informatics Vol. 3*, page 359. North-Holland, Amsterdam, 1983.

G. Celeux and J. Diebolt. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Comp. Stat. Quart*, 2:73, 1985.

G. Celeux and J. Diebolt. L'Algorithme SEM: un Algorithme d'Apprentissage Probabiliste pour la Reconnaisance de Mélanges de Densités. *Rev. de Stat. Appl.*, 34:35, 1986.

N. E. Day. Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, 56:463, 1969.

W. Dehnen. The Distribution of Nearby Stars in Velocity Space Inferred from HIPPARCOS Data. AJ, 115:2384, 1998.

W. Dehnen and J. J. Binney. Local Stellar Kinematics from HIPPARCOS Data. MNRAS, 298:387, 1998.

A. P. Dempster, N. M. Laird, and D. B Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1, 1977.

J. Diebolt and C. P. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:363, 1994.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7:1, 1979.

B. Efron and R. Tibshirai. *An Introduction to the Bootstrap.* Chapman & Hall, London, 1993.

ESA. *The* Hipparcos *and Tycho Catalogues.* ESA SP-1200, Noordwijk: ESA, 1997.

B. Famaey, A. Jorissen, X. Luri, M. Mayor, S. Udry, H. Dejonghe, and C. Turon. Local

Kinematics of K and M Giants from CORAVEL/Hipparcos/Tycho-2 Data. Revisiting the Concept of Superclusters. A&A, 430:165, 2005.

E. D. Feigelson and G. J. Babu. Linear Regression in Astronomy. II. ApJ, 397:55, 1992. .

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2000.

Z. Ghahramani and M. J. Beal. Variational Inference for Bayesian Mixtures of Factor Analysers. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, page 449. MIT Press, 2000.

Z. Ghahramani and M. I. Jordan. Supervised Learning from Incomplete Data via an EM approach. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, San Francisco, 1994a. Morgan Kaufman Publishers.

Z. Ghahramani and M. I. Jordan. Learning from Incomplete Data. Technical report, CBCL Technical Report # 108. Center for Biological and Computational Learning. MIT, 1994b.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

R. J. Hathaway. Another Interpretation of EM Algorithm for Mixture Distributions. *Statist. and Probab. Letters*, 4:53, 1986.

A. Helmi, S. D. M. White, P. T. de Zeeuw, and H. Zhao. Debris Streams in the Solar Neighbourhood as Relicts from the Formation of the Milky Way. Nature, 402:53–55, 1999.

D. W. Hogg, M. R. Blanton, S. T. Roweis, and K. V. Johnston. Modeling Complete Distributions with Incomplete Observations: The Velocity Ellipsoid from Hipparcos Data. ApJ, 629:268, 2005.

T. Isobe, E. D. Feigelson, M. G. Akritas, and G. J. Babu. Linear Regression in Astronomy. ApJ, 364:104, 1990. .

M. Jamshidian and R. I. Jennrich. Conjugate Gradient Acceleration of the EM Algorithm. *Journal of the Americal Statistical Association*, 88:221, 1993.

P. N. Jones and G. J. McLachlan. Improving the Convergence Rate of the EM Algorithm for a Mixture Model Fitted to Grouped Truncated Data. *Journal of Statistical Computation and Simulation*, 43:31, 1992.

B. C. Kelly. Some Aspects of Measurement Error in Linear Regression of Astronomical Data. ApJ, 665:1489, 2007.

K. Lange. A Quasi-Newton Acceleration of the EM Algorithm. *Statistica Sinica*, 5:1, 1995.

C. Liu and D. B. Rubin. The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika*, 81:633, 1994.

T. A. Louis. Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:226, 1982.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

G. A. McIntyre, C. Brooks, W. Compston, and A. Turek. The Statistical Assessment of Rb-Sr Isochrons. J. Geophys. Res., 71:5459, 1966.

G. J. McLachlan and K. Basford. *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, 1988.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1997.

X. L. Meng and D. B. Rubin. Obtaining Asymptotic Variance-Covariance Matrices for Missing-Data Problems using EM. In *Proceedings of the Statistical Computing Section, American Statistical Association*, page 140, Alexandria, VA, 1989. American Statistical Association.

X. L. Meng and D. B. Rubin. Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm. *Journal of the Americal Statistical Association*, 86:899, 1991.

X. L. Meng and D. B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80:267, 1993.

R. N. Neal and G. E. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse and Other Variants. In M. I. Jordan, editor, *Learning in Graphical Models*, page 195, Dordrecht, 1998. Kluwer Academic Publishers.

B. Nordström, M. Mayor, J. Andersen, J. Holmberg, F. Pont, B. R. Jørgensen, E. H. Olsen, S. Udry, and N. Mowlavi. The Geneva-Copenhagen Survey of the Solar Neighbourhood. Ages, Metallicities, and Kinematic Properties of ∼14 000 F and G Dwarfs. A&A, 418: 989, 2004.

J. J. Oliver, R. A. Baxter, and C. S. Wallace. Unsupervised Learning Using MML. In *In Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96*, page 364. Morgan Kaufmann Publishers, 1996.

D. Ormoneit and V. Tresp. Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Terms and Network Averaging. In D. S. Touretzky, M. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8, NIPS, Denver,CO, November 27-30, 1995*. MIT Press, 1996.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 2007.

L. Rabiner and J. Biing-Hwang. *Fundamentals of Speech Recognition*. PTR Prentice-Hall, 1993.

C. Rasmussen. The Infinite Gaussian Mixture Model. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, page 554. MIT Press, 2000.

R. A. Redner and H. F. Walker. Mixture Densities, Maximum Likelihood and the EM algorithm. *SIAM Review*, 26:195, 1984.

S. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59:731, 1997.

J. Rissanen. Modeling by Shortest Data Description. *Automatica*, 14:465, 1978.

S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian Approaches to Gaussian Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1133, 1998.

S. Sakai et al. The Hubble Space Telescope Key Project on the Extragalactic Distance Scale. XXIV. The Calibration of Tully-Fisher Relations and the Value of the Hubble Constant. ApJ, 529:698, 2000. .

G. Schwartz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6:461, 1978.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

J. Skuljan, J. B. Hearnshaw, and P. L. Cottrell. Velocity Distribution of Stars in the Solar Neighbourhood. MNRAS, 308:731, 1999.

M. Stone. Cross-validation Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:111, 1974.

R. B. Tully and J. R. Fisher. A New Method of Determining Distances to Galaxies. A&A, 54:661, 1977.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Split and Merge EM Algorithm for Improving Gaussian Mixture Density Estimates. In *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, page 274, 1998.

F. van Leeuwen. Validation of the New Hipparcos Reduction. A&A, 474:653, 2007a.

F. van Leeuwen. *Hipparcos, the New Reduction of the Raw Data*, volume 250 of *Astrophysics and Space Science Library*. Springer, 2007b.

C. S. Wallace and D. M. Boulton. An Information Measure for Classification. *Computer Journal*, 11:185, 1968.

J. H. Wolfe. Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research*, 5:329, 1970.

C. F. J Wu. On the Convergence Properties of the EM Algorithm. *Ann. Statist.*, 11:95, 1983.

D. York. Least-squares Fitting of a Straight Line. *Canadian J. Phys.*, 44:1079, 1966.

D. York. Least Squares Fitting of a Straight Line with Correlated Errors. *Earth Planet. Sci. Lett.*, 5:320, 1969.

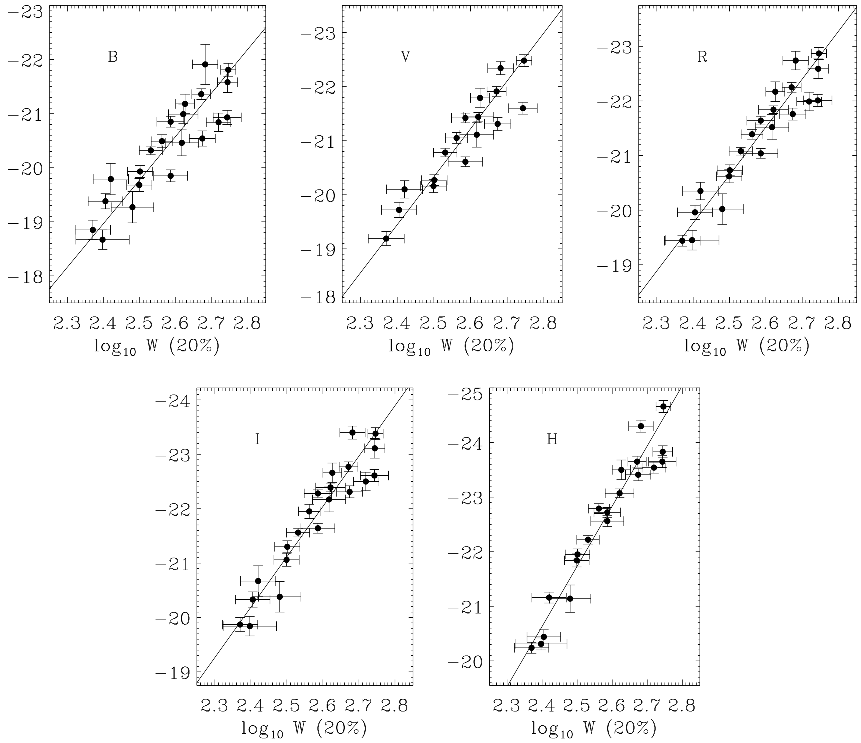

FIG 1. *The Tully–Fisher relation from* HST *data: The absolute magnitude in five bands (B, V, R, I, and H) as a function of the velocity width W is well fit by a power-law for spiral galaxies. The relation is found by fitting a single Gaussian distribution to the distribution of points in the absolute magnitude–*$\log_{10} W$ *plane for each bandpass.*

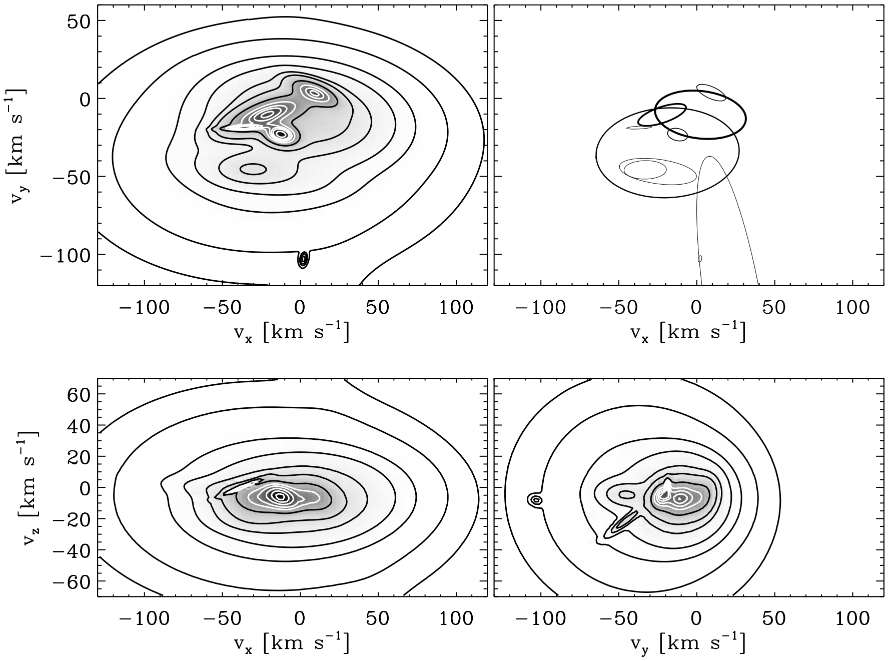

FIG 2. *Two-dimensional projections of the three-dimensional velocity distribution of* Hipparcos *stars using 10 Gaussians and $w = 4$ km^2 s^{-2}. The top right plot shows 1-sigma covariance ellipses around each individual Gaussian in the v_x–v_y plane; the thickness of each covariance ellipse is proportional to the natural logarithm of its amplitude α_j. In the other three panels the density grayscale is linear and contours contain, from the inside outward, 2, 6, 12, 21, 33, 50, 68, 80, 90, 95, 99, and 99.9 percent of the distribution. 50 percent of the distribution is contained within the innermost dark contour. The feature at $v_y \approx -100$ km s^{-1} is real and corresponds to a known feature in the velocity distribution: the Arcturus moving group; Indeed, all the features that appear in these projections are real and correspond to known features.*

Center for Cosmology and Particle Physics
Department of Physics
New York University
4 Washington Place
New York, NY 10003
E-mail: jo.bovy@nyu.edu; david.hogg@nyu.edu

Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012
E-mail: roweis@cs.nyu.edu

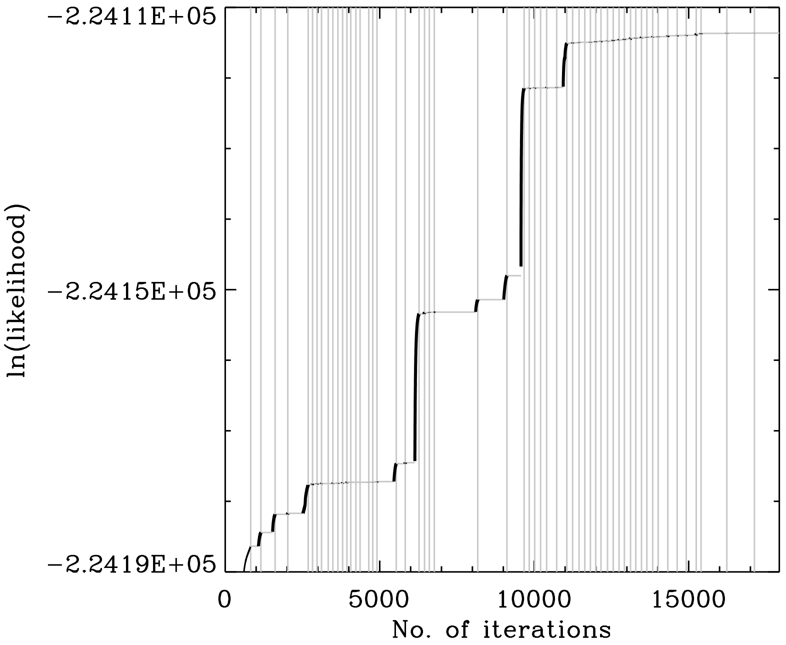

FIG 3. *Convergence of the full algorithm: total log likelihood at each iteration step. Shown are only split-and-merge steps that improve the likelihood; each vertical gray line corresponds to a point at which a successful split and merge is performed. For clarity's sake, we show in black only the parts of the split-and-merge steps at which the likelihood is larger than the likelihood right before that split-and-merge procedure; the log likelihoods of the steps in a split-and-merge procedure in which the likelihood is still climbing back up to the previous maximum in likelihood have been replaced by horizontal gray segments. The y-axis has been cut off for display purposes: The log likelihood of the initial condition was -2.39E5.*