# Robust Estimation of Gaussian Mixtures from Noisy Input Data

Shaobo Hou Aphrodite Galata

School of Computer Science
The University of Manchester

## Abstract

*We propose a variational bayes approach to the problem of robust estimation of gaussian mixtures from noisy input data. The proposed algorithm explicitly takes into account the uncertainty associated with each data point, makes no assumptions about the structure of the covariance matrices and is able to automatically determine the number of the gaussian mixture components. Through the use of both synthetic and real world data examples, we show that by incorporating uncertainty information into the clustering algorithm, we get better results at recovering the true distribution of the training data compared to other variational bayesian clustering algorithms.*

## 1. Introduction

Standard EM-based clustering algorithms [7] assume that input data points are all equally important and the effect of noise or measurement errors is often ignored and not explicitly modeled during model estimation. However, this is not always a valid assumption since the input data can be corrupted by measurement errors. Uncertainties can also be introduced by additional transformations on the data such as dimensionality reduction. In particular, in the case of non-linear transformations, it is no longer safe to assume that the uncertainties are uniform across the data. If the level of uncertainties can be quantified, it makes sense to incorporate them into the clustering algorithm to improve the estimation of the true data distribution.

In this paper we propose a novel algorithm for learning a mixture of Gaussians that takes into account the uncertainties of the input data. In our formulation we assume the uncertainty on a data point can be modelled by a multivariate Gaussian distribution and is independent from the other data points. Intuitively, this allows a data point with large uncertainty to exert less influence on the mixture components, than a data point with smaller uncertainty. The optimal mixture model that represents the data with uncertainties is found using a variational bayesian algorithm that automatically chooses the appropriate number of components in the mixture model. We show that by taking into account the uncertainty of information, our algorithm performs better at estimating the correct number of clusters and recovering the true distribution of the training data compared to other variational bayesian clustering algorithms [6, 3]. The proposed algorithm is evaluated on a number of synthetic and real data sets and is shown to improve the results of various pattern recognition tasks such as motion segmentation and partitioning of microarray gene expressions.

## 2. Related Work

Previously, researchers have looked into the problem of incorporating uncertainty information into the field of model fitting and have developed algorithms such as Total Least Square [1] which assumes the noise on all data points are drawn from the same uncertainty distribution, and Fundamental Numerical Scheme [5] which allows different data points to be associated with different uncertainty distributions. It has also been studied in the context of support vector machine classification [2].

Various researchers have also investigated the problem of unsupervised clustering of data with uncertainty. Chaudhuri and Bhowmik [4] proposed a modified K-means algorithm which assumes uniform uncertainties such that the true position of a data point can be anywhere within a hypersphere centred on its observed position. Kumar and Patel [11] also proposed generalisations of K-means and hierarchical clustering to handle zero-mean Gaussian measurement uncertainty. However their formulation is simply based on intuition and is not probabilistically well principled. Handman and Govaert [8] proposed an EM [7] clustering algorithm which modelled non-identically distributed uncertainty as rectangular error zones.

More recently in Bioinformatics, in order to group genes with similar expression patterns from microarray experiments, Liu et al. [12] proposed a probabilistic clustering algorithm for estimating the maximum likelihood mixture of Gaussians with spherical covariance matrices from data with zero-mean Gaussian measurement errors represented

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

by diagonal covariance matrices. In their method, the parameters of the mixture components are updated using gradient descent based optimisation, and Bayesian Information Criterion (BIC) [14] is used to determine the appropriate number of components. Our method is more general and does not make any assumption about the structure of the covariance matrices of the mixture components and the measurement uncertainties. The variational bayesian model selection used by our method is well principled and handles datasets of various sizes whereas BIC is an asymptotic result which may not be able to handle small datasets well.

Sun et al.[16] proposed a generalised Expectation Maximisation (GEM) algorithm for estimating the maximum likelihood mixture of Student t-distributions from astrophysical datasets with measurement errors and thus improving the detection of peculiar quasars as statistical outliers. Our algorithm is similar to theirs in that we both solve the intractable integration in the resulting formulation using variational approximation. However, while their algorithm returns the maximum likelihood solution for a fixed number of components, our algorithm finds the optimal number of components during a single clustering run. We also place priors on the parameters of the mixture components to prevent singularity and components with degenerate shapes from forming.

As the level of uncertainties on the input data increases, maximum likelihood algorithms such as [12] [16] tend to return mixtures with some very narrow components which are collapsing onto a point or a hyperplane. The t-distribution is somewhat more robust to this effect than the Gaussian distribution, it can still be affected by it. Although this is not an issue if the estimated mixture model is only going to be used in some discriminative tasks, it does however, pose a serious problem if it is meant for generative tasks which requires sampling and synthesising from the learnt distribution. This is not a problem in our clustering algorithm because we place priors on the parameters of the mixture components which constrains the shape of the components to prevent them from collapsing.

## 3. Learning mixture of Gaussians in the presence of errors

Our aim is to estimate a mixture of $K$ Gaussian components which best represents the distribution of a set of $\mathbf{N}$ data points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that have been observed in a $D$ dimensional space, where the optimal number of component $K$ is unknown. It is assumed that each data point $\mathbf{x}_n$ is a noisy measurement of its true position and is drawn from a Gaussian distribution $\mathcal{N}(\mathbf{x}_n|\mathbf{t}_n, \mathbf{C}_n)$. The mean $\mathbf{t}_n$ is the unknown true position of the data point and the covariance matrix $\mathbf{C}_n$ is the known uncertainty on the measurement.

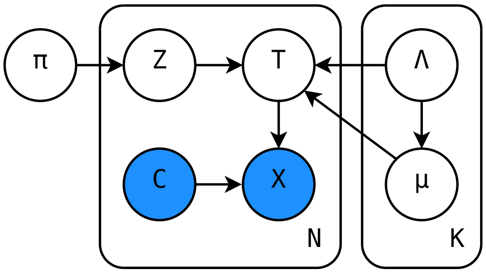


Figure 1. Mixture of Gaussians with Uncertainty

Under a latent variable model, we associate each data point $\mathbf{x}_n$ with a binary latent variable $\mathbf{z}_n = \{z_{n1}, \ldots, z_{nk}\}_{k=1}^{K}$, in which only one element is set to one to indicate that a data point $\mathbf{t}_n$ was generated from that component and all other elements of the latent variable are equal to zero. Given the latent variables $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ and $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$, the conditional distribution of $\mathbf{X}$ can be written as:

$$p(\mathbf{X}|\mathbf{C}, \mathbf{T}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{t}_n, \mathbf{C}_n) \tag{1}$$

And the conditional distributions of $\mathbf{T}$ and $\mathbf{Z}$ are given by:

$$p(\mathbf{T}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}} \tag{2}$$

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{3}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ are the mean vector and precision matrix of the $k$th Gaussian component and $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ are the mixing coefficients for the components.

We place a Normal-Wishart prior on the mean and precision of the Gaussians components as shown by Bishop [3]:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0) \tag{4}$$

where $\mathbf{m}_0$ is usually set to zero and $\beta_0$ is set to a very small value. $\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu)$ is a Wishart distribution with scale matrix $\mathbf{W}$ and $\nu$ degrees of freedom.

The model described can be represented as a directed graph in figure 3. Assuming uniform prior on $\mathbf{C}$, the joint distribution over all variables conditioned on $\boldsymbol{\pi}$ is:

$$\begin{aligned} p(\mathbf{X}, \mathbf{C}, \mathbf{T}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\pi}) &= p(\mathbf{X}|\mathbf{C}, \mathbf{T})p(\mathbf{T}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \times \\ & \quad p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \end{aligned} \tag{5}$$

### 3.1. Variational Inference

Variational inference is a framework for computing analytical approximation to naturally intractable posterior distribution, by restricting the range of functions used for approximation. Given that $\mathbf{X}$ denotes the set of observable variables, $\mathbf{H}$ denotes the set of latent variables and parameters and $q(\mathbf{H})$ is a variational distribution over $\mathbf{H}$, the log marginal probability of $\mathbf{X}$ is:

$$\begin{aligned} \ln p(\mathbf{X}) &= \ln \int p(\mathbf{X}, \mathbf{H}) d\mathbf{H} \\ &= \ln \int q(\mathbf{H}) \frac{p(\mathbf{X}, \mathbf{H})}{q(\mathbf{H})} d\mathbf{H} \\ &\geq \int q(\mathbf{H}) \ln \frac{p(\mathbf{X}, \mathbf{H})}{q(\mathbf{H})} d\mathbf{H} \\ \mathcal{L}(q) &= \int q(\mathbf{H}) \ln \frac{p(\mathbf{X}, \mathbf{H})}{q(\mathbf{H})} d\mathbf{H} \qquad (6) \end{aligned}$$

where $\mathcal{L}(q)$ is a strict lower bound on $\ln p(\mathbf{X})$ and is derived using Jensen's inequality. The difference between $\ln p(\mathbf{X})$ and $\mathcal{L}(q)$ is the Kullback-Leibler divergence between $q$ and $p$, which is always positive.

Assuming that the $q$ distribution factorises with respect to the disjoint groups of variables in $\mathbf{H}$, such that $q(\mathbf{H}) = \prod_{i=1}^{M} q_i(\mathbf{H}_i)$, then the optimal solution for the $j$th factorised distribution is given by the expectation of the log of the joint distribution over all the other $q$ distributions [3]:

$$\ln q_j^*(\mathbf{H}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{H})] + const. \qquad (7)$$

### 3.2. Variational Bayes mixture of Gaussians with Uncertainty

In our case, the set of latent variables is $\{\mathbf{T}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ and its variational distribution can be factorised as:

$$q(\mathbf{H}) = q(\mathbf{T}|\mathbf{Z}) q(\mathbf{Z}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \qquad (8)$$

which can be further factorised:

$$\begin{aligned} q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \prod_{k=1}^{K} q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k) \qquad (9) \\ q(\mathbf{T}|\mathbf{Z}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} q(\mathbf{t}_n | z_{nk} = 1)^{z_{nk}} \qquad (10) \end{aligned}$$

The optimal solution for $q^*(\mathbf{t}_n | z_{nk} = 1)$ is the posterior distribution of the true position $\mathbf{t}_n$ with uncertainty $\mathbf{C}_n$ and conditioned on the $k$th mixture component:

$$q^*(\mathbf{t}_n | z_{nk} = 1) = \mathcal{N}(\mathbf{t}_{n|k} | \mathbf{r}_{n|k}, \mathbf{E}_{n|k}) \qquad (11)$$

where the mean and the covariance are computed as:

$$\begin{aligned} \mathbf{r}_{n|k} &= \mathbf{E}_{n|k} (\mathbf{C}_n^{-1} \mathbf{x}_n + \nu_k \mathbf{W}_k \mathbf{m}_k) \qquad (12) \\ \mathbf{E}_{n|k} &= (\mathbf{C}_n^{-1} + \nu_k \mathbf{W}_k)^{-1} \qquad (13) \end{aligned}$$

The optimal solution for $q^*(\mathbf{Z})$ is computed as:

$$\begin{aligned} q^*(\mathbf{Z}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \qquad (14) \\ r_{nk} &= \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} = \mathbb{E}[z_{nk}] \qquad (15) \end{aligned}$$

where $r_{nk}$ is the responsibility of the $k$th component for the $n$th data point and:

$$\begin{aligned} &\ln \rho_{nk} = \ln \pi_k + \frac{1}{2} \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2} \ln(2\pi) \\ &- \frac{1}{2} \{ D \beta_k^{-1} + \nu_k (\mathbf{r}_{n|k} - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{r}_{n|k} - \mathbf{m}_k) \\ &+ \nu_k \mathrm{Tr}(\mathbf{E}_{\mathrm{n|k}} \mathbf{W}_{\mathrm{k}}) \} - \frac{\mathrm{D}}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}_{\mathrm{n}}| \\ &- \frac{1}{2} \{ (\mathbf{r}_{n|k} - \mathbf{t}_n)^T \mathbf{C}_n^{-1} (\mathbf{r}_{n|k} - \mathbf{t}_n) + \mathrm{Tr}(\mathbf{E}_{\mathrm{n|k}} \mathbf{C}_{\mathrm{n}}^{-1}) \} \\ &+ \frac{D}{2} + \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathbf{E}_{n|k}| \end{aligned} \qquad (16)$$

and $\psi(.)$ is the digamma function and $\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |\mathbf{W}_k|$.

The optimal solution for $q^*(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k)$ is similar to the derivation in [3] and is computed as:

$$q^*(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \qquad (17)$$

where:

$$\begin{aligned} \beta_k &= \beta_0 + N_k \qquad (18) \\ \mathbf{m}_k &= \frac{1}{\beta_k} (\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{r}}_k) \qquad (19) \\ \bar{\mathbf{r}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \mathbf{r}_{n|k} \qquad (20) \\ N_k &= \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \qquad (21) \end{aligned}$$

and the optimal solution for $q^*(\boldsymbol{\Lambda}_k)$ is computed as:

$$q^*(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_k) \qquad (22)$$

where:

$$\begin{aligned} \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_k \\ &+ \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{r}}_k - \mathbf{m}_0)(\bar{\mathbf{r}}_k - \mathbf{m}_0)^T \qquad (23) \\ \nu_k &= \nu_0 + N_k \qquad (24) \\ \mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] (\mathbf{r}_{n|k} - \bar{\mathbf{r}}_k)(\mathbf{r}_{n|k} - \bar{\mathbf{r}}_k)^T \\ &+ \frac{1}{N_k} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \mathbf{E}_{n|k} \qquad (25) \end{aligned}$$

Finally, since no priors are placed on the mixing coefficients $\boldsymbol{\pi}$, it is not considered as latent variable and is instead optimised as parameters:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \qquad (26)$$

Following the approach suggested by Corduneanu and Bishop [6] for choosing the correct number of mixture components, we initialise the mixture with a large number of components and remove any whose mixing coefficients have been driven to zero during the optimisation.

### 3.2.1 Variational Lower Bound

The optimal mixture model is found by repeatedly updating the variational distributions of $\mathbf{T}$, $\mathbf{Z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ and then optimising the mixture coefficients $\boldsymbol{\pi}$. The convergence of the iterative update is monitored by computing the variational lower bound $\mathcal{L}(q)$, which cannot decrease after each update. For our mixture model, the lower bound defined by equation (6) is computed as:

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}[\ln p(\mathbf{X},\mathbf{C},\mathbf{T},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\pi})] - \mathbb{E}[\ln q(\mathbf{T},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] \\ &= \mathbb{E}[\ln p(\mathbf{X}|\mathbf{C},\mathbf{T})] + \mathbb{E}[\ln p(\mathbf{T}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] \\ &\quad + \mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\mathbf{T}|\mathbf{Z})] \\ &\quad - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})]\end{aligned}$$

where:

$$\begin{aligned}&\mathbb{E}[\ln p(\mathbf{X}|\mathbf{C},\mathbf{T})] = \\ &-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{(\mathbf{r}_{n|k}-\mathbf{x}_n)^T\mathbf{C}_n^{-1}(\mathbf{r}_{n|k}-\mathbf{x}_n)\right\} \\ &-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{\mathrm{Tr}(\mathbf{E}_{\mathrm{n|k}}\mathbf{C}_{\mathrm{n}}^{-1}) + \ln|\mathbf{C}_{\mathrm{n}}| + \mathrm{D}\ln(2\pi)\right\}\end{aligned} \tag{27}$$

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{T}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}N_k\{&\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - D\beta_k^{-1} \\ &-\nu_k\mathrm{Tr}(\mathbf{S}_{\mathrm{k}}\mathbf{W}_{\mathrm{k}}) - \nu_{\mathrm{k}}\mathrm{Tr}(\bar{\mathbf{E}}_{\mathrm{k}}\mathbf{W}_{\mathrm{k}}) \\ -\nu_k(\bar{\mathbf{r}}_k-\mathbf{m}_k)^T&\mathbf{W}_k(\bar{\mathbf{r}}_k-\mathbf{m}_k) - D\ln(2\pi)\}\end{aligned} \tag{28}$$

$$\mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\pi_k \tag{29}$$

$$\begin{aligned}\mathbb{E}[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{k=1}^{K}\{D\ln(\frac{\beta_0}{2\pi}) + \mathbb{E}\ln|\boldsymbol{\Lambda}_k| - \frac{D\beta_0}{\beta_k} \\ &-\beta_0\nu_k(\mathbf{m}_k-\mathbf{m}_0)^T\mathbf{W}_k(\mathbf{m}_k-\mathbf{m}_0)\} + K\{-\frac{\nu_0}{2}\ln|\mathbf{W}_0| \\ &-\frac{\nu_0 D}{2}\ln 2 - \frac{D(D-1)}{4}\ln\pi - \sum_{i=1}^{D}\ln\Gamma(\frac{\nu_0-1+i}{2})\} \\ &+\frac{\nu_0-D-1}{2}\sum_{k=1}^{K}\mathbb{E}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\sum_{k=1}^{K}\nu_k\mathrm{Tr}(\mathbf{W}_0^{-1}\mathbf{W}_{\mathrm{k}})\end{aligned} \tag{30}$$

$$\begin{aligned}\mathbb{E}[\ln q(\mathbf{T}|\mathbf{Z})] = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\{&D + D\ln(2\pi) \\ &+\ln|\mathbf{E}_{n|k}|\}\end{aligned} \tag{31}$$

$$\mathbb{E}[\ln q(\mathbf{Z})] = \sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln r_{nk} \tag{32}$$

$$\begin{aligned}\mathbb{E}[\ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{k=1}^{K}\{\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] + D\ln(\frac{\beta_k}{2\pi}) - D \\ &-\frac{\nu_k}{2}\ln|\mathbf{W}_k| - \frac{\nu_k D}{2}\ln 2 - \frac{D(D-1)}{4}\ln\pi \\ -\sum_{i=1}^{D}\ln\Gamma(\frac{\nu_k-1+i}{2}) &+ \frac{\nu_k-D-1}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \frac{\nu_k D}{2}\}\end{aligned} \tag{33}$$

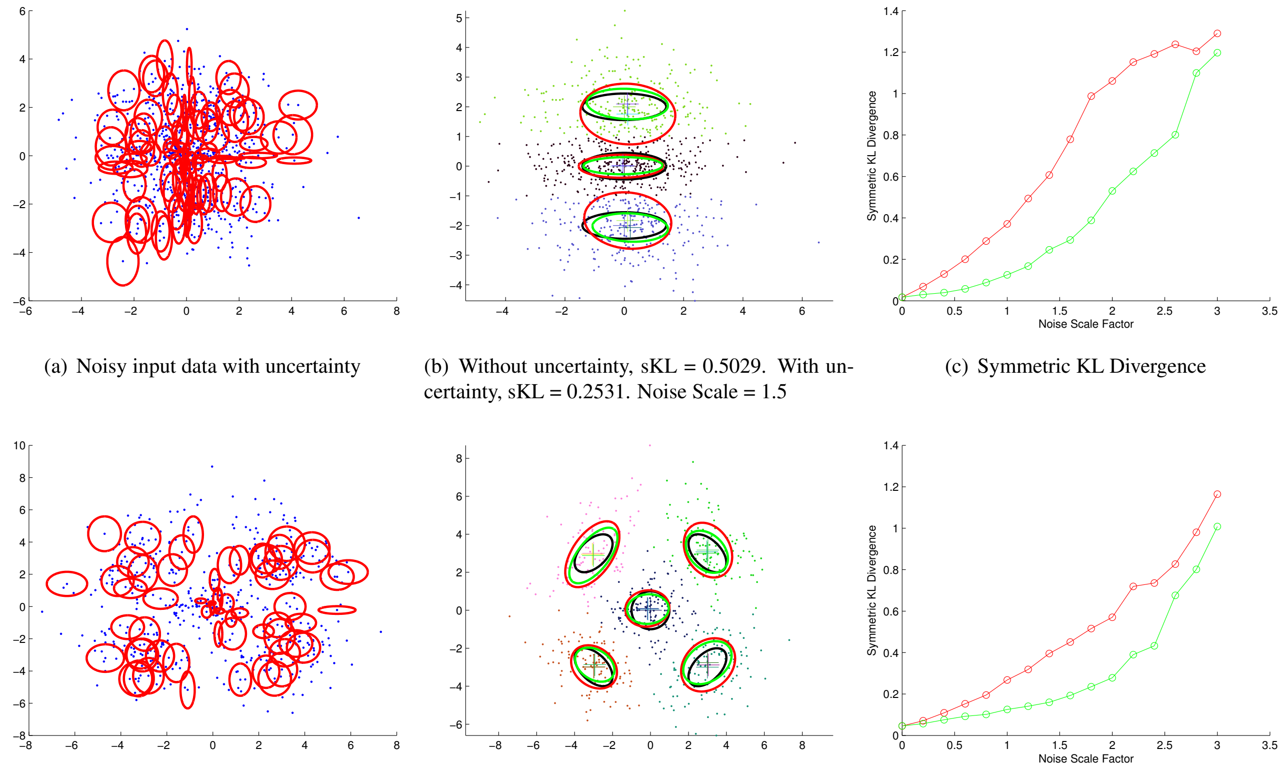
## 4. Results

We have tested our clustering algorithm on both synthetic and real data sets to demonstrate the benefit gained by taking into account of the uncertainty information.

### 4.1. Synthetic Examples

Two synthetic datasets proposed by Corduneanu and Bishop [6] were used to test the performance of our clustering algorithm. The first data set contains 900 data points sampled from a mixture of three Gaussian components with means $[0,-2],[0,0]$ and $[0,2]$ and the same covariance $[2,0;0,0.2]$. The second data set contains 600 data points sampled from a mixture of five Gaussian components with means $[0,0],[3,-3],[3,3],[-3,3]$ and $[-3,-3]$ and covariances $[1,0;0,1],[1,0.5;0.5,1],[1,-0.5;-0.5,1],[1,0.5;0.5,1]$ and $[1,-0.5;-0.5,1]$.

For each data point $\mathbf{t}_n$, we generate the covariance matrix of uncertainty about the data point as $\mathbf{C}_n = [|x_n|/10,0;0,|y_n|/5]*\lambda$ and sample a noisy data point $\mathbf{x}_n$ from $\mathcal{N}(\mathbf{t}_n,\mathbf{C}_n)$. $\lambda$ is an experiment parameter for controlling the overall scale of noise. The noisy data points are then clustered using the variational bayesian clustering algorithm from [6] and the results compared with our clustering algorithm which also takes into account the uncertainty information. The noisy input data is shown in figure 2(a) and 2(d) and the corresponding clustering results are shown in figure 2(b) and 2(e). The clustering algorithms were initialised with a mixture of 20 components and while both were able to find the correct number of mixture components, the mixtures estimated by our algorithm were closer to the ground truth, as can be seen from visual inspection of the clusters and from the symmetric KL divergence scores computed with the ground truth mixtures. Figure 2(c) and



(a) Noisy input data with uncertainty

(b) Without uncertainty, sKL = 0.5029. With uncertainty, sKL = 0.2531. Noise Scale = 1.5

(c) Symmetric KL Divergence

(d) Noisy input data with uncertainty

(e) Without uncertainty, sKL = 0.4413. With uncertainty, sKL = 0.1665. Noise Scale = 1.5

(f) Symmetric KL Divergence

Figure 2. In the left column, red ellipses represent the uncertainty on the point it is centred on; for clarity, only the uncertainty in one of every ten data points is shown. In the middle column, the black ellipses represent the ground truth, the red ellipses represent clustering without uncertainty and the green ellipses represent clustering with uncertainty. In the right column, the red plot represents clustering without uncertainty and the green plot represents clustering with uncertainty.

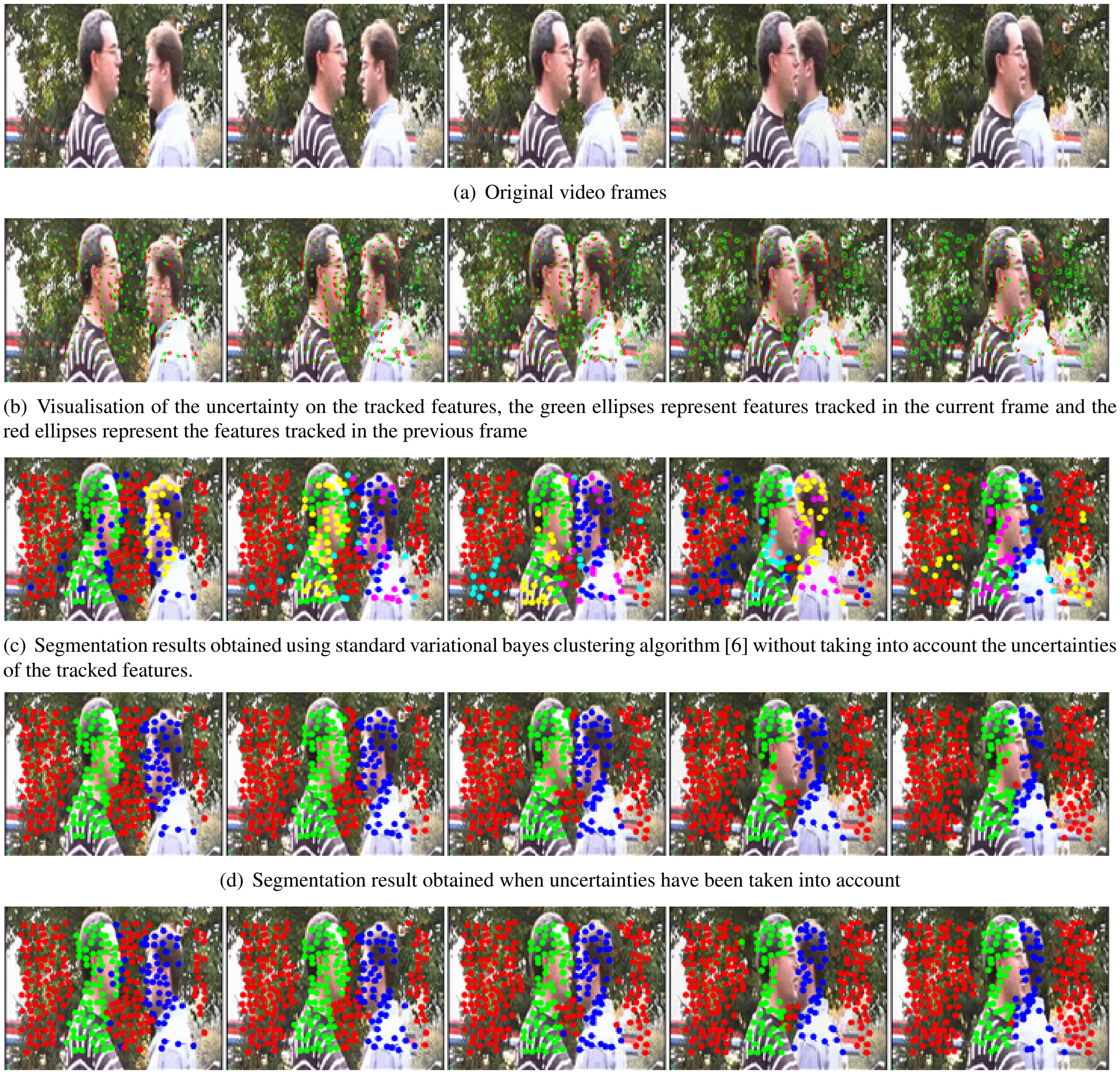
2(f) show our clustering algorithm gives consistently lower symmetric KL score as $\lambda$ varies from 0 to 3. Even in the simplifying case of negligible uncertainty and the extreme case of very large uncertainty, the resulting mixture is at least as good as the ones obtained without taking uncertainty into account.

## 4.2. Motion Segmentation

We applied our clustering algorithm to the problem of multi-body motion segmentation, which is an important preprocessing step for many computer vision applications. We tracked a number of feature points through a video sequence using the Kanade-Lucas-Tomasi (KLT) tracker [15]. The uncertainty of each tracked feature point was evaluated using the method described by Nicklels and Hutchinson [13] for estimating uncertainty in SSD based feature tracking, for which KLT tracking is an example.

We perform motion segmentation for the $t$th frame by clustering the velocities of tracked features at frame $t$ and $t-1$, therefore $\mathbf{x}_n = [[\mathbf{f}_n^t - \mathbf{f}_n^{t-1}], [\mathbf{f}_n^{t-1} - \mathbf{f}_n^{t-2}]]$, where $\mathbf{f}_n^t$ denotes the position of the $n$th tracked feature at the $t$th frame. The uncertainty on $\mathbf{f}_n^t$ is denoted by the matrix $\mathbf{C}_n^t$ and it follows that the uncertainty on $\mathbf{x}_n$ is $\mathbf{C}_n = [\mathbf{C}_n^t, \mathbf{0}; \mathbf{0}, \mathbf{C}_n^{t-1}]$.

Figure 3 shows the motion segmentation result on a well known sequence of two persons walking past each other. For this sequence, the features tracked on the cloth of the person on the left is often not well localised along the Y axis due to its stripey nature and similarly some features tracked on the cloth of the person on the right is not well localised due to its lack of texture. The figure shows that our algorithm was able to appropriately take into account the uncertainty on feature positions and find a mixture with three components to represent the motion vectors, corresponding to the two moving persons and the background. Occasionally, a fourth component will be found which usually corresponds to some outliers such as features that have latched onto the wrong image patches because the original features have disappeared due to occlusion. In comparison, if the uncertainty information were not taken into account,


(a) Original video frames

(b) Visualisation of the uncertainty on the tracked features, the green ellipses represent features tracked in the current frame and the red ellipses represent the features tracked in the previous frame

(c) Segmentation results obtained using standard variational bayes clustering algorithm [6] without taking into account the uncertainties of the tracked features.

(d) Segmentation result obtained when uncertainties have been taken into account

(e) Segmentation results obtained using Kanatani and Sugaya's method in [10]

Figure 3. Comparison of motions segmentation results on a well known sequence of two persons walking past each other. Frame 5, 7, 9, 11 and 13 are shown here. The coloured circles only serves to distinguish between different clusters in the same frame and although the visualisation try to make the assignment of colours to clusters consistent from frame to frame, this is not guaranteed. The figure is best viewed in colour and please refer to the supplemental material for the full size version and videos

then a lot of features that were not well localised will be incorrectly segmented.

We also compared our method to the multi-stage optimisation based segmentation method by Kanatani and Sugaya [10] which is known to give very accurate results. Although their method is designed to operate on the trajectory of tracked features over the entire sequence, this is not possible for the sequence in figure 3 as one person is occluded by the other person for part of the sequence. So in our experiment, we test their method at each frame on features that have been successfully tracked for the last four frames, therefore the input data consists of feature positions from frame $t-4$ to frame $t$. We also provided their method with the information that there are three objects in the scene. It can be seen that although our approach only uses a simple transformation model and velocity information from the last two frames, there are very few feature points that are incorrectly segmented and even they can be eliminated if a longer temporal window is used.

We also tested our algorithm on the three datasets used in [10], though in this case, because only features that are consistently tracked are provided, i.e. those with low tracking uncertainty, the improvement to motion segmentation is not as significant. For these experiments we try to use feature velocity information from as many frames as possible by concatenating velocity vectors from multiple frames into a single feature vector and cluster that instead.

One particular problem with our approach to motion segmentation is that sometimes the moving objects are much smaller than the background and consequently the background features significantly outnumbers the foreground features, causing too few components to be selected or too many features assigned to the background component which has a very large prior probability. This can be mitigated by placing a strong prior on the mixing coefficients $\boldsymbol{\pi}$ and encouraging them to remain close to $\frac{1}{K}$.

### 4.3. Clustering Microarray Gene Expression

Clustering algorithms are important tools in Bioinformatics for analysing microarray gene expression data and grouping genes with similar expression patterns [12]. Since microarray experiments are complex multiple stages procedures, the acquired data can exhibit high levels of measurement errors which are introduced in the various stages of the experiments. Therefore it is important to explicitly take into account of these measurement errors during clustering to make the algorithm more robust to noise. Liu et al. [12] presented an clustering algorithm which estimates a mixture of Gaussians with spherical covariance matrix and takes into account of zero-mean Gaussian measurement errors represented by diagonal covariance matrices.

We applied our clustering algorithm to the six group dataset with 10 conditions and seven group dataset with 10 conditions from [12]. Table 1 and 2 shows the comparison of average adjusted rand index [9] of data partitioning obtained by our algorithm and the algorithm described in [12], the top row shows the variance of the zero-mean Gaussian noise added to the datasets:

In all experiments, we assume that the number of classes is unknown and need to estimated along with the mixture. Tables 1 and 2 shows that when our algorithm is restricted to estimating spherical Gaussian components, it achieves comparable results to those obtained by Liu et al [12]. If diagonal Gaussian components were estimated instead, then significant improvements can be obtained. Our algorithm also has the advantage that the appropriate number of mixture components is found during the clustering process which is more efficient than estimating mixtures with different number of components and then choose the best one using BIC.

## 5. Conclusion

In this paper, we presented a variational bayes solution to the problem of estimating a mixture of Gaussians from noisy input data with known uncertainty distributions. The algorithm was tested on both synthetic datasets and real datasets. We have shown that by incorporating uncertainty information into the clustering algorithm, we can improve the results of pattern recognition tasks such as multi-body motion segmentation of feature point trajectories and partitioning of microarray gene expressions. In all experiments, we also assumed that the correct number of mixture components is unknown and model selection is done within the variational bayesian framework [6].

## Acknowledgements

Thanks to Xuejun Liu and Magnus Rattray for providing the Microarray Gene Expression data.

## References

[1] In S. V. Huffel and P. Lemmerling, editors, *Total Least Squares and Errors-in-Variables Modeling*. Kluwer Academic Publishers, Dordrecht, 2002.

[2] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *Advances in Neural Information Processing Systems 17*, pages 161–168, 2005.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] B. B. Chaudhuri and P. R. Bhowmik. An approach of clustering data with noisy or imprecise feature measurement. *Pattern Recogn. Lett.*, 19(14):1307–1317, 1998.

[5] W. Chojnacki, M. J. Brooks, A. van den Hengel, and D. Gawley. On the fitting of surfaces to data with covariances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1294–1303, 2000.

[6] A. Corduneanu and C. M. Bishop. Variational bayesian model selection for mixture distribution. In *Proc. Artifical Intelligence and Statistics*, 2001.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[8] H. Handman and G. Govaert. Mixture model clustering of uncertain data. In *Proc. Fuzzy Systems*, pages 879–884, 2005.

[9] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

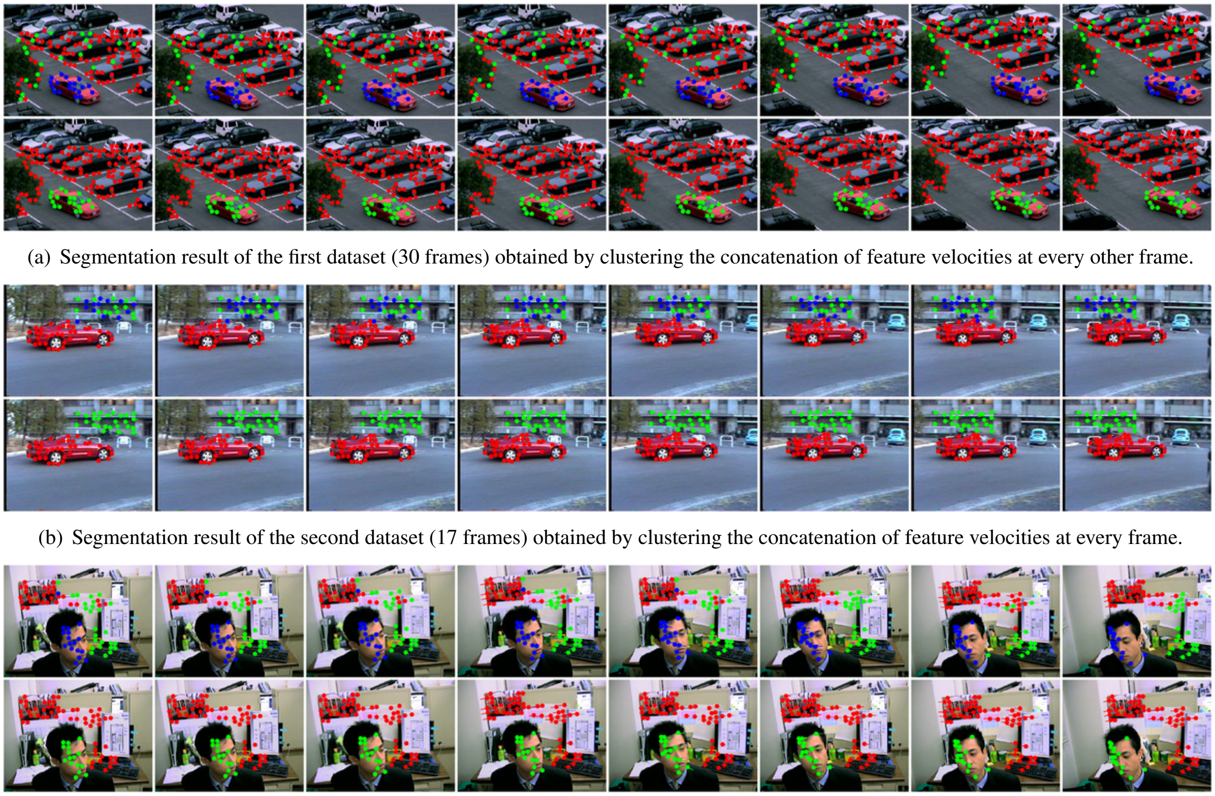
[10] K. Kanatani and Y. Sugaya. Multi-stage optimization for multi-body motion segmentation. In *Australia-Japan Advanced Workshop on Computer Vision*, 2003.

[11] M. Kumar and N. R. Patel. Clustering data with measurement errors. Technical report, 1998.

[12] X. Liu, K. K. Lin, B. Andersen, and M. Rattray. Including probe-level uncertainty in model-based gene expression clustering. *BMC Bioinformatics*, 8:98, 2007.

[13] K. Nickels and S. Hutchinson. Estimating uncertainty in ssd-based feature tracking. *Image Vision Comput.*, 20(1):47–58, 2002.

[14] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.



(a) Segmentation result of the first dataset (30 frames) obtained by clustering the concatenation of feature velocities at every other frame.

(b) Segmentation result of the second dataset (17 frames) obtained by clustering the concatenation of feature velocities at every frame.

(c) Segmentation result of the third dataset (100 frames) obtained by clustering the concatenation of feature velocities at every 5th frame for each 50 frame segment. The velocity of a feature at frame $t$ is also computed with respect to its position at $t - 5th$ frame

Figure 4. Motion segmentation results on the three datasets used by Kanatani and Sugaya [10]. For each dataset, the top row shows the result when tracking uncertainty is not taken into account and bottom row shows the result when it is.

Table 1. Average Adjusted Rand Index for 6 group dataset

| | 0.01 | 0.1 | 0.2 |
|---|---|---|---|
| WITHOUT UNCERTAINTY | $0.6048 \pm 0.0462$ | $0.6173 \pm 0.0398$ | $0.6167 \pm 0.0204$ |
| PUMACLUST | $0.7474 \pm 0.0308$ | $0.7123 \pm 0.0482$ | $0.6882 \pm 0.0301$ |
| OUR ALGORITHM (SPHERICAL) | $0.7527 \pm 0.0363$ | $0.7065 \pm 0.0406$ | $0.6905 \pm 0.0291$ |
| OUR ALGORITHM (DIAGONAL) | $0.8036 \pm 0.0428$ | $0.7543 \pm 0.0558$ | $0.7488 \pm 0.0307$ |

Table 2. Average Adjusted Rand Index for 7 group dataset

| | 0.0 | 0.01 | 0.1 |
|---|---|---|---|
| WITHOUT UNCERTAINTY | $0.5440 \pm 0.0428$ | $0.5463 \pm 0.0441$ | $0.5552 \pm 0.0473$ |
| PUMACLUST | $0.6980 \pm 0.0464$ | $0.6866 \pm 0.0428$ | $0.6532 \pm 0.0570$ |
| OUR ALGORITHM (SPHERICAL) | $0.6990 \pm 0.0415$ | $0.6982 \pm 0.0475$ | $0.6467 \pm 0.0442$ |
| OUR ALGORITHM (DIAGONAL) | $0.7380 \pm 0.0451$ | $0.7357 \pm 0.0456$ | $0.6824 \pm 0.0557$ |

[15] J. Shi and C. Tomasi. Good features to track. In *Proc. CVPR*, pages 593 – 600, 1994.

[16] J. Sun, A. Kabán, and S. Raychaudhury. Robust mixtures in the presence of measurement errors. In *Proc. ICML*, pages 847–854, 2007.