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

Nguyen D. Pham

Department of Computer Science
University of Houston

September 15, 2015

Introduction

- Parametric Density Estimation: Maximum Likelihood
- Application:
  - Alternative to non-parametric method.
  - Clustering and classification.

Discussion

The Model

- Multivariate random variable $\boldsymbol{X}$ 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: $\Psi = \{\boldsymbol{\pi}, \boldsymbol{\vartheta}\}$

- 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 $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$. The density function defined by set of parameters Θ, $p(x|\Theta)$. The density of the observations:

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

  Finding Θ that maximizes the likelihood function $\mathcal{L}$

$$\Theta^* = \arg\max_{\Theta} \mathcal{L}\left(\Theta|\boldsymbol{X}\right)$$

- Assume the observation $\boldsymbol{X}$ is incomplete, and latent variable $\boldsymbol{Y}$. The "true" density function:

$$p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta)$$

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

$$\mathcal{L}\left(\Theta|\boldsymbol{Z}\right) = \mathcal{L}\left(\Theta|\boldsymbol{X}, \boldsymbol{Y}\right)$$

- The likelihood is a function of the random variable $\boldsymbol{Y}$, thus we define its expectation over the domain of $\boldsymbol{Y}$

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

  and take logarithm, the problem becomes:

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

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.

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

  - 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:

$$p\left(x|\Theta\right) = \sum_{i=1}^{k} \omega_i f_i \left(x|\theta_i\right)$$

  The parameters set $\Theta = \left(\omega_1, ..., \omega_k, \theta_1, ..., \theta_k\right)$, $\sum_{i=1}^{k} \omega_i = 1$ and each f_i is a density function.
- Incomplete-data log-likelihood from n observations $\boldsymbol{X}$:

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

  log of sum is difficult to optimize.
- Introduce the latent variable $\boldsymbol{Y} = \{y_i\}_{i=1}^{n}$ where $y_i \in 1, ..., k$ indicating the membership of an observation in k components.

EM for Mixture Model

- If we can observe $\boldsymbol{Y}$, the complete-data log-likelihood is:

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

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

  which is easier to process.

- However $\boldsymbol{Y}$ is a random variable, we must compute the expectation as shown in equation (1). This is more involved...

EM for Mixture Model

» Skip to the results: Mixture of Gaussians, the parameters set is $\Theta = \{\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ which are the weights, the means, and the covariances.

$$\omega_j^{new} = \frac{1}{n} \sum_{i=1}^{n} p\left(j \middle| x_i, \hat{\Theta}\right)$$

$$\mu_j^{new} = \frac{\sum_{i=1}^{n} x_i p\left(j \middle| x_i, \hat{\Theta}\right)}{\sum_{i=1}^{n} p\left(j \middle| x_i, \hat{\Theta}\right)}$$

$$\Sigma_j^{new} = \frac{\sum_{i=1}^{n} p\left(j \middle| x_i, \hat{\Theta}\right) \left(x_i - \mu_j^{new}\right) \left(x_i - \mu_j^{new}\right)^T}{\sum_{i=1}^{n} p\left(j \middle| x_i, \hat{\Theta}\right)}$$

where $j = 1..k$ (number of mixture components).

Model Selection

- Bayesian information criterion (BIC): Maximum log-likelihood with minimal complexity:

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

  ρ is the number of free parameters.
- Others: ICL, DIC, AIC, all are related to BIC by some approximation.

ECM Algorithm

- Expectation Conditional Maximization (ECM): A subclass of GEM algorithm. M-step is replaced by multipe CM-steps.
- Idea: finding multivariate Θ is difficult, it is easier to maximize by one parameter, assuming the others are constants. In general, devide Θ 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 $\rightarrow p(p+1)/2$ free parameters for each covariance matrix $\rightarrow$ parsimonious models.
- Eigen decomposition of the covariance matrix:

$$\Sigma_j = \lambda_j \Gamma_j \Delta_j \Gamma_j^T$$

  λ_j: Volume.
  Γ_j: matrix where columns are normalized eigenvectors: Orientation.
  Δ_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*.

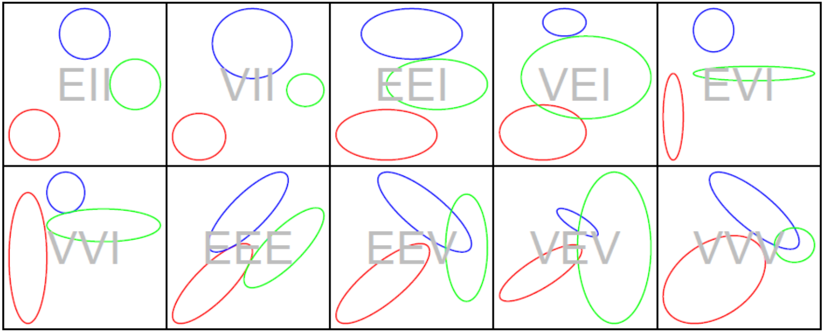
Parsimonious Mixtures of Contaminated Gaussian Distribution models

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.

Family	Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_j$	ML	Free covariance parameters
Spherical	EII	Equal	Spherical	-	$\lambda \boldsymbol{I}$	CF	1
	VII	Variable	Spherical	-	$\lambda_j \boldsymbol{I}$	CF	k
Diagonal	EEI	Equal	Equal	Axis-Aligned	$\lambda \boldsymbol{\Delta}$	CF	p
	VEI	Variable	Equal	Axis-Aligned	$\lambda_j \boldsymbol{\Delta}$	IP	$k+p-1$
	EVI	Equal	Variable	Axis-Aligned	$\lambda \boldsymbol{\Delta}_j$	CF	$1+k(p-1)$
	VVI	Variable	Variable	Axis-Aligned	$\lambda_j \boldsymbol{\Delta}_j$	CF	kp
General	EEE	Equal	Equal	Equal	$\lambda \boldsymbol{\Delta}\boldsymbol{\Gamma}\boldsymbol{\Delta}'$	CF	$p(p+1)/2$
	VEE	Variable	Equal	Equal	$\lambda_j \boldsymbol{\Delta}\boldsymbol{\Gamma}\boldsymbol{\Delta}'$	IP	$k+p-1+p(p-1)/2$
	EVE	Equal	Variable	Equal	$\lambda \boldsymbol{\Delta}_j\boldsymbol{\Gamma}\boldsymbol{\Delta}_j'$	IP	$1+k(p-1)+p(p-1)/2$
	EEV	Equal	Equal	Variable	$\lambda \boldsymbol{\Delta}\boldsymbol{\Gamma}_j\boldsymbol{\Delta}'$	CF	$p+kp(p-1)/2$
	VVE	Variable	Variable	Equal	$\lambda_j \boldsymbol{\Delta}_j\boldsymbol{\Gamma}\boldsymbol{\Delta}_j'$	IP	$kp+p(p-1)/2$
	VEV	Variable	Equal	Variable	$\lambda_j \boldsymbol{\Delta}\boldsymbol{\Gamma}_j\boldsymbol{\Delta}'$	IP	$k+p-1+kp(p-1)/2$
	EVV	Equal	Variable	Variable	$\lambda \boldsymbol{\Delta}_j\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j'$	CF	$1+k(p-1)+kp(p-1)/2$
	VVV	Variable	Variable	Variable	$\lambda_j \boldsymbol{\Delta}_j\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j'$	CF	$kp(p+1)/2$

(Punzo and McNicholas)

Bivariate Model Illustration



(Brendan Murphy)

Parsimonious Mixture of Contaminated Gaussian Distributions

- The paper introduce the model as described in slide 4
- The parameters set partitions: $\Psi = \{\Psi_1, \Psi_2\}$ where $\Psi_1 = \{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and $\Psi_2 = \{\boldsymbol{\eta}\}$.
- 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 VVV 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 unecessary, as it is a known result.

Outlier Detection

- Scan through the data set, for each x_i, perform outlier detection in two steps.
  - Determine group membership: MAP (maximal a posteriori).

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

  - Determine if x_i is an outlier:

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

- The parameter α_j sets the a priori for the good parameters, and is allowed to define with a lower bound or a constant.

Experiment Result

- Compare BIC and ICL, using synthesized data sets: same performance.
- On two real data sets: clustering and detect outliers.
- The resuls show the strength of the proposed model to detect outliers. But only with 2-d data sets (somewhat quite simple).

Further Reading

» Dempster A.P., 1976. Maximum Likelihood from Incomplete Data via the EM Algorithm.

» Donald B.R., 1993. Maximum Likelihood Estimation via the ECM Algorithm: A Genaral Framework.

» Celeux G., 1993. Gaussian Parsimonious Clustering Models.

» Steele R.J., 2009. Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models.