Finite Singular Multivariate Gaussian Mixture

Khalil MASMOUDI

21/06/2016

Plan

1. Basic definitions
   - Singular Multivariate Normal Distribution
2. Finite Singular Multivariate Normal Distributions Mixture
   - The model
   - Parameters estimation
3. Perspectives

Plan

1. Basic definitions
   - Singular Multivariate Normal Distribution
2. Finite Singular Multivariate Normal Distributions Mixture
   - The model
   - Parameters estimation
3. Perspectives

Multivariate Normal Distribution

Gaussian Vector Properties

Let $X = (X_1, X_2, ..., X_d)^T \sim N_d(\mu, \Sigma)$.

- $\forall \beta \in \mathbb{R}^d, \beta^T X \sim N(\beta^T \mu, \beta^T \Sigma \beta)$
- Laplace Transform :$L(\theta) = E(e^{\theta^T X}) = e^{\theta^T \mu + \frac{1}{2}\theta^T \Sigma \theta}$

Density function

if Σ is positive-definite $(\Sigma > 0)$:

$\forall x \in \mathbb{R}^d, f(x) = (2\pi)^{\frac{-d}{2}} |\Sigma|^{\frac{-1}{2}} exp(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$

Singular Multivariate Normal Distribution

Definition

Let $X = (X_1, X_2, ..., X_d)^T \sim N_d(\mu, \Sigma)$

- If $r = rank(\Sigma) < d$, X has a singular multivariate normal distribution.
- Laplace Transform :$L(\theta) = E(e^{\theta^T X}) = e^{\theta^T \mu + \frac{1}{2}\theta^T \Sigma \theta}$
- $\Sigma = (P_1, P_2) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}^T$ where $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}$
- The density function is defined on an affine subspace of $\mathbb{R}^d$:
  $\begin{cases} f(x) = (2\pi)^{\frac{-r}{2}} |\Lambda|^{\frac{-1}{2}} exp(\frac{-1}{2}(x-\mu)^T \Sigma^+ (x-\mu)) \\ P_2^T x = P_2^T \mu \end{cases}$
  Where Σ^+ is the Moore-Penrose pseudo-inverse of Σ.

Plan

1. Basic definitions
   - Singular Multivariate Normal Distribution
2. Finite Singular Multivariate Normal Distributions Mixture
   - The model
   - Parameters estimation
3. Perspectives

Mixture Model

Definition

Consider a mixture of K d-variate singular normal distributions with density :

$$f(x|\Theta) = \sum_{k=1}^{K} \pi_k f_k(x|\mu_k, \Sigma_k) \quad (1)$$

- $\pi_1, \pi_2, ..., \pi_K$ are the mixing weights :$(\sum_{k=1}^{K} \pi_k = 1)$
- $f_k(x|\mu_k, \Sigma_k)$ is the density function of multivariate degenerate Gaussian distribution with mean μ_k and covariance matrix Σ_k
- $\Theta = \{\pi_k, \mu_k, \Sigma_k; k = 1..K\}$ the parameters vector
- All components are concentrated on an affine subspace E.

Concrete Situation

Clustering data

- Y is drawn form a population which consists of K groups G_1, G_2,...,G_K in proportions $\pi_1, ..., \pi_K$.
- Z is K-dimensional component-label vector

$$Z_i = (Z_{i1}, ..., Z_{iK}) : Z_{ik} = (Z_i)_k = \begin{cases} 1 & \text{if } Y_i \in G_k \\ 0 & \text{if } Y_i \notin G_k \end{cases}$$

- Z_i is distributed according to a multinomial distribution with probabilities vector $\pi = (\pi_1, ..., \pi_K)$

$$Z_i \sim Mult_K(1, \pi) \text{ and } P(Z_i = z_i) = \prod_{k=1}^{K} \pi_k^{z_i k}$$

Parameters estimation

Data in the EM framework (Dempster et al., 1977)

- $y = (y_1, ..., y_n)$: the observed data.
- $z = (z_1, ..., z_n)$: the hidden data.
- The complete data is $(y^T, z^T)^T$

Likelihood function

- The complete likelihood function is given by :

$$l(y_1, ..., y_n, z_1, ..., z_n|\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{K} [\pi_j^{z_{ij}} f_j^{z_{ij}}(y_i|\mu_j, \Sigma_j)]$$

- The complete-data log-likelihood can be written as :

$$L(y_1, ..., y_n, z_1, ..., z_n|\Theta) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} log(\pi_j) + z_{ij} f_j(y_i|\mu_j, \Sigma_j)$$

EM Algorithm : 3 steps

Initialization

- $\pi_k^{(0)} = \frac{1}{K}$
- $\mu_k^{(0)} \in E$
- $\Sigma_k^{(0)}$ positive semi-definite verifying : $\varphi(\Sigma_k^{(0)}) = \bar{E}$ where $\varphi(A)$ denotes the column space of a matrix A and $\bar{E}$ is the vector subspace associated to E.

Expectation Step : E-Step

After l iterations, the conditional expectation of the complete-data log likelihood given the observed data using the current fit can be written as follows :

$$Q(\Theta||\Theta^{(l)}) = E_{\Theta^{(l)}}(L(y, z, \Theta)|y)$$

EM Algorithm : 3 steps

Expectation Step : E-Step

- The posterior probability that y_i belongs to the jth component of the mixture :

$$\tau_{ij}^{(l)} = E_{\Theta^{(l)}}(Z_{ij}|y) = P_{\Theta^{(l)}}(Z_{ij} = 1|y) = \frac{\pi_j^{(l)} f_j(y_i|\mu_j^{(l)}, \Sigma_j^{(l)})}{\sum_{k=1}^{K} \pi_k^{(l)} f_k(y_i|\mu_k^{(l)}, \Sigma_k^{(l)})}$$

- The conditional expectation of the complete-data log likelihood given y is :

$$Q(\Theta||\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_{ij}^{(l)}[log(\pi_j) + log(f_j(y_i|\mu_j, \Sigma_j))]$$

$$Q(\Theta||\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_{ij}^{(l)}[log(\pi_j) - \frac{r_j}{2}log(2\pi) - \frac{1}{2}log|\Lambda_j| - \frac{1}{2}(y_i - \mu_j)^T \Sigma_j^{+}(y_i - \mu_j)]$$

EM Algorithm : 3 steps

Maximization Step : M-Step

The global maximization of $Q(\Theta||\Theta^{(l)})$ with respect to Θ over the parameter space to give the updated estimate $\Theta^{(l+1)}$.

$$\Theta^{(l+1)} = \text{argmax}_{\Theta}\, Q(\Theta||\Theta^{(l)})$$

- $\pi_j^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}^{(l)}$
- $\mu_j^{(l+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(l)} y_i}{n_j}$
- $n_j = \sum_{i=1}^{n} \tau_{ij}^{(l)}$

EM Algorithm : 3 steps

Maximization Step : Covariance matrices

Let

$$S_j^{(l+1)} = \sum_{i=1}^{n} \tau_{ij}^{(l)} (y_i - \mu_j^{(l+1)})(y_i - \mu_j^{(l+1)})^T \quad (2)$$

and H_j be the $d \times r_j$ matrix of eigenvectors corresponding to the r_j largest eigenvalues of $S_j^{(l+1)}$, denoted by $L_j = diag(l_{j1}, ..., l_{jr_j})$.
Then the maximum of Q with respect to Σ_j is achieved for :

$$\Sigma_j^{(l+1)} = \frac{H_j L_j H_j^T}{n_j} \text{ where } n_j = \sum_{i=1}^{n} \tau_{ij}^{(l)}.$$

Plan

1. Basic definitions
   - Singular Multivariate Normal Distribution
2. Finite Singular Multivariate Normal Distributions Mixture
   - The model
   - Parameters estimation
3. Perspectives

Perspectives

Singular Multivariate Normal Distribution

1. Simulation study in progress.
2. Clustering real data.
3. Convergence problems : Choosing 'good' Starting values.

Thank you !

Moore-Penrose Pseudo-inverse

Definition

Let $A \in \mathbb{M}_{n,m}$, then there exists a unique $A^+ \in \mathbb{M}_{n,m}$ that satisfies :

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $A^+A = (A^+A)^*$ Hermitian
4. $AA^+ = (AA^+)^*$ Hermitian

Where M^* is the conjugate transpose of matrix M.

Remark

If A is square and non-singular, it is clear that $A^+ = A^{-1}$.
For Σ we can check that : $\Sigma^+ = P_1 \Lambda^{-1} P_1^T$

Maximum Likelihood Estimation

Theorem (M.S. Srivastava & D. von Rosen)

Let $Y = (y_1 : y_2 : ... : y_n)$ be a set of i.i.d data drawn from $N_d(\mu, \Sigma)$, where Σ is of rank $r(\Sigma) = r$, and $\Sigma = P_1 \Lambda P_1^T$. Let $S = Y(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T)Y$ and H be the $p \times r$ matrix of eigenvectors corresponding to the r largest eigenvalues of S, denoted by $L = diag(l_1, ..., l_r)$. $\mathbb{1}$ stands for $n \times 1$ vector of ones. Then the MLE of Λ and P_1 are respectively given by :

$$\hat{\Lambda} = \frac{1}{n} L$$

$$\hat{P}_1 = H$$

$$\hat{\Sigma} = \frac{1}{n} HLH^T$$