The Astrophysical Journal, 629:268–275, 2005 August 10

MODELING COMPLETE DISTRIBUTIONS WITH INCOMPLETE OBSERVATIONS: THE VELOCITY ELLIPSOID FROM *HIPPARCOS* DATA

David W. Hogg,[1] Michael R. Blanton,[1] Sam T. Roweis,[2] and Kathryn V. Johnston[3]

Received 2004 August 30; accepted 2005 May 3

ABSTRACT

An algorithm is developed to model the three-dimensional velocity distribution function of a sample of stars using only measurements of each star's two-dimensional tangential velocity. The algorithm works with "missing data": it reconstructs the three-dimensional distribution from data (velocity measurements) that all have one dimension that is unmeasured (the radial direction). It also accounts for covariant measurement uncertainties on the tangential velocity components. The algorithm is applied to tangential velocities measured in a kinematically unbiased sample of 11,865 stars taken from the *Hipparcos* catalog, chosen to lie on the main sequence and have well-measured parallaxes. The local stellar velocity distribution function of each of a set of 20 color-selected subsamples is modeled as a mixture of two three-dimensional Gaussian ellipsoids of arbitrary relative responsibility. In the fitting, one Gaussian (the "halo") is fixed at the known mean velocity and velocity variance tensor of the Galaxy halo, and the other (the "disk") is allowed to take an arbitrary mean and an arbitrary variance tensor. The mean and variance tensors (commonly known as the "velocity ellipsoid") of the disk velocity distribution are both found to be strong functions of stellar color, with long-lived populations showing larger velocity dispersion, slower mean rotation velocity, and smaller vertex deviation than short-lived populations. The local standard of rest (LSR) is inferred in the usual way, and the Sun's motion relative to the LSR is found to be $(U, V, W)_{\odot} = (10.1, 4.0, 6.7) \pm (0.5, 0.8, 0.2)$ km s^{-1}. Artificial data sets are made and analyzed, with the same error properties as the *Hipparcos* data, to demonstrate that the analysis is unbiased. The results are shown to be insensitive to the assumption that the velocity distributions are Gaussian.

Subject headings: Galaxy: fundamental parameters — Galaxy: kinematics and dynamics — methods: statistical — solar neighborhood — stars: kinematics

1. INTRODUCTION

The classical picture of the evolution of the velocity structure in the Galactic disk is that stars are born within low-dispersion clusters from cool gas on near-circular orbits. These clusters evaporate, and the stellar orbit distribution is heated through gravitational perturbations to the smooth disk potential. Over time a stellar population's velocity dispersion grows, and its mean motion lags behind that of pure circular orbits at the same galactocentric radius. Thus, the velocity distribution of stars in the solar neighborhood has been characterized as an ellipsoid whose centroid, size, and orientation varies systematically with the lifetimes (and hence colors) of the stars under investigation (e.g., Dehnen & Binney 1998).

This field has undergone a recent renaissance with the release of the *Hipparcos* data set of proper motions and parallaxes, measured with accuracies of a few milliarcseconds. Studies using these data to analyze the local velocity distribution of stars can be broadly split into two categories.

First, there are determinations of the moments of the velocity distribution as a function of color, assuming (as above) that it can be described by a mean velocity and a single velocity dispersion tensor (Dehnen & Binney 1998; Bienaymé 1999). These have led to more stringent limits on the solar motion relative to a (hypothetical) zero-dispersion population (the local standard of rest [LSR]), the age of the Galactic disk, and rate of heating of stellar populations (Binney et al. 2000).

Second, there are nonparametric derivations of the full three-dimensional velocity distribution function (Dehnen 1998; Skuljan et al. 1999; Chereul et al. 1998). These have revealed that the velocity distribution is poorly described by a single ellipsoid; in fact it contains significant structures on smaller velocity scales. Importantly, the structures do not seem to be dominated by short-lived stars. These structures can be variously interpreted as trails from evaporating clusters (Chereul et al. 1999) or overdensities induced by resonances in the disk associated with the bar (Dehnen 2000; Fux 2001) and/or spiral arms (Quillen 2003).

Our long-term goal is to pursue the latter category of project, i.e., to develop algorithms to locate, understand the significance of, and characterize nontrivial structures in the velocity distribution, not just in the local Galaxy but in the Galaxy halo. These projects will require new space-based astrometry data (e.g., what we expect from the upcoming *Gaia* mission) in combination with large ground-based surveys (e.g., Sloan Digital Sky Survey, York et al. 2000; Two Micron All Sky Survey, Skrutskie et al. 1997; Grid Giant Star Survey, Majewski et al. 2000). In the short term, we have begun by analyzing the *Hipparcos* data set with a very general algorithm for fitting distribution functions to data measured with nontrivial error covariances and missing information.

From a computer science or nonlinear statistics perspective, these problems fall into the category of "missing data" problems, in which one constructs a model of an object (here the velocity distribution function) using data points (here tangential velocities), every one of which is incomplete (because, in this case, there is no radial information). We present a framework for a large set of algorithms for solving such problems and the details of the specific restriction to the velocity distribution function as measured with velocities projected onto the sphere.

In this paper we further restrict our attention on the trial problem of rederiving the properties of the velocity ellipsoid near the

[1] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003; david.hogg@nyu.edu.

[2] Department of Computer Science, University of Toronto, Toronto, ON M5S 364, Canada.

[3] Department of Astronomy, Wesleyan University, Middletown, CT 06459.

Sun. In what follows, it is assumed that any color-selected, kinematically unbiased sample of stars has a velocity distribution function that can be modeled (for the purposes of measuring its velocity variance) by a sum of two Gaussian ellipsoids, one for "halo stars" and one for "disk stars," and later by a sum or mixture of $K > 2$ Gaussian ellipsoids. Model parameters are chosen to maximize the total likelihood of the *Hipparcos* measurements (which, in this case, are two-dimensional tangential velocity vectors), given their uncertainties (which are two-dimensional covariance tensors); i.e., the results presented here represent the optimization of an explicit, justified, scalar objective function. Our work differs from previous work in several respects: we have the scalar objective function, we present tests of the algorithm with relatively realistic artificial data, and we relax the Gaussian assumption (i.e., expand the space of allowed distribution functions).

In future papers in this series, we intend to generalize our parameterization (to multimodal disk distributions), locate velocity-space structures, measure their statistical significance, and characterize their properties. This phenomenology will be essential for distinguishing the various pictures for the origin of the velocity substructure in the Galaxy disk.

2. MODEL AND ALGORITHM

In what follows, the standard Galactic velocity coordinate system is used, with directions x, y, and z (and associated unit vectors $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$) pointing toward the Galactic center, in the direction of circular orbital motion, and toward the north Galactic pole, respectively. Vectors are implicitly defined to be column matrices, so $\boldsymbol{a}^T\boldsymbol{b}$ is the scalar product and $\boldsymbol{a}\boldsymbol{b}^T$ is a rank 2 tensor. The components $\hat{\boldsymbol{x}}^T\boldsymbol{v}$, $\hat{\boldsymbol{y}}^T\boldsymbol{v}$, and $\hat{\boldsymbol{z}}^T\boldsymbol{v}$ of a velocity $\boldsymbol{v}$ are conventionally named U, V, and W.

We treat any color-selected population of stars from *Hipparcos* as being composed of two kinematically distinct populations of stars, a "halo" population with velocity distribution described by a Gaussian ellipsoid in velocity space with a mean velocity $\boldsymbol{v}_{\rm halo}$ with respect to the Sun and velocity dispersion (variance) tensor $\boldsymbol{V}_{\rm halo}$, with these parameters fixed at (Sirko et al. 2004)

$$\boldsymbol{v}_{\rm halo} = \left(-220 \text{ km s}^{-1}\right)\hat{\boldsymbol{y}},$$
$$\boldsymbol{V}_{\rm halo} = \left(100 \text{ km s}^{-1}\right)^2\left(\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^T + \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T + \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^T\right), \qquad (1)$$

plus a "disk" population described by another Gaussian ellipsoid with mean $\boldsymbol{v}_{\rm disk}$ and dispersion tensor $\boldsymbol{V}_{\rm disk}$, both of which are allowed to vary arbitrarily. The relative amplitude $\alpha_{\rm halo}$ of the halo Gaussian (i.e., the fraction of stars in the halo) is also allowed to vary arbitrarily. Sensitivity of the results to the assumed halo velocity dispersion of 100 km s^{-1} is discussed below.

The vast majority ($\sim$99%) of the sample is expected to be members of the disk population. However, the inclusion of a halo Gaussian prevents halo stars from distorting the measurement of the disk velocity variance. In effect, the halo Gaussian "clips out" velocity outliers in a responsible way.

Almost all the difficulty in inferring the parameters of this model, i.e., $\boldsymbol{v}_{\rm disk}$ (three parameters), $\boldsymbol{V}_{\rm disk}$ (six parameters), and the relative responsibility of the halo Gaussian (one parameter), comes from the fact that *Hipparcos* does not measure the total three-space velocity $\boldsymbol{v}$ of each star, but only its two-dimensional tangential projection.

2.1. *Model Generalities*

The approach developed here is extremely general and can be applied to many different density estimation tasks in the presence of partially observed data. The assumption is that there are low-dimension observations $\boldsymbol{w}_i$, which are noisy projections of higher dimension "true values" $\boldsymbol{v}_i$,

$$\boldsymbol{w}_i = \boldsymbol{R}_i\boldsymbol{v}_i + \text{noise}, \qquad (2)$$

where the $\boldsymbol{R}_i$ are known, nonsquare (or zero determinant) projection matrices and the noise is drawn from a Gaussian with zero mean and known (low dimension) covariance tensor $\boldsymbol{S}_i$. It is also assumed that the $\boldsymbol{v}_i$ are drawn independently and identically distributed from a probability distribution function $p(\boldsymbol{v})$ in the higher dimension space. The goal is to fit a model for $p(\boldsymbol{v})$ using only the incomplete observations ($\boldsymbol{w}_i$), their covariances ($\boldsymbol{S}_i$), and the projection matrices ($\boldsymbol{R}_i$).

Note that there is no assumption that all data points have similar nonsquare projection matrices; in fact the projection matrices (and thus the observations) may have different dimensionalities.

The density model $p(\boldsymbol{v})$ is parameterized as a mixture of K Gaussians:

$$p(\boldsymbol{v}) = \sum_{j=1}^{K}\alpha_j N(\boldsymbol{v}|\boldsymbol{m}_j, \boldsymbol{V}_j), \qquad (3)$$

where the amplitudes or "rates" α_j sum to unity and the function $N(\boldsymbol{v}|\boldsymbol{m}, \boldsymbol{V})$ is the normal (Gaussian) distribution with mean $\boldsymbol{m}$ and variance tensor $\boldsymbol{V}$.

For a known projection matrix $\boldsymbol{R}_i$ and noise covariance $\boldsymbol{S}_i$ in $\boldsymbol{w}$-space (the lower dimensional space of the observations), each component of the mixture marginalizes to a lower dimensional Gaussian, and so the induced density is a conditional mixture of Gaussians on $\boldsymbol{w}$:

$$p(\boldsymbol{w}, \boldsymbol{v}, j) = p(\boldsymbol{w}|\boldsymbol{v})p(\boldsymbol{v}|j)p(j),$$
$$p(\boldsymbol{w}|\boldsymbol{R}, \boldsymbol{S}) = \sum_j\int_{\boldsymbol{v}} p(\boldsymbol{w}|\boldsymbol{v})p(\boldsymbol{v}|j)p(j)\,d\boldsymbol{v},$$
$$p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{R}, \boldsymbol{S}) = N(\boldsymbol{w}|\boldsymbol{R}\boldsymbol{v}, \boldsymbol{S}),$$
$$p(\boldsymbol{w}_i|\boldsymbol{R}_i, \boldsymbol{S}_i) = \sum_{j=1}^{K}\alpha_j N\left(\boldsymbol{w}_i|\boldsymbol{R}_i\boldsymbol{m}_j, \boldsymbol{T}_{ij}\right),$$
$$\boldsymbol{T}_{ij} = \boldsymbol{R}_i\boldsymbol{V}_j\boldsymbol{R}_i^T + \boldsymbol{S}_i, \qquad (4)$$

where functions like $p(x, y, z)$ are joint probability distribution functions of x, y, and z and functions like $p(x|y)$ are probability distribution functions of x given (or at a specific value of) y. All other symbols are described above, except $\boldsymbol{T}_{ij}$, which is the combined variance for each measurement i under the assumption that it is drawn from Gaussian j, with part of the variance coming from the (projected) variance $\boldsymbol{V}_j$ of the Gaussian and part coming from the measurement uncertainty variance $\boldsymbol{S}_i$. This model is called the "projected mixture of Gaussians" model hereafter.

The objective of the fitting procedure is to maximize the conditional likelihood of the entire set of low-dimensional projected observations, given the nonsquare matrices and the error covariances. In particular, we are fitting for the means ($\boldsymbol{m}_j$), variance tensors ($\boldsymbol{V}_j$), and amplitudes (α_j) of the mixture of Gaussians in the high-dimensional (unobserved) space. Assuming the noise on each observation is independent of other observation noises, this (log) likelihood is

$$\phi = \sum_i \ln p(\boldsymbol{w}_i|\boldsymbol{R}_i, \boldsymbol{S}_i) = \sum_i \ln\sum_{j=1}^{K}\alpha_j N\left(\boldsymbol{w}_i|\boldsymbol{R}_i\boldsymbol{m}_j, \boldsymbol{T}_{ij}\right). \qquad (5)$$

The model parameters can be optimized in several ways. One approach is to directly compute gradients and use a generic optimizer to ascend the objective; this is complicated by restrictions on the parameters (e.g., the variances must be symmetric and positive definite, the amplitudes must be nonnegative and sum to unity). Another approach is to view the high-dimensional quantities as hidden variables and use the expectation-maximization (EM) algorithm (Dempster et al. 1977) to iteratively maximize the likelihood function. We take the latter approach; for details see the Appendix.

2.2. *Hipparcos Measurements and Their Uncertainties*

The sample used in this study is a kinematically unbiased sample of 11,865 nearby main-sequence stars (Dehnen & Binney 1998) from the *Hipparcos* catalog (ESA 1997), all chosen to have parallaxes measured at a signal-to-noise ratio of $\pi/\sigma_\pi > 10$. We made no corrections for Galactic rotation (since this study is simply of stellar velocities relative to the Sun); indeed we did no processing or correction of the *Hipparcos* data beyond making the sample cut.

The "low-dimension" data $\boldsymbol{w}_i$ referred to above are the measured tangential components of each star's "high-dimension" true three-dimensional velocity $\boldsymbol{v}_i$, with

$$\boldsymbol{w}_i = \boldsymbol{R}_i \boldsymbol{v}_i,$$
$$\boldsymbol{R}_i = \left(\hat{\boldsymbol{l}}_i \hat{\boldsymbol{l}}_i^T + \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^T\right), \qquad (6)$$

where the $\boldsymbol{R}_i$ are nonsquare or zero-determinant matrices and $\hat{\boldsymbol{l}}_i$ and $\hat{\boldsymbol{b}}_i$ are the tangential unit vectors pointing in the Galactic latitude and longitude directions for each star.

The $\boldsymbol{w}_i$ are constructed from the *Hipparcos* measurements (parallax and proper motion) as

$$\boldsymbol{w}_i = \frac{r_0}{\pi_i}\left(\cos b_i \frac{dl_i}{dt}\hat{\boldsymbol{l}}_i + \frac{db_i}{dt}\hat{\boldsymbol{b}}_i\right), \qquad (7)$$

where r_0 is the radius of the Earth's orbit, π_i is the star's parallax, l_i and b_i are its Galactic latitude and longitude, and the cos b_i factor takes into account the spherical geometry. This calculation ignores the Lutz-Kelker bias (Lutz & Kelker 1973), but this is small for the sample used here. Since the *Hipparcos* catalog reports proper motions in equatorial rather than Galactic coordinates, the above requires a rotation depending on each star's angular position on the sky and the epoch (1991.25) of the catalog positions.

The *Hipparcos* catalog entries, which can be represented by some vector $\boldsymbol{c}_i$ for each star, come with single-star uncertainty covariance matrices $\boldsymbol{C}_i$. If we can represent the derivative of the $\boldsymbol{w}_i$ with respect to the catalog entries $\boldsymbol{c}_i$ by a matrix $\boldsymbol{Q}_i$,

$$d\boldsymbol{w}_i = \boldsymbol{Q}_i d\boldsymbol{c}_i, \qquad (8)$$

then the measurement uncertainty covariances $\boldsymbol{S}_i$ for the $\boldsymbol{w}_i$ are given by

$$\boldsymbol{S}_i = \boldsymbol{Q}_i \boldsymbol{C}_i \boldsymbol{Q}_i^T. \qquad (9)$$

This is accurate only in the limit of small parallax errors, which is fine for this sample. In addition, this whole procedure ignores star-to-star covariances, which could be significant but are not reported in the *Hipparcos* catalog.

3. RESULTS

Following the general approach of Dehnen & Binney (1998), 20 color-selected subsamples of stars (≈594 stars each) were made by cutting the color-sorted star list into equal-sized pieces (as closely as possible), and, for each subsample, the 10 parameters $\boldsymbol{v}_{\rm disk}$, $\boldsymbol{V}_{\rm disk}$, and $\alpha_{\rm halo}$ were found by the optimization described above. Figure 1 shows the 10 parameters for each of the 20 subsamples. The vertical error bars on the points indicate uncertainties computed with 20 independent bootstrap resamplings of the data in each of the subsamples. Redder (and therefore longer lived) stellar populations have larger velocity dispersions.

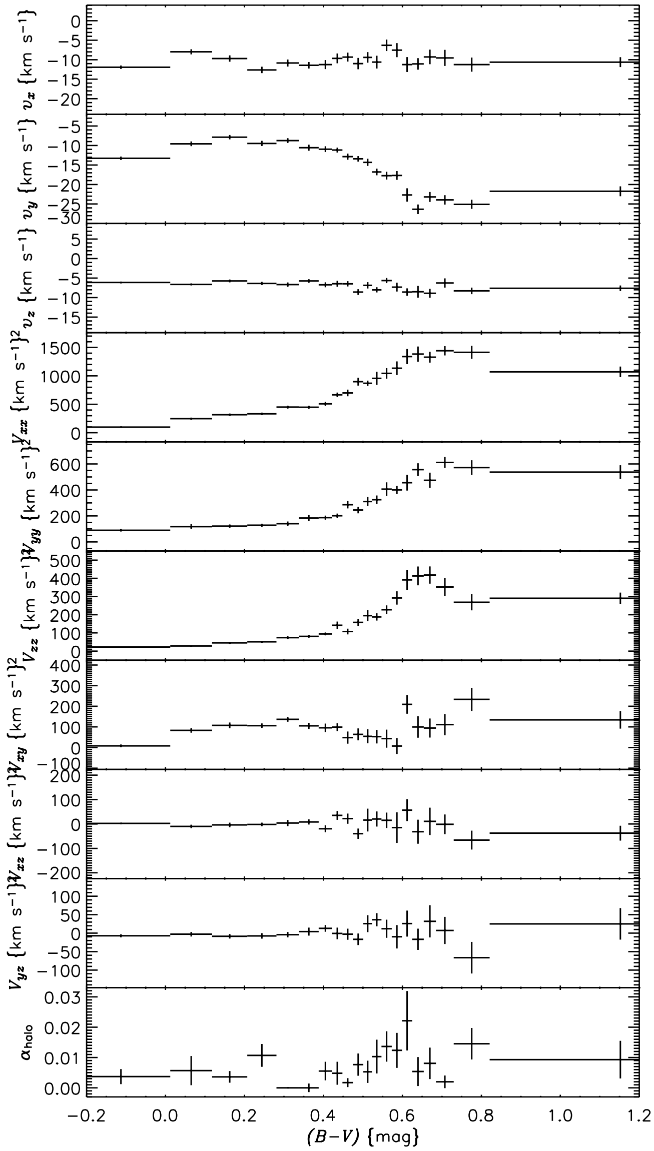

FIG. 1.—Best-fit parameters of the model as a function of stellar color for the 20 color-selected subsamples. The off-diagonal elements of the velocity variance tensor $\boldsymbol{V}_{\rm disk}$ have been scaled by square roots of products of the diagonal elements.

Table 1 gives the 10 parameters, uncertainties, and uncertainty correlation matrix for one of the subsamples. Figure 2 shows the mean $\boldsymbol{v}_{\rm disk}$ of the disk velocity distribution as a function of the trace Tr($\boldsymbol{V}_{\rm disk}$) of its variance tensor $\boldsymbol{V}_{\rm disk}$. The mean velocity in the $\hat{\boldsymbol{y}}$-direction is a strong function of velocity dispersion; this linear dependence justifies the standard methodology for determination of the LSR.

Operationally, the LSR is defined to be the mean velocity for a hypothetical population of stars with zero velocity distribution, i.e., the extrapolation to $\mathrm{Tr}(\boldsymbol{V}_{\rm disk}) = 0$ of the trend shown in Figure 2. The points in Figure 2 have significant uncertainties in both dimensions, so fitting a line responsibly is not trivial. For

TABLE 1

Example Parameter Set for the Subsample with $0.654 < (B - V) < 0.685$

Parameter	Value	Units	Correlation Matrix of the (Squared) Uncertainties									
$\hat{x}^T v_{\rm disk}$	−9.3 ± 1.9	km s^{-1}	1.00	0.20	0.05	0.10	−0.14	0.38	−0.49	−0.30	0.16	−0.53
$\hat{y}^T v_{\rm disk}$	−23.2 ± 1.3	km s^{-1}	0.20	1.00	0.39	−0.23	−0.37	−0.08	0.12	0.18	−0.07	0.24
$\hat{z}^T v_{\rm disk}$	−8.9 ± 1.1	km s^{-1}	0.05	0.39	1.00	−0.42	−0.17	0.03	0.09	0.28	−0.30	−0.02
$\hat{x}^T V_{\rm disk}\hat{x}$	1329.0 ± 95.0	km^2 s^{-2}	0.10	−0.23	−0.42	1.00	0.37	−0.04	−0.10	−0.19	0.36	−0.24
$\hat{y}^T V_{\rm disk}\hat{y}$	474.0 ± 58.0	km^2 s^{-2}	−0.14	−0.37	−0.17	0.37	1.00	−0.30	0.21	0.07	0.53	−0.02
$\hat{z}^T V_{\rm disk}\hat{z}$	418.0 ± 47.0	km^2 s^{-2}	0.38	−0.08	0.03	−0.04	−0.30	1.00	0.04	0.02	−0.05	−0.47
$\hat{x}^T V_{\rm disk}\hat{y}$	95.0 ± 46.0	km^2 s^{-2}	−0.49	0.12	0.09	−0.10	0.21	0.04	1.00	0.27	0.00	0.15
$\hat{x}^T V_{\rm disk}\hat{z}$	11.0 ± 56.0	km^2 s^{-2}	−0.30	0.18	0.28	−0.19	0.07	0.02	0.27	1.00	−0.54	0.19
$\hat{y}^T V_{\rm disk}\hat{z}$	32.0 ± 44.0	km^2 s^{-2}	0.16	−0.07	−0.30	0.36	0.53	−0.05	0.00	−0.54	1.00	0.05
$\alpha_{\rm halo}$	0.0081 ± 0.0052	...	−0.53	0.24	−0.02	−0.24	−0.02	−0.47	0.15	0.19	0.05	1.00

this purpose we again use the projected mixtures of Gaussians procedure described above, but now there are 19 two-dimensional data points $\boldsymbol{w}_i$, the $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ are the same (i.e., the $\boldsymbol{R}_i$ are the identity matrices), and we only fit a single Gaussian ellipsoid. The straight line shown in the $\hat{y}^T v_{\rm disk}$ (v_y) panel of Figure 2 is the principal eigenvector of the best-fit Gaussian. The $\hat{x}^T v_{\rm disk}$ and $\hat{z}^T v_{\rm disk}$ (v_x and v_z) panels show simply weighted averages. The errors in the fit are computed by bootstrap resampling the 19 samples themselves.

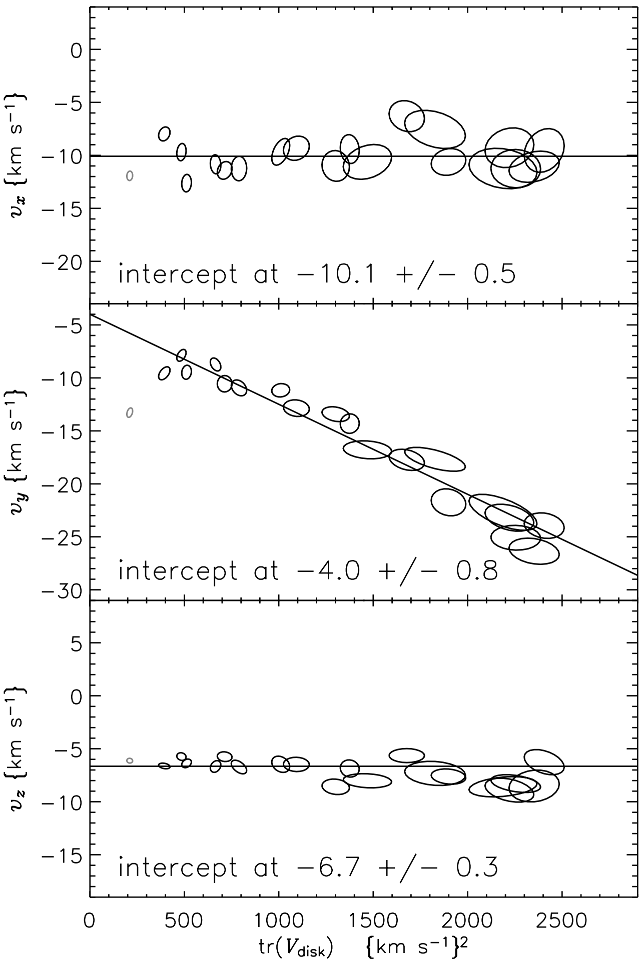

Fig. 2.—Mean velocity $\boldsymbol{v}_{\rm disk}$ as a function of total velocity variance Tr($\boldsymbol{V}_{\rm disk}$) for the determination of the LSR. In each panel, the ellipses indicate the 1 σ uncertainty regions (from bootstrap resampling—see text) of the measurements. The linear fit of V vs. S^2 was performed with the projected Gaussian mixtures algorithm because it accounts correctly for the finite errors in both dimensions (see text). The point shown in gray was excluded from the fit because the stars in that subsample are very short lived (see text). The reported uncertainties on the intercepts are from 20 bootstrap resamplings of the ellipsoidal points shown.

The fits shown in Figure 2 provide an intercept corresponding to the estimated velocity relative to the Sun of a hypothetical population with vanishing velocity dispersion. The velocity of the Sun relative to this LSR is therefore

$$\boldsymbol{v}_\odot = -\boldsymbol{v}_{\rm LSR}, \tag{10}$$

$$\boldsymbol{v}_\odot = \left(10.1 \pm 0.5 \text{ km s}^{-1}\right)\hat{\boldsymbol{x}} + \left(4.0 \pm 0.8 \text{ km s}^{-1}\right)\hat{\boldsymbol{y}} + \left(6.7 \pm 0.2 \text{ km s}^{-1}\right)\hat{\boldsymbol{z}}. \tag{11}$$

Recall that rejection of halo stars was accomplished by fitting the velocity field with two Gaussians, one of which was fixed at the halo parameters given in equation (1). Refitting with the halo velocity dispersion increased to 150 km s^{-1} changes the inferred LSR by much less than the magnitude of its uncertainty.

The vertex deviation is defined to be the angle between the x-axis and the projection onto the x-y plane of the eigenvector corresponding to the largest eigenvalue of the velocity variance tensor $\boldsymbol{V}_{\rm disk}$. The vertex deviation is shown as a function of stellar color for the 20 subsamples in Figure 3.

4. ALGORITHM TESTS

To test the algorithm, 20 subsamples of artificial three-dimensional stellar velocities $\boldsymbol{v}_i$ were generated with a mixtures of Gaussians random sample generator. The distribution was made with two Gaussians, one for the halo, with parameters as assumed above, and one for the disk, with mean $\boldsymbol{v}_{\rm disk}$ and variance tensor $\boldsymbol{V}_{\rm disk}$ different for each subsample. The artificial $\boldsymbol{v}_i$ were projected into artificial measurements $\boldsymbol{w}_i$ using the same $\boldsymbol{R}_i$ as in the real 20 subsamples, and errors were added, drawn from two-dimensional Gaussian ellipsoids with the same variances as

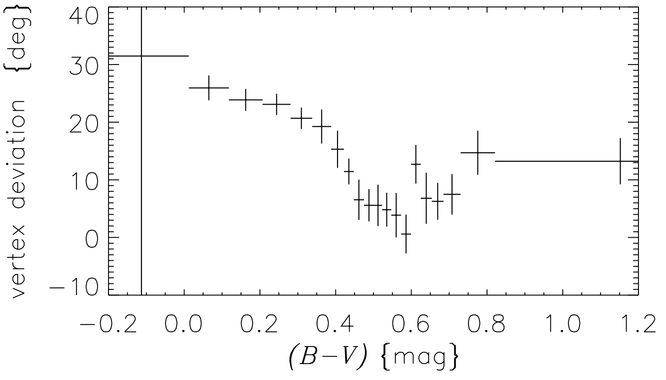

Fig. 3.—Vertex deviation (see text for definition) as a function of color.

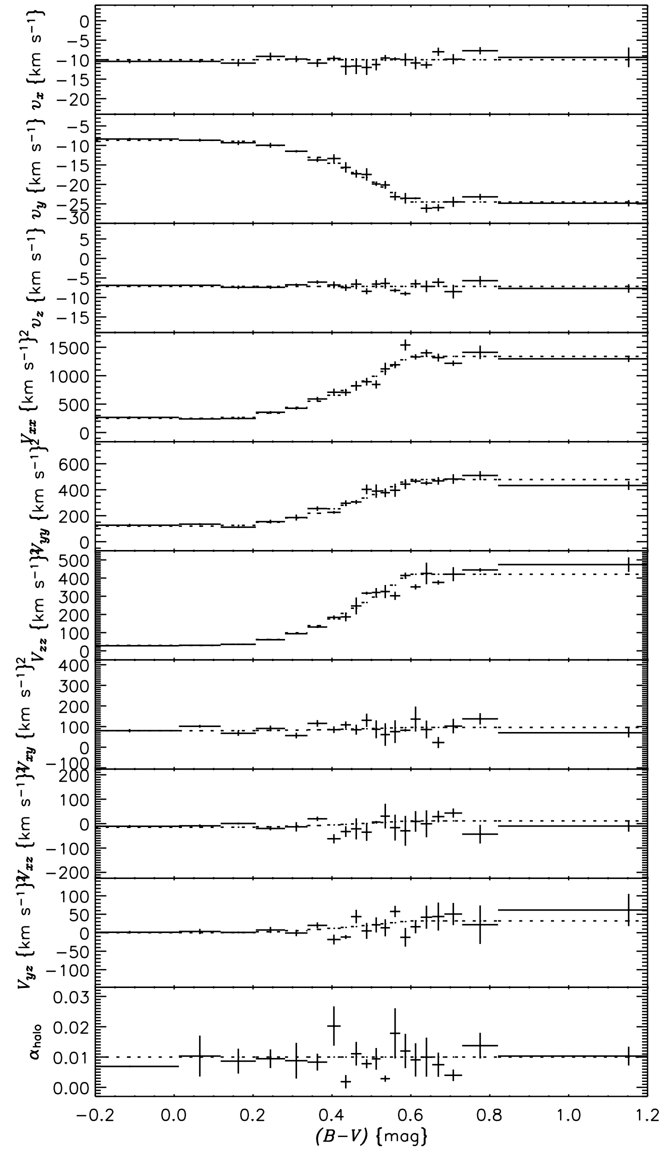

FIG. 4.—Same as Fig. 1, but for the artificial data (see text). The dotted lines indicate the input values used to make the artificial data.

the measurement uncertainty covariance tensors $\boldsymbol{S}_i$; i.e., the artificial data were given all the observational properties of the real data (modulo the assumptions of this study).

For subsamples with mean color $(B-V) < 0.1$ mag, the artificial variance tensor $\boldsymbol{V}_{\rm disk}$ was set to the measured value (shown in Fig. 1) for the subsample with $(B-V) \approx 0.05$ mag, and for subsamples with mean $(B-V) > 0.6$ mag, the artificial tensor $\boldsymbol{V}_{\rm disk}$ was set to the measured value for the subsample with $(B-V) \approx 0.67$ mag. In between, i.e., for artificial subsamples with mean color 0.1 mag $< (B-V) < 0.6$ mag, the variance tensor was made to vary quadratically with color, so as to approximate the appearance of the true observations (and span the range of observed variance tensors. The artificial mean $\boldsymbol{v}_{\rm disk}$ was set to a linear function of the trace $\mathrm{Tr}(\boldsymbol{V}_{\rm disk})$ of the variance.

Exactly the same fitting code and bootstrap analysis were applied to the artificial data as were applied to the real data. The results are shown in Figures 4 and 5, along with the input values used to make the artificial data. The best-fit parameters are, except for a couple of samples, within one standard deviation of the input parameters. More importantly for present purposes, the LSR is very well determined; the algorithm returns the correct LSR velocity to well within one standard deviation. We conclude that the algorithm is not significantly biased.

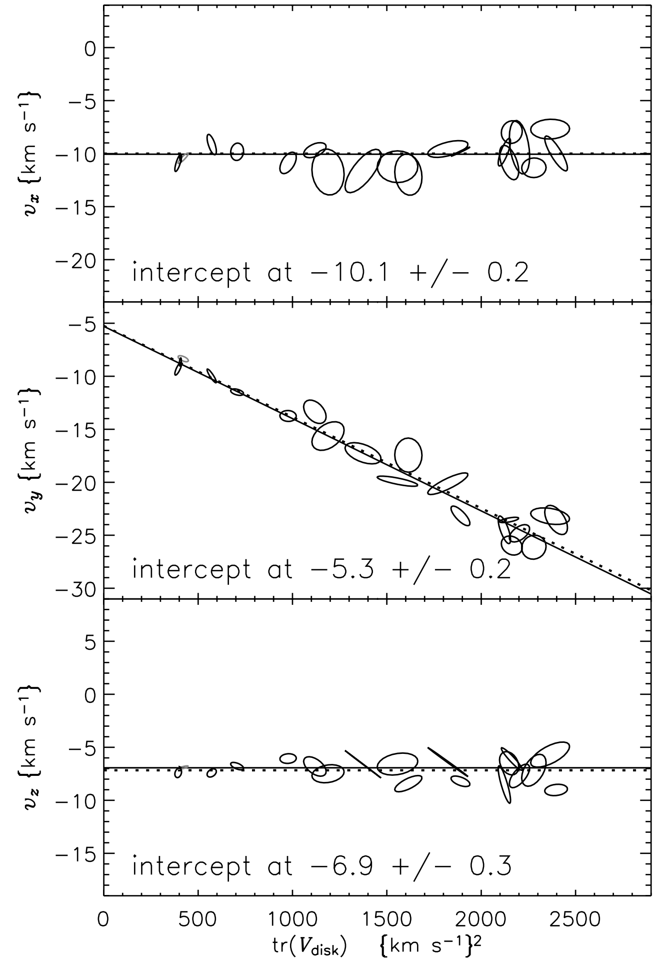

FIG. 5.—Same as Fig. 2, but for the artificial data (see text). In each panel there is a dotted line indicating the input values used to make the artificial data.

5. GENERALIZED MULTI-GAUSSIAN DISK

Perhaps the greatest limitation of LSR measurements like this one is that the disk velocity distribution function is far from Gaussian; it is not even unimodal (Dehnen 1998; Skuljan et al. 1999; Chereul et al. 1998). One of the primary goals of our future work is to explore the complexities of disk star velocities. As a baby step toward checking the influence of disk velocity non-Gaussianity on the LSR determination, the model was generalized to allow for not just one Gaussian ellipsoid to fit the each color subsample's disk velocity distribution function but $K_{\rm disk} > 1$ Gaussians, all constrained to have the same mean. Models with $K_{\rm disk} > 1$ have the freedom to have larger "tails" to the velocity distribution and for those tails to be rotated or twisted in velocity space relative to the core of the velocity distribution.

The generalized model is optimized by an algorithm constructed exactly parallel to that of the $K_{\rm disk} = 1$ model, but now there are $4 + 6K_{\rm disk}$ free parameters for each subsample.

Increasing $K_{\rm disk}$ increases the goodness of fit (of course), but at very large $K_{\rm disk}$, the data will be "overfit." To determine the optimal value of $K_{\rm disk}$ for each of the 20 stellar subsamples (the optimal $K_{\rm disk}$ will, in general, be different for different subsamples, of course), a "jackknife likelihood" was computed: For each of five iterations, a randomly selected 10% of each subsample was

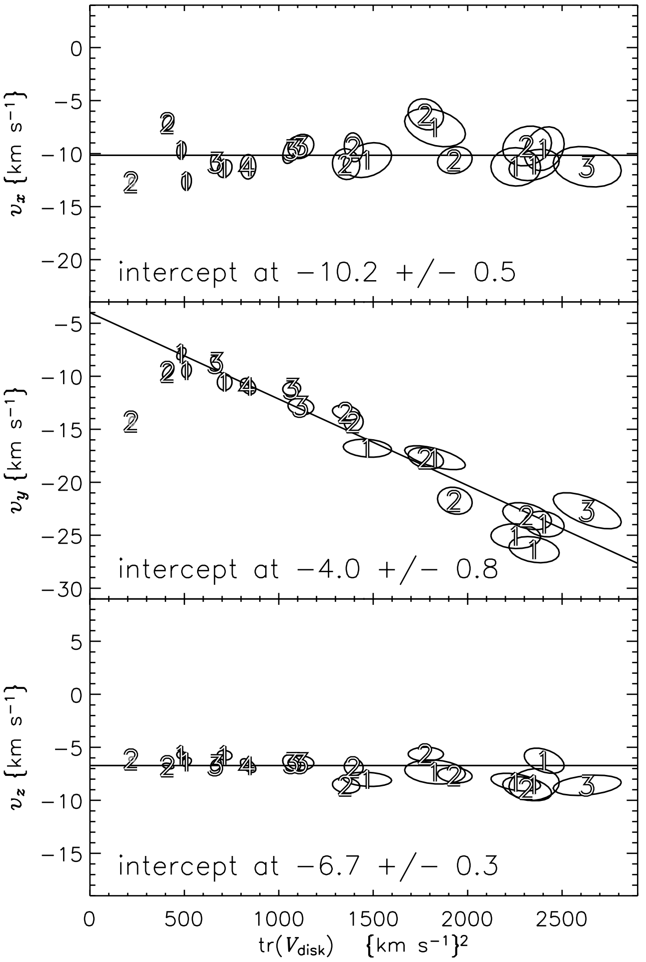

Fig. 6.—Same as Fig. 2, but for the generalized model in which the disk velocity distribution is fit with a mixture of $K_{\rm disk}$ Gaussian ellipsoids with a common mean. Each data point shows the result for that subsample for the optimal jackknife value of $K_{\rm disk}$ (see text) and is labeled by that value of $K_{\rm disk}$.

removed and put aside as a "test set." Fitting (i.e., parameter determination by maximum likelihood) was performed on the remaining 90%, and the likelihood of the test set was tested within the context of the best-fit model. The logarithms of the jackknife likelihoods for the five iterations were averaged, and the $K_{\rm disk}$ with the best jackknife likelihood was chosen for each subsample.

Figure 6 shows the LSR determination when the generalized model is used and each sample is fit with the optimal jackknife likelihood value of $K_{\rm disk}$. The velocity of the Sun relative to the LSR that we find when using the optimal $K_{\rm disk}$ values is

$$\boldsymbol{v}_\odot = \left(10.2 \pm 0.5 \text{ km s}^{-1}\right)\hat{\boldsymbol{x}} + \left(4.0 \pm 0.8 \text{ km s}^{-1}\right)\hat{\boldsymbol{y}} + \left(6.7 \pm 0.2 \text{ km s}^{-1}\right)\hat{\boldsymbol{z}}. \qquad (12)$$

This is extremely similar to (much closer than one standard deviation away from) that found using $K_{\rm disk} = 1$ (Fig. 2). This suggests that the assumption of Gaussianity is not strongly affecting the results.

6. DISCUSSION

Above we developed and used a novel algorithm to infer the three-dimensional velocity distribution from a kinematically unbiased sample of *Hipparcos* stars.

The local velocity dispersion is a strong function of stellar color, and the mean velocity of a color-selected stellar population is a linear function of its velocity variance; this confirms previous results (e.g., Dehnen & Binney 1998). The extrapolation of this relation to zero velocity dispersion provides an estimate of the LSR, which is found to be

$$\boldsymbol{v}_\odot = \left(10.1 \pm 0.5 \text{ km s}^{-1}\right)\hat{\boldsymbol{x}} + \left(4.0 \pm 0.8 \text{ km s}^{-1}\right)\hat{\boldsymbol{y}} + \left(6.7 \pm 0.2 \text{ km s}^{-1}\right)\hat{\boldsymbol{z}}, \qquad (13)$$

where $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$ are unit vectors pointing in the directions of the standard Galactic velocity components *U*, *V*, and *W*. This result is similar to previous LSR determinations; we compare this result to one previous study below. Our answer did not change much when we relaxed the assumption that the disk star velocity distributions can be modeled as Gaussians.

We also showed that it is possible to robustly and reliably solve a missing data problem in astrophysics: the reconstruction of aspects of the three-dimensional velocity distribution function from individual velocity measurements, every one of which is missing data in the radial direction.

Dehnen & Binney (1998), using the same subsample of the same data set, find a somewhat different solar velocity relative to the LSR; they find

$$\boldsymbol{v}_\odot = \left(10.00 \pm 0.36 \text{ km s}^{-1}\right)\hat{\boldsymbol{x}} + \left(5.25 \pm 0.62 \text{ km s}^{-1}\right)\hat{\boldsymbol{y}} + \left(7.17 \pm 0.38 \text{ km s}^{-1}\right)\hat{\boldsymbol{z}}. \qquad (14)$$

They also find a lower mean velocity variance for the long-lived disk stars. These two differences are probably related, since the LSR is determined by fitting the relationship between velocity and velocity variance. Although the studies agree to within about one standard deviation, better agreement might be expected, since both studies are using identical data subsets of the same data set. Both studies have made the incorrect assumption that the stellar velocity distribution is Gaussian; Dehnen & Binney (1998) did so in subtracting a measurement uncertainty variance from the measured velocity variance. The method presented here has been shown (using the generalized multi-Gaussian disk model) to be insensitive to the Gaussianity of the velocity distribution. The incorrect assumption of Gaussianity, entering differently in the Dehnen & Binney (1998) investigation, may account for the difference between the results. Because our method involves the optimization of a well-defined objective, because we have tested our method successfully with artificial data, and because we have been able to relax the Gaussian assumption, we prefer our result. Certainly the differences show that stellar velocity studies have become precise enough that algorithms matter. However, it must be emphasized that the true velocity distribution is far from a unimodal Gaussian (Dehnen 1998; Skuljan et al. 1999; Chereul et al. 1998), so it is not clear whether it is possible to make a "correct" LSR determination at all.

We thank the *Hipparcos* team for the generous release of the catalog. We also thank Walter Dehnen and James Binney for use of their kinematically unbiased sample and Scott Tremaine and Matias Zaldarriaga for useful discussions. In addition, D. W. H. particularly thanks Ken Banas for important lessons in statistics. This research made use of the NASA Astrophysics Data System. M. R. B. and D. W. H. are partially supported by NASA Long-Term Space Astrophysics (LTSA) grant NAG5-11669 and NSF grant PHY-0101738; K. V. J. is partially supported by NASA LTSA grant NAG5-9064 and NSF Faculty Early Career Development (CAREER) award AST-0133617. S. T. R. is funded in part by the LEARN project of the Institute for Robotics and Intelligent Systems and by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program. K. V. J. gratefully acknowledges the hospitality of the New York University Physics Department during the summer visit when this work was initialized.

APPENDIX

FITTING MIXTURES WITH INCOMPLETE DATA USING THE EM ALGORITHM

The EM algorithm (Dempster et al. 1977) can be used to optimize the likelihood function of a probabilistic model involving incomplete observation data (hidden variables). Starting from user-supplied starting parameters, its iterations generate a sequence of parameters that monotonically increase the likelihood of a fixed data set under the model; thus, it finds locally maximum likelihood parameters.

EM proceeds by optimizing, at each point in parameter space, a new function, which is a strict lower bound on the data likelihood. This new function depends on the original model parameters as well as on some extra auxiliary quantities introduced by EM. The EM algorithm iteratively increases the lower bound by coordinate ascent: first (the "M step") the original model parameters are optimized (holding the auxiliary quantities fixed), and then (the "E step") the auxiliary quantities are optimized (with the parameters fixed). After the optimization of the auxiliary quantities, the new function actually becomes equal to the true model likelihood (the bound becomes tight); thus, at each iteration this true likelihood is nondecreasing. The M and E steps are iterated to convergence, which, here as usually, is identified by extremely small incremental improvement in the logarithm of the likelihood per iteration.

In particular, for any auxiliary distribution $q(\boldsymbol{v}, j|\boldsymbol{w})$ we can lower bound the model likelihood $\ln p(\boldsymbol{w})$ by a functional $F(q)$. (In what follows, we slightly abuse notation by using j both as an index and as a random variable representing the identity of the mixture component responsible for generating a particular data point.) Thus,

$$\begin{aligned}
\ln p(\boldsymbol{w}|\theta) &= \ln \sum_j \int_{\boldsymbol{v}} p(\boldsymbol{w}, \boldsymbol{v}, j|\theta)\, d\boldsymbol{v} \\
&\geq \sum_j \int_{\boldsymbol{v}} q \ln \frac{p(\boldsymbol{w}, \boldsymbol{v}, j|\theta)}{q}\, d\boldsymbol{v} = F(\boldsymbol{w}|q, \theta) \\
&\geq F(\boldsymbol{w}|q, \theta) = \sum_j \int_{\boldsymbol{v}} q(\boldsymbol{v}, j|\boldsymbol{w})[\ln p(\boldsymbol{w}, \boldsymbol{v}, j|\theta) - \ln q(\boldsymbol{v}, j|\boldsymbol{w})]\, d\boldsymbol{v} \\
\ln p(\boldsymbol{w}|\theta) &\geq F(\boldsymbol{w}|q, \theta) = \langle \ln p(\boldsymbol{w}, \boldsymbol{v}, j|\theta)\rangle_q + H(q),
\end{aligned} \tag{A1}$$

where θ represents the set of model parameters and H is the entropy of the distribution q.

Our strategy is now coordinate maximization of F. In the E step we maximize F with respect to the auxiliary distribution q. It is easy to show (for example, by checking that it saturates the bound on F) that the maximizing distribution q is the conditional distribution of $\boldsymbol{v}$ and j, given the observations $\boldsymbol{w}$ and the current parameters:

$$\textbf{E-step}: \quad q(\boldsymbol{v}, j) \leftarrow \mathrm{argmax}_q F(\boldsymbol{w}|q, \theta) = p(\boldsymbol{v}, j|\boldsymbol{w}, \theta). \tag{A2}$$

In the M step we maximize F with respect to the parameters θ. This maximization reduces to maximization of the expected complete log likelihood under the current variational distribution (since the entropy of q does not depend on θ):

$$\textbf{M-step}: \quad \theta \leftarrow \mathrm{argmax}_\theta F(\boldsymbol{w}|q, \theta) = \mathrm{argmax}_\theta \sum_j \int_{\boldsymbol{v}} q(\boldsymbol{v}, j) \ln p(\boldsymbol{w}, \boldsymbol{v}, j|\theta)\, d\boldsymbol{v}. \tag{A3}$$

A1. EM ALGORITHM FOR PROJECTED MIXTURES

For the projected mixtures model, the parameters consist of the mixture component amplitudes α_j, means $\boldsymbol{m}_j$, and variances $\boldsymbol{V}_j$. The auxiliaries are posterior distributions for each star i: $q(j|\boldsymbol{w}_i)$ over mixture components and $q(\boldsymbol{v}|j, \boldsymbol{w}_i)$ over the true velocity (given the observed projected velocity, assuming it came from a specific component). In the terminology of this appendix, the parameters of the probability distribution for the true three-space velocity $\boldsymbol{v}$ of each star, especially including the probabilities that the star was drawn from each of the Gaussian velocity components j, are the "auxiliary quantities," and the parameters (amplitudes, means, and velocity variance tensors) of the velocity components j are the "model parameters."

First we consider the E step, in which the auxiliary distributions $q(\boldsymbol{v}, j|\boldsymbol{w}_i, \theta)$ are optimized. In the case of projected mixtures, the posterior over $\boldsymbol{v}, j$ given $\boldsymbol{w}$ and the model parameters θ is itself a conditional mixture of Gaussians:

$$\begin{aligned}
\text{E step:} \quad p(\boldsymbol{v}, j|\boldsymbol{w}) &= p(j|\boldsymbol{w})p(\boldsymbol{v}|j, \boldsymbol{w}) \\
p(j|\boldsymbol{w}_i) &= \frac{\alpha_j N\left(\boldsymbol{w}|\boldsymbol{R}\boldsymbol{m}_j, \boldsymbol{T}_j\right)}{\sum_k \alpha_k N(\boldsymbol{w}|\boldsymbol{R}\boldsymbol{m}_k, \boldsymbol{T}_k)} \\
p(\boldsymbol{v}|\boldsymbol{w}_i, j) &= N\left(\boldsymbol{v}|\boldsymbol{b}_{ij}, \boldsymbol{B}_{ij}\right) \\
\boldsymbol{b}_{ij} &= \boldsymbol{m}_j + \boldsymbol{V}_j \boldsymbol{R}_i^T \boldsymbol{T}_{ij}^{-1}\left(\boldsymbol{w} - \boldsymbol{R}_i \boldsymbol{m}_j\right) \\
\boldsymbol{B}_{ij} &= \boldsymbol{V}_j - \boldsymbol{V}_j \boldsymbol{R}_i^T \boldsymbol{T}_{ij}^{-1} \boldsymbol{R}_i \boldsymbol{V}_j^T.
\end{aligned} \tag{A4}$$

Thus, the optimal choice for $q(\boldsymbol{v}, j|\boldsymbol{w}_i, \theta)$ is

$$\begin{aligned} q(\boldsymbol{v}, j|\boldsymbol{w}_i) &= q(j|\boldsymbol{w}_i)q(\boldsymbol{v}|j, \boldsymbol{w}_i), \\ q(j|\boldsymbol{w}_i) &= q_{ij} = p(j|\boldsymbol{w}_i), \\ q(\boldsymbol{v}|\boldsymbol{w}_i, j) &= N\big(\boldsymbol{v}|\boldsymbol{b}_{ij}, \boldsymbol{B}_{ij}\big). \end{aligned} \tag{A5}$$

For the M step updates, we must explicitly write out the form of the functional F and take its partial derivatives with respect to each set of model parameters:

$$\begin{aligned} F &= \sum_i \langle \ln p(\boldsymbol{w}_i|\boldsymbol{v}_i) + \ln p(\boldsymbol{v}_i|j) + \ln p(j)\rangle_{q_i} + H(q_i) \\ &= -\frac{1}{2}\sum_i \sum_j q_{ij}\left[\Big\langle(\boldsymbol{w}_i - \boldsymbol{R}_i\boldsymbol{v})^T\boldsymbol{S}_i^{-1}(\boldsymbol{w}_i - \boldsymbol{R}_i\boldsymbol{v})(\boldsymbol{v} - \boldsymbol{m}_j)^T\boldsymbol{V}_j^{-1}(\boldsymbol{v} - \boldsymbol{m}_j)\Big\rangle_{q(\boldsymbol{v}|\boldsymbol{w}_i, j)} + \ln\det\boldsymbol{S}_i + \ln\det\boldsymbol{V}_j + \ln\alpha_i\right] \\ &= -\frac{1}{2}\sum_i \sum_j q_{ij}\Big\{\boldsymbol{w}_i^T\boldsymbol{S}_i^{-1}\boldsymbol{w}_i + \boldsymbol{m}_j^T\boldsymbol{V}_j^{-1}\boldsymbol{m}_j - 2\boldsymbol{w}_i^T\boldsymbol{S}_i^{-1}\boldsymbol{R}_i\boldsymbol{b}_{ij} - 2\boldsymbol{m}_j^T\boldsymbol{V}_j^{-1}\boldsymbol{b}_{ij} \\ &\quad + \mathrm{Tr}\Big[(\boldsymbol{R}_i^T\boldsymbol{S}_i^{-1}\boldsymbol{R}_i + \boldsymbol{V}_j^{-1})(\boldsymbol{b}_{ij}\boldsymbol{b}_{ij}^T + \boldsymbol{B}_{ij})\Big] + \ln\det\boldsymbol{S}_i + \ln\det\boldsymbol{V}_j + \ln\alpha_i\Big\}. \end{aligned} \tag{A6}$$

Deriving the E step and M step for the particular projected mixtures model given above leads to the following updated equations:

$$\begin{aligned} \text{M step:}\quad \alpha_j &\leftarrow \frac{1}{N}\sum_i q_{ij} \\ \boldsymbol{m}_j &\leftarrow \frac{1}{q_j}\sum_i q_{ij}\boldsymbol{b}_{ij} \\ \boldsymbol{V}_j &\leftarrow \frac{1}{q_j}\sum_i q_{ij}\Big[\big(\boldsymbol{m}_j - \boldsymbol{b}_{ij}\big)\big(\boldsymbol{m}_j - \boldsymbol{b}_{ij}\big)^T + \boldsymbol{B}_{ij}\Big] \\ \boldsymbol{T}_{ij} &\leftarrow \boldsymbol{R}_i\boldsymbol{V}_j\boldsymbol{R}_i^T + \boldsymbol{S}_i. \end{aligned} \tag{A7}$$

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. Notice that we do not 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 updated equation for $\boldsymbol{V}_j$ is guaranteed by its form to produce a symmetric nonnegative definite matrix.

REFERENCES

Bienaymé, O. 1999, A&A, 341, 86
Binney, J., Dehnen, W., & Bertelli, G. 2000, MNRAS, 318, 658
Chereul, E., Crézé, M., & Bienaymé, O. 1998, A&A, 340, 384
———. 1999, A&AS, 135, 5
Dehnen, W. 1998, AJ, 115, 2384
———. 2000, AJ, 119, 800
Dehnen, W., & Binney, J. J. 1998, MNRAS, 298, 387
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. R. Stat. Soc. B, 39, 1
ESA. 1997, The *Hipparcos* and Tycho Catalogues (ESA SP-1200; Noordwijk: ESA)
Fux, R. 2001, A&A, 373, 511
Lutz, T. E., & Kelker, D. H. 1973, PASP, 85, 573
Majewski, S. R., Ostheimer, J. C., Kunkel, W. E., & Patterson, R. J. 2000, AJ, 120, 2550
Quillen, A. C. 2003, AJ, 125, 785
Sirko, E., et al. 2004, AJ, 127, 914
Skrutskie, M. F., et al. 1997, in The Impact of Large Scale Near-IR Sky Surveys, ed. F. Garzon et al. (Dordrecht: Kluwer), 25
Skuljan, J., Hearnshaw, J. B., & Cottrell, P. L. 1999, MNRAS, 308, 731
York, D., et al. 2000, AJ, 120, 1579