Robust Mixtures in the Presence of Measurement Errors

Jianyong Sun\textsuperscript{1,2}, Ata Kabán\textsuperscript{1} & Somak Raychaudhury\textsuperscript{2}

\textsuperscript{1}School of Computer Science
\textsuperscript{2}School of Physics and Astronomy
The University of Birmingham

This work is funded by PPARC, UK
Overview

- Introduction
- An outlier-robust mixture model for data with known measurement errors
- Solving by a structured-variational EM
- Which are the outliers?
- Experiments & results
- Conclusions
Introduction

• Robust density modeling aims at capturing the structure of typical observations while dealing with outliers
  – required to avoid biases of parameter estimates
  – peculiar objects are of interest in domains such as Astrophysics, for identifying candidates of possibly new types of objects (e.g. from archives of multi-wavelength astronomical images) that deserve follow-up study (e.g. using spectroscopy).

• Bottleneck: ‘the likely overabundance of interesting objects found’ (Djorgowski) – the interpretation and understanding of which will necessitate costly detailed analysis.
• “Here is the list of top outlying objects.”
• <Astronomer>: Ah, of course, many of these have large measurement errors. Tell us something new!
• “?”
• <Astronomer>: The errors depend on instrumental limitations and measurement conditions and are unavoidable with astrophysical measurements.
• “Um... OK, then give us your measurement errors.”
Introduction

• Goal:
• Given measurements from N objects, over d features, together with their associated measurement errors

\[ t_{in} \pm \sqrt{s_{in}} \quad (i = 1, \ldots, d; \ n = 1, \ldots, N) \]

• find the peculiar objects whose peculiarity is not due to measurement errors (‘genuine’, potentially interesting outliers)
• along with a model of the density of non-outliers.
Robust density modeling


It assumes that data are points.

Instead, the real scientific data also contains error estimates.
Can knowledge of the errors help us infer the ‘genuine’ outliers?
Can knowledge of the errors help us infer the ‘genuine’ outliers?
Can knowledge of the errors help us infer the ‘genuine’ outliers?
A robust mixture for data with errors

- A heteroscedastic Gaussian error model (with unknown mean and known variance) will account for the measurement errors

\[ p(t_n | w_n) = \mathcal{N}(t_n | w_n, S_n) \]

- A mixture of Student-t densities will model the density of ‘w’.

\[ p(w_n) = \sum_{k=1}^{K} \pi_k S_t(w_n | \mu_k, \Sigma_k, \nu_k) \]

- Putting these together, the data likelihood is:

\[
p(t) = \sum_k \pi_k \int \int \mathcal{N}(t|w,S) \mathcal{N}(w|\mu_k, \frac{\Sigma_k}{u}) \mathcal{G} \left( u | \frac{\nu_k}{2}, \frac{\nu_k}{2} \right) dudw \]

\[ N(t | \mu_k, S + \frac{\Sigma_k}{u}) \]
• The joint likelihood of all variables:

\[
\prod_{n} \prod_{k} \left[ p(t_n | w_n) p(w_n | u_n, z_n = k) \right] \delta(z_n = k) \times \\
\prod_{n} \prod_{k} \left[ p(u_n | z_n = k) p(z_n = k | \pi) \right] \delta(z_n = k)
\]
Solving by variational EM

• Lower bounding the log likelihood – for any \( q() \), we have:

\[
\log p(t_n|\theta) = \int q(h_n) \log \frac{p(h_n, t_n|\theta)}{q(h_n)} \frac{q(h_n)}{p(h_n|t_n, \theta)} dh_n
\]

\[\geq \int q(h_n) \log \frac{p(h_n, t_n|\theta)}{q(h_n)} dh \equiv \mathcal{F}(t_n|q, \theta)\]

where

\[h_n = (w_n, u_n, z_n)\] the hidden variables in the model

\[\theta = (\{\mu_k\}, \{\Sigma_k\}, \{v_k\}, \pi)\] the model parameters

• E-M with variational E-step at (k+1)-th iteration:
  
  - Var E step:
    \[q^{k+1}(h_n) = \arg \max_q \mathcal{F}(t_n|q, \theta^k)\]

  - M-step
    \[\theta^{k+1} = \arg \max_{\theta} \sum_n \mathcal{F}(t_n|q^{k+1}, \theta)\]
Choice of the variational distribution

- Option 1: Fully factorial form
  \[
  q(w, u, z) \equiv q(w)q(u)q(z)
  \]

- Option 2: Tree-structured factorial form
  \[
  q(w, u, z) \equiv \prod_{k=1}^{K} [q(w | z = k)q(u | z = k)q(z = k)]^{\delta(z = k)}
  \]

In both cases, the optimal functional form of the terms of \(q()\) is tractable to compute, by taking the functional derivatives and equating them to the identically null function.
The Var-E-step expressions

\[ q(w|k) = \frac{\exp\langle \log [p(t|w)p(u|k)] \rangle_{u|k}}{\int \exp\langle \log [p(t|w)p(u|k)] \rangle_{u|k} dw} = \mathcal{N}(w|\langle w \rangle_k, \Sigma_{w|k}) \]

\[ q(u|k) = \frac{\exp\langle \log [p(w|u,k)p(u|k)] \rangle_{w|k}}{\int \exp\langle \log [p(w|u,k)p(u|k)] \rangle_{w|k} du} = \mathcal{G}(u|a_k, b_k) \]

\[ q(k) = \frac{\exp(A_{t,k})}{\sum_{k'} \exp(A_{t,k'})} \]

where

\[ \Sigma_{w|k} = S \left[ \frac{\Sigma_k}{\langle u \rangle_{u|k}} + S \right]^{-1} \frac{\Sigma_k}{\langle u \rangle_{u|k}} \]

\[ \langle w \rangle_k = \Sigma_{w|k} \left( \langle u \rangle_{u|k} \Sigma_k^{-1} \mu_k + S^{-1} t \right) \]

\[ a_k = \frac{\nu_k + d}{2}; \quad b_k = \frac{\nu_k + C_k}{2} \]

\[ C_k = (\langle w \rangle_k - \mu_k)^T \Sigma_k^{-1} ((\langle w \rangle_k - \mu_k) + \text{Tr} \left( \Sigma_k^{-1} \Sigma_{w|k} \right) \]

and the expression \( A_{t,k} \) depends on all other variables and will be available from the likelihood bound computations.
The expression denoted by $A_{t,k}$ earlier is the following:

$$A_{t,k} = \langle \log p(t|\textbf{w}) \rangle_{\textbf{w}|k} + \langle \log p(u|k) \rangle_{u|k} + \log \pi_k + \langle \log p(\textbf{w}|u, k) \rangle_{\textbf{w},u|k} - \langle \log q(\textbf{w}|k) \rangle_{\textbf{w}|k} - \langle \log q(u|k) \rangle_{u|k}$$

$$= Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6$$

$$Q_1 = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\textbf{S}| - \frac{1}{2} \text{Tr}(\Sigma_{\textbf{w}|k}^{-1}) - \frac{1}{2} \left[ (\langle \textbf{w} \rangle_{k} - \textbf{t})^T \Sigma_{\textbf{w}|k}^{-1} (\langle \textbf{w} \rangle_{k} - \textbf{t}) \right]$$

$$Q_2 = \left( \frac{\nu_k}{2} - 1 \right) \langle \log u \rangle_{u|k} - \frac{\nu_k}{2} \frac{a_k}{b_k} + \frac{\nu_k}{2} \log \left( \frac{\nu_k}{2} \right) - \log \Gamma \left( \frac{\nu_k}{2} \right)$$

$$Q_3 = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| + \frac{d}{2} \frac{a_k}{b_k}$$

$$- \frac{1}{2} \frac{a_k}{b_k} \left[ (\langle \textbf{w} \rangle_{k} - \mu_k)^T \Sigma_k^{-1} (\langle \textbf{w} \rangle_{k} - \mu_k) \right] - \frac{1}{2} \frac{a_k}{b_k} \text{Tr}(\Sigma_{\textbf{w}|k} \Sigma_k^{-1})$$

$$Q_4 = \log \pi_k;$$

$$Q_5 = \frac{d}{2} + \frac{d}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{\textbf{w}|k}|$$

$$Q_6 = -[(a_k - 1) \langle \log u \rangle_{u|k} + a_k \log b_k - a_k - \log \Gamma(a_k)]$$

and where $\langle \log u \rangle_{u|k} = \psi(a_k) - \log b_k$ and $\psi(\cdot)$ is the di-gamma function
The variational bound

• With the previous expression, then the likelihood bound for a set of N data points is computed as the following:

\[ F = \sum_n \sum_k q(z_n = k) [A_{tn,k} - \log q(z_n = k)] \]

• At each iteration of the variational E-step and the M-step updates, the variational log likelihood bound is guaranteed not to decrease, and alternating these two steps will converge to a local optimum of the bound.

• Monitoring the likelihood bound is useful for assessing convergence of the iterative algorithm.
The M-step expressions

- Solving the stationary equations w.r.t. the model parameters yields:

\[
\mu_k = \frac{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n|k} \langle w_n \rangle_{w_n|k}}{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n|k}}
\]

\[
\Sigma_k = \frac{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n|k} \Sigma_{n,k}}{\sum_{n=1}^{N} q(z_n = k)}
\]

\[
\pi_k = \frac{1}{N} \sum_{n=1}^{N} q(z_n = k)
\]

where \( \tilde{\Sigma}_{n,k} = [(\mu_k - \langle w_n \rangle_k)(\mu_k - \langle w_n \rangle_k)^T + \Sigma_{w_n|k}] \).

\[
\sum_{n} q(z_n = k) \left[ \log\left( \frac{\nu_k}{2} \right) + 1 + \log u_n \right]_{k} - \frac{a_{nk}}{b_{nk}} - \psi\left( \frac{\nu_k}{2} \right) = 0
\]
• q(u|k) and q(w|k) will depend on their Markov blanket, but q(z) will depend on all variables. However the required expression is available from the likelihood bound computation.

• In the limit of zero measurement errors, the structured variational EM reduces to the exact EM for t-mixtures --- which is naturally expected since the marginal likelihood reduces to the t-mixture likelihood.
Computational scaling

• The time complexity per iteration is $O(d^3KN)$.
• The same for the fully factorial posterior choice. Though the CPU time to convergence (evaluated empirically) was observed to take about 1.5 times longer in the case of the structured posterior.
• The most expensive operation is the $dxd$ matrix inversion. However, $d$ is expected to be relatively small – otherwise $\Sigma_k$ would need to be assumed diagonal, in which case the cubic operation disappears.
Determining the number of components

- We used a lower bound to the Minimum Message Length in conjunction with our likelihood bound. Other methods are possible.
- The MML criterion is to maximise:

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

where the last term is the data log likelihood – which we replace by our variational lower bound \( \sum_n \mathcal{F}(v_n | q, \theta) \)

and where

- \( \hat{n} \) = dimensionality of the free parameters (in our case, \( \{\mu_k, \Sigma_k\} \) have \( d + d(d - 1)/2 \) free params)
- \( k_{nz} \) = \# nonzero mixture components

This maximisation task is similar to that presented before, except the update of the mixture priors, which is now:

\[
\pi_k = \frac{\max \left\{ 0, \sum_{n=1}^{N} q(z_n = k) - \frac{\hat{n}}{2} \right\}}{\sum_{j=1}^{K} \max \left\{ 0, \sum_{n=1}^{N} q(z_n = j) - \frac{\hat{n}}{2} \right\}}
\]
Which are the outliers?

- The quantity of interest is the posterior expectation of $u_n$. We use the variational expectation, computed wrt $\sum q(u|k)q(k)$. (The lower this expression is, the most outlying the associated observation.)

$$e \equiv \sum_k q(k) \frac{\nu_k + d}{\nu_k + \text{Tr}(\Sigma_k^{-1} \Sigma_w | k) + \Delta^2_{w|k}}$$

where

$$\Delta^2_{w|k} = (\langle w \rangle_{w|k} - \mu_k)^T \Sigma_k^{-1} (\langle w \rangle_{w|k} - \mu_k)$$

- For t-mixtures (i.e. no measurement error info), in turn, the posterior mean expression for $u_n$ is:

$$e_{MoT} \equiv \sum_k p(k|t) \frac{\nu_k + d}{\nu_k + (t - \mu_k)^T \Sigma_k^{-1} (t - \mu_k)}$$

- So, knowledge about the measurement errors re-defines the outliers.
Experiments and Results

1) How accurate is the structured variational approximation employed?
   - How it compares to a fully factorial approximation?
   - How it compares to the ‘ground truth’ Markov-Chain-EM?

2) To what extent does knowledge of measurement errors help us to infer the outliers that are not due to these errors?
   - How it compares to ignoring the knowledge about the errors?
   - How it compares to the idealised case of having the clean data with no measurement errors?

3) Application in Astronomy: Finding peculiar objects of interest from the SDSS quasar catalogue (high-redshift quasars).
1) Comparison of alternative approximate EM methods

Log likelihood (bound) over consecutive iterations, is a run on a synthetic data set. For MCMC using Gibbs sampling, 10,000 samples were used for computing the posterior estimates and the first 2,000 samples were discarded as burn-in.

All algorithms were started from the same parameter values.
• Comparison of the fully factorial and the structured variational EM, in terms of correct outlier detection rate (AUC), computed from 20 repeats

<table>
<thead>
<tr>
<th>Data sample size</th>
<th>Fully factorial</th>
<th>Structured</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.909 ± 0.054</td>
<td>0.923 ± 0.030</td>
<td>0.0071</td>
</tr>
<tr>
<td>600</td>
<td>0.920 ± 0.15</td>
<td>0.963 ± 0.022</td>
<td>0.0423</td>
</tr>
</tbody>
</table>

- The structured variational EM takes cca. 1.5 times longer than the fully factorial variational EM
- However the AUC values were found to be significantly superior (at 0.05% level)
2) Accuracy of detecting ‘genuine’ outliers

- Synthetic data sets sampled from the model, starting from 3 well separated Gaussians and genuine outliers from a uniform distribution.
- Five different error levels defined: Diagonals of $S$ will range between $[0,0.01]$, $[0,0.1]$, $[0,1]$, $[0,10]$ and $[0,100]$ respectively.
- Computed the Area Under the ROC curve (AUC) achieved by the inferred outlierness against the true outliers, averaged over 10 independent repeats.
Accuracy of detecting "genuine" outliers: Results on the synthetic data sets
Further tests on semi-synthetic data, derived from real data with more realistic density structure. It still allows us to evaluate the benefits of having measurement error information, in a controlled manner.

- The ‘lymphography’ (148 data points, 18 dimensions). Initially 4 classes, of which two are too small (6 points in total) and will be considered outliers.
- We simulate measurement errors by adding hetero-schedastic Gaussian noise with variances ranging in [0,0.1]
Accuracy of detecting “genuine” outliers: Results on semi-synthetic data sets

ROC curves, measured on out-of-sample data, averaged over 10 runs
3) Application to Astrophysical Data Analysis

- We analysed a data set of 10,000 quasars extracted from a well studied survey in astrophysics: the SDSS quasar catalogue.
- Features: 5 magnitudes measured with five different optical filters (u,g,i,r,z). From these, to avoid bias with brightness, we construct 4 features, each related to a color, by subtracting ‘r’ from the others.
- Spectroscopic redshift estimates are available, which we use for validating the interpretability of our results. There is physical reason for high redshift objects to be perceived as outliers in the density of quasars in the color space.
Application: Detecting high-redshift quasars from the SDSS quasar catalog

AUC vs. different redshift thresholds. A large fraction of quasars at redshift 2.5 or higher are detected with high probability.

This is not possible with 2D projections as currently used by astronomers.
Conclusions

• Knowledge of errors is useful for making the hunt for the ‘interesting’ outliers more directed
• Keeping some of the dependencies in the variational posterior density can increase accuracy
• The method can detect high redshift quasars from the SDSS photometric quasar catalogue
• We are extending these findings to combined statistical and visual analysis of data with outliers and error information.
• Comparison with a Maximum A Posteriori (MAP) estimator (for $u_n$) in terms of outlier finding and cluster finding, computed from 10 independent realizations of the synthetic data

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>Var-EM</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.915 ± 0.13</td>
<td>0.964 ± 0.015</td>
<td>0.0890</td>
</tr>
<tr>
<td>Clust accuracy(%)</td>
<td>0.920 ± 0.15</td>
<td>0.931 ± 0.014</td>
<td>0.0079</td>
</tr>
</tbody>
</table>

- MAP is more variable and less accurate than var-EM
Extension to constrained mixtures

- The non-outliers are assumed to lie on a 2D manifold. The outliers have no structure assumed.
- Constrained mixtures can model the underlying manifold in the same spirit as in Generative Topographic Mapping (Bishop & Svensen & Williams)

\[ p(t; W, \sigma, \nu) = \frac{1}{K} \sum_k \int \int \mathcal{N}(t|w, S) \mathcal{N}(w|W\phi(x_k), \frac{1}{u\sigma}) \mathcal{G}(u|\nu_k, \nu_k) du dw \]

- The measurement errors distort the overall configuration, but their variances are known.
- Similar variational treatment applies
Some early results

• Synthetic data in 10-D: 3 clusters & outliers
• It is beneficial to represent the outlierness in a 3rd dimension.