Paper Reading: Outlier Detection via Parsimonious Mixtures of Contaminated Gaussian Distributions

Nguyen D. Pham

Department of Computer Science
University of Houston

September 15, 2015
» Parametric Density Estimation: Maximum Likelihood

» Application:
  * Alternative to non-parametric method.
  * Clustering and classification.
Discussion

Introduction
Methodology
  EM Algorithm
  EM Variants
Main Contribution
Experiment Result
Further Reading
The Model

» Multivariate random variable $\mathbf{X}$ as a mixture model of $k$ components.

$$p(x; \Psi) = \sum_{j=1}^{k} \pi_j f(x; \vartheta_j)$$

weights $\{\pi_j\}_{j=1}^{k}$, $\pi_j > 0$, $\sum_{j=1}^{k} \pi_j = 1$

parameters of the density functions: $\{\vartheta_j\}_{j=1}^{k}$

parameters set: $\Psi = \{\pi, \vartheta\}$

» Contaminated Gaussian distribution:

$$f(x; \vartheta_j) = \alpha_j \phi(x; \mu_j, \Sigma_j) + (1 - \alpha_j) \phi(x; \mu_j, \eta_j \Sigma_j)$$

$\alpha_j \in [0, 1]$, $\eta_j > 0$, $\vartheta_j = \{\alpha_j, \mu_j, \Sigma_j, \eta_j\}$

Multivariate Gaussian:

$$\phi(x; \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - \mu) \Sigma^{-1} (x - \mu)^T \right\}$$
EM Algorithm

Set of observation \( \mathbf{X} = \{x_1, x_2, \ldots, x_n\} \). The density function defined by set of parameters \( \Theta \), \( p(x|\Theta) \). The density of the observations:

\[
p(X|\Theta) = \prod_{i=1}^{n} p(x_i|\Theta) = \mathcal{L}(\Theta|X)
\]

Finding \( \Theta \) that maximizes the likelihood function \( \mathcal{L} \)

\[
\Theta^* = \arg \max_{\Theta} \mathcal{L}(\Theta|X)
\]

Assume the observation \( \mathbf{X} \) is incomplete, and latent variable \( \mathbf{Y} \). The “true” density function:

\[
p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta)
\]

\( \mathbf{Z} = (X, Y) \) is the complete data set. Define the complete-data likelihood function:

\[
\mathcal{L}(\Theta|Z) = \mathcal{L}(\Theta|X, Y)
\]

The likelihood is a function of the random variable \( \mathbf{Y} \), thus we define its expectation over the domain of \( \mathbf{Y} \)

\[
E[\mathcal{L}(\Theta|Z)] = E[p(x, y|\Theta) | x, \Theta]
\]

and take logarithm, the problem becomes:

\[
\Theta^* = \arg \max_{\Theta} E[\log p(x, y|\Theta) | x, \Theta]
\]
EM Algorithm

In very simple case, we can solve the problem analytically. In most cases, we need numerical method: EM algorithm, an iterative process where each iteration includes two steps: E-step and M-step.

- **E-step**: fix the parameters, calculate the expectation.

\[
Q \left( \Theta, \Theta^{(i-1)} \right) = E \left[ \log p(x, y|\Theta) | x, \Theta \right]
\]  

(1)

- **M-step**: find the parameters to maximize E-step result.

\[
\Theta^{(i)} = \arg \max_{\Theta} Q \left( \Theta, \Theta^{(i-1)} \right)
\]

Generalized EM (GEM): relaxing the M-step, only need to find \( \Theta^{(i)} \) to increase \( Q \)
EM for Mixture Model

» Assume a mixture of density functions:

\[
p (x|\Theta) = \sum_{i=1}^{k} \omega_i f_i (x|\theta_i)
\]

The parameters set \( \Theta = (\omega_1, ..., \omega_k, \theta_1, ..., \theta_k) \), \( \sum_{i=1}^{k} \omega_i = 1 \) and each \( f_i \) is a density function.

» Incomplete-data log-likelihood from \( n \) observations \( X \):

\[
\log (L (\Theta|X)) = \log \prod_{i=1}^{n} p (x_i, \Theta) = \sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} \omega_j p_j (x_i|\theta_j) \right)
\]

log of sum is difficult to optimize.

» Introduce the latent variable \( Y = \{y_i\}_{i=1}^{n} \) where \( y_i \in 1, ..., k \) indicating the membership of an observation in \( k \) components.
» If we can observe $Y$, the complete-data log-likelihood is:

$$\log (L (\Theta | X, Y)) = \log (p (X, Y|\Theta)) = \sum_{i=1}^{n} \log (p (x_i|y_i) p(y_i))$$

$$= \sum_{i=1}^{k} \log (\omega_{y_i} p_{y_i} (x_i|\theta_{y_i}))$$

which is easier to process.

» However $Y$ is a random variable, we must compute the expectation as shown in equation (1). This is more involved...
Skip to the results: Mixture of Gaussians, the parameters set is \( \Theta = \{ \omega, \mu, \Sigma \} \) which are the weights, the means, and the covariances.

\[
\begin{align*}
\omega_j^{new} &= \frac{1}{n} \sum_{i=1}^{n} p \left( j \mid x_i, \hat{\Theta} \right) \\
\mu_j^{new} &= \frac{\sum_{i=1}^{n} x_i p \left( j \mid x_i, \hat{\Theta} \right)}{\sum_{i=1}^{n} p \left( j \mid x_i, \hat{\Theta} \right)} \\
\Sigma_j^{new} &= \frac{\sum_{i=1}^{n} p \left( j \mid x_i, \hat{\Theta} \right) \left( x_i - \mu_j^{new} \right) \left( x_i - \mu_j^{new} \right)^T}{\sum_{i=1}^{n} p \left( j \mid x_i, \hat{\Theta} \right)}
\end{align*}
\]

where \( j = 1 \ldots k \) (number of mixture components).
Bayesian information criterion (BIC): Maximum log-likelihood with minimal complexity:

\[ BIC = -2 \ln \left( p \left( X | \hat{\Theta} \right) \right) + \rho \ln(n) \]

\( \rho \) is the number of free parameters.

Others: ICL, DIC, AIC, all are related to BIC by some approximation.
Expectation Conditional Maximization (ECM): A subclass of GEM algorithm. M-step is replaced by multiple CM-steps.

Idea: finding multivariate $\Theta$ is difficult, it is easier to maximize by one parameter, assuming the others are constants. In general, divide $\Theta$ into subsets, find the parameters in each subset while putting constrain on the rest.

How many CM-steps? Depend on the analysis of the log-likelihood function.
Parsimonious Variants

» p-variate domain → \( p(p + 1)/2 \) free parameters for each covariance matrix → parsimonious models.

» Eigen decomposition of the covariance matrix:

\[
\Sigma_j = \lambda_j \Gamma_j \Delta_j \Gamma_j^T
\]

\( \lambda_j \): Volume.
\( \Gamma_j \): matrix where columns are normalized eigenvectors: Orientation.
\( \Delta_j \): diagonal matrix of eigenvalues in decreasing order: Shape.

» Constraints on the three components yield fourteen parsimonious models, grouped into three categories: spherical, diagonal, and general.
Parsimonious Mixtures of Contaminated Gaussian Distribution models

Table 1: Nomenclature, covariance structure, type of ML solution in the first CM-step of the ECM algorithm (CF=closed form and IP=iterative procedure), and number of free covariance parameters for each member of the PMCGD family.

<table>
<thead>
<tr>
<th>Family</th>
<th>Model</th>
<th>Volume</th>
<th>Shape</th>
<th>Orientation</th>
<th>$\Sigma_j$</th>
<th>ML</th>
<th>Free covariance parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spherical</td>
<td>EII</td>
<td>Equal</td>
<td>Spherical</td>
<td>-</td>
<td>$\lambda I$</td>
<td>CF</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>VII</td>
<td>Variable</td>
<td>Spherical</td>
<td>-</td>
<td>$\lambda_j I$</td>
<td>CF</td>
<td>$k$</td>
</tr>
<tr>
<td>Diagonal</td>
<td>EEI</td>
<td>Equal</td>
<td>Equal</td>
<td>Axis-Aligned</td>
<td>$\lambda \Delta$</td>
<td>CF</td>
<td>$p$</td>
</tr>
<tr>
<td></td>
<td>VEI</td>
<td>Variable</td>
<td>Equal</td>
<td>Axis-Aligned</td>
<td>$\lambda_j \Delta$</td>
<td>IP</td>
<td>$k + p - 1$</td>
</tr>
<tr>
<td></td>
<td>EVI</td>
<td>Equal</td>
<td>Variable</td>
<td>Axis-Aligned</td>
<td>$\lambda \Delta_j \Lambda_j$</td>
<td>CF</td>
<td>$1 + k (p - 1)$</td>
</tr>
<tr>
<td></td>
<td>VVI</td>
<td>Variable</td>
<td>Variable</td>
<td>Axis-Aligned</td>
<td>$\lambda_j \Delta_j \Lambda_j$</td>
<td>CF</td>
<td>$kp$</td>
</tr>
<tr>
<td>General</td>
<td>EEE</td>
<td>Equal</td>
<td>Equal</td>
<td>Equal</td>
<td>$\lambda \Delta \Gamma \Delta'$</td>
<td>CF</td>
<td>$p (p + 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>VEE</td>
<td>Variable</td>
<td>Equal</td>
<td>Equal</td>
<td>$\lambda_j \Delta \Gamma \Delta'$</td>
<td>IP</td>
<td>$k + p - 1 + p (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>EVE</td>
<td>Equal</td>
<td>Variable</td>
<td>Equal</td>
<td>$\lambda \Delta_j \Gamma \Delta_j \Lambda_j$</td>
<td>IP</td>
<td>$1 + k (p - 1) + p (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>EEV</td>
<td>Equal</td>
<td>Equal</td>
<td>Variable</td>
<td>$\lambda \Delta \Gamma_j \Delta_j$</td>
<td>CF</td>
<td>$p + kp (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>VVE</td>
<td>Variable</td>
<td>Variable</td>
<td>Equal</td>
<td>$\lambda_j \Delta_j \Gamma \Delta_j$</td>
<td>IP</td>
<td>$kp + p (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>VEV</td>
<td>Variable</td>
<td>Equal</td>
<td>Variable</td>
<td>$\lambda \Delta_j \Gamma_j \Delta_j$</td>
<td>IP</td>
<td>$k + p - 1 + kp (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>EVV</td>
<td>Equal</td>
<td>Variable</td>
<td>Variable</td>
<td>$\lambda \Delta_j \Gamma_j \Delta_j$</td>
<td>CF</td>
<td>$1 + k (p - 1) + kp (p - 1) / 2$</td>
</tr>
<tr>
<td></td>
<td>VVV</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>$\lambda_j \Delta_j \Gamma_j \Delta_j$</td>
<td>CF</td>
<td>$kp (p + 1) / 2$</td>
</tr>
</tbody>
</table>

(Punzo and McNicholas)
Bivariate Model Illustration

(Brendan Murphy)
The paper introduce the model as described in slide 4.

The parameters set partitions: \( \Psi = \{ \Psi_1, \Psi_2 \} \) where \( \Psi_1 = \{ \pi, \alpha, \mu, \Sigma \} \) and \( \Psi_2 = \{ \eta \} \).

The authors use **mixture**, a **R** package implementing the persimonious mixture Gaussians.

They implement the 2-step ECM, and give the analytic equations for the VVVV model (in section 3.1).

They detail the initialization of the contaminated components: \( \alpha_j = \eta_j = 1 \), meaning: no contamination.

The discussion on convergence is somewhat unnecessary, as it is a known result.
Scan through the data set, for each $x_i$, perform outlier detection in two steps.

- Determine group membership: MAP (maximal a posteriori).

$$j^* = \arg \max_j \pi_j (\alpha_j \phi (x_j ; \mu_j, \Sigma_j) + (1 - \alpha_j) \phi (x_j ; \mu_j, \eta_j \Sigma_j))$$

- Determine if $x_i$ is an outlier:

$$\max (\alpha_j \phi (x_j ; \mu_j, \Sigma_j) , (1 - \alpha_j) \phi (x_j ; \mu_j, \eta_j \Sigma_j))$$

The parameter $\alpha_j$ sets the a priori for the good parameters, and is allowed to define with a lower bound or a constant.
» Compare BIC and ICL, using synthesized data sets: same performance.
» On two real data sets: clustering and detect outliers.
» The results show the strength of the proposed model to detect outliers. But only with 2-d data sets (somewhat quite simple).
Further Reading


