Adaptive Estimation for Mixture Parameters

Jiayang Sun[1], Peng Liu[1] and Jiahua Chen[2]

Case Western Reserve University[1]

and

U of Waterloo[2]

jiayang@case.edu

SSC 2005, Saskatoon, June 12-15, 2005

Abstract

daptive estimators for parameters in m
are especially useful when data come in
arge to be inputed into a standard algori
quentially overtime (ie. they are on-line
tors are compared with the standard
n the EM algorithm can be use
test we develop a data

Motivation

- Mixture models:

$$f(x) = \sum_{i=1}^{k} \pi_i \phi(x|\theta_i)$$

are *useful* in modeling populations with heterogeneity, which occurs often in practice. Determining the number of components is challenging.

Motivation

- Mixture models:

$$f(x) = \sum_{i=1}^{k} \pi_i \phi(x|\theta_i)$$

  are *useful* in modeling populations with heterogeneity, which occurs often in practice. Determining the number of components is challenging.

- Recently Charnigo and Sun (04a, 04b) developed a D-test for testing 1 versus k components.

Motivation

- Mixture models:

$$f(x) = \sum_{i=1}^{k} \pi_i \phi(x|\theta_i)$$

are *useful* in modeling populations with heterogeneity, which occurs often in practice. Determining the number of components is challenging.

- Recently Charnigo and Sun (04a, 04b) developed a D-test for testing 1 versus k components.

The D-test deviates from the LRT paradigm. It has good power, good asymptotic properties, and closed-form solutions in terms of $(\hat{\pi}_i, \hat{\theta})$. It can be generalized to test k versus $k+1$ components.

Question

Hence, if the adaptive estimates are available, the D-test has great potential in data mining and online data applications.

Question

Hence, if the adaptive estimates are available, the D-test has great potential in data mining and online data applications.

Q: How to estimate (π, θ) adaptively and efficiently?

First set data: $Y^{(1)} = \{Y_1, \ldots, Y_{n_1}\}$

Second set data: $Y^{(2)} = \{Y_{n_1+1}, \ldots, Y_n\}$

Question

Hence, if the adaptive estimates are available, the D-test has great potential in data mining and online data applications.

Q: How to estimate (π, θ) adaptively and efficiently?

First set data: $Y^{(1)} = \{Y_1, \ldots, Y_{n_1}\} \Longrightarrow \pi^{(1)}, \theta^{(1)}$

Second set data: $Y^{(2)} = \{Y_{n_1+1}, \ldots, Y_n\}$

Question

Hence, if the adaptive estimates are available, the D-test has great potential in data mining and online data applications.

Q: How to estimate (π, θ) adaptively and efficiently?

First set data: $Y^{(1)} = \{Y_1, \ldots, Y_{n_1}\} \Longrightarrow \pi^{(1)}, \theta^{(1)}$

Second set data: $Y^{(2)} = \{Y_{n_1+1}, \ldots, Y_n\} \Longrightarrow \pi^{(2)}, \theta^{(2)}$

based on $Y^{(2)}$ and $\pi^{(1)}, \theta^{(1)}$.

Question

Hence, if the adaptive estimates are available, the D-test has great potential in data mining and online data applications.

Q: How to estimate (π, θ) adaptively and efficiently?

First set data: $Y^{(1)} = \{Y_1, \ldots, Y_{n_1}\} \Longrightarrow \pi^{(1)}, \theta^{(1)}$

Second set data: $Y^{(2)} = \{Y_{n_1+1}, \ldots, Y_n\} \Longrightarrow \pi^{(2)}, \theta^{(2)}$

based on $Y^{(2)}$ and $\pi^{(1)}, \theta^{(1)}$.

How?

Notes:

- Two cases: (1) k to k (large p or n, or online data)
  (2) k to $k+1$ (online data)
- $n = n_1 + n_2$
- $n = \sum_i^{k'} m_i,\ m_i = \#\{j : Y_j \in G_i\}$, G_i the i-th component

Solutions and Road Map

- Partial EM (PEM)
- Bayesian (Sampling) MAP estimate (B)
  - ♣ Comparisons
- Bayesian Variants
  - Bayesian Partial EM (BPEM)
  - Complete Data Augmentation (AGM)

- Choice of Hyperparameters
  - ♣ Use EB and Asymptotics

- Final Performance
- Discussions

Formulation: k to k

- Consider a finite mixture of k normal[1] distributions with a pdf:

$$f(x) = \sum_{i=1}^{k} \pi_i \frac{1}{\sigma_i} \phi(\frac{x - \mu_i}{\sigma_i})$$

  where $\phi \sim N(0,1)$, $\sigma_i > 0, \pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$.

- Let $\mu = (\mu_1, \ldots, \mu_k), \sigma = (\sigma_1, \ldots, \sigma_k), \pi = (\pi_1, \ldots, \pi_k)$, then $\theta = (\mu, \sigma, \pi)$.

- Write $Y = (Y^{(1)}, Y^{(2)}) = (Y_1, \ldots, Y_n) \overset{iid}{\sim} f$

---

[1] The methods we develop can be applied more generally

EM on Y, if we can

Start from some *good* $\theta^{(0)}$, compute

E-step:

$$\tau_{ij} = P(Y_j \in G_i | Y) = \frac{\pi_i \frac{1}{\sigma_i} \phi(\frac{Y_j - \mu_i}{\sigma_i})}{\sum_{t=1}^{k} \pi_t \frac{1}{\sigma_t} \phi(\frac{Y_j - \mu_t}{\sigma_t})}, \qquad \pi_i = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij},$$

M-step

$$\mu_i = \frac{\sum_{j=1}^{n} \tau_{ij} Y_j}{n\pi_i}, \qquad \sigma_i^2 = \frac{\sum_{j=1}^{n} \tau_{ij} (Y_j - \mu_i)^2}{n\pi_i}$$

and then iterate until convergence.

Partial EM

Treat $\pi_i^{(1)}, \mu_i^{(1)}, \sigma_i^{2(1)}$ as initial values. Compute

$$\tau_{ij}^{(2)} = \frac{\pi_i^{(1)} \frac{1}{\sigma_i^{(1)}} \phi\left(\frac{Y_j - \mu_i^{(1)}}{\sigma_i^{(1)}}\right)}{\sum_{t=1}^{k} \pi_t^{(1)} \frac{1}{\sigma_t^{(1)}} \phi\left(\frac{Y_j - \mu_t^{(1)}}{\sigma_t^{(1)}}\right)}, \quad Y_j \in Y^{(2)}$$

Then

$$\tilde{\pi}_i = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij} = \frac{1}{n} \left(n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} \tau_{ij}^{(2)} \right)$$

$$\tilde{\mu}_i = \frac{\sum_{j=1}^{n} \tau_{ij} Y_j}{n \tilde{\pi}_i} = \frac{n_1 \pi_i^{(1)} \mu_i^{(1)} + \sum_{Y_j \in Y^{(2)}} \tau_{ij}^{(2)} Y_j}{n \tilde{\pi}_i}$$

Partial EM - cont

$$\tilde{\sigma}_i^2 = \frac{n_1 \pi_i^{(1)} [\sigma_i^{2(1)} + (\mu_i^{(1)} - \tilde{\mu}_i)^2] + \sum_{Y^{(2)}} \tau_{ij}^{(2)} (Y_j - \tilde{\mu}_i)^2}{n \tilde{\pi}_i}$$

With $\tilde{\pi}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2$, we could again compute

$$\tilde{\tau}_{ij}^{(2)} = \frac{\tilde{\pi}_i \frac{1}{\tilde{\sigma}_i} \phi(\frac{Y_j - \tilde{\mu}_i}{\tilde{\sigma}_i})}{\sum_{t=1}^{k} \tilde{\pi}_t \frac{1}{\tilde{\sigma}_t} \phi(\frac{Y_j - \tilde{\mu}_t}{\tilde{\sigma}_t})}, \quad Y_j \in Y^{(2)}$$

Then the updated estimates can be obtained by iteratively computing above steps, until convergence.

Bayesian (Sampling) MAP

Idea: Use $\pi_i^{(1)}, \mu_i^{(1)}, \sigma_i^{2(1)}, i = 1, \ldots, k$ to build priors and then use the priors, data $Y^{(2)}$ and a surrogate to find posterior distributions and MAP estimates.

Setup: We have

$$Y_1, \ldots, Y_n \quad \text{iid} \quad f(x) = \sum_{i=1}^{k} \pi_i \frac{1}{\sigma_i} \phi(\frac{x - \mu_i}{\sigma_i})$$

The natural conjugate priors on $\theta = (\pi, \mu, \sigma)$ are

$$(\pi_1, \ldots, \pi_k) \sim Dirichlet(\alpha_1, \ldots, \alpha_k),$$

$$\mu_i | \sigma_i \sim N(\xi_i, \sigma_i^2/\eta_i); \quad \beta_i = 1/\sigma_i^2 \sim Gamma(r_i/2, \lambda_i/2)$$

where $\alpha_i, \xi_i, \eta_i, r_i$ and λ_i are hyperparameters.

complete log-posterior

Also define z_{ij} as the *component indicator variable* (surrogate) for Y_j, i.e.,

$$z_{ij} = \begin{cases} 1 & \text{if } Y_j \sim \phi(x; \mu_i, \sigma_i), \\ 0 & \text{otherwise,} \end{cases}$$

and $\Sigma_i z_{ij} = 1$. Then the complete log-posterior is

$$L(\theta, \pi|Y) = \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \left[\ln \pi_i + \ln \sqrt{\beta_i} - \frac{(Y_j - \mu_i)^2}{2\sigma_i^2} \right] + \ln [\text{priors}] + C$$

1-step update

MAP-step:

$$\pi_i = \frac{m_i + \alpha_i - 1}{n + \sum_{i=1}^{k} \alpha_i - k} \qquad \mu_i = \frac{\sum_{j=1}^{n} z_{ij} Y_j + \eta_i \xi_i}{\sum_{j=1}^{n} z_{ij} + \eta_i}$$

$$\frac{1}{\sigma_i^2} = \beta_i = \frac{\sum_{j=1}^{n} z_{ij} + r_i - 1}{\sum_{j=1}^{n} z_{ij}(Y_j - \mu_i)^2 + \eta_i(\mu_i - \xi_i)^2 + \lambda_i}$$

E-step: Estimate z_{ij} by

$$\tau_{ij} = \frac{\pi_i^{(1)} \frac{1}{\sigma_i^{(1)}} \phi\left(\frac{Y_j - \mu_i^{(1)}}{\sigma_i^{(1)}}\right)}{\sum_{t=1}^{k} \pi_t^{(1)} \frac{1}{\sigma_t^{(1)}} \phi\left(\frac{Y_j - \mu_t^{(1)}}{\sigma_t^{(1)}}\right)}$$

Comparisons

mu=(0,4), n1:n2=1:1

$Pi:

	n=1k	2.5k	5k	10k	1k(s)	2.5
EM	3.107036	1.259436	0.6289870	0.3115171	2.708380	1.12
PEM	3.425930	1.364331	0.6820706	0.3380298	2.747352	1.13
B1	5.281034	2.125327	1.1015659	0.5424845	4.471163	1.84
B2	3.633327	1.433654	0.7156936	0.3550152	2.821426	1.15
B3	5.616947	2.214934	1.1352690	0.5541706	4.807809	1.93
Y1	6.242600	2.538097	1.2460644	0.6200195	5.441384	2.26
Y2	6.290869	2.500923	1.2735599	0.6174719	5.432249	2.21

$$EM > PEM > B2 > (Y1, Y2)$$

B1: $\alpha = \pi_i^{(1)}\sqrt{n}$ –> sometimes okay

B2: $\alpha = \pi_i^{(1)} n$ –> best, should have used n_1.

B1: $\alpha = \pi_i^{(1)}$ –> uniformly bad

Variants of Bayes Estimates

1. Iterate MAP (inside too) and E steps through $Y^{(2)}$
   $\Longrightarrow$ Baysian Partial EM

2. Similar to Baysian Partial EM but with m_i and n estimated using $(\pi^{(1)}, \theta^{(1)})$ and $Y^{(2)}$
   $\Longrightarrow$ Data Augumentation

Data Augmentation

$$\pi_i = \frac{m_i + \alpha_i - 1}{n + \sum_{i=1}^{k} \alpha_i - k} = \frac{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + \alpha_i - 1}{n_1 + n_2 + \sum_{i=1}^{k} \alpha_i - k}$$

$$\mu_i = \frac{n_1 \pi_i^{(1)} \mu_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} Y_j + \eta_i * \xi_i}{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + \eta_i}$$

$$\frac{1}{\sigma_i^2} = \beta_i = \frac{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + r_i - 1}{\sigma_i^{(1)} n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} (Y_j - \mu_i)^2 + \eta_i (\mu_i - \xi_i)^2 + \lambda_i}$$

Data Augmentation

$$\pi_i = \frac{m_i + \alpha_i - 1}{n + \sum_{i=1}^{k} \alpha_i - k} = \frac{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + \alpha_i - 1}{n_1 + n_2 + \sum_{i=1}^{k} \alpha_i - k}$$

$$\mu_i = \frac{n_1 \pi_i^{(1)} \mu_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} Y_j + \eta_i * \xi_i}{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + \eta_i}$$

$$\frac{1}{\sigma_i^2} = \beta_i = \frac{n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} + r_i - 1}{\sigma_i^{(1)} n_1 \pi_i^{(1)} + \sum_{Y_j \in Y^{(2)}} z_{ij} (Y_j - \mu_i)^2 + \eta_i (\mu_i - \xi_i)^2 + \lambda_i}$$

(Hyperparameters can be estimated once for all or updated too.)

Choices of Hyperparameters α

Note that

$$\hat{\pi}_i = \frac{m_i/n + (\alpha_i - 1)/n}{1 + (\sum_{j=1}^{k} \alpha_j - k)/n}$$

and the MSE of the estimator $\hat{\pi}_i$ is

$$E(\hat{\pi}_i - \pi_i)^2 \quad = \quad E\left(\frac{Z_n + B}{1 + \frac{\sum \alpha_j - k}{n}}\right)^2 \approx \frac{\frac{\pi_i(1-\pi_i)}{n} + B^2}{(1 + \frac{\sum \alpha_j - k}{n})^2}$$

where $Z_n = m_i/n - \pi_i, B = \frac{\alpha_i - 1}{n} - \frac{\sum \alpha_j - k}{n}\pi_i$. So, we'll choose α_i such that the asymptotic MSE is minimized and $\hat{\pi}_i$ is consistent.

Choices of Hyperparameters α

Note that

$$\hat{\pi}_i = \frac{m_i/n + (\alpha_i - 1)/n}{1 + (\sum_{j=1}^{k} \alpha_j - k)/n}$$

and the MSE of the estimator $\hat{\pi}_i$ is

$$E(\hat{\pi}_i - \pi_i)^2 \quad = \quad E\left(\frac{Z_n + B}{1 + \frac{\sum \alpha_j - k}{n}}\right)^2 \approx \frac{\frac{\pi_i(1-\pi_i)}{n} + B^2}{(1 + \frac{\sum \alpha_j - k}{n})^2}$$

where $Z_n = m_i/n - \pi_i, B = \frac{\alpha_i - 1}{n} - \frac{\sum \alpha_j - k}{n}\pi_i$. So, we'll choose α_i such that the asymptotic MSE is minimized and $\hat{\pi}_i$ is consistent.

The solution is $\alpha_i = \pi_i^{(1)} n_1$.

ξ and η

We let

$$\xi_i = \mu_i^{(1)}, \quad \eta_i = \pi_i^{(1)} n_2$$

r and λ

- Using MME and asymptotic expansion of $\hat{\sigma}$: Choose a large λ_i say 10, and set

$$r_i = \frac{1}{\sigma_i^{2(1)}} \lambda_i$$

- Using mode: Set the mode at $1/\sigma_i^{2(1)}$ to get

$$r_i = \frac{1}{\sigma_i^{2(1)}} \lambda_i + 2$$

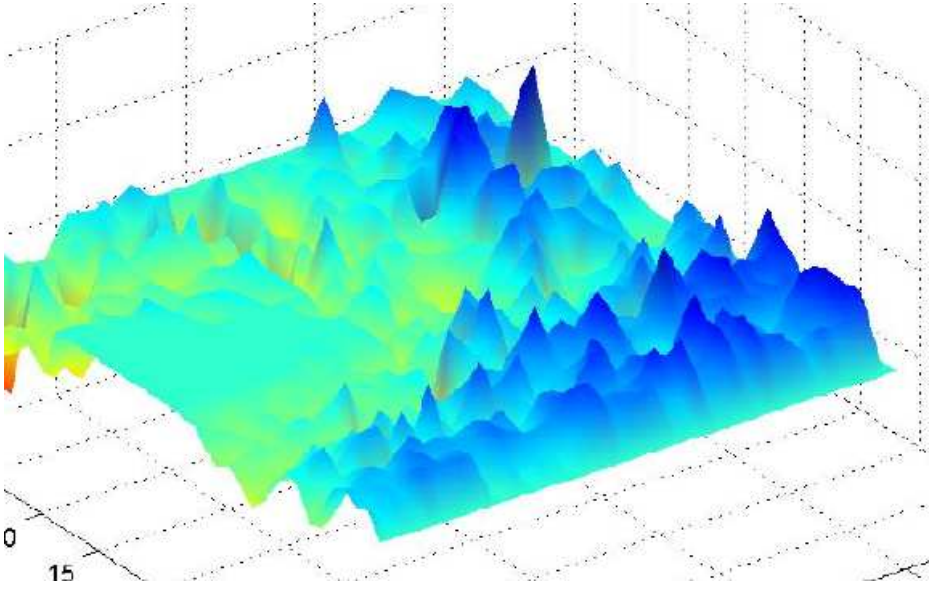
k **to** $k+1$

Set:

$$(\mu_1, \mu_2, \pi_1^{(1)}, \pi_2^{(1)}, \sigma_1, \sigma_2) \longrightarrow (\mu_1, \mu_2, \mu_3, \pi_1, \pi_2, \pi_3, \sigma_1, \sigma_2, \sigma_3)$$

where

$$\pi_1^{(1)} = \frac{\pi_1}{\pi_1 + \pi_2}, \quad \pi_2^{(1)} = \frac{\pi_2}{\pi_1 + \pi_2}$$



Performance

k to k: PEM, Agm1 and Agm2 (see details here)

k to $k+1$: PEM

n=n1+n2=10000, m=1000

$Y^{(1)}$: Size$= n_1, \pi = (0.5, 0.5), \mu = (0, 4), \sigma^2 = (1, 1)$

$Y^{(2)}$: Size$= n_2, \pi = (0.3, 0.3, 0.4), \mu = (0, 4, 8)\sigma^2 = (1, 1, 1)$

Y2(EM): EM algorithm based only on Y2

PEM: Partial EM

n1:n2=1:1

	Pi	Mu	Sigma
Y2(EM)	0.05758397	1.0151787	3.299722
PEM	0.03679747	0.5099786	1.262925

n1:n2=4:1

Y2(EM)	0.15024636	2.5769655	8.596389
PEM	0.07677624	0.7715693	1.783635

n1:n2=1:4

Y2(EM)	0.03729397	0.6248261	2.005826
PEM	0.03160076	0.5002733	1.349916

Discussion

Wasserman (1998): Using data dependent prior in mixtures is important!
Asymptotics: PEM, Augm?

Tests: Adaptation