

# A Fast Algorithm for Robust Mixtures in the Presence of Measurement Errors

Jianyong Sun and Ata Kabán

**Abstract**—In experimental and observational sciences, detecting atypical, peculiar data from large sets of measurements has the potential of highlighting candidates of interesting new types of objects that deserve more detailed domain-specific followup study. However, measurement data is nearly never free of measurement errors. These errors can generate false outliers that are not truly interesting. Although many approaches exist for finding outliers, they have no means to tell to what extent the peculiarity is not simply due to measurement errors. To address this issue, we have developed a model-based approach to infer *genuine* outliers from multivariate data sets when measurement error information is available. This is based on a probabilistic mixture of hierarchical density models, in which parameter estimation is made feasible by a tree-structured variational expectation-maximization algorithm. Here, we further develop an algorithmic enhancement to address the scalability of this approach, in order to make it applicable to large data sets, via a  $K$ -dimensional-tree based partitioning of the variational posterior assignments. This creates a non-trivial tradeoff between a more detailed noise model to enhance the detection accuracy, and the coarsened posterior representation to obtain computational speedup. Hence, we conduct extensive experimental validation to study the accuracy/speed tradeoffs achievable in a variety of data conditions. We find that, at low-to-moderate error levels, a speedup factor that is at least linear in the number of data points can be achieved without significantly sacrificing the detection accuracy. The benefits of including measurement error information into the modeling is evident in all situations, and the gain roughly recovers the loss incurred by the speedup procedure in large error conditions. We analyze and discuss in detail the characteristics of our algorithm based on results obtained on appropriately designed synthetic data experiments, and we also demonstrate its working in a real application example.

**Index Terms**— $K$ -dimensional (KD)-tree, measurement errors, outlier detection, robust mixture modeling, variational expectation-maximization (EM) algorithm.

## I. INTRODUCTION

**R**OBUST DATA modeling is an important task in scientific data mining, with the goal to capture the structure of

Manuscript received April 15, 2009; revised December 24, 2009, March 1, 2010, and March 12, 2010. Date of publication July 15, 2010; date of current version August 6, 2010. This work was supported by PPARC, under Grant PP/C503138/1. The work of J. Sun was supported by BBSRC/EPSRC, under Grant BB/D019613/1.

J. Sun is with the Center for Plant Integrative Biology, School of Bioscience, The University of Nottingham, Sutton Bonington LE12 5RD, U.K., and also with Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300191, China (e-mail: j.sun@cpib.ac.uk).

A. Kabán is with the School of Computer Science, University of Birmingham, Birmingham B15 2TT, U.K. (e-mail: axk@cs.bham.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNN.2010.2048219

the typical observations while dealing with atypical or outlying observations in an automated manner. In the same time, detecting outliers has the potential of highlighting candidates of interesting new types of objects that deserve more detailed domain-specific followup study.

An outlier is an object that lies far away from the bulk of data density, therefore it may represent some peculiar, possibly interesting object. However, measurement data from experimental sciences, such as biochemistry, meteorology or astrophysics, most often contains measurement errors, due to a variety of inevitable factors. For example, in astrophysics, known measurement errors arise from uncertainties in instrument calibration and physical limitations of devices and experimental conditions. These are recorded and are available as part of the data. Although many efficient data mining approaches exist for finding outliers [2], [7], [25], [27], [35], [38], there is nothing to prevent them from confusing erroneous measurements with potentially interesting rare or peculiar ones. They have no means to tell to what extent the peculiarity is not simply due to measurement errors. Since these errors can also generate outliers, a great deal of effort could be saved for the domain scientists if the detection procedure could take principled account of the error information to avoid detecting outliers that are not truly interesting.

We have been therefore developing a model-based approach to infer *genuine* outliers from multivariate data sets in the presence of measurement error information [41], by explicitly incorporating knowledge of measurement errors into the model. This is based on a probabilistic mixture of hierarchical density models, building on work from statistics and statistical machine learning regarding the use of heavy-tail Student's  $t$ -distributions in the model construction. A brief review of these follows in the next section. The probabilistic framework allows us to incorporate additional information in a straightforward manner, and makes it feasible to develop further extensions as appropriate, which we shall exploit in this paper.

Although this provides a principled solution to the problem, and statistically significant improvements have been demonstrated in high error conditions as a proof of concept, the basic algorithm inherits the computational scaling difficulty of other expectation-maximization (EM)-based mixture model estimation algorithms [15], namely in each training step all data points need to be visited. The training process is therefore linear with the number of data points, and the EM-based algorithm is thus too time-consuming to be applied to large

data sets. In this paper, we set up therefore to develop a scalable version of this principled model-based approach.

To accelerate the training process of the EM-based algorithm, many techniques have been proposed in the literature, e.g., [32], [33], [37]. The main idea of these techniques is to first partition the data set into several groups (cells), and cache sufficient statistics of the data points in the cells, such as the mean and covariance. The subsequent training process is then only carried out on these sufficient statistics rather than the entire data set. Since the posteriors of the latent variables in the EM-based algorithm need to be inferred, most of these existing speedup algorithms are unable to make exact inference due to the partitioning. As a result, these techniques have to inevitably approximate the data likelihood. Moreover, convergence of most of these algorithms is not guaranteed. The algorithm recently developed by Verbeek *et al.* is claimed to be the first convergent accelerated-EM-clustering algorithm [48]. It takes a principled approach by using the idea of data partitioning for making the variational posterior assignments.

We are not aware of accelerated algorithms available specifically for robust mixtures, let alone in the presence of measurement errors. This is what we propose and develop in this paper, by combining our model-based approach [41] with the speedup technique devised in [48], and some further technical enhancements. The algorithm so obtained also produces an accelerated mixture of student  $t$ -distributions in the special case of zero noise limit.

After giving a brief overview of recent and related work in the next section, the remainder of this paper is organized as follows. Section II describes the underlying model, first proposed in [41], along with the tree-structured variational approximation that makes parameter estimation feasible but still costly for large data sets. The accelerated version of the variational EM algorithm, using  $K$ -dimensional (KD)-tree partitioning of the posterior assignments is presented in Section III. Extensive experimental evaluation and analysis are provided in Section IV, thoroughly assessing both the speed/accuracy tradeoffs in a variety of data conditions and the gains achieved by including the measurement error information. Finally, Section V concludes the paper.

### A. Related Work

Since our initial work in [41] motivated by astrophysical data mining problems, the usefulness of explicitly accounting for known measurement error information has been recognized in a diversity of other areas too [24], such as in computer vision and bioinformatics. In [24], a “partial” Bayesian Gaussian mixture has been developed, which includes a measurement error layer in the same spirit as [41]. The density estimator layers differ in that, unlike our use of  $t$ -mixtures, they choose to employ a Bayesian mixture of Gaussians, which is also robust against singularities. However, this model design was made neither with the aim of detecting outliers, nor does it scale up to large data sets, which represent the focus of our work.

The work of [49] adopts the variable kernel density estimation approach [44] to deal with noisy data, using Gaussian mixtures to approximate the underlying true densities. Still, these methods are likely to be sensitive to the outlying

observations due to the use of a finite number of Gaussian building blocks.

In the statistics and statistical machine learning communities, there has been ample evidence for the beneficial effects of the heavy-tailed  $t$ -distribution, as a robust building block. It has been adopted in various applications, such as clustering [5], [11], [36], [43], visualization [42], [47], robust projections—robust probabilistic principal component analysis [52] and robust canonical correlation analysis (CCA) [3] and factor analyzers [4], [12], [30], in signal processing [10], sequential data modeling [9], image processing [8], [46], to name just a few. The  $t$ -distribution has heavy tails, hence it gives non-zero probability to observations that are far away from the bulk of the density.

It has been found in these applications that the  $t$ -distribution based approach was superior to methods based on Gaussian distributions. Moreover, the maximum likelihood estimate of  $t$ -mixtures is robust against outliers [36]. Since achieving such robustness by means of full Bayesian methods can be computationally quite demanding, the use of  $t$ -distributions in conjunction with maximum likelihood estimation procedures (or approximations thereof) represents an attractive basis for our effort of developing scalable yet principled means of detecting genuine outliers from large data sets. But besides the practical computational issue, we should note that in the massive data sample problem that we deal with in this paper, Bayesian priors on the model parameters would have little benefit if any. Indeed, one should be aware that in certain cases a misspecified prior can lead to over-fitting even in infinite sample size conditions [21]. However, in other settings that would justify a full-Bayesian treatment, our algorithm could be adapted with little effort.

## II. ROBUST MIXTURE MODELING OF DATA WITH ERROR INFORMATION

This section describes the probabilistic model that underlies our approach to robust modeling of multivariate data sets in the presence of measurement errors, along with a tree-structured variational EM algorithm for parameter estimation. A preliminary version appears in [41], where slightly more technical details can be found.

### A. Model

To obtain a data set in experimental and/or observational sciences, often a series of experiments and/or observations are carried out for the sake of accuracy. According to the central limit theorem, it is reasonable to assume that these observations follow a normal distribution. We can assume that each individual measurement in the data set has the form  $t_{in} \pm \sqrt{s_{in}}$ ,  $n = 1, \dots, N$ ;  $i = 1, \dots, d$ , where  $N$  is the number of data instances,  $d$  is the number of features, and  $s_{in}$  is the measurement error variance estimates that are known and available. The variance reflects the cumulative effects of the uncertainties in instrument calibration and physical limitations of devices and experimental conditions. We can express this compactly by writing a heteroscedastic noise model, so for

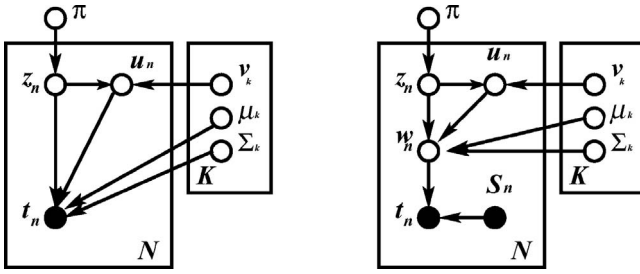


Fig. 1. (a) Plate diagrams of the mixture of  $t$ -distribution. (b) Proposed model for data with errors.

each observed datum  $\mathbf{t}_n \in \mathfrak{R}^d$ , we have a known diagonal covariance matrix  $\mathbf{S}_n$

$$p(\mathbf{t}_n|\mathbf{w}_n) = \mathcal{N}(\mathbf{t}_n|\mathbf{w}_n, \mathbf{S}_n) \quad (1)$$

where  $\mathcal{N}(\mathbf{t}_n|\mathbf{w}_n, \mathbf{S}_n)$  denotes the normal distribution with unknown mean  $\mathbf{w}_n$  and known diagonal covariance matrix  $\mathbf{S}_n$ .

The unknown mean  $\mathbf{w}_n$  represents the clean, error-free version of the data, and it may have a quite peculiar density which contains outliers. We call these “true” or “genuine” outliers of the data density. The observed data density, i.e., the density of  $\mathbf{t}_n$  may have other outliers, as a result of the measurement errors, and it is the former that is of interest to try to infer.

For this purpose, we model the density of  $\mathbf{w}_n$  as a robust mixture of  $K$  Student  $t$ -distributions (MoT)

$$p(\mathbf{w}_n) = \sum_{k=1}^K \pi_k S_t(\mathbf{w}_n|\mu_k, \Sigma_k, \nu_k) \quad (2)$$

where  $S_t(\mathbf{w}_n|\mu_k, \Sigma_k, \nu_k)$  is the Student’s  $t$ -distribution. This can be rewritten as a scale-mixture [28]

$$S_t(\mathbf{w}_n|\mu_k, \Sigma_k, \nu_k) = \int_0^\infty \mathcal{N}\left(\mathbf{w}_n|\mu_k, \frac{\Sigma_k}{u_n}\right) \mathcal{G}\left(u_n|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) du_n \quad (3)$$

and  $\mathcal{G}(u|a, b) = b^a \exp(-bu)u^{a-1} / \Gamma(a)$  is the gamma distribution with shape parameter  $a$  and scale parameter  $b$ ,  $\Gamma(\cdot)$  is the gamma function.

The distribution of the observed data  $\mathbf{t}_n$  is then obtained by integration over  $\mathbf{w}_n$ , leading to the following likelihood function:

$$p(\mathbf{t}_n|\theta) = \sum_k \pi_k \iint \left[ \mathcal{N}(\mathbf{t}_n|\mathbf{w}_n, \mathbf{S}_n) \mathcal{N}\left(\mathbf{w}_n|\mu_k, \frac{\Sigma_k}{u_n}\right) \mathcal{G}\left(u_n|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \right] du_n d\mathbf{w}_n \quad (4)$$

where  $\theta = \{\mu_k, \Sigma_k, \nu_k, 1 \leq k \leq K\}$  are the model parameters.

Fig. 1 gives the plate diagram of this model (reproduced from [41]). In the plate diagram, the filled circles ( $\mathbf{t}_n$  and  $\mathbf{S}_n$ ) denote the observed values, while the unfilled circles ( $\mathbf{w}_n, u_n, z_n$ ) indicate the latent random variables except the model parameters.

Thus, the complete probability model, i.e., the joint probability of all variables (the observed variables  $\mathbf{t}$  and the latent variables  $\mathbf{w}, u, z$ ) has the following factorized form:

$$\mathcal{L}_C = \prod_n \prod_k [p(\mathbf{t}_n|\mathbf{w}_n)p(\mathbf{w}_n|u_n, z_n = k)p(u_n|z_n = k) p(z_n = k|\pi)]^{\delta(z_n=k)} \quad (5)$$

$$\equiv \prod_n \prod_k [p(\mathbf{t}_n, \mathbf{w}_n, u_n, z_n = k)]^{\delta(z_n=k)} \quad (6)$$

where  $\delta(\cdot)$  is the Kronecker delta,  $p(\mathbf{t}_n, \mathbf{w}_n, u_n, z_n = k)$  is the full likelihood, and

$$p(\mathbf{w}_n|u_n, z_n = k) = \mathcal{N}\left(\mathbf{w}_n|\mu_k, \frac{\Sigma_k}{u_n}\right) \quad (7)$$

$$p(u_n|z_n = k) = \mathcal{G}\left(u_n|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \quad (8)$$

$$p(z_n = k|\pi) = \pi_k. \quad (9)$$

In order to derive maximum likelihood parameter estimates, we should be maximizing the likelihood function of the training data  $\mathcal{Y}$ . However, since the integration in (4) is intractable, a tree-structured variational approximation<sup>1</sup> has been found effective [41]. The associated variational EM algorithm works by maximizing the so-called free-energy function [26] with respect to an auxiliary distribution  $q(\mathbf{w}_n, u_n, z_n)$  (E-step) and the model parameters (M-step), respectively. The free energy function is a lower bound on the log likelihood

$$\mathcal{L}(\theta|\mathcal{Y}) \geq H(q) + \sum_{n,k} \int \int q(\mathbf{w}_n, u_n, z_n = k) \log p(\mathbf{t}_n, \mathbf{w}_n, u_n, z_n = k) d\mathbf{w}_n du_n \quad (10)$$

$$\equiv \mathcal{F}_O(\theta|\mathcal{Y}) \quad (11)$$

where  $H(q)$  is the entropy of the variational distribution  $q$  of the form

$$H(q) = - \sum_{n,k} \left[ \int \int q(\mathbf{w}_n, u_n, z_n = k) \log(q(\mathbf{w}_n, u_n, z_n = k)) d\mathbf{w}_n du_n \right] \quad (12)$$

As a result, the following algorithm is obtained [41].

### B. Tree-Structured Variational Algorithm

In the variational E-step, the auxiliary posterior distribution  $q(\mathbf{w}_n, u_n, z_n = k)$  is sought in a tree-structured factorized form

$$q(\mathbf{w}_n|z_n = k)q(u_n|z_n = k)q(z_n = k). \quad (13)$$

Note that the same form of factorization,  $q(u, z) = q(u|z)q(z)$ , has been adopted recently in [5] and suggested in [31]. Moreover, the experimental results in [5] and [42] favor the tree-structured factorization over the full factorization  $q(u, z) = q(u)q(z)$ .

Due to the conjugacy properties of the distributions used, through maximizing the free-energy function  $\mathcal{F}_O$  with respect

<sup>1</sup>This should not be confused with variational Bayes inference. The variational approximation is employed here to make a maximum likelihood estimation tractable.

to the factorized auxiliary distribution  $q$ , the following posterior distributions were obtained:

$$q(\mathbf{w}_n | z_n = k) = \mathcal{N}(\mathbf{w}_n | \langle \mathbf{w} \rangle_{nk}, \Sigma_{\mathbf{w}_n | k}) \quad (14)$$

$$q(u_n | z_n = k) = \mathcal{G}(u_n | a_{nk}, b_{nk}) \quad (15)$$

$$q(z_n = k) = \frac{\exp(A_{\mathbf{t}_n, k})}{\sum_{k'} \exp(A_{\mathbf{t}_n, k'})} \quad (16)$$

where  $\langle \cdot \rangle$  denotes expectation, and

$$\Sigma_{\mathbf{w}_n | k} = \mathbf{S}_n \left[ \frac{\Sigma_k}{\langle u_n \rangle_k} + \mathbf{S}_n \right]^{-1} \frac{\Sigma_k}{\langle u_n \rangle_k} \quad (17)$$

$$\langle \mathbf{w} \rangle_{nk} = \Sigma_{\mathbf{w}_n | k} \left( \langle u_n \rangle_k \Sigma_k^{-1} \mu_k + \mathbf{S}_n^{-1} \mathbf{t}_n \right) \quad (18)$$

$$a_{nk} = \frac{v_k + d}{2} \quad (19)$$

$$b_{nk} = \frac{v_k + C_{nk}}{2} \quad (20)$$

and

$$C_{nk} = (\langle \mathbf{w}_n \rangle_k - \mu_k)^T \Sigma_k^{-1} (\langle \mathbf{w}_n \rangle_k - \mu_k) + \text{Tr}(\Sigma_{\mathbf{w}_n | k}^{-1} \Sigma_k) \quad (21)$$

$$\langle u_n \rangle_k = \frac{a_{nk}}{b_{nk}} \quad (22)$$

$$A_{\mathbf{t}_n, k} = \langle \log p(\mathbf{t}_n, \mathbf{w}_n, u_n, z_n = k) \rangle_{\mathbf{w}_n, u_n | k} - \langle \log q(u_n | z_n = k) \rangle_{u_n | k} - \langle \log q(\mathbf{w}_n | z_n = k) \rangle_{\mathbf{w}_n | k}. \quad (23)$$

The parameter re-estimation is completed in the M-step, with the following re-estimation formulae for  $\mu_k$ ,  $\Sigma_k$ , and  $\pi_k$ :

$$\mu_k = \frac{\sum_{n=1}^N q(z_n = k) \langle u_n \rangle_{u_n | k} \langle \mathbf{w}_n \rangle_{\mathbf{w}_n | k}}{\sum_{n=1}^N q(z_n = k) \langle u_n \rangle_{u_n | k}} \quad (24)$$

$$\Sigma_k = \frac{\sum_{n=1}^N q(z_n = k) \langle u_n \rangle_{u_n | k} \tilde{\Sigma}_{n, k}}{\sum_{n=1}^N q(z_n = k)} \quad (25)$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^N q(z_n = k) \quad (26)$$

where  $\tilde{\Sigma}_{n, k} = [(\mu_k - \langle \mathbf{w}_n \rangle_k)(\mu_k - \langle \mathbf{w}_n \rangle_k)^T + \Sigma_{\mathbf{w}_n | k}]$ . The optimization process can be monitored by evaluating the free energy  $\mathcal{F}_O(\theta | \mathcal{Y})$ .

The accuracy of this variational approximation was experimentally assessed in [41], and it was found to compare favorably to a simpler, fully factorial approximation scheme, approaching the accuracy of a Markov-Chain-EM while maintaining computational simplicity.

However, the time complexity of this algorithm is  $\mathcal{O}(d^3 KN)$ , which means that, unfortunately, its utility is limited to relatively small to medium-size data sets. Therefore, in the next section, we devise a more scalable alternative.

### III. ACCELERATED VARIATIONAL ALGORITHM

To accelerate the estimation of the model proposed in the previous section, it is useful to note first, that in variational EM, *any* auxiliary distribution  $q$  is acceptable in the E-step, provided that the free-energy is non-decreasing [34]. The intuition then is that the maxima of the variational free energy

function would be a good approximation to those of the data likelihood  $\mathcal{L}$  [20]. This observation allows us to assign the variational posteriors to the data points in such a manner that the resulting algorithm scales better. In particular, we can assume equal auxiliary posteriors for groups of data points that are nearby in the input space. This is the underlying idea of the method given in the sequel.

Similarly to [48], we consider a partitioning  $\mathcal{P}$  of the data set as a collection of non-overlapping cells  $\mathcal{P} = \{\mathcal{A}_1, \dots, \mathcal{A}_m\}$ . To the data points in a cell  $\mathcal{A}$  we will assign the same auxiliary posterior probabilities of the latent variables  $\mathbf{w}$ ,  $u$ ,  $z$ . Thus, all the data points  $\mathbf{t}_n \in \mathcal{A}$  will share the same latent variables, which we now denote by  $\mathbf{w}_\mathcal{A}$ ,  $u_\mathcal{A}$ ,  $z_\mathcal{A}$ , and accordingly, the auxiliary distribution is denoted by  $q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A})$ . Now, the data likelihood function can be approximated as follows:

$$\mathcal{L}(\theta | \mathcal{Y}) \geq \sum_{\mathcal{A} \in \mathcal{P}} \mathcal{F}_\mathcal{A} \equiv \mathcal{F}_\mathcal{P} \quad (27)$$

where

$$\mathcal{F}_\mathcal{A} = \sum_{\mathbf{t}_n \in \mathcal{A}} \sum_k \left( \langle \log p(\mathbf{t}_n, \mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k) \rangle_q - \langle \log q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k) \rangle_q \right) \quad (28)$$

and  $\langle \cdot \rangle_q$  is the expectation with respect to  $q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A})$ . Note that on this occasion, the full likelihood  $p(\mathbf{t}_n, \mathbf{w}_n, u_n, z_n)$  is replaced by  $p(\mathbf{t}_n, \mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A})$  since all  $\mathbf{w}_n$ ,  $n \in \mathbf{I}_\mathcal{A}$  (where  $\mathbf{I}_\mathcal{A} = \{n : \mathbf{t}_n \in \mathcal{A}\}$ ) take the same value  $\mathbf{w}_\mathcal{A}$ , and  $u_n$ ,  $z_n$ ,  $n \in \mathbf{I}_\mathcal{A}$  will equal to  $u_\mathcal{A}$  and  $z_\mathcal{A}$ , respectively.

Using this, we can then employ the same line of thought as in Section II, to derive a better scaling algorithm. As in (13), we factorize the posterior  $q$  with respect to each cell  $\mathcal{A}$  in a tree-form, as follows:

$$\begin{aligned} q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k) &= q(\mathbf{w}_\mathcal{A}, u_\mathcal{A} | z_\mathcal{A} = k) q(z_\mathcal{A} = k) \\ &= q(\mathbf{w}_\mathcal{A} | z_\mathcal{A} = k) q(u_\mathcal{A} | z_\mathcal{A} = k) q(z_\mathcal{A} = k). \end{aligned} \quad (29)$$

Note that here the factorization  $q(u_\mathcal{A}, z_\mathcal{A}) = q(u_\mathcal{A} | z_\mathcal{A}) q(z_\mathcal{A})$  is applied. Then, the free-energy function for each cell  $\mathcal{A}$  can be evaluated as follows:

$$\begin{aligned} \mathcal{F}_\mathcal{A} &= \sum_{\mathbf{t}_n \in \mathcal{A}} \sum_k \left[ \langle \log p(\mathbf{t}_n, \mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k) \rangle_{\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k} \right] \\ &\quad - \sum_{\mathbf{t}_n \in \mathcal{A}} \sum_k \left[ \langle \log q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k) \rangle_{\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A} = k} \right]. \end{aligned} \quad (30)$$

It should be observed also that the resulting tradeoff between speed and accuracy is achieved gracefully, since in the case of the finest possible partition  $\mathcal{P}$ , i.e., the one in which each cell  $\mathcal{A}$  only has one data point, the approximated free-energy function  $\mathcal{F}_\mathcal{P}$  recovers the free energy function  $\mathcal{F}_O$ , used in the previous section.

#### A. Variational E-Step: Locally Shared Posteriors

Analogously to the variational E-step in Section II, we now need to infer the locally shared posteriors  $q(\mathbf{w}_\mathcal{A}, u_\mathcal{A}, z_\mathcal{A})$  in the E-step. To this end, we maximize  $\mathcal{F}_\mathcal{P}$  with respect to the posterior  $q$  for each cell. Due to the conjugacy properties

used, we obtain the following (see Appendix A for detailed derivation):

$$q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) = \mathcal{N}(\mathbf{w}_{\mathcal{A}}|\langle \mathbf{w} \rangle_{\mathcal{A}|k}, \Sigma_{\mathbf{w}_{\mathcal{A}}|k}) \quad (31)$$

$$q(u_{\mathcal{A}}|z_{\mathcal{A}} = k) = \mathcal{G}(u_{\mathcal{A}}|a_{\mathcal{A}|k}, b_{\mathcal{A}|k}) \quad (32)$$

and

$$\frac{q(z_{\mathcal{A}} = k) \propto \pi_k \exp \left\{ \sum_{\mathbf{t}_n \in \mathcal{A}} \langle \log p(\mathbf{t}_n, \mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_q \right\}}{\exp \left\{ |\mathcal{A}| \langle \log q(\mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_q \right\}} \quad (33)$$

where

$$\Sigma_{\mathbf{w}_{\mathcal{A}}|k} = \left[ \left( \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{S}_n^{-1} \right) + |\mathcal{A}| \langle u \rangle_{\mathcal{A}|k} \Sigma_k^{-1} \right]^{-1} \quad (34)$$

$$\langle \mathbf{w} \rangle_{\mathcal{A}|k} = \Sigma_{\mathbf{w}_{\mathcal{A}}|k} \left[ \left( \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{S}_n^{-1} \mathbf{t}_n \right) + |\mathcal{A}| \Sigma_k^{-1} \mu_k \langle u \rangle_{\mathcal{A}|k} \right] \quad (35)$$

and

$$a_{\mathcal{A}|k} = \frac{\nu_k + d}{2} \quad (36)$$

$$b_{\mathcal{A}|k} = \frac{\nu_k}{2} + \frac{\langle \langle \mathbf{w} \rangle_{\mathcal{A}|k} - \mu_k \rangle^T \Sigma_k^{-1} (\langle \mathbf{w} \rangle_{\mathcal{A}|k} - \mu_k) + \text{Tr}(\Sigma_{\mathbf{w}_{\mathcal{A}}|k} \Sigma_k^{-1})}{2}. \quad (37)$$

Here and in the following,  $|\mathcal{A}|$  denotes the cardinality of cell  $\mathcal{A}$ , and

$$\langle u \rangle_{\mathcal{A}|k} = \frac{a_{\mathcal{A}|k}}{b_{\mathcal{A}|k}} \quad \langle \log u \rangle_{\mathcal{A}|k} = \psi(a_{\mathcal{A}|k}) - \log b_{\mathcal{A}|k}. \quad (38)$$

### B. Variational M-Step

In the variational M-step, maximizing  $\mathcal{F}_{\mathcal{P}}$  with respect to the model parameters and solving the stationary equations, we obtain the following updated equations:

$$\mu_k = \frac{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \langle u \rangle_{\mathcal{A}|k} \langle \mathbf{w} \rangle_{\mathcal{A}|k}}{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \langle u \rangle_{\mathcal{A}|k}} \quad (39)$$

$$\Sigma_k = \frac{\sum_{\mathcal{A}} q(z_{\mathcal{A}} = k) |\mathcal{A}| \langle u \rangle_{\mathcal{A}|k} \widetilde{\Sigma}_{\mathbf{w}_{\mathcal{A}}|k}}{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k)} \quad (40)$$

$$\pi_k = \frac{1}{N} \sum_{\mathcal{A} \in \mathcal{P}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \quad (41)$$

where

$$\widetilde{\Sigma}_{\mathbf{w}_{\mathcal{A}}|k} = \left[ (\mathbf{w}_{\mathcal{A}|k} - \mu_k) (\mathbf{w}_{\mathcal{A}|k} - \mu_k)^T + \Sigma_{\mathbf{w}_{\mathcal{A}}|k} \right]. \quad (42)$$

Further, we update the degrees of freedom parameters  $\nu_k$ ,  $1 \leq k \leq K$  by maximizing the log-likelihood bound  $\mathcal{F}_{\mathcal{P}}$ , i.e., solving the following nonlinear equation:

$$\sum_{\mathbf{t}_n \in \mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \left[ \log \left( \frac{\nu_k}{2} \right) + 1 + \langle \log u \rangle_{\mathcal{A}|k} - \frac{a_{\mathcal{A}|k}}{b_{\mathcal{A}|k}} - \psi \left( \frac{\nu_k}{2} \right) \right] = 0. \quad (43)$$

To find the root of this equation, we employed the secant method, though any numerical solver could be used. Similarly to [45], we update the degrees of freedom in every fifth iteration of the overall iterative algorithm.

Finally, we can use the evaluation of the log-likelihood bound to monitor the optimization process, the details of which are given in Appendix B. Usefully, some values can be cached without needing to be computed at each iteration.

### C. Special Case: Accelerated Estimation for Mixtures of $t$ -Distributions

It may be worth pointing out that an accelerated estimation algorithm for  $t$ -mixtures [36] (MoT) can be obtained with no effort, as a special case of the previous analysis, simply by considering the zero-noise limit. In case  $\mathbf{S}_n = 0$ , where  $1 \leq n \leq N$ , the model in (4) degenerates to the MoT [41]. Consequently, zero measurement errors will result in an accelerated version of the MoT, as follows.

In case  $\mathbf{S}_n = 0$ , the auxiliary posterior  $q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)$ ,  $1 \leq k \leq K$  is no longer required; while the posterior  $q(u_{\mathcal{A}}|z_{\mathcal{A}} = k)$  is still a gamma distribution with the shape parameter  $a_{\mathcal{A}|k}$  as before, and the scale parameter  $b_{\mathcal{A}|k}$  is modified as

$$b_{\mathcal{A}|k} = \frac{\nu + (\langle \mathbf{t} \rangle_{\mathcal{A}} - \mu_k)^T \Sigma_k^{-1} (\langle \mathbf{t} \rangle_{\mathcal{A}} - \mu_k)}{2}. \quad (44)$$

In the variational M-step, the model parameters get updated as

$$\mu_k = \frac{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \langle u \rangle_{\mathcal{A}|k} \langle \mathbf{t} \rangle_{\mathcal{A}}}{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \langle u \rangle_{\mathcal{A}|k}} \quad (45)$$

$$\Sigma_k = \frac{\sum_{\mathcal{A}} q(z_{\mathcal{A}} = k) |\mathcal{A}| \langle u \rangle_{\mathcal{A}|k} \Xi_{\mathcal{A}k}}{\sum_{\mathcal{A}} |\mathcal{A}| q(z_{\mathcal{A}} = k)} \quad (46)$$

$$\pi_k = \frac{1}{N} \sum_{\mathcal{A} \in \mathcal{P}} |\mathcal{A}| q(z_{\mathcal{A}} = k) \quad (47)$$

where  $\Xi_{\mathcal{A}k} = [(\langle \mathbf{t} \rangle_{\mathcal{A}} - \mu_k) (\langle \mathbf{t} \rangle_{\mathcal{A}} - \mu_k)^T]$ .

### D. Partitioning by KD-Tree

So far, we did not specify the partitioning method for deciding the membership of points into cells. Indeed, any partitioning is applicable for the speed-enhancement scheme just described, provided that the values in (67) are cached in the partition nodes. We can thus choose any convenient partitioning method. One of the popular and scalable partitioning methods is the KD-tree [6], [32]. A KD-tree is a binary tree. In the constructed tree, the root node contains all data points. Each parent node has two children nodes, and each child node contains a subset of the data points in the parent node, and the union of the points in the children nodes consists of the points in the parent node. Usually, the axis-aligned hyperplane is used to split the parent nodes. In our experiments, the hyperplane that cuts along the dimension with maximum length of the parent node is applied to build the KD-tree. The values in (67) (see Appendix B) are stored in each node.

The proposed variational E and M steps can be iteratively carried out till convergence at a certain partition level  $\mathcal{P}$ ,

resulting in a log-likelihood  $\mathcal{F}_P$ . Since the proposed variational EM algorithm does not sweep over all the data points, but only the cached statistics, its running time will be much less than that of the original algorithm. This is because  $|\mathcal{P}| \ll N$ , where  $|\mathcal{P}|$  is the number of the nodes in the partition  $\mathcal{P}$ . However, since we only use the cached statics, the performance of the algorithm will be inevitably inferior to that of the original algorithm. In other words, the difference between the log-likelihood of the data set  $\mathcal{L}$  and  $\mathcal{F}_O$  will be smaller than  $\mathcal{L} - \mathcal{F}_P$  after training. To increase the data log-likelihood, we can split the current partition into a finer partition and apply the proposed EM algorithm on the finer partition. This procedure can iterate until some criterion is satisfied. At first sight one may think the iterative procedure may greatly increase the computational time due to the slow convergence of the EM algorithm [29]. However, this procedure will in fact not increase the computational cost significantly, since the previously estimated model parameters and posteriors will help to accelerate the EM convergence in the finer partition level. Previously estimated parameters impose a better initialization in the current EM execution: the variational EM will need less iterations to converge.

The next questions are to decide which cells to split and how to split the cells. We adopted an idea similar to that used in [48]. For each cell  $\mathcal{A}$  in the current partition, suppose its child cells are  $\mathcal{A}_a$  and  $\mathcal{A}_b$ , then we compute the increase in data log-likelihood  $\mathcal{F}_\Delta = \mathcal{F}_\mathcal{A} - (\mathcal{F}_{\mathcal{A}_a} + \mathcal{F}_{\mathcal{A}_b})$  by assigning the same locally shared posteriors in cell  $\mathcal{A}$  to  $\mathcal{A}_a$  and  $\mathcal{A}_b$ . A certain percentage (50% in our experiments) of the cells with maximal  $\mathcal{F}_\Delta$  values are further partitioned. This can be done in time complexity of  $\mathcal{O}(|\mathcal{P}|)$ .

### E. Detecting Outliers

Following [36] and [41], the expectation of the latent variable  $u$  is of interest to determine which points are outliers. A sufficiently small  $\langle u \rangle$  value corresponds to an outlier. However, in our accelerated version of the algorithm, this criterion will now refer not to one single data point but to a cell with locally shared posteriors, that is

$$\langle u \rangle_{\mathcal{A}} = \sum_{k=1}^K q(z_{\mathcal{A}} = k) \langle u \rangle_{\mathcal{A}|k} \quad (48)$$

so this would suggest considering as outliers the cells with small  $\langle u \rangle_{\mathcal{A}}$  values. However, this detection approach is dubious: outliers and non-outliers could be split in the same cell. As a result, some non-outliers will be wrongly detected as true outliers and viceversa. To overcome this problem, we propose to use a heuristic approach with concepts borrowed from [13], as follows.

The training process of the algorithm is carried out first, until convergence. Then, we look at the density of each cell, i.e., the ratio between the number of objects in the cell and the volume of it. Those that are less dense are more likely to contain outliers, because the sparsity of a cell indicates that it is likely to be situated far away from the bulk of the density. So, we split a certain percentage (10% in our experiments) of the less dense cells (those having minimal density) to form a

finer partition. This splitting is performed using the special cut developed in [13], that is, the axis-aligned hyperplane in the *skewness* dimension is used to split the parent nodes. If we let  $C_m$  denote the center of mass of the objects in the cell and let  $C_v$  denote the center of volume of the cell, the “skewness” is defined as in [13]

$$\max_{\text{dimension } i} \left\{ \frac{\text{distance of } C_m \text{ from } C_v \text{ along } i}{\text{length of region along } i} \right\}. \quad (49)$$

Here we point out that only the concepts (the cell density and the skewness) proposed in [13] are borrowed into our algorithm, the actual implementation differs—in [13], the skewness is used to build the tree, while the inverse of density is used to flag the outliers in this paper.

### F. Algorithm

Now we are ready to describe the enhanced, accelerated algorithm in detail as follows.

- 1) Construct a KD-tree on the data set by means of classical axis-aligned hyperplane, and cache the values in (67) in each cell.
- 2) Expand the KD-tree to some depth to form the initial partition level  $\mathcal{P}_0$ .
- 3) Initialize the model parameters  $\mu_k$ ,  $\Sigma_k$  and  $\pi_k$ ,  $1 \leq k \leq K$  by estimating a mixture of  $t$ -distributions (MoT) on the values  $\langle t \rangle_{\mathcal{A}}$  [see (68)] of the cells in  $\mathcal{P}_0$ ; set  $i = 0$ ,  $\mathcal{F}_{\mathcal{P}_i} = -\infty$ .
- 4) Variational E-step. For each cell  $\mathcal{A}$  in the current partition  $\mathcal{P}_i$ , compute the locally shared posteriors using (31), (32), and (33). Calculate the free-energy function  $\mathcal{F}$  (see the lower bound evaluation in Appendix B).
- 5) Variational M-step. Update the model parameters using (39), (40), and (41).
6. If  $|\frac{\mathcal{F}}{\mathcal{F}_{\mathcal{P}_i}} - 1| > 1.0^{-5}$ ,  $\mathcal{F}_{\mathcal{P}_i} = \mathcal{F}$ , go to step 4.
7. Expand the current partition  $\mathcal{P}_i$  to create  $\mathcal{P}_{i+1}$  based on the heuristic described in Section III-D.
8. If  $|\frac{\mathcal{F}_{\mathcal{P}_i}}{\mathcal{F}_{\mathcal{P}_{i-1}}} - 1| > 1.0^{-5}$ , set  $i = i + 1$ , go to step 4.
9. Split the cells with the smallest cell-density values by using the special cut based on the skewness, as described in Section III-E.

**Remark 1:** In step 2, the tree is expanded to some depth. In theory, we can start the algorithm from any depth level of the built tree. However, due to the existence of outliers, the developed variational algorithm will not perform well if the expansion is not deep enough. The obvious reason is that the number of the cells is so small in a low-depth partition level that the proposed algorithm cannot differentiate the bulk part of the data set from the outlying parts. On the other hand, it is not necessary to start from the very beginning (depth level = 2) since the estimation of the model parameters at this level does not necessarily improve the convergence speed of the whole algorithm. Based on these considerations, in our experiments, the depth level is set to 10.

**Remark 2:** In step 3, the MoT is adopted to initialize the configurations of the model parameters. We can also use other approaches, such as k-means or randomization. The use

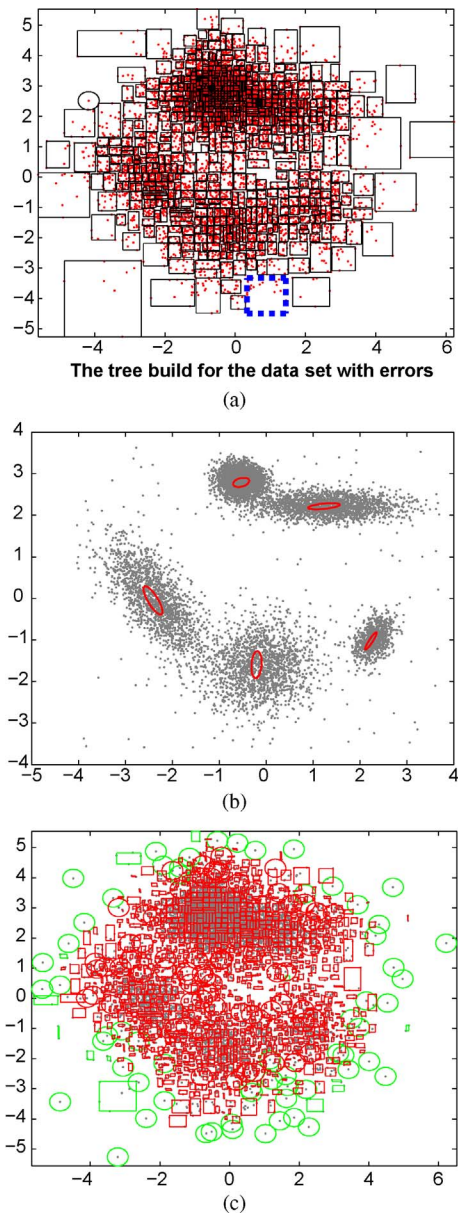


Fig. 2. Example to demonstrate the (a) KD-tree building, (b) model parameters estimation, and (c) outlier detection. (a) Dashed box shows the non-outliers and outliers are located in the same cell. (c) Green circles and boxes are the outliers found by the developed algorithm after splitting those boxes that mix the non-outliers and outliers (i.e., the result of Step 9 in the developed algorithm, see Section III-F).

of MoT can make the initialization close to the true model parameters, and thereafter make the variational algorithm in steps 4 and 5 converge in less iterations.

**Remark 3:** In step 9, the heuristic described in Section III-E is applied. This step is only used to improve the outlier detection accuracy, and is not a part of the learning procedure. The working of this heuristic is shown in Fig. 2(c). Comparing this with Fig. 2(a), we see that the outliers are isolated from cells (e.g., observe the cell with dashed edges) as a result of the heuristic.

**Remark 4:** If the equations formulated in Section III-C are applied in steps 4 and 5, we obtain an accelerated estimation algorithm for mixtures of  $t$ -distributions—which is of course

useful in the error-free case or when measurement errors are not available.

In the sequel, the accelerated algorithm that includes measurement error information will be referred to as “FMoT/Error,” its original (non-accelerated) version as “MoT/Error,” the  $t$ -mixture will be referred to as “MoT,” and its accelerated version as “FMoT.”

### G. Determining the Number of Mixture Components

One of the important issues in mixture modeling and clustering is to determine the optimal number of components. The optimal number can be automatically determined either by a Bayesian approach, such as in [5], [8], [43], [46] or based on information theory, such as minimum message length (MML) [50], minimum description length [39], Bayesian information criterion [40], Akaike information criterion [1], or by cross validation [19]. Among these methods, the Bayesian approach is currently most popular, and there are well known connections between the Bayesian approach and information theoretic ones [23].

Since we have (approximated) maximum likelihood estimates from our fast estimation algorithms developed in the previous sections, it is most convenient to employ the information-theoretically justified MML criterion to determine the optimal number of components. According to the MML theory, the optimal number of components can be determined by maximizing the following objective function [41]:

$$\mathcal{L}_{\mathcal{M}}(\theta|\mathcal{Y}) = \log p(\mathcal{Y}|\theta) - \frac{\hat{n}}{2} \sum_{k:\pi_k>0} \log \left( \frac{N\pi_k}{12} \right) - \frac{k_{nz}}{2} \log \left( \frac{N}{12} \right) - \frac{k_{nz}(\hat{n}+1)}{2} \quad (50)$$

where  $p(\mathcal{Y}|\theta)$  is the data log-likelihood,  $\hat{n}$  is the dimensionality of the parameters, and  $k_{nz}$  is the size of the set of components with non-zero mixing proportions, i.e.,  $k_{nz} = |\{k, \pi_k > 0, 1 \leq k \leq K\}|$ . The free parameters involved in the proposed algorithm are the means and the full covariance matrices of  $\mathcal{N}(\mathbf{w}_n|u_n, z_n = k)$ . Thus, the dimensionality of the  $k$ th parameter  $\theta_k = (\mu_k, \Sigma_k)$ , is  $d + d(d-1)/2$ .

In the speedup algorithm, we approximate the data likelihood by (27). Hence, replacing this in (50), leads to maximizing

$$\mathcal{L}_{\mathcal{M}}(\theta, \mathcal{Y}) \geq \sum_{t_n \in \mathcal{A}} \mathcal{F}_{\mathcal{A}} - \frac{\hat{n}}{2} \sum_{k:\pi_k>0} \log \left( \frac{N\pi_k}{12} \right) - \frac{k_{nz}}{2} \log \left( \frac{N}{12} \right) - \frac{k_{nz}(\hat{n}+1)}{2}. \quad (51)$$

This maximization is similar to the variational EM procedures presented in Sections III-A–III-B and algorithmically the only difference is in computing the mixing proportions  $\pi_k$  in the variational M-steps, which is now

$$\pi_k = \frac{\max \{0, \sum_{\mathcal{A} \in \mathcal{P}} |\mathcal{A}| q(z_{\mathcal{A}} = k) - \frac{\hat{n}}{2}\}}{\sum_{j=1}^K \max \{0, \sum_{\mathcal{A} \in \mathcal{P}} |\mathcal{A}| q(z_{\mathcal{A}} = j) - \frac{\hat{n}}{2}\}}. \quad (52)$$

Of course, only the non-zero-probability components of the mixtures will contribute to  $q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}})$ ,  $q(u_{\mathcal{A}}|z_{\mathcal{A}})$  and  $q(z_{\mathcal{A}})$ .

### H. Scaling

We now consider the time complexity of the proposed algorithm, per iteration, at each partition level  $\mathcal{P}_i$ . In step 1, the KD-tree is built. It is well-known that building such a tree has the time complexity  $\mathcal{O}(N \log N)$ , but the building needs to be done only once at the beginning of the algorithm.

In the variational E-step (step 4), we need to compute the posterior mean and covariance ( $\langle \mathbf{w} \rangle_{\mathcal{A}|k}$  and  $\Sigma_{\mathbf{w}_{\mathcal{A}|k}}$ ) for each cell  $\mathcal{A}$ . This takes  $\mathcal{O}(d^3 K)$  operations. The computation of the parameters of  $q_{\mathcal{A}}(u|k)$ ,  $a_{\mathcal{A}|k}$  and  $b_{\mathcal{A}|k}$  takes  $\mathcal{O}(d^3 K)$ , and the responsibility  $q_{\mathcal{A}}(k)$  takes  $\mathcal{O}(d^3 K)$  time as well. Therefore, in total, the time complexity of the proposed algorithm is  $\mathcal{O}(d^3 K |\mathcal{P}_i|)$ .

In step 5, the variational M-step, the time complexity per iteration is  $\mathcal{O}(d^3 K |\mathcal{P}_i|)$ . The time complexity of step 7, for deciding the cells to be split is  $\mathcal{O}(d^3 K |\mathcal{P}_i|)$ .

The most expensive operation in the variational E and M steps appears to be the matrix inversion which is  $\mathcal{O}(d^3)$  for each component, however it should be noted this is costly only when  $\Sigma_k$  are modeled as full covariances, which is feasible in relatively low-dimensional problems ( $d \ll N$ ). If this model was to be used on high dimensional data, then a diagonal form  $\Sigma_k$  could need to be taken—in which case the cubic operation is no longer required since the matrices to be inverted become diagonal. Another way to alleviate the problem caused by using full covariance is to replace it by a low rank approximation which can capture the covariances between the local principal directions, i.e., factor analyzers, as resolved in [4].

Since at each partition  $\mathcal{P}_i$ , the number of cells  $|\mathcal{P}_i|$  is much smaller than  $N$ , i.e., the size of the data set, we can see that the time complexity of the proposed algorithm at each iteration is much less than that of the original algorithm, which means the proposed algorithm is indeed more time-efficient than the original algorithm.

## IV. EXPERIMENTAL RESULTS

In this section, we extensively assess the performance of our method, and study the tradeoff between speed and accuracy achievable, as a function of a variety of data conditions. As we have seen from the technical constructions in the previous sections, this tradeoff originates from formulating a more detailed noise model to enhance the detection accuracy, while in the same time making the posterior representation coarser to obtain computational savings. The behavior of this tradeoff can be non-trivial, thus in the sequel we carry out a thorough assessment through a series of experiments on appropriately designed synthetic data sets.

The first set of experiments are meant to evaluate the performance of the proposed speedup procedure against the more time-consuming original estimation algorithm developed in [41]. This is carried out in measurement noise conditions of medium size, and varying the structure of the data density. We then assess the outlier detection accuracy in conditions of increasing measurement uncertainty in a second set of experiments to see the benefits of being able to take into account knowledge of measurement errors. Moreover, we carry

out experiments to investigate the effect of non-uniformly distributed outliers on the performance of the developed algorithm. Further, the performance of the developed MML criterion for the speedup algorithm to find the optimal number of components is investigated. Finally, we also illustrate a real application scenario, namely detecting high redshift quasars given photometric measurements with errors.

### A. Synthetic Datasets

In creating synthetic data sets for controlled experiments, we adopt the procedure described in [48] in the first instance. This is based on theoretical results by [14], regarding the identifiability of Gaussian mixture data densities. Data points are sampled from a randomly chosen  $K$ -component Gaussian mixture in  $d$ -dimensions with a component separation<sup>2</sup> of  $c$ . In addition, we simulate outliers by inserting uniformly generated data points in the data sample. These are the true (“genuine”) outliers in the data, that we try to recover. Finally, we add Gaussian noise with zero mean and unit variance to all sample points, in order to simulate measurement errors. The resulting data set is the input of our algorithm. The task is to recover the genuine outliers (along with the density of non-outliers), despite the Gaussian noise added.

### B. Assessment of the Proposed Speedup Procedure

The first extensive set of experiments is aimed at assessing the proposed speedup scheme (referred to as “FMoT/Error”) against the original, more costly algorithm presented in [41] (referred to as “MoT/Error”), in terms of the following criteria:

- 1) the speedup factor, i.e., the ratio between the time used by FMoT/Error and MoT/Error;
- 2) the out-of-sample log-likelihood, i.e., the ability of the algorithm to infer structure that generalizes to previously unseen data from the same density;
- 3) the correct outlier detection rates measured by the area under the receiver operating characteristics (AUC) values [17], [18] that are computed from in-sample data.

To quantify the accuracy of outlier detections in a compact manner, we compute, for each experiment, the AUC [17], [18]. One of the advantages of this measure is that it only requires the relative outlieriness values inferred for each data point (see Section III-E), which are positive reals, without needing us to decide on a threshold or hard-partition the points into outliers and non-outliers. More precisely, the AUC gives us the probability that a true genuine outlier is detected by our algorithm.

For computing out-of-sample likelihoods, test sets are created, of 100 test points each, following the same sampling as the associated training sets. The out-of-sample log-likelihood is then computed by carrying out the E-step of the variational EM algorithms with fixed model parameters

<sup>2</sup>As stated in [14], a Gaussian mixture is said to be  $c$ -separated if, for each pair  $(i, j)$  of component densities  $\|\mu_i - \mu_j\| \geq c \sqrt{d \max\{\lambda_{\max}(\Sigma_i), \lambda_{\max}(\Sigma_j)\}}$ , where  $\lambda_{\max}(\Sigma)$  denotes the maximum eigenvalue of  $\Sigma$ , the covariance matrix. Moreover, the dispersion of the data set sampled from a Gaussian with covariance  $\Sigma$  is determined by the eccentricity, which is defined as  $\sqrt{\lambda_{\max}(\Sigma)}/\sqrt{\lambda_{\min}(\Sigma)}$  (see [14] for details). In the experiments, we fix the eccentricity value to be 10 for all the components.

$\mu_k, \Sigma_k, \pi_k, \nu_k, 1 \leq k \leq K$  obtained from the training process.

In the experiments, data sets with varying  $N, c, K$  and  $d$  are created. As an example, Fig. 2(a) shows such a data set, superimposed with a KD-tree partitioning. In this example, the training data set consisted of 10000 points drawn from a 5-component 2-separated Gaussian mixture in two dimensions. Plot (b) shows the obtained clustering recovered by our FMoT/Error algorithm, and plot (c) highlights the outliers detected. We can see from plot (a) that, after the inclusion of measurement errors, the estimation for the “true” model parameters becomes much harder. However, from plot (b), we see the estimated grouping is very close to the true grouping.

Extensive and detailed comparison results are summarized in Fig. 3. These are averages over 30 independent runs. The speedup factor was measured based on our MATLAB 7.0 implementation of both MoT/Error [41] and FMoT/Error. The plots in the first column in Fig. 3 show that the speedup of the accelerated variational EM algorithm versus the original one is at least linear in the number of data points, and the separation  $c$  has a positive effect over the original algorithm. The larger the  $c$ , the higher the speedup factor. On the other hand, in Fig. 3 we also see that the data dimensionality  $d$  and the number of mixing components  $K$  have negative effects on the speedup factor. The effect of data dimensions  $d$  is not as bad as that of  $K$ , however. From dimension two to ten, the speedup factor decreases very slowly, but from five components to 30 components, the speedup factor decreases fairly quickly—most probably due to that the time complexity of the most computationally expensive step (step 2) is  $\mathcal{O}(d^3 K |\mathcal{P}_i|)$ , which depends on the number of clusters  $K$  and the number of cells in the  $i$ th splitting. At  $K = 30$ , one can observe that the ratio between the average number of data points in each cluster and the number of cells used is fairly low. This indicates that, in the accelerated training step, although we do not need to visit all the data points, still there is considerable computational effort involved. Consequently, the speedup factor of the accelerated algorithm drops below 10 in this case.

It should be pointed out that the experimental results reported in [48] display a much more drastic detrimental effect of the data dimensionality, which is due to the bad performance of their algorithm in the coarse partition level. However, as we discussed earlier, in our case, running the accelerated algorithm in the coarse partition level is not necessary (see Remark 1 in Section III-F).

In terms of the outlier detection rate, we see from the plots in the second column (Fig. 3) that in all cases the outliers can still be detected to a high accuracy rate, not much worse than the accuracy achieved by the original algorithm. This demonstrates the efficiency of the proposed heuristic for improving the outlier detection accuracy. In terms of the out-of-sample log-likelihood, we also obtained approximately the same values as the original algorithm for different number of data points. That shows that the performance of the accelerated algorithm in terms of density estimation and generalization is almost as good as that of the original algorithm.

### C. Assessment of the Benefits of Incorporating Knowledge of the Measurement Errors

In this section, we carry out experiments to assess to what extent knowledge of measurement errors helps us to infer the “genuine” outliers, i.e., the outliers that are not due to these errors. In particular, we are interested to find out the following.

- 1) How much accuracy gain can we achieve by the explicit inclusion of known measurement errors in our model.
- 2) How much accuracy we necessarily lose for not having access to the error-free data.
- 3) What is the tradeoff between speed and accuracy achievable in the above terms.

To this end, we use synthetic data sets derived from mixtures of Gaussian, as before, and vary the extent of measurement errors while fixing other data characteristics. The data sets we tested have 100000 data points, five clusters, five dimensions, separation parameter  $c = 2$ , and the largest eigenvalue of class covariances is three. Against these fixed data characteristics, five different measurement error levels are defined for this set of experiments, i.e., the diagonal elements of the error variance matrix  $\mathbf{S}$  will range between  $[0, 0.01]$ ,  $[0, 0.1]$ ,  $[0, 1]$ ,  $[0, 10]$ , and  $[0, 100]$ , respectively. In other words, the elements of  $\mathbf{S}$  will be uniformly sampled from the range at the data generation.

Since the mixture of  $t$ -distributions model ignores any knowledge of measurement errors, this is what we employ in our comparisons, both in its original form (MoT) and accelerated form (Fast-MoT, see Section III-C), in two different instances.

- 1) MoT and Fast-MoT applied to the data contaminated with observation errors, called “MoT/Base” and “FMoT/Base,” respectively, provide baselines against of which we measure our improvements achieved by including measurement error information in our modeling.
- 2) MoT and Fast-MoT applied to the clean data, referred to as “MoT/Idealized” and “FMoT/Idealized,” respectively, provides idealized upper limits of the achievable accuracy. In the same time, inspecting all these methods comparatively to their fast versions will reveal the extent of detection rate accuracy sacrificed for speed, in a variety of measurement noise conditions. Note that in all these applications, the data sets do contain genuine outliers.

Fig. 4 summarizes the results obtained. For each of the five error conditions, the mean and one standard deviation of the AUC values over 30 runs on independent realizations of the data are shown on these plots. Fig. 4(a) shows the in-sample performance whereas Fig. 4(b) shows the out of sample performance, i.e., the ability to detect genuine outliers in previously unseen data from the same density model. In the figures, the solid lines indicate the results obtained by the accelerated methods, while the dashed lines are obtained by the original methods.

The results are quite intuitive, and several conclusions may be traced from these figures.

- 1) It can be seen that the differences between MoT/Idealized and FMoT/Idealized is very small

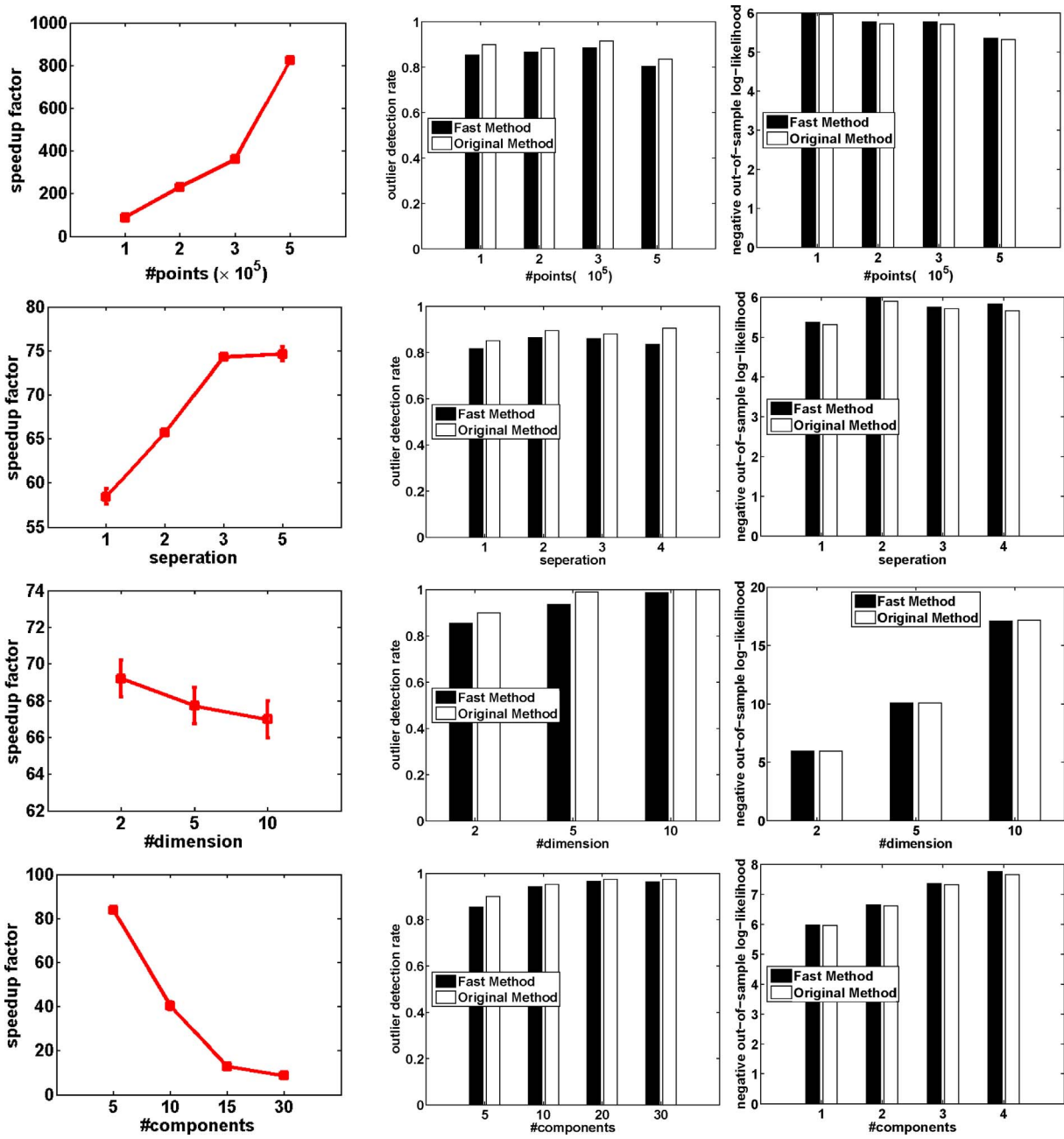


Fig. 3. Comparative results of the accelerated algorithm (FMoT/Error against MoT/Error) on data sets with varying number of data points (row 1), varying extent of component separation (row 2), varying data dimensionality (row 3), and varying number of clusters (row 4). The assessment criteria are: the speedup factor (column 1), the out-of-sample log-likelihood (column 2), and the correct outlier detection rate (column 3). The bar charts show averages of 30 runs on independent realizations of the data sets.

for both the in-sample and out-of-sample performance. This indicates the success of the proposed speedup approach in noise-free scenarios.

- 2) Among the five algorithms, FMoT/Base (i.e., the accelerated MoT on the contaminated data) deteriorates fastest with increasing error levels. This was expected, since the measurement error information is not taken into account, while further loss is incurred for the variational speedup scheme.
- 3) MoT/Error exhibits a systematic and increasingly statistically significant improvement with respect to

MoT/Base, as the measurement uncertainty increases, both on in-sample data and on out-of-sample data.

- 4) The picture is similar between FMoT/Base and FMoT/Error—though, of course the fast version performs below the non-accelerated one.
- 5) It is worth observing that the performance of the fast FMoT/Error method is about the same (and even better at high noise levels) as that of MoT/Base. Clearly, the extra complication in the formalism pays off here, as the accuracy loss incurred by the speedup is recovered by the modeling of measurement errors.

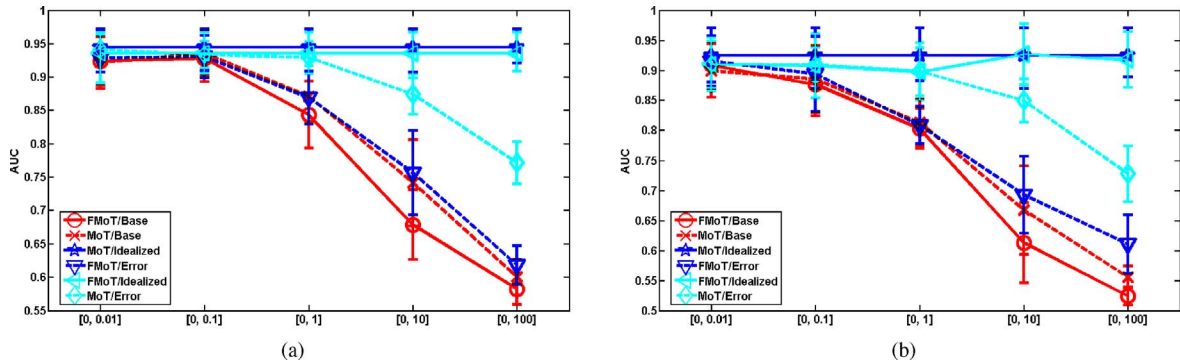


Fig. 4. Comparison results in varying measurement error conditions, between the fast MoT algorithm on clean data (FMoT/Idealized), the fast MoT on data with measurement errors (FMoT/Base), the original MoT algorithm on clean data (MoT/Idealized), the original MoT on data with measurement error (MoT/Base), the fast MoT with measurement errors on data with measurement errors (FMoT/Error), and the original MoT with measurement error (MoT/Error). Comparison on (a) in-sample and (b) out-of-sample data.

We can conclude, therefore, that FMoT/Error represents a desirable tradeoff between accuracy and speed for analyzing large data sets, while MoT/Error should be the procedure of choice when speed is not an issue and accuracy is at a premium. This is increasingly so, as the error levels get larger. As we have seen, at small-to-medium error levels, the accuracy traded off for speed is much smaller.

#### D. Non-Uniformly Sampled Outliers and Further Comparisons

One of the notable advantages of modeling data that contains outliers by a  $t$ -mixture (as opposed to, e.g., a mixture of Gaussians and a uniform distribution) is that the distribution of outliers is not pre-specified and not restricted [36]. In the previous experiments, the outliers have been uniformly sampled from the data space—however the conclusions hold also for non-uniform outlier distributions. To convince ourselves, we tested the performance of the developed speedup algorithm (FMoT/Error) when the outliers are sampled from half of the data space—and an example of this is shown in Fig. 5(a). On this plot, “x” denotes the outliers, and the figure displays a set of 10 000 points drawn from a 5-component 2-separated Gaussian mixtures in two dimensions with unit variance Gaussian errors added on each data point. A comprehensive set of results on data sets having such non-uniform outlier distributions, while varying the cluster-separation parameter and the number of clusters, confirmed indeed the same behavior as seen before in the previous section.

We also compared these results with a recent Bayesian approach developed in [5], the robust Bayesian clustering (RBC) via type-2 Student Mixture Model, in a low-error setting (variance of errors in the data was set to 1). One should note that unlike our algorithm, this method has not been designed to take advantage of measurement error information, and has not been the subject of computational speedup procedures either. Nevertheless it represents the state of the art, hence it is indicative from a user’s point of view to see comparative performance test results. The Bayesian method [5] is computationally quite intensive, however one should note that similar developments to the ones presented in this paper could be made straightforwardly to this method too.

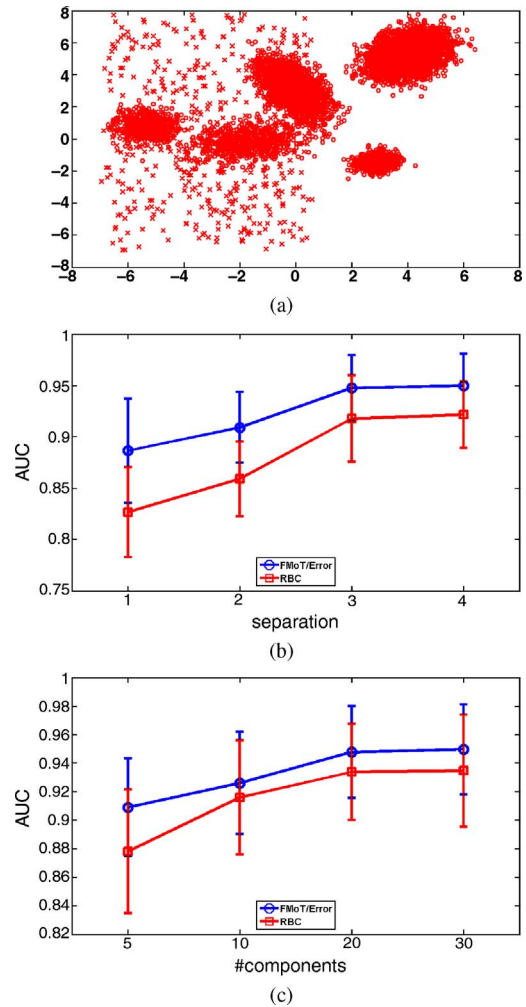


Fig. 5. Experimental results for data sets with non-uniform outliers. (a) Exemplar of data set where the outliers sampled from half of the data space. (b) and (c) AUC values obtained by FMoT/Error and RBC, when varying the data characteristics: (b) separation parameter  $c = 1, 2, 3, 4$ , and (c) number of data clusters  $K = 5, 10, 20, 30$ .

Structurally, RBC maintains the posterior correlations between the indicator variables  $z$  and the scale variables  $u$ , just as our method [41] and FMoT/Error, and in addition RBC also posits priors on the model parameters and takes a fully Bayesian approach inferring variational posteriors over them.

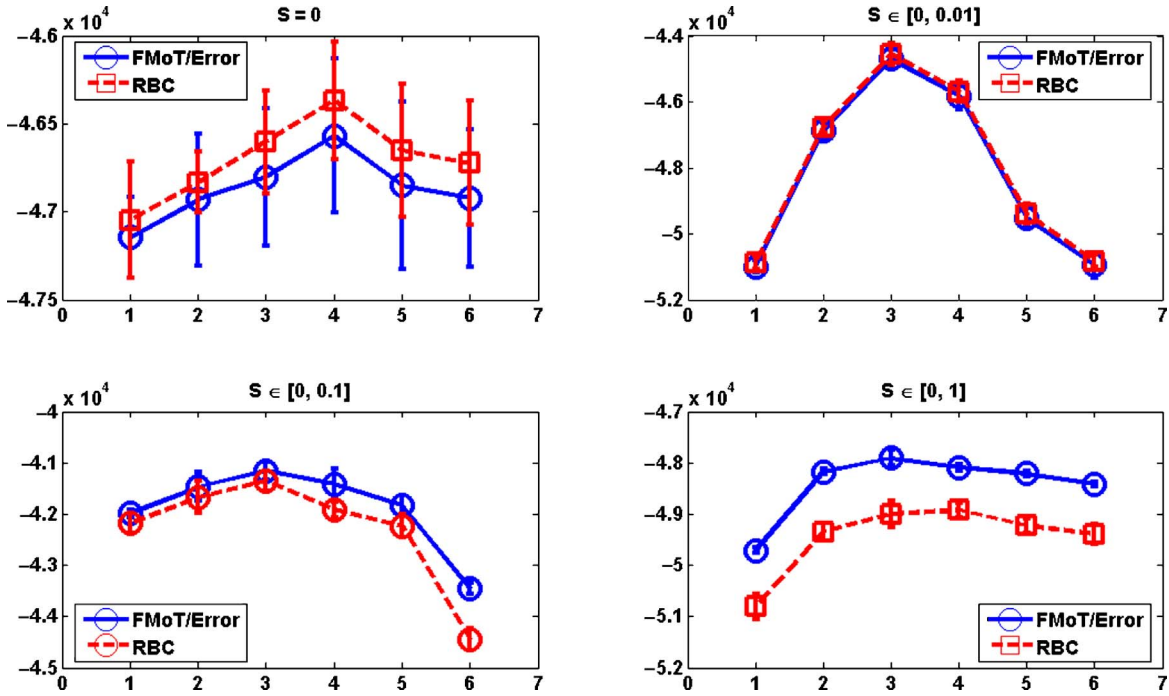


Fig. 6. Model selection experiments for finding the optimal number of clusters. The true number of clusters is three, and  $c = 2$ . The  $x$ -axis represents the range of model orders tested and the  $y$ -axis is the minimum message length of our model FMoT/Error/MML (solid lines) compared with the Bayesian log evidence bound in RBC (dashed lines). We see the two methods behave similarly, and in most cases they both pick the true number of clusters.

Fig. 5(b) and (c) summarizes the variation of the AUC performances of FMoT/Error when varying the cluster-separation of the data sets  $c = 1, 2, 3, 4$  [Fig. 5(b)] and when varying the true number of clusters  $K = 5, 10, 20, 30$  [Fig. 5(c)] comparatively with the existing version of the RBC of [5]. We see that both methods behave similarly in terms of the variation of the accuracy with the varied parameter values, and the actual performance of the two methods is also comparable in this setting. The slightly higher performance of our FMoT/Error is most likely due to its capability of taking advantage of the knowledge of measurement errors. Still, it is quite remarkable that this advantage keeps its performance above of that of RBC despite the computational speedup procedure.

#### E. Assessment of the Proposed Procedure for Determining the Optimal Number of Components

In this section, we carry out controlled experiments to assess the MML-based model selection procedure for our speedup algorithm, called FMoT/Error/MML, as described in Section III-G. We demonstrate results on a data set of 100 000 data points, three clusters, two dimensions, and separation parameter  $c = 2$ , which has been created using the same sampling mechanism as described in Section IV-C. We test four different measurement error levels, i.e., the diagonal elements of the error variance matrix  $S$  will range between 0 (error-free),  $[0, 0.01]$ ,  $[0, 0.1]$ , and  $[0, 1]$ , respectively.

In these experiments, we test the model orders from one to six. For each of these, the average log-likelihood bound over 30 runs is summarized and shown in Fig. 6. For comparison we use the RBC via type-2 Student mixture model developed by Archambeau *et al.* [5], in which the full Bayesian treat-

ment allows for model selection, via the log evidence bound. As somewhat expected on the basis of known relationships between Bayesian and information-theoretic model selection [23], we see the two methods behave similarly indeed, and in most cases they both pick the true number of clusters. In the error-free case we can observe that the inaccuracy incurred by the speedup is negligible, while as the error level increases the advantage of our model's ability of taking into account these errors becomes apparent despite the speedup.

#### F. Application Example: Detecting High-Redshift Quasars From the Sloan Digital Sky Survey (SDSS) Quasar Catalogue

In [41], the MoT/Error approach has been applied to a sample of quasars from the SDSS quasar photometric catalogue [51]. The benefits of this model-based approach was demonstrated against existing naive approaches. Here, we show that we can now achieve statistically similar detection rates with a significantly decreased computational effort. The actual domain-specific astronomical interpretation of the results is outside the scope of this paper (and will be presented elsewhere).

As in [41], the initial features are five observed magnitudes:  $u_{\text{mag}}$ ,  $g_{\text{mag}}$ ,  $i_{\text{mag}}$ ,  $r_{\text{mag}}$ , and  $z_{\text{mag}}$ , from which we created four features relating to color, by subtracting  $r_{\text{mag}}$  from each, to avoid discriminating on the basis of brightness. The measurement errors are known for each feature and each object, and spectroscopic redshift estimates are available that we use for validating our results. The redshift relates to the distance of the object from us, and with higher redshift the entire spectral pattern is more severely shifted, therefore high redshift quasars may be perceived as outliers in the color space [16]. However,

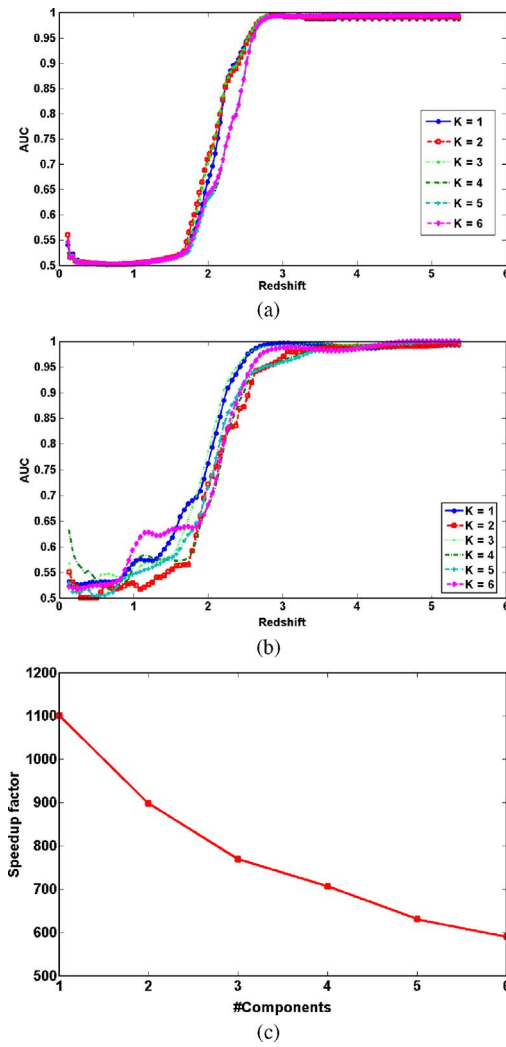


Fig. 7. AUC versus possible redshift thresholds, as obtained by (a) MoT/Error [41] and (b) FMoT/Error from the SDSS quasar sample. (c) CPU time taken by FMoT/Error versus MoT/Error.

as already mentioned in [41], a comprehensive model based approach that both: 1) works in the multivariate feature space; and 2) takes principled account of the measurement errors has not been available to astronomers.

We now apply both our MoT/Error and FMoT/Error to the same sample of 10000 quasars and compute the AUC values against a varying redshift threshold. The resulting graphs are shown on Fig. 7, with different number of clusters  $K$ . The  $y$ -coordinate of each point on this curve indicates the probability of detecting quasars of redshift greater than its  $x$ -coordinate.

We can see the detection rates are nearly identical with both methods. However, the time required by the accelerated algorithm FMoT/Error is just a fraction of that required by MoT/Error.

Similarly to MoT/Error, FMoT/Error is able to identify as outliers an overwhelming fraction of quasars at redshifts as low as 2.5, whereas the 2-D projection based method of [16] can manage only those with  $z > 3.5$ , which are extremely rare, and obvious from naive projections. By being able to identify the latter category, when the SDSS galaxy catalogue is complete

with four-color magnitudes, our principled approach is well placed to accurately identifying an order of magnitude more quasars than existing methods would.

## V. CONCLUSION

We presented a robust mixture based approach to infer the genuine outliers from erroneous data, by exploiting the availability of measurement errors. We developed a fast variational EM algorithm for parameter estimation in this model, based on a tree-structured variational approximation and partitioned posterior assignments. We conducted extensive experiments to study the accuracy/speed tradeoffs achievable in a variety of data conditions, and analyzed in detail the characteristics of our algorithm based on results obtained on appropriately designed synthetic data experiments. Among findings, at low-to-moderate error levels, the speedup factor was found to be at least linear in the number of data points, without significantly sacrificing the detection accuracy. We demonstrated clear benefits of including measurement error information into the modeling, and the accuracy gain was able to recover the loss incurred by the speedup procedure in large error conditions. Finally, we have also shown the working of our approach on a real application example, demonstrating both effectiveness and efficiency. In the future, this paper could be applied to the exciting area of sensor networks [22].

## ACKNOWLEDGMENT

The authors would like to thank S. Raychaudhury for early work on the astronomy application. They also would like to thank anonymous reviewers for their constructive comments and suggestions.

## APPENDIX A

### DERIVATION OF AUXILIARY POSTERIOR DISTRIBUTION

This section details the derivation of (31). We need to maximize the free-energy functional  $\mathcal{F}_{\mathcal{A}}$  with respect to  $q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)$ , subject to the constraint  $\int q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)d\mathbf{w} = 1$ . Discarding terms that are independent of the random variable  $\mathbf{w}_{\mathcal{A}}$ , and using the Lagrange multiplier method, the functional to be maximized is the following:

$$\begin{aligned} \mathcal{F}_{\mathbf{w}_{\mathcal{A}}|k} &= q(z_{\mathcal{A}} = k) \cdot \\ &\sum_{\mathbf{t}_n \in \mathcal{A}} \langle \log [p(\mathbf{t}_n|\mathbf{w}_{\mathcal{A}})p(\mathbf{w}_{\mathcal{A}}|u_{\mathcal{A}}, z_{\mathcal{A}} = k)] \rangle_{\mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|z_{\mathcal{A}}=k} + \\ &\lambda \left( \int q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)d\mathbf{w} - 1 \right) - \\ &q(z_{\mathcal{A}} = k) \langle \log q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{\mathbf{w}_{\mathcal{A}}|k}. \end{aligned} \quad (53)$$

If we let

$$E(\mathcal{A}, k) = \sum_{\mathbf{t}_n \in \mathcal{A}} \langle \log [p(\mathbf{t}_n|\mathbf{w}_{\mathcal{A}})p(\mathbf{w}_{\mathcal{A}}|u_{\mathcal{A}}, z_{\mathcal{A}} = k)] \rangle_{u_{\mathcal{A}}|k}$$

then taking derivatives of  $\mathcal{F}_{\mathbf{w}_{\mathcal{A}}|k}$  with respect to  $q_{\mathcal{A}}(\mathbf{w}_{\mathcal{A}}|k)$  and  $\lambda$ , we have

$$\begin{aligned} \frac{\partial \mathcal{F}_{\mathbf{w}_{\mathcal{A}}|k}}{\partial q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)} &= -q(z_{\mathcal{A}} = k) [1 + \log q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)] \\ &+ q(z_{\mathcal{A}} = k)E(\mathcal{A}, k) + \lambda \end{aligned} \quad (54)$$

$$\frac{\partial \mathcal{F}_{\mathbf{w}|k}}{\partial \lambda} = \int q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) d\mathbf{w} - 1. \quad (55)$$

Equating these to zero, from the Karush-Kuhn-Tucker conditions we have

$$q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) = \frac{\exp\{E(\mathcal{A}, k)\}}{\exp\left\{1 - \frac{\lambda}{q(z_{\mathcal{A}}=k)}\right\}}. \quad (56)$$

Taking integral with respect to  $q(\mathbf{w}_{\mathcal{A}}|k)$  on both sides, we get

$$\lambda = q(z_{\mathcal{A}} = k) \left(1 - \log \left[ \int \exp(E(\mathcal{A}, k)) d\mathbf{w} \right]\right). \quad (57)$$

Finally, replacing (57) into (54), we get

$$q(\mathbf{w}_{\mathcal{A}}|k) = \frac{\exp(E(\mathcal{A}, k))}{\int \exp(E(\mathcal{A}, k)) d\mathbf{w}} \quad (58)$$

which leads to (31) given in the main text.

The derivation of (32) and (33) follows the same line of thought. This is straightforward, we omit the details.

## APPENDIX B

### LOG-LIKELIHOOD BOUND AND CACHED VALUES

The approximated free-energy function  $\mathcal{F}_{\mathcal{P}}$  can be used to monitor the variational optimization process. Moreover, the evaluation is also needed to compute the responsibilities  $q_{\mathcal{A}}(k)$ , with respect to the following terms:

$$\begin{aligned} \langle \log p(\mathbf{t}_n, \mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_q &= \langle \log p(\mathbf{t}_n|\mathbf{w}_{\mathcal{A}}) \rangle_{\mathbf{w}_{\mathcal{A}}|k} + \\ &\langle \log p(u_{\mathcal{A}}|k) \rangle_{u_{\mathcal{A}}|k} + \langle \log p(\mathbf{w}_{\mathcal{A}}|u_{\mathcal{A}}, z_{\mathcal{A}} = k) \rangle_{\mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|k} \end{aligned} \quad (59)$$

$$\begin{aligned} \langle \log q(\mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_q &= \langle \log q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{\mathbf{w}_{\mathcal{A}}|k} + \\ &\langle \log q(u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{u_{\mathcal{A}}|k} \end{aligned} \quad (60)$$

where  $\langle \cdot \rangle_{\mathbf{w}_{\mathcal{A}}|k}$  denotes the expectation with respect to  $q(\mathbf{w}_{\mathcal{A}}|k)$ , and  $\langle \cdot \rangle_{u_{\mathcal{A}}|k}$  denotes the expectation with respect to  $q(u_{\mathcal{A}}|z_{\mathcal{A}} = k)$ . The terms can be evaluated as follows:

$$\begin{aligned} \langle \log p(\mathbf{t}_n|\mathbf{w}_{\mathcal{A}}) \rangle_{\mathbf{w}_{\mathcal{A}}|k} &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{S}_n| - \\ &\frac{1}{2} \left[ (\mathbf{t}_n - \mathbf{w}_{\mathcal{A}|k})^T \mathbf{S}_n^{-1} (\mathbf{t}_n - \mathbf{w}_{\mathcal{A}|k}) + \text{Tr}(\Sigma_{\mathbf{w}_{\mathcal{A}|k}} \mathbf{S}_n^{-1}) \right] \end{aligned} \quad (61)$$

$$\begin{aligned} \langle \log p(\mathbf{w}_n|u_{\mathcal{A}}, z_{\mathcal{A}} = k) \rangle_{\mathbf{w}_{\mathcal{A}}, u_{\mathcal{A}}|k} &= -\frac{d}{2} \log(2\pi) - \\ &\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \langle u \rangle_{\mathcal{A}|k} \left[ (\mathbf{w}_{\mathcal{A}|k} - \mu_k)^T \Sigma_k^{-1} (\mathbf{w}_{\mathcal{A}|k} - \mu_k) \right] \\ &- \frac{1}{2} \langle u \rangle_{\mathcal{A}|k} \left[ \text{Tr}(\Sigma_{\mathbf{w}_{\mathcal{A}|k}} \Sigma_k^{-1}) \right] + \frac{d}{2} \langle \log u \rangle_{\mathcal{A}|k} \end{aligned} \quad (62)$$

$$\begin{aligned} \langle \log p(u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{u_{\mathcal{A}}|k} &= \frac{v_k}{2} \log \frac{v_k}{2} - \frac{v_k}{2} \langle u \rangle_{\mathcal{A}|k} + \\ &\left( \frac{v_k}{2} - 1 \right) \langle \log u \rangle_{\mathcal{A}|k} - \log \Gamma \left( \frac{v_k}{2} \right) \end{aligned} \quad (63)$$

$$\langle \log q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{\mathbf{w}_{\mathcal{A}}|k} = \frac{1}{2} \log |\Sigma_{\mathbf{w}_{\mathcal{A}|k}}| + \frac{d}{2} + \text{const.} \quad (64)$$

$$\begin{aligned} \langle \log q(u_{\mathcal{A}}|z_{\mathcal{A}} = k) \rangle_{u_{\mathcal{A}}|k} &= -[a_{\mathcal{A}|k} \log b_{\mathcal{A}|k} + (a_{\mathcal{A}|k} - 1) \langle \log u \rangle_{\mathcal{A}|k}] \\ &+ a_{\mathcal{A}|k} + \log \Gamma(a_{\mathcal{A}|k}). \end{aligned} \quad (65)$$

As we will see, the responsibility  $q(z_{\mathcal{A}} = k)$  for each cell  $\mathcal{A}$ , and the posterior  $q(\mathbf{w}_{\mathcal{A}}|z_{\mathcal{A}} = k)$  can be efficiently computed at each iteration if some statistics of the points in  $\mathcal{A}$  are stored in advance. Note that in evaluating  $\sum_{\mathbf{t}_n \in \mathcal{A}} \langle \log p(\mathbf{t}_n|\mathbf{w}_{\mathcal{A}}) \rangle_{\mathbf{w}_{\mathcal{A}}|k}$ , we have

$$\begin{aligned} &\sum_{\mathbf{t}_n \in \mathcal{A}} (\mathbf{t}_n - \mathbf{w}_{\mathcal{A}|k})^T \mathbf{S}_n^{-1} (\mathbf{t}_n - \mathbf{w}_{\mathcal{A}|k}) \\ &= \sum_{\mathbf{t}_n \in \mathcal{A}} [\mathbf{t}_n^T \mathbf{S}_n^{-1} \mathbf{t}_n - 2\mathbf{t}_n^T \mathbf{S}_n^{-1} \mathbf{w}_{\mathcal{A}|k} + \mathbf{w}_{\mathcal{A}|k}^T \mathbf{S}_n^{-1} \mathbf{w}_{\mathcal{A}|k}] \\ &= \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{t}_n^T \mathbf{S}_n^{-1} \mathbf{t}_n - 2 \left( \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{t}_n^T \mathbf{S}_n^{-1} \right) \mathbf{w}_{\mathcal{A}|k} + \\ &\quad \text{Tr} \left( \mathbf{w}_{\mathcal{A}|k} \mathbf{w}_{\mathcal{A}|k}^T \left[ \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{S}_n^{-1} \right] \right). \end{aligned} \quad (66)$$

Since the following values:

$$\sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{t}_n^T \mathbf{S}_n^{-1} \mathbf{t}_n, \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{t}_n^T \mathbf{S}_n^{-1} \text{ and } \sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{S}_n^{-1} \quad (67)$$

are independent of the training process, we can cache them in the node for cell  $\mathcal{A}$ . Moreover, we also cache

$$\langle \mathbf{t} \rangle_{\mathcal{A}} = \frac{\sum_{\mathbf{t}_n \in \mathcal{A}} \mathbf{t}_n}{|\mathcal{A}|} \quad (68)$$

i.e., the mean value of the noisy data in cell  $\mathcal{A}$ .

## REFERENCES

- [1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
- [2] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 203–215, Feb. 2005.
- [3] C. Archambeau, N. Delannay, and M. Verleysen, "Robust probabilistic projections," in *Proc. 23rd Int. Conf. Mach. Learning*, Jun. 2006, pp. 33–40.
- [4] C. Archambeau, N. Delannay, and M. Verleysen, "Mixtures of robust probabilistic principal component analyzers," *Neurocomputing*, vol. 71, nos. 7–9, pp. 1274–1282, Mar. 2008.
- [5] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Netw.*, vol. 20, no. 1, pp. 129–138, Jan. 2007.
- [6] J. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.
- [7] M. M. Breuning, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2000, pp. 93–104.
- [8] G. Chantas, N. Galasanos, A. Likas, and M. Saunders, "Variational Bayesian image restoration based on a product of t-distributions image prior," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1795–1805, Oct. 2008.
- [9] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "A robust approach toward sequential data modeling and its application in automatic gesture recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2008, pp. 1937–1940.
- [10] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "Signal modeling and classification using a robust latent space model based on t-distributions," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 949–963, Mar. 2008.
- [11] S. Chatzis and T. Varvarigou, "Robust fuzzy clustering using mixtures of student's t-distributions," *Pattern Recognit. Lett.*, vol. 29, no. 13, pp. 1901–1905, Oct. 2008.
- [12] S. Chatzis and T. Varvarigou, "Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributions," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 505–517, Jun. 2009.
- [13] A. Chaudhary, A. Szalay, and A. Moore, "Very fast outlier detection in large multidimensional data sets," in *Proc. ACM SIGMOD Workshop Res. Issues DMKD*, Jun. 2002, pp. 45–52.

- [14] S. Dasgupta, "Learning mixtures of Gaussians," in *Proc. IEEE Symp. Foundations Comput. Sci.*, vol. 40, Oct. 1999, pp. 634–644.
- [15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [16] X. Fan, M. Strauss, G. Richards, J. Hennawi, R. Becker, and *et al.*, "A survey of  $z > 5.7$  quasars in the Sloan Digital Sky Survey. IV. Discovery of seven additional quasars," *Astron. J.*, vol. 131, pp. 1203–1209, Mar. 2006.
- [17] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," HP Labs, Palo Alto, CA, Tech. Rep. HPL-2003-4, 2003.
- [18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
- [19] S. Geisser, *Predictive Inference*. New York: Chapman and Hall, 1993.
- [20] Z. Ghahramani and M. I. Jordan, "Learning from incomplete data," MIT Press, Cambridge, MA, Tech. Rep. AIM-1509, Oct. 1994.
- [21] P. Grünwald and J. Langford, "Suboptimal behavior of Bayes and MDL in classification under misspecification," *Mach. Learning*, vol. 66, nos. 2–3, pp. 119–149, Mar. 2007.
- [22] D. Gu, "Distributed em algorithm for Gaussian mixtures in sensor networks," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1154–1166, Jul. 2008.
- [23] A. Honkela and H. Valpola, "Variational learning and bits-back coding: An information theoretic view to Bayesian learning," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 800–810, Jul. 2004.
- [24] S. Hou and A. Galata, "Robust estimation of Gaussian mixtures from noisy input data," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2008, pp. 1–8.
- [25] T. Johnson, I. Kwok, and R. Ng, "Fast computation of 2-D depth contours," in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining*, Aug. 1998, pp. 224–228.
- [26] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Mach. Learning*, vol. 37, no. 2, pp. 183–233, Nov. 1999.
- [27] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proc. 24th Int. Conf. Very Large Data Bases*, Aug. 1998, pp. 392–403.
- [28] C. Liu and D. Rubin, "ML estimation of the  $t$  distribution using EM and its extensions, ECM and ECME," *Stat. Sin.*, vol. 5, no. 1, pp. 19–39, Jan. 1995.
- [29] C. Liu, D. Rubin, and Y. Wu, "Parameter expansion to accelerate EM: The PX-EM algorithm," *Biometrika*, vol. 85, no. 4, pp. 755–770, Dec. 1998.
- [30] G. McLachlan, R. Bean, and L.-T. Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate  $t$ -distribution," *Comput. Stat. Data Anal.*, vol. 51, no. 11, pp. 5327–5338, Jul. 2007.
- [31] T. Minka and J. Winn, "Gates," in *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press, Dec. 2008, pp. 1073–1080.
- [32] A. Moore, "Very fast EM-based mixture model clustering using multiresolution KD-tree," in *Advances in Neural Information Processing Systems*, vol. 11. Denver, CO: MIT Press, Nov. 1999, pp. 543–549.
- [33] A. Moore and M. Lee, "Cached sufficient statistics for efficient machine learning with large data sets," *J. Artif. Intell. Res.*, vol. 8, pp. 67–91, Mar. 1998.
- [34] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Boston, MA: Kluwer Academic, Sep. 1998, pp. 355–368.
- [35] S. Papadimitriou and C. Faloutsos, "Cross-outlier detection," in *Proc. 8th Int. Symp. SSTD*, LNCS 2750. Jul. 2003, pp. 199–213.
- [36] D. Peel and G. McLachlan, "Robust mixture modeling using the  $t$  distribution," *Stat. Comput.*, vol. 10, no. 4, pp. 339–348, Oct. 2000.
- [37] D. Pelleg and A. Moore, "Accelerating exact k-means algorithms with geometric reasoning," in *Proc. 5th ACM SIGKDD*, Aug. 1999, pp. 277–281.
- [38] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, Jun. 2000.
- [39] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1998.
- [40] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
- [41] J. Sun, A. Kabán, and S. Raychaudhury, "Robust mixtures in the presence of measurement errors," in *Proc. 24th Int. Conf. Mach. Learning*, Oct. 2007, pp. 847–854.
- [42] J. Sun, A. Kabán, and S. Raychaudhury, "Robust visual mining of data with error information," in *Proc. 11th Eur. Conf. PKDD*, Sep. 2007, pp. 573–580.
- [43] M. Svensen and C. Bishop, "Robust Bayesian mixture modeling," *Neurocomputing*, vol. 64, pp. 235–252, Mar. 2005.
- [44] G. Terrell and D. Scott, "Variable kernel density estimation," *Ann. Stat.*, vol. 20, no. 3, pp. 1236–1265, Sep. 1992.
- [45] M. Tipping and N. Lawrence, "Variational inference for student- $t$  models: Robust Bayesian interpolation and generalized component analysis," *Neurocomputing*, vol. 69, nos. 1–3, pp. 123–141, Dec. 2005.
- [46] D. Tzikas, A. Likas, and N. Galasanos, "Variational Bayesian sparse kernel-based blind image deconvolution with Student's  $t$ -distribution," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 753–764, Apr. 2009.
- [47] A. Vellido, P. Lisboa, and D. Vicente, "Handling outliers and missing data in brain tumor clinical assessment using  $t$ -GTM," *Comput. Biol. and Medicine*, vol. 36, no. 10, pp. 1049–1063, Oct. 2005.
- [48] J. Verbeek, J. Nunnink, and N. Vlassis, "Accelerated EM-based clustering of large data sets," *Data Mining Knowl. Discovery*, vol. 13, no. 3, pp. 291–307, Nov. 2006.
- [49] N. Vlassis and J. Verbeek, "Gaussian mixture learning from noisy data," Informatics Instit., Faculty Sci., Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep. IAS-UVA-04-01, 2004.
- [50] C. Wallace and D. Dowe, "Minimum message length and Kolmogorov complexity," *Comput. J.*, vol. 42, no. 4, pp. 270–283, 1999.
- [51] D. York, "The Sloan digital sky survey: Technical summary," *Astron. J.*, vol. 120, no. 3, pp. 1579–1587, Sep. 2000.
- [52] J. Zhao and Q. Jiang, "Probabilistic PCA for  $t$ -distributions," *Neurocomputing*, vol. 69, nos. 16–18, pp. 2217–2226, Oct. 2006.

**Jianyong Sun** received the B.S. degree in computational and applied mathematics from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree in computer science from the University of Essex, Essex, U.K., in 2006.

He is currently a Research Fellow with the Center for Plant Integrative Biology, School of Bioscience, The University of Nottingham, Sutton Bonington, U.K. His current research interests include optimization, statistical machine learning, and data mining.

Dr. Sun is a member of the Intelligent Modeling and Analysis Research Group, School of Computer Science, University of Nottingham.

**Ata Kabán** received the B.S. degree in computer science from the Babeş-Bolyai University, Cluj-Napoca, Romania, in 1999, the Ph.D. degree in musicology from the Gheorghe Dima Music Academy, Cluj-Napoca, Romania, in 1999, and the Ph.D. degree in computer science from the University of Paisley, Paisley, U.K., in 2001.

She is currently a Lecturer with the School of Computer Science, University of Birmingham, Birmingham, U.K. Her current research interests include statistical machine learning and data mining.