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

BY JO BOVY[1], DAVID W. HOGG[1,2] AND SAM T. ROWEIS[3]

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 heteroskedastic uncertainty distribution of the data point and projecting out the missing data directions. We show how this basic algorithm can be extended with conjugate priors on all of the model parameters and a "split-and-merge" procedure designed to avoid local maxima of the likelihood. We demonstrate the full method by applying it to the problem of inferring the three-dimensional velocity distribution of stars near the Sun from noisy two-dimensional, transverse velocity measurements from the *Hipparcos* satellite.

1. Introduction. Inferring a distribution function given a finite set of samples from this distribution function and the related problem of finding clusters and/or overdensities in the distribution is a problem of significant general interest [e.g., McLachlan and Basford (1988); Rabiner and Biing-Hwang (1993); Dehnen (1998); Helmi et al. (1999); Skuljan, Hearnshaw and

Cottrell (1999); Hogg et al. (2005)]. Performing this inference given only a noisy set of measurements is a problem commonly encountered in many of the sciences [see examples in Carroll et al. (2006)]. In many cases of interest, the noise properties of the observations are different from one measurement to the next (i.e., they are heteroskedastic), even though the uncertainties are well characterized for each observation. This is, for example, often the case in astronomy, where in many cases the dominant source of uncertainty is due to well-characterized photon-counting statistics, while spatial and temporal variations in the atmosphere cause the uncertainties to significantly vary even for sources of the same apparent brightness observed with the same telescope.

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, that is, data with vanishingly small uncertainties and with all of the dimensions measured. In the low signal-to-noise regime 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 to 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 uncertainty convolution is different for each data point, each data point is a sample of a different distribution, that is, 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.

Most existing approaches to density estimation only apply in the high signal-to-noise regime [e.g., McLachlan and Basford (1988); Silverman (1986); Diebolt and Robert (1994)], and most approaches to density estimation from noisy samples are nonparametric techniques that assume that the noise is homoskedastic [e.g., Stefanski and Carroll (1990); Zhang (1990)]. The case of heteroskedastic uncertainties has only recently attracted attention [e.g., Delaigle and Meister (2008); Staudenmayer, Ruppert and Buonaccorsi (2008)], and all of the approaches that have been developed so far are nonparametric. None of these approaches can be used when only incomplete data are available, although parametric techniques that properly account for incomplete, but noiseless, data have been developed [Ghahramani and Jordan (1994a, 1994b)].

In this paper we show that the frequently used Gaussian-mixture-model approach to density estimation 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 these; in the expectation step these actual values are estimated based on the noisy and incomplete 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, 1994b).

We also show how prior distributions for a Bayesian version of the calculation reporting a MAP estimate 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 briefly discuss the practical issues having to do with model selection in the mixture model approach.

An application to a real data set is given in Section 5, where we fit the distribution of stellar velocities near the Sun. The observed velocities of stars that we use for this purpose have all of the properties that the approach developed in this paper handles correctly: The velocity measurements are noisy, and since we only use observations of the velocity components in the plane of the sky, the data are incomplete, and this incompleteness is different for each velocity measurement, which covers the full sky. Nevertheless, we are able to obtain good agreement with other fits of the velocity distribution based on complete data.

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)].

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 $\mathbf{v}_i$,

$$\text{(1)} \qquad\qquad\qquad \mathbf{w}_i = \mathbf{R}_i \mathbf{v}_i + \text{noise},$$

where the noise is drawn from a Gaussian with zero mean and known covariance matrix $\mathbf{S}_i$. 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, that is, 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 probability density $p(\mathbf{v})$ of the true values $\mathbf{v}$ as a mixture of K Gaussians,

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

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

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

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

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

(4)

where

$$p(\mathbf{w}_i|\mathbf{v}) = \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{v}, \mathbf{S}_i),$$

(5)
$$p(\mathbf{v}|j, \theta) = \mathcal{N}(\mathbf{v}|\mathbf{m}_j, \mathbf{V}_j),$$

$$p(j|\theta) = \alpha_j.$$

This likelihood works out to be itself a mixture of Gaussians

$$(6) \qquad p(\mathbf{w}_i|\theta) = \sum_j \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_j, \mathbf{T}_{ij}),$$

where

$$(7) \qquad \mathbf{T}_{ij} = \mathbf{R}_i\mathbf{V}_j\mathbf{R}_i^\top + \mathbf{S}_i.$$

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 (log) likelihood of the model given the data, that is,

$$(8) \qquad \phi = \sum_i \ln p(\mathbf{w}_i|\theta) = \sum_i \ln \sum_{j=1}^K \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_j, \mathbf{T}_{ij}).$$

This function can be optimized in several ways, one of which is to 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 nonnegative and add up to one, the variance matrices must be positive definite and symmetric) and multimodality of the likelihood surface. In what follows we will describe a different approach that is natural in this setting: An EM algorithm that iteratively and monotonically maximizes the likelihood, while naturally respecting the restrictions on the parameters.

3. Fitting mixtures with heterogeneous, noisy data using an EM algorithm. To optimize the likelihood in equation (8), we can extend the standard EM approach to Gaussian mixture estimation. In the case of complete and precise observations, the problem is framed as a tractable missing-data problem by positing that the labels or indicator variables q_{ij} indicating which Gaussian j a data point i was drawn from are missing [Dempster, Laird and Rubin (1977)]. We extend this approach by including the true values $\mathbf{v}_i$ as extra missing data. This is a well-known approach for handling measurement uncertainty in latent variable or random effects models [e.g., Schafer (1993); Schafer and Purdy (1996)].

We write down the "full data" log likelihood—the likelihood we would write down if we had the true values $\mathbf{v}_i$ and the labels q_{ij},

$$(9) \qquad \Phi = \sum_i \sum_j q_{ij} \ln \alpha_j \mathcal{N}(\mathbf{v}_i|\mathbf{m}_j, \mathbf{V}_j).$$

We will now show how we can use the EM methodology to find straight-forward update steps that maximize the full data likelihood of the model. In Appendix A we 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 (9), we find

$$
\Phi = \sum_i \sum_j q_{ij} \left[\ln \alpha_j - \frac{d}{2} \ln(2\pi) \right.
$$

(10)

$$
\left. - \frac{1}{2} \ln \det \mathbf{V}_j - \frac{1}{2} (\mathbf{v}_i - \mathbf{m}_j)^\top \mathbf{V}_j^{-1} (\mathbf{v}_i - \mathbf{m}_j) \right],
$$

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 equal to the posterior probability that a data point $\mathbf{w}_i$ was drawn from the component j. The expectation of the $\mathbf{v}_i$ and the $\mathbf{v}_i\mathbf{v}_i^\top$ can be found as follows: Consider the joint distribution for the true and observed velocities, denoted by the expanded 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

(11)
$$
\mathbf{m}' = \begin{bmatrix} \mathbf{m}_j \\ \mathbf{R}_i\mathbf{m}_j \end{bmatrix}
$$

and covariance matrix

(12)
$$
\mathbf{V}' = \begin{bmatrix} \mathbf{V}_j & \mathbf{V}_j\mathbf{R}_i^\top \\ \mathbf{R}_i\mathbf{V}_j & \mathbf{T}_{ij} \end{bmatrix}.
$$

The conditional distribution of the $\mathbf{v}_i$ given the data $\mathbf{w}_i$ is normal with mean

(13)
$$
\mathbf{b}_{ij} \equiv \mathbf{m}_j + \mathbf{V}_j\mathbf{R}_i^\top \mathbf{T}_{ij}^{-1}(\mathbf{w}_i - \mathbf{R}_i\mathbf{m}_j)
$$

and covariance matrix

(14)
$$
\mathbf{B}_{ij} \equiv \mathbf{V}_j - \mathbf{V}_j\mathbf{R}_i^\top \mathbf{T}_{ij}^{-1}\mathbf{R}_i\mathbf{V}_j.
$$

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

$$
\langle \Phi \rangle = \sum_{i,j} q_{ij} \left[\ln \alpha_j - \frac{d}{2} \ln(2\pi) \right.
$$

(15)

$$
\left. - \frac{1}{2} \text{Trace}[\ln \mathbf{V}_j + (\mathbf{B}_{ij} + (\mathbf{m}_j - \mathbf{b}_{ij})(\mathbf{m}_j - \mathbf{b}_{ij})^\top)\mathbf{V}_j^{-1}] \right].
$$

Straightforward optimization of this with respect to the model parameters gives the following algorithm:

$$E\text{-step:} \quad q_{ij} \leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_j, \mathbf{T}_{ij})}{\sum_k \alpha_k \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_k, \mathbf{T}_{ik})},$$

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

(16)
$$\mathbf{B}_{ij} \leftarrow \mathbf{V}_j - \mathbf{V}_j\mathbf{R}_i^\top \mathbf{T}_{ij}^{-1}\mathbf{R}_i\mathbf{V}_j,$$

$$M\text{-step:} \quad \alpha_j \leftarrow \frac{1}{N}\sum_i q_{ij},$$

$$\mathbf{m}_j \leftarrow \frac{1}{q_j}\sum_i q_{ij}\mathbf{b}_{ij},$$

$$\mathbf{V}_j \leftarrow \frac{1}{q_j}\sum_i q_{ij}[(\mathbf{m}_j - \mathbf{b}_{ij})(\mathbf{m}_j - \mathbf{b}_{ij})^\top + \mathbf{B}_{ij}],$$

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

In Appendix A we prove that this procedure for maximizing the full data likelihood also monotonically increases the likelihood of the data $\mathbf{w}_i$ given the model, as is the case for the EM algorithm for noiseless and complete measurements [Dempster, Laird and Rubin (1977); Wu (1983)].

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 (16) becomes singular; the EM updates might get stuck in a local maximum because of the monotonic increase in likelihood ensured by the EM algorithm. The latter can be avoided through the use of a stochastic EM procedure [Broniatowski, Celeux and Diebolt (1983); Celeux and Diebolt (1985); Celeux and Diebolt (1986)] or through the split and merge procedure described below.

4.1. *Bayesian-inspired regularization.* The problem of singular covariances can be mitigated through the use of priors on the model parameters in a Bayesian setting [Ormoneit and Tresp (1996)]. It should be emphasized here that this calculation is only Bayesian in the sense of producing a maximum a posteriori (MAP) point estimate rather than a maximum likelihood estimate, and that this is no different than penalized maximum likelihood. We briefly show here that this procedure can be applied here as well.

The regularization scheme of Ormoneit and Tresp (1996) introduces conjugate priors on the Gaussian mixtures parameters space $\theta = (\alpha_j, \mathbf{m}_j, \mathbf{V}_j)$ as penalty terms. These conjugate priors are the following: A normal density $\mathcal{N}(\mathbf{m}_j | \hat{\mathbf{m}}, \eta^{-1} \mathbf{V}_j)$ for the mean of each Gaussian, a Wishart density $\mathcal{W}(\mathbf{V}_j^{-1} | \omega, \mathbf{W})$ [Gelman et al. (2000)],

$$(17) \quad \mathcal{W}(\mathbf{V}_j^{-1} | \omega, \mathbf{W}) = c(\omega, \mathbf{W}) | \mathbf{V}_j^{-1} |^{\omega - (d+1)/2} \exp[- \operatorname{Trace}[\mathbf{W} \mathbf{V}_j^{-1}]],$$

with $c(\omega, \mathbf{W})$ a normalization constant, for the covariance of each Gaussian, and a Dirichlet density $\mathcal{D}(\alpha | \gamma)$, given by

$$(18) \qquad \mathcal{D}(\alpha | \gamma) = b \prod_j \alpha_j^{\gamma_j - 1},$$

where b is a normalizing factor, for the amplitudes $\{\alpha_j\}$. Optimizing the posterior distribution for the model parameters replaces the M-step of equation (16) with

$$\alpha_j \leftarrow \frac{\sum_i q_{ij} + \gamma_j - 1}{N + \sum_k \gamma_k - K}, \qquad \mathbf{m}_j \leftarrow \frac{\sum_i q_{ij} \mathbf{b}_{ij} + \eta \hat{\mathbf{m}}}{q_j + \eta},$$

$$\mathbf{V}_j \leftarrow \left(\sum_i q_{ij} [(\mathbf{m}_j - \mathbf{b}_{ij})(\mathbf{m}_j - \mathbf{b}_{ij})^\top + \mathbf{B}_{ij}] \right.$$

(19)

$$\left. + \eta (\mathbf{m}_j - \hat{\mathbf{m}})(\mathbf{m}_j - \hat{\mathbf{m}})^\top + 2 \mathbf{W} \right)$$

$$\Big/ q_j + 1 + 2(\omega - (d+1)/2).$$

Hyperparameters can be set by leave-one-out cross-validation. Vague priors on the amplitude and the means are obtained by setting

$$(20) \qquad \gamma_j = 1 \qquad \forall j, \qquad \omega = (d+1)/2, \qquad \eta = 0.$$

Since we are only interested in the MAP estimate, propriety of the resulting posterior is not an issue with the improper prior resulting from this choice.

The label-switching problem in Bayesian mixtures [Jasra, Holmes and Stephens (2005)] is not an issue for the maximization of the posterior distribution here.

4.2. *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 and two of these Gaussians are merged, while the third Gaussian is split into two Gaussians

[Ueda et al. ([1998](#))]. 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 approaches [Ghahramani and Beal ([2000](#)); Beal ([2003](#))] 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.

Full details of the split and merge algorithm are given in Appendix B.

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 $\mathbf{W}$. If we further assume that $\mathbf{W} = w\mathbf{I}$, then this covariance regularization parameter basically sets the square of the smallest scale of the distribution function on which we can reliably 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. Other techniques include methods based on Bayesian model selection [Roberts et al. ([1998](#))] as well as approaches based on minimum encoding inference [Wallace and Boulton ([1968](#)); Oliver, Baxter and Wallace ([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 1), 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, Hogg and Roweis ([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. These methods also address uncertainty quantification throughout the model. 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
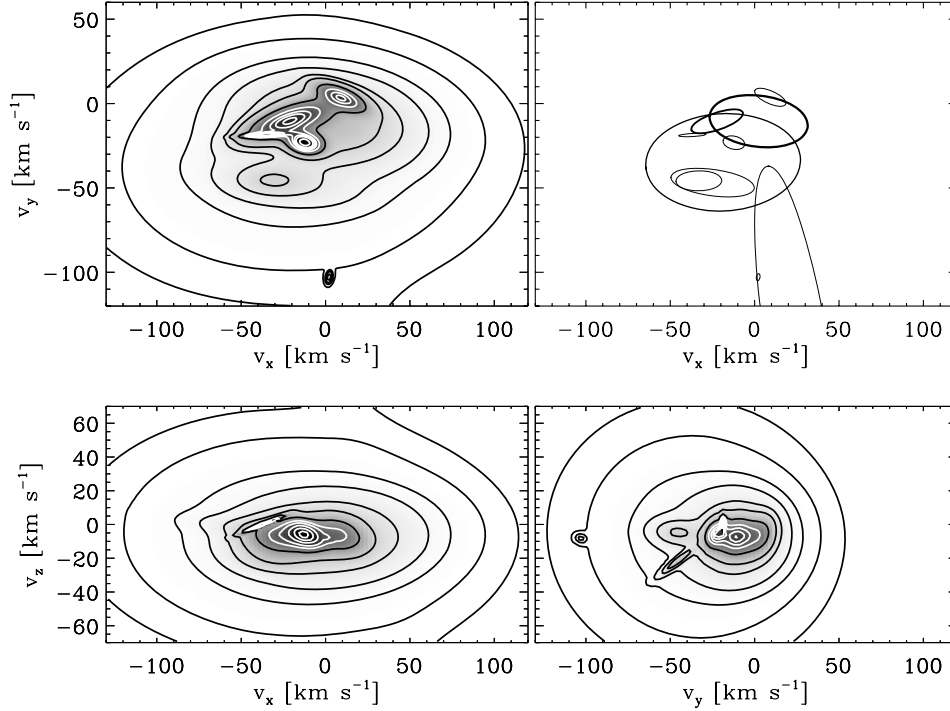
FIG. 1. *Two-dimensional projections of the three-dimensional velocity distribution of Hipparcos stars using 10 Gaussians and $w = 4$ km$^2 \cdot s^{-2}$. The top right plot shows 1-sigma covariance ellipses around each individual Gaussian in the v_x–v_y plane; the thickness of each covariance ellipse 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 $\cdot s^{-1}$ 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.*

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 $\mathbf{v}_i$—known in the Bayesian MCMC literature as data augmentation—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 (13) and variance given in equation (14).

5. 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, Hogg and Roweis (2009)].

The local velocity distribution is interesting because we can learn about the structure and evolution of the Galactic disk from deviations from the smooth, close to Gaussian velocity distribution expected in simple, axisymmetric models of the disk. It has been shown that the dynamical effects of a bar-shaped distribution of stars in the central region of our Galaxy can resonate in the outer parts of the disk and give rise to structure in the velocity distribution [e.g., Dehnen (2000); Bovy (2010)]. Similarly, steady-state or transient spiral structure can effect the velocities of stars in a coherent way, such that we can see this effect locally [e.g., Quillen and Minchev (2005); De Simone, Wu and Tremaine (2004)]. Inferring the local velocity distribution from observational data is therefore necessary to assess whether these dynamical signatures are observed.

Velocities of stars are not directly observable. Rather, they need to be patched together from observations of the stars' directions on the sky at different times—the branch of astronomy known as astrometry—and spectroscopic observations to determine the velocity along the line of sight. The annual motion of the Earth around the Sun gives rise to an apparent displacement of a star relative to background objects that is inversely proportional to the distance to the star. Measurements of this apparent shift, or parallax, can thus be used to determine the distance to stars. Parallaxes are traditionally reported in units of arcseconds; a star with a parallax of 1 arcsecond is defined to be at a distance of 1 parsec (pc), which equals 3×10^{16} m. The intrinsic motion of a star also gives rise to a systematic shift in its position relative to background sources, such that its angular motion—known as its proper motion—can be measured. Combining the distance and angular velocity gives the components of the space velocity of a star that are perpendicular to the line of sight.

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 milli-arcseconds [ESA (1997)]. From this catalogue of ∼100,000 stars, kinematically unbiased samples of stars with accurate positions and velocities can be extracted [Dehnen and Binney (1998)]. Since this was a purely astrometric mission, and the only components of a star's velocity that can be measured astrometrically are the components perpendicular to the line of sight, the line-of-sight velocities of the stars in the *Hipparcos* sample were not obtained during the mission.

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; the diameter of the Galactic

disk is about 30,000 pc). In addition to this, since distances are measured as inverse distances (parallaxes), only distances that are measured relatively precisely 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 ∼10-percent, such that the velocities perpendicular to the line of sight that are obtained from the proper motions and the distances have low signal-to-noise. Since the dominant source of noise is due to simple photon-counting statistics [van Leeuwen (2007a)], the uncertainties are well characterized and can be assumed to be known, as is necessary for the technique developed in this paper to apply. Star-to-star correlations are negligible and can be ignored [van Leeuwen (2007a)].

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 toward the Galactic center, the y-axis points in the direction of Galactic rotation, and the z-axis points toward the North Galactic Pole [Blaauw et al. (1960); Binney and Merrifield (1998)]. The measured velocities perpendicular to the line of sight 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.

Observations of celestial objects are expressed in the equatorial coordinate system, in which the Earth's geographic poles and equator are projected onto the celestial sphere. The components of the three-dimensional velocities $\mathbf{v}$ of the stars in the Galactic coordinate system in terms of the observed quantities in the equatorial coordinate frame—angular position on the sky (α, δ), inverse distance (π), angular motion on the sky (μ_α, μ_δ), and line-of-sight velocity (v_r)—are given by

$$(21) \qquad \mathbf{v} \equiv \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \mathbf{TA} \begin{bmatrix} v_r \\ \dfrac{k}{\pi}\mu_\alpha \cos\delta \\ \dfrac{k}{\pi}\mu_\delta \end{bmatrix},$$

where $k = 4.74047$, $[v_r] = \mathrm{km \cdot s^{-1}}$, $[\pi] = \mathrm{arcsec}, [\mu_\alpha] = [\mu_\delta] = \mathrm{arcsec} \cdot \mathrm{yr^{-1}}$. The matrix $\mathbf{T}$ transforms the velocities from the equatorial reference frame in which the observations are made to the Galactic coordinate frame; it depends on a few parameters defining this coordinate transformation and is given by

$$(22) \qquad \mathbf{T} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ \sin\theta & -\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\sin\delta_{\mathrm{NGP}} & 0 & \cos\delta_{\mathrm{NGP}} \\ 0 & 1 & 0 \\ \cos\delta_{\mathrm{NGP}} & 0 & \sin\delta_{\mathrm{NGP}} \end{bmatrix}$$

$$(23) \qquad \times \begin{bmatrix} \cos\alpha_{\mathrm{NGP}} & \sin\alpha_{\mathrm{NGP}} & 0 \\ -\sin\alpha_{\mathrm{NGP}} & \cos\alpha_{\mathrm{NGP}} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The matrix $\mathbf{T}$ depends on the epoch that the reduced data are referred to (1991.25 for *Hipparcos*) through the values of α_{NGP}, δ_{NGP} and θ (the position in equatorial coordinates of the north Galactic pole, and the Galactic longitude of the north Celestial pole, respectively). These quantities were defined for the epoch 1950.0 as follows: [Blaauw et al. (1960)]: $\alpha_{\mathrm{NGP}} = 12$ h 49 m, $\delta_{\mathrm{NGP}} = 27^\circ.4$, and $\theta = 123^\circ$.

The matrix $\mathbf{A}$ depends on the position of the source on the sky,

$$(24) \qquad \mathbf{A} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\delta & 0 & -\sin\delta \\ 0 & 1 & 0 \\ \sin\delta & 0 & \cos\delta \end{bmatrix}.$$

In the context of the deconvolution technique described above, the observations are $\mathbf{w} \equiv [\, v_r, \quad \frac{k}{\pi}\mu_\alpha \cos\delta, \quad \frac{k}{\pi}\mu_\delta \,]^\top$ and the projection matrix is $\mathbf{R}^{-1} \equiv \mathbf{TA}$. Since we do not use the radial velocities of the stars, we set v_r to zero in $\mathbf{w}$ and use a large uncertainty-variance for this component of the uncertainty variance $\mathbf{S}_i$; equivalently, we could remove v_r from $\mathbf{w}$ and restrict $\mathbf{R}$ to the projection on the sky.

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 11,865 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; 2007b)]. A particular reconstruction of the underlying velocity distribution of the stars is shown in Figure 1, in which 10 Gaussians are used, a prior on the variances was used (the prior was restricted to $\mathbf{W} = w\mathbf{I}$), and this regularization parameter w is set to 4 km$^2 \cdot$ s^{-2}.

These values for the hyperparameters were set using an external data set rather than any of the other methods described in Section 4.3. For this we use a set of 7682 stars from the Geneva–Copenhagen Survey [Nordström et al. (2004); Holmberg, Nordström and Andersen (2009)] for which the line-of-sight velocity (perpendicular to the plane of the sky) has been measured spectroscopically. This is a separate data set from the one considered above. It partially overlaps with the previous data set and it is also kinematically unbiased. We then fit the velocity distribution for different choices of the hyperparameters and evaluate the probability of the line-of-sight velocities for the best-fit velocity distribution based on tangential velocities. The values of the hyperparameters that maximize this probability are $K = 10$ and $w = 4$ km$^2 \cdot$ s^{-2}.

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, Hearnshaw and Cottrell (1999); Nordström et al. (2004); Famaey et al. (2005); Antoja et al. (2008)]. In particular, this means that the main shape of the velocity distribution agrees with that found in previous studies and that the number and location of the peaks in the distribution, all real and known features, are consistent with those found before. This includes the very sparsely populated feature at $v_y \approx -100$ km $\cdot$ 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. In contrast to previous determinations of the velocity distributions, our method allows us to study the structures found quantitatively, since it turns out that individual structures in the velocity distribution are well represented by individual components in the mixture model. Thus, we were able to conclude that these structures are not the remnants of a large group of stars that formed together in a cluster, but rather that they are probably caused by dynamical effects related to the bar at the center of the Milky Way or spiral structure [Bovy and Hogg (2010)].

The convergence of the algorithm is shown in Figure 2. 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.

6. 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.[4] 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 IDL[5] or Python[6] wrapper functions supplied with the C code.

The code can do everything described above. The convergence of a single run is quick, but when including split-and-merge iterations the convergence

---

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

[5]http://www.ittvis.com/ProductServices/IDL.aspx.
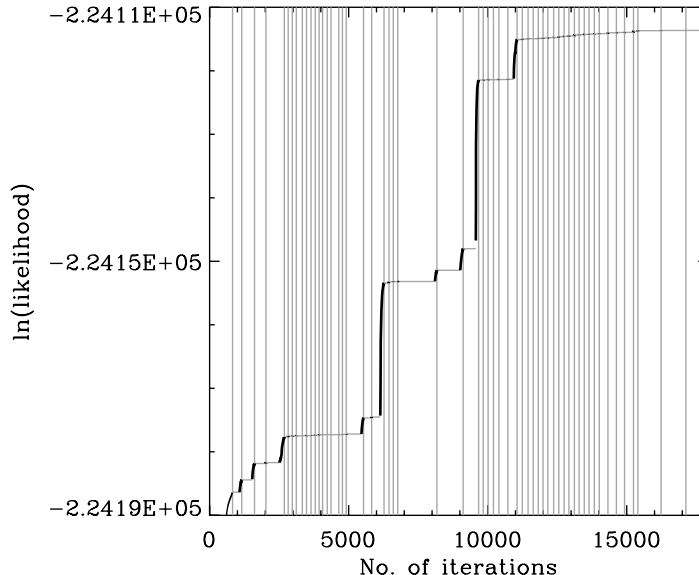
[6]http://www.python.org/.

FIG. 2. *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 $-2.39E-5$.*

is rather slow because of 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).

7. 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 conjugate priors on all of the

model parameters without losing any of its analytical attractiveness and that the algorithm can accommodate the split-and-merge algorithm to deal 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 nonparametric approaches to density modeling, for example, in mixture models with an infinite number of components based on the Dirichlet Process [e.g., Rasmussen (2000)]. In this way current advances in nonparametric 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: PROOF THAT THE PROPOSED ALGORITHM MAXIMIZES THE LIKELIHOOD

We use Jensen's inequality in the continuous case for a concave function f and a nonnegative integrable function q, where we have assumed that q is normalized, that is, q is a probability distribution. For each observation $\mathbf{w}$ we can then introduce a function $q(\mathbf{v}, j)$ such that

$$\ln p(\mathbf{w}|\theta) = \ln \sum_j \int_{\mathbf{v}} d\mathbf{v}\, p(\mathbf{w}, \mathbf{v}, j|\theta)$$

$$(25) \qquad\qquad \geq \sum_j \int_{\mathbf{v}} d\mathbf{v}\, q(\mathbf{v}, j) \ln \frac{p(\mathbf{w}, \mathbf{v}, j|\theta)}{q(\mathbf{v}, j)} = F(\mathbf{w}|q, \theta),$$

$$\ln p(\mathbf{w}|\theta) \geq F(\mathbf{w}|q, \theta) = \langle \ln p(\mathbf{w}, \mathbf{v}, j|\theta) \rangle_q + \mathcal{H}(q),$$

where $\mathcal{H}$ is the entropy of the distribution $q(\mathbf{v}, j)$. This inequality becomes an equality when we take

$$(26) \qquad\qquad\qquad q(\mathbf{v}, j) = p(\mathbf{v}, j|\mathbf{w}, \theta).$$

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

$$(27) \qquad\qquad p(\mathbf{v}, j|\mathbf{w}_i, \theta) = p(\mathbf{v}|\mathbf{w}_i, \theta, j)p(j|\mathbf{w}_i, \theta).$$

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)

$$F = \sum_{i,j} q_{ij} \int_{\mathbf{v}} d\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j) \ln p(\mathbf{w}_i, \mathbf{v}, j|\theta)$$

$$= \sum_{i,j} q_{ij} \left[\ln \alpha_j + \int_{\mathbf{v}} d\mathbf{v}\, p(\mathbf{v}|\mathbf{w}_i, \theta, j) \right.$$

$$\times \left[-\frac{1}{2} \ln \det \mathbf{V}_j - \frac{1}{2} (\mathbf{v} - \mathbf{m}_j)^\top \mathbf{V}_j^{-1} (\mathbf{v} - \mathbf{m}_j) \right] \right].$$

This shows that this reduces exactly to the procedure described above, that is, 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.

APPENDIX B: SPLIT AND MERGE ALGORITHM

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. 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

(28)
$$\alpha_{j_1'} = \alpha_{j_1} + \alpha_{j_2},$$
$$\theta_{j_1'} = \frac{\theta_{j_1} q_{j_1} + \theta_{j_2} q_{j_2}}{q_{j_1} + q_{j_2}},$$

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:

(29)
$$\alpha_{j_2'} = \alpha_{j_3'} = \alpha_{j_3}/2,$$
$$\mathbf{V}_{j_2'} = \mathbf{V}_{j_3'} = \det(\mathbf{V}_{j_3})^{1/d} \mathbf{I}.$$

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 $\mathbf{m}_{j_3}$, for example,

(30)
$$\mathbf{m}_{j_m'} = \mathbf{m}_{j_3} + \varepsilon_{j_m},$$

where $\|\varepsilon_{j_m}\|^2 \ll \det(\mathbf{V}_{j_3})^{1/d}$ 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 (16) for the parameters of the three affected Gaussians, while keeping the parameters of the other Gaussians fixed, including 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 does not 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 reach 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:

$$(31) \qquad J_{\mathrm{merge}}(j, k|\theta) = \mathbf{P}_j(\theta)^\top \mathbf{P}_k(\theta),$$

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)]

$$(32) \qquad D_{\mathrm{KL}}(P \parallel Q) = \int \mathrm{d}x\, p(x) \ln \frac{p(x)}{q(x)}.$$

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

$$(33) \qquad J_{\mathrm{split}}(l|\theta) = \frac{1}{q_l} \sum_i q_{il} \left[\ln\left(\frac{q_{il}}{q_l}\right) - \ln \mathcal{N}(\mathbf{w}_i|\mathbf{m}_l, \mathbf{V}_l) \right].$$

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

$$(34) \qquad J_{\mathrm{split}}(l|\theta) = \frac{1}{q_l} \sum_i q_{il} \left[\ln\left(\frac{q_{il}}{q_l}\right) - \ln \mathcal{N}(\mathbf{w}_i | \mathbf{R}_i \mathbf{m}_l, \mathbf{R}_i \mathbf{V}_l \mathbf{R}_i^\top) \right].$$

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 $J_{\mathrm{split}}(l|\theta)$.

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

1. Run the EM algorithm as specified in equations (16) and (19). Store the resulting model parameters θ^* and the corresponding model log likelihood ϕ^*.
2. Compute the merge criterion $J_{\mathrm{merge}}(j, k|\theta^*)$ for all pairs j, k and the split criterion $J_{\mathrm{split}}(l|\theta^*)$ 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 (28) and the parameters of the two Gaussian resulting from splitting the third Gaussian using equations (29) and (30). Then run the partial EM procedure on the parameters of the three affected Gaussians, that is, 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.

REFERENCES

ANTOJA, T., FIGUERAS, F., FERNÁNDEZ, D. and TORRA, J. (2008). Origin and evolution of moving groups. I. Characterization in the observational kinematic-age-metallicity space. *Astron. Astrophys.* **490** 135.

Baxter, R. A. (1995). Finding overlapping distributions with MML. Technical Report No. 244, Dept. Computer Science, Monash Univ., Clayton, Australia.

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, Univ. College London.

Binney, J. and Merrifield, M. (1998). *Galactic Astronomy*. Princeton Univ. Press, Princeton, NJ.

Blaauw, A., Gum, C. S., Pawsey, J. L. and Westerhout, G. (1960). The new IAU system of galactic coordinates (1958 revision). *Mon. Not. R. Astron. Soc.* **121** 123.

Bovy, J. (2010). Tracing the Hercules stream around the galaxy. *Astrophys. J.* **725** 1676.

Bovy, J., Hogg, D. W. and Roweis, S. T. (2009). The velocity distribution of nearby stars from Hipparcos data I. The significance of the moving groups. *Astrophys. J.* **700** 1794.

Bovy, J. and Hogg, D. W. (2010). The velocity distribution of nearby stars from Hipparcos data II. The nature of the low-velocity moving groups. *Astrophys. J.* **717** 617.

Broniatowski, M., Celeux, G. and Diebolt, J. (1983). Reconaissance de Densités par un Algorithme d'Apprentissage Probabiliste. In *Data Analysis and Informatics, Vol. 3* 359–373. North-Holland, Amsterdam. MR0787647

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL. MR2243417

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quart* **2** 73.

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

De Simone, R., Wu, X. and Tremaine, S. (2004). The stellar velocity distribution in the solar neighbourhood. *Mon. Not. R. Astron. Soc.* **350** 627.

Dehnen, W. (1998). The distribution of nearby stars in velocity space inferred from Hipparcos data. *Astron. J.* **115** 2384.

Dehnen, W. (2000). The effect of the outer Lindblad resonance of the galactic bar on the local stellar velocity distribution. *Astron. J.* **119** 800.

Dehnen, W. and Binney, J. J. (1998). Local stellar kinematics from Hipparcos data. *Mon. Not. R. Astron. Soc.* **298** 387.

Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli* **14** 562–579. MR2544102

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol. Stat.* **39** 1–38. MR0501537

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. Ser. B Methodol. Stat.* **56** 363–375. MR1281940

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

Famaey, B., Jorissen, A., Luri, X., Mayor, M., Udry, S., Dejonghe, H. and Turon, C. (2005). Local kinematics of K and M giants from CORAVEL/Hipparcos/ Tycho-2 data. Revisiting the concept of superclusters. *Astron. Astrophys.* **430** 165.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2000). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL. MR1385925

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12* (S. A. Solla, T. K. Leen and K. R. Muller, eds.) 449. MIT Press, Cambridge, MA.

Ghahramani, Z. and Jordan, M. I. (1994a). Learning from incomplete data. CBCL Technical Report No. 108. Center for Biological and Computational Learning, MIT.

Ghahramani, Z. and Jordan, M. I. (1994b). Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems 6* (J. D. Cowan, G. Tesauro and J. Alspector, eds.) 120–127. Morgan Kaufman, San Francisco.

Helmi, A., White, S. D. M., de Zeeuw, P. T. and Zhao, H. (1999). Debris streams in the solar neighbourhood as relicts from the formation of the milky way. *Nature* **402** 53–55.

Hogg, D. W., Blanton, M. R., Roweis, S. T. and Johnston, K. V. (2005). Modeling complete distributions with incomplete observations: The velocity ellipsoid from Hipparcos data. *Astrophys. J.* **629** 268.

Holmberg, J., Nordström, B. and Andersen, J. (2009). The Geneva–Copenhagen survey of the solar neighbourhood III. Improved distances, ages, and kinematics. *Astron. Astrophys.* **501** 941.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge Univ. Press, Cambridge. MR2012999

McLachlan, G. J. and Basford, K. (1988). *Mixture Models: Inference and Application to Clustering.* Dekker, New York. MR0926484

Nordström, B., Mayor, M., Andersen, J., Holmberg, J., Pont, F., Jørgensen, B. R., Olsen, E. H., Udry, S. and Mowlavi, N. (2004). The Geneva–Copenhagen survey of the solar neighbourhood. Ages, metallicities, and kinematic properties of $\sim$14 000 F and G dwarfs. *Astron. Astrophys.* **418** 989.

Oliver, J. J., Baxter, R. A. and Wallace, C. S. (1996). Unsupervised learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96)* 364. Morgan Kaufmann, San Francisco.

Ormoneit, D. and Tresp, V. (1996). Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging. In *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, November 27–30, 1995* (D. S. Touretzky, M. Mozer and M. E. Hasselmo, eds.) 542–548. MIT Press, Cambridge.

Quillen, A. C. and Minchev, I. (2005). The effect of spiral structure on the stellar velocity distribution in the solar neighborhood. *Astron. J.* **130** 576.

Rabiner, L. and Biing-Hwang, J. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall, New York.

Rasmussen, C. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12* (S. A. Solla, T. K. Leen and K. R. Muller, eds.) 554–560. MIT Press, Cambridge.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B Methodol. Stat.* **59** 731–792. MR1483213

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14** 465.

Roberts, S. J., Husmeier, D., Rezek, I. and Penny, W. (1998). Bayesian approaches to Gaussian mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 1133.

Schafer, D. W. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrika* **80** 899.

Schafer, D. W. and Purdy, K. G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika* **83** 813–824. MR1440046

22   J. BOVY, D. W. HOGG AND S. T. ROWEIS

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, Boca Raton, FL. MR0848134

SKULJAN, J., HEARNSHAW, J. B. and COTTRELL, P. L. (1999). Velocity distribution of stars in the solar neighbourhood. *Mon. Not. R. Astron. Soc.* **308** 731.

STAUDENMAYER, J., RUPPERT, D. and BUONACCORSI, J. (2008). Density estimation in the presence of heteroscedastic measurement error. *J. Amer. Statist. Assoc.* **103** 726–736. MR2524005

STEFANSKI, L. A. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861

STONE, M. (1974). Cross-validation choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol. Stat.* **36** 111–147. MR0356377

UEDA, N., NAKANO, R., GHAHRAMANI, Z. and HINTON, G. E. (1998). 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* 274–283. IEEE.

VAN LEEUWEN, F. (2007a). *Hipparcos, the New Reduction of the Raw Data. Astrophysics and Space Science Library* **250**. Springer, Dordrecht.

VAN LEEUWEN, F. (2007b). Validation of the new Hipparcos reduction. *Astron. Astrophys.* **474** 653.

WALLACE, C. S. and BOULTON, D. M. (1968). An information measure for classification. *Comput. J.* **11** 185.

WU, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

ZHANG, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806–831. MR1056338

J. BOVY
D. W. HOGG
CENTER FOR COSMOLOGY
  AND PARTICLE PHYSICS
DEPARTMENT OF PHYSICS
NEW YORK UNIVERSITY
4 WASHINGTON PLACE
NEW YORK, NEW YORK 10003
USA
E-MAIL: jo.bovy@nyu.edu
     david.hogg@nyu.edu

S. T. ROWEIS
COURANT INSTITUTE
  OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
251 MERCER STREET
NEW YORK, NEW YORK 10012
USA
E-MAIL: roweis@cs.nyu.edu