Advances in EM-test for Finite Mixture Models

Jiahua Chen
Canada Research Chair, Tier I

Department of Statistics University of British Columbia

International Workshop on Perspectives on High-dimensional Data Analysis

Outline

1. Finite mixture models
   - Genetic Example
   - Finite mixture models
2. Hypothesis test
   - Test of homogeneity
   - Advances toward realistic solution
3. EM-test
   - Further advances
   - Limiting distribution

A genetic example: trait

- Geneticists often study Sodium-lithium countertransport (SLC) activity in red blood cells, since it
    - relates to blood pressure and the prevalence of hypertension;
    - is relatively easier to study than blood pressure.
- A search of “Sodium-lithium countertransport” shows up 12,400 results. The leading one is cited 676 times.

Population heterogeneity

- One genetic hypothesis is that the SLC activity is determined by a simple model of inheritance compatible with the action of a single gene with two alleles.
- Each observation (of SLC value) was composed of the sum of the effect of a genetic component and a normally distributed fluctuation.
- Thus, a general population may be divided into three subpopulations: (1) those has two copies of the allele that elevates the SLC activity; (2) those have one copy; and (3) those have 0 copies
- Hence, a random sample from the population should behave as a finite mixture of up to three components.

Population heterogeneity

- One genetic hypothesis is that the SLC activity is determined by a simple model of inheritance compatible with the action of a single gene with two alleles.
- Each observation (of SLC value) was composed of the sum of the effect of a genetic component and a normally distributed fluctuation.
- Thus, a general population may be divided into three subpopulations: (1) those has two copies of the allele that elevates the SLC activity; (2) those have one copy; and (3) those have 0 copies
- Hence, a random sample from the population should behave as a finite mixture of up to three components.

Population heterogeneity

- One genetic hypothesis is that the SLC activity is determined by a simple model of inheritance compatible with the action of a single gene with two alleles.
- Each observation (of SLC value) was composed of the sum of the effect of a genetic component and a normally distributed fluctuation.
- Thus, a general population may be divided into three subpopulations: (1) those has two copies of the allele that elevates the SLC activity; (2) those have one copy; and (3) those have 0 copies
- Hence, a random sample from the population should behave as a finite mixture of up to three components.

Population heterogeneity

- One genetic hypothesis is that the SLC activity is determined by a simple model of inheritance compatible with the action of a single gene with two alleles.
- Each observation (of SLC value) was composed of the sum of the effect of a genetic component and a normally distributed fluctuation.
- Thus, a general population may be divided into three subpopulations: (1) those has two copies of the allele that elevates the SLC activity; (2) those have one copy; and (3) those have 0 copies
- Hence, a random sample from the population should behave as a finite mixture of up to three components.

Heterogeneity leads to mixture model

- There are two competing genetic models: simple dominance model and additive model.
  - If one allele is dominant, then the data are a random sample from a two-component normal mixture model;
  - If the genetic effect is additive, then the data are a random sample from a three-component normal mixture model.

The data will be shown in the next slide.

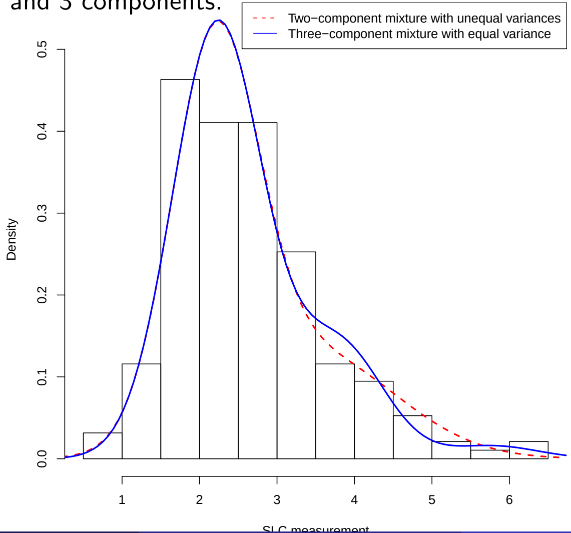
Heterogeneity leads to mixture model

- There are two competing genetic models: simple dominance model and additive model.
  - If one allele is dominant, then the data are a random sample from a two-component normal mixture model;
  - If the genetic effect is additive, then the data are a random sample from a three-component normal mixture model.

The data will be shown in the next slide.

SLC DATA

FIGURE: Histogram of 190 SLC measurements and suggestive normal mixture models with 2 and 3 components.



Reading from the histogram and fits

- It is not apparent whether a 2-component or a 3-component model is the "correct model".
- A rigorous statistical analysis would be helpful to shed light to the preference of the two competing models.
- One may take model selection approach, diagnostic approach and so on to answer this question.
- A statistical hypothesis test is likely the most desired approach.

Density function of a finite mixture

- Let $\{f(x;\theta) : \theta \in \Theta\}$ be a parametric distribution family where Θ is parameter space for θ.
- A finite mixture model is a class of distributions with density function in the form of

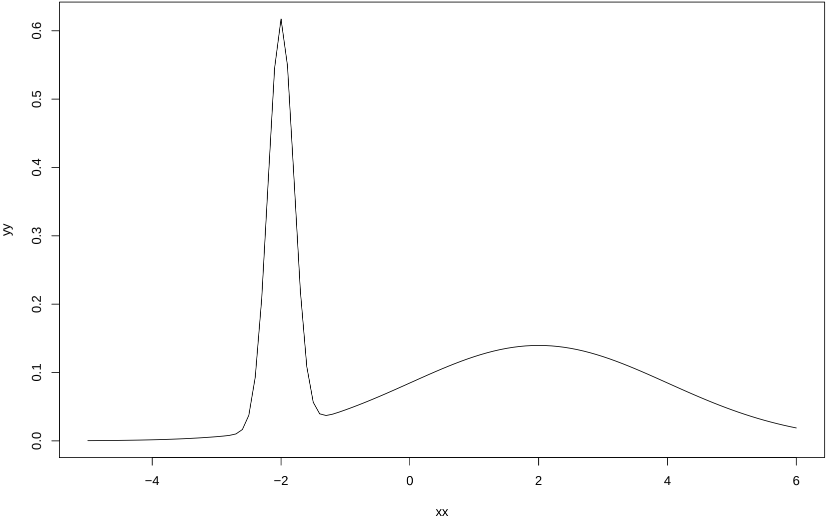
$$f(x;\Psi) = \sum_{h=1}^{m} \alpha_h f(x;\theta_h).$$

  - $f(x;\theta)$: kernel/component density function.
  - m: order of the finite mixture model.
  - θ_h: the parameter of the hth sub-population.
  - α_h: the proportion of the hth sub-population.

Mixing distribution

- One may put all parameters into a mixing distribution:
  - $\Psi(\theta) = \sum_{h=1}^{m} \alpha_h I(\theta_h \leq \theta)$.
  - $\Psi(\theta)$ is a distribution on Θ with m support points.

Density function of a 2-component normal mixture



Incomplete data structure

- A random variable X from a finite mixture model can be regarded as generated in two steps.
  - In the first step, a value of θ is generated from the mixing distribution Ψ.
  - When Ψ is discrete, this θ is labelled by h, the hth subpopulation.
  - Given θ_h, X is a random outcome from sub-population $f(x; \theta_h)$.
- Thus, the data from mixture models are "by definite" incomplete observations.

Genetic example and the mixture model

- An individual can have genotypes AA, Aa or aa.
- The SLC activity level of a randomly selected individual has density function

$$f(x;\Psi) = \sum_{h \in \{AA, Aa, aa\}} \alpha_h \phi(x; \mu_h, \sigma_h^2).$$

  where $\phi(x; \mu_h, \sigma_h^2)$ is the normal density with mean μ_h and variance σ_h^2.
- The genotype of the sample individual is generally unknown, particularly in this case.

Genetic question in statistical terminology

- Ignore some details, the statistical problem on the existence of a major gene is to test the null hypothesis of $m = 1$ against $m > 1$.
  - This is homogeneity test.
- To determine whether the major gene (allele) is additive or dominate, the statistical problem is to test the null hypothesis of $m = 2$ against $m = 3$.
  - This is to test the order of the mixture model.

Two-component model

- Given an iid sample $X_1, \ldots, X_n$ from a two-component mixture,
- the log-likelihood function of the mixing distribution is given by

$$\ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) = \sum_i \log\{\alpha_1 f(x_i; \theta_1) + \alpha_2 f(x_i; \theta_2)\}.$$

- Is the underlying population in fact homogeneous?
- That is, does $\theta_1 = \theta_2$?

Likelihood ratio test (LRT) for homogeneity

- The standard approach is to compute likelihood ratio test statistic:

$$R_n = 2\{\sup \ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) - \sup \ell_n(\alpha_1, \alpha_2, \theta, \theta)\}.$$

- Reject H_0 if R_n is larger than some threshold value.
- It only leaves a technical issue of computing the proper threshold value.

Likelihood ratio test (LRT) for homogeneity

- The standard approach is to compute likelihood ratio test statistic:

$$R_n = 2\{\sup \ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) - \sup \ell_n(\alpha_1, \alpha_2, \theta, \theta)\}.$$

- Reject H_0 if R_n is larger than some threshold value.
- It only leaves a technical issue of computing the proper threshold value.

The technical issue is challenging

- For regular models, R_n has an asymptotic chisquared distribution under the null hypothesis.
- Chisquared distributions are well documented and easily computed numerically.
- Hence, a proper threshold value can be easily determined based on chisquared distribution for hypothesis testing under regular models.

Finite mixture model is not regular

- Use $\alpha_1 f(x;\theta_1) + \alpha_2 f(x;\theta_2)$ for illustration:
  - When $\alpha_1 = 0$, any θ_1 value parameterizes the same distribution. There is a loss of identifiability (type I).
  - When $\theta_1 = \theta_2$, any (α_1, α_2) parameterize the same distribution. There is again a loss of identifiability (type II).
  - The null model is not an interior point in the set of alternative models.
- All of these violate the "regularity conditions" for "good behaviors" of classical likelihood approaches.

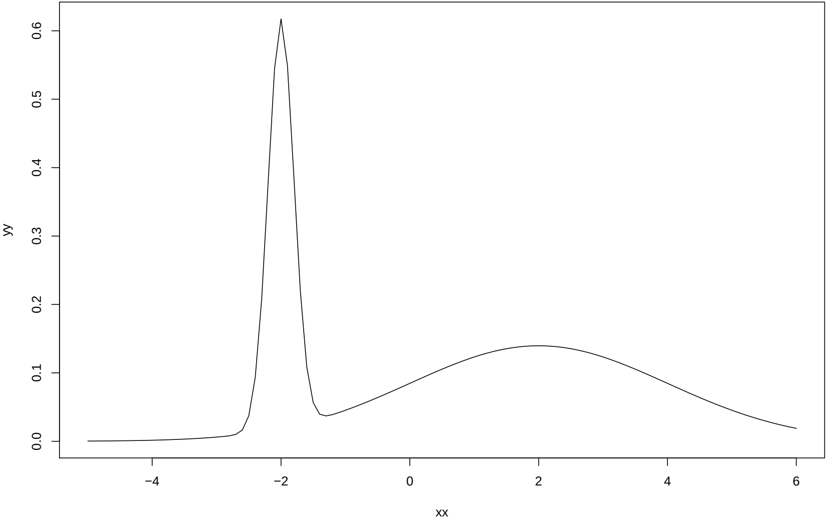
Surprises on LRT, I

- Researchers/geneticists believed the limiting distribution of R_n is still chisquare, except the degree of freedom needs more research.
- However,
  - For $(1-\alpha)N(0,1)+\alpha N(\theta,1)$ and when $\Theta = R$ Hartigan (1985) found that $R_n \to \infty$ as $n \to \infty$.
  - If the LRT statistics R_n is used, no finite threshold value is appropriate from asymptotic point of view.

Surprises on LRT, II

- For $(1-\alpha)N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2)$, the likelihood function is unbounded (based on an iid sample).
- See the plot of the density function of the two-component normal mixture model again.

Density function of a 2-component normal mixture



Breakthroughs starts from a Binomial mixture

- Suppose we have iid observations from a 2-component binomial distribution:

$$\alpha_1 \mathrm{Bin}(m, \theta_1) + \alpha_2 \mathrm{Bin}(m, \theta_2).$$

- Using parameter transformation and for homogeneity test, Chernoff and Lander (1995) obtained limiting distributions of the LRT statistics R_n.
    - This is the first result without requiring "separation condition" $|\theta_1 - \theta_2| > \epsilon$.

Immediate follow-up successes

- The limiting distribution of R_n was derived without separation condition by many authors soon after.
  - key conditions include
    (1) Θ is compact,
    (2) $E\{f(X;\theta)/f(X;\theta_0)\}^2 < \infty$ for any $\theta \in \Theta$.
  - drawbacks of the limiting distribution include
    (1) being a functional of Gaussian process,
    (2) dependent on Θ and θ_0.
- So what? the limiting distribution is not too useful for determining the threshold value.

A meaningful step toward a statistical solution

- Let

$$p\ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) = \ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) + C \log\{4\alpha_1\alpha_2\}.$$

- Similar to usual LRT, define

$$\tilde{R}_n = 2\{\max_{H_1} p\ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2) - \max_{H_0} p\ell_n(\alpha_1, \alpha_2, \theta_1, \theta_2)\}.$$

- Chen (1995, CJS) shows that the limiting distribution of $\tilde{R}_n$ is $0.5\chi_0^2 + 0.5\chi_1^2$.

What is the significance?

- The modified likelihood ratio statistic $\tilde{R}_n$ is an asymptotic pivot: its distribution does not depend the null distribution.
- The quantiles of $0.5\chi_0^2 + 0.5\chi_1^2$ (rather than a functional of a Gaussian process) can be easily computed.
- Significance of this result: practically the first implementable likelihood-based homogeneity test.

Why properties make $p\ell_n$ work?

- The first helpful property is that ℓ_n is bounded under binomial mixture model.
- The second helpful property is $C \log\{4\alpha_1\alpha_2\} \to -\infty$ as $\alpha_1\alpha_2 \to 0$.
  - Thus, $p\ell_n$ does not attain its maximum at small $\alpha_1\alpha_2$.
- Because of these, the $\tilde{R}_n$ is practically confined on $\alpha_1 \in [\epsilon, 1-\epsilon]$.
- On $[\epsilon, 1-\epsilon]$, the mixture model is almost "regular" which leads a simple limiting behavior.

Advance to homogeneity test to non-binomial mixtures

- The idea works for general homogeneity tests if ℓ_n is stochastically bounded.
- Boundedness comes under key conditions:

  (1) Θ is compact,

  (2) $E\{f(X;\theta)/f(X;\theta_0)\}^2 < \infty$ for any $\theta \in \Theta$.

Modified likelihood ratio test

- As long as (1) and (2) hold, the MLRT idea works and the limiting distributions are useful in applications:
  - Chen, Chen and Kalbfleisch (2001, JRSS, B) give the result for general homogeneity tests.
  - Chen, Chen and Kalbfleisch (2004, JRSS, B) succeed at finding the limiting distribution of $\tilde{R}_n$ for testing $m = 2$ against some $m > 2$.
- Regretfully, these results are obtained when Θ is compact and is one-dim.

Something new is still desirable

- Neither Chen, et al. (2001, 2004) is applicable to the genetic problem on SLC activity data because:
  - its $\theta = (\mu, \sigma)$ is 2-dimensional.
  - under normal mixture models, condition $E\{f(X;\theta)/f(X;\theta_0)\}^2 < \infty$ is not satisfied for all θ.
- Moving MLRT forward is vital. How?

An insight to the test of homogeneity, I

- Suppose the data are from a homogeneous model $f(x;\theta_0)$ and we want to examine the possibility that the actual model is a mixture with $m = 2$.
- Both LRT and MLRT let $f(x;\theta_0)$ compete against all potential models with $m = 2$.

An insight to the test of homogeneity, II

- In particular, a model such as

$$(1-\epsilon)f(x;\theta_0)+\epsilon f(x;\theta)$$

  is a competitor.

  - Without compact assumption on Θ, there are "too many" competitors.
  - A competitor with θ-value such that

$$E\{f(X;\theta)/f(X;\theta_0)\}^2=\infty$$

    has, in addition, unfair advantage!

- They explain the two undesirable conditions behind LRT and MLRT.

An insight to the test of homogeneity, II

- In particular, a model such as

$$(1-\epsilon)f(x;\theta_0)+\epsilon f(x;\theta)$$

is a competitor.

  - Without compact assumption on Θ, there are "too many" competitors.
  - A competitor with θ-value such that

$$E\{f(X;\theta)/f(X;\theta_0)\}^2=\infty$$

has, in addition, unfair advantage!

- They explain the two undesirable conditions behind LRT and MLRT.

An insight to the test of homogeneity, II

- In particular, a model such as

$$(1-\epsilon)f(x;\theta_0)+\epsilon f(x;\theta)$$

  is a competitor.

  - Without compact assumption on Θ, there are “too many” competitors.
  - A competitor with θ-value such that

$$E\{f(X;\theta)/f(X;\theta_0)\}^2=\infty$$

    has, in addition, unfair advantage!

- They explain the two undesirable conditions behind LRT and MLRT.

An insight to the test of homogeneity, II

- In particular, a model such as

$$(1-\epsilon)f(x;\theta_0)+\epsilon f(x;\theta)$$

  is a competitor.

  - Without compact assumption on Θ, there are “too many” competitors.

  - A competitor with θ-value such that

$$E\{f(X;\theta)/f(X;\theta_0)\}^2=\infty$$

    has, in addition, unfair advantage!

- They explain the two undesirable conditions behind LRT and MLRT.

EM-test for homogeneity test, I

- The key behind EM-test is to initially confine the range of H_a.
- Here is a simplified illustration:
  - initially test $H_0 : f(x;\theta)$ against $H_a' : 0.30f(x;\theta_1) + 0.70f(x;\theta_2)$.
  - Under H_0, this R_n has a simple $0.5\chi_0^2 + 0.5\chi_1^2$ limiting distribution.
- This test is not sensible, because the actual distribution of the data could be $0.45f(x;\theta_1) + 0.55f(x;\theta_2)$.

EM-test for homogeneity test, II

- If the sample is from H_0, both $0.45f(x;\theta_1) + 0.55f(x;\theta_2)$ and $0.30f(x;\theta_1) + 0.70f(x;\theta_2)$ will fit data well.

- If the sample is from $0.45f(x;\theta_1) + 0.55f(x;\theta_2)$, fitting $0.30f(x;\theta_1) + 0.70f(x;\theta_2)$ should leave a lot of room for further improvement.

EM-test for homogeneity test, III

- Thus, whether the data is from H_0 or not can be judged on how big a room there still is for improvement from the initially fit of a restrictive model $0.30f(x;\theta_1) + 0.70f(x;\theta_2)$.
- Our additional trick:
  use EM-iteration to improve the initial fit gradually.
- If a fixed number of EM-iteration increases the value of R_n substantially, H_0 is rejected.
- Further enhancement: use multiple initial fits $\beta f(x;\theta_1) + (1-\beta)f(x;\theta_2)$, such as $\beta \in \{0.1, 0.3, 0.5\}$.

The EM-test statistic for homogeneity

- Find the MLE of θ under the null hypothesis $\hat{\theta}_0$.
- Define two intervals $I_1 = (-\infty, \hat{\theta}_0)$ and $I_2 = [\hat{\theta}_0, \infty)$.
- Find $\hat{\theta}_1 \in I_1$ and $\hat{\theta}_2 \in I_2$ that maximizes $p\ell_n(0.3, 0.7, \theta_1, \theta_2)$.
- Let $(\alpha_1, \alpha_2, \theta_1, \theta_2)^{(0)} = (0.3, 0.7, \hat{\theta}_1, \hat{\theta}_2)$
- Perform EM-iteration k times.
- Define

$$EM_n^{(k)}(0.3) = 2\{p\ell_n((\alpha_1, \alpha_2, \theta_1, \theta_2)^{(K)}) - p\ell_n(0.5, 0.5, \hat{\theta}_0, \hat{\theta}_0)\}.$$

- Finally, let $EM_n^{(k)} = \max\{EM_n^{(k)}(0.1), EM_n^{(k)}(0.3), EM_n^{(k)}(0.5)\}$.

Ugly definition, beautiful limiting distribution

Theorem (Li, Chen and Marriott, 2008, Biometrika)

- Given a random sample of size n from $\alpha_1 f(x;\theta_1) + \alpha_2 f(x;\theta_2)$.
- Assume that $f(x;\theta)$ is smooth enough, makes the mixture model identifiable, and so on.
- Under the null distribution $f(x;\theta_0)$, and for any fixed finite k, $EM_n^{(k)} \to 0.5\chi_0^2 + 0.5\chi_1^2$ in distribution as $n \to \infty$.

- This result is obtained without $E\{f(X;\theta)/f(X;\theta_0)\}^2 < \infty$ nor compact Θ.
- Yet it is still for one-dim θ, and for homogeneity test only.
- We cannot stop at this point!

EM-TEST FOR $H_0 : m = m_0$

- From homogeneity test to $H_0 : m = m_0$ can be technical challenging.
- Li and Chen (2010, JASA) employed some special tricks to ensure the success of generalizing the result.

Define EM-test for $H_0 : m = m_0$, I

- Consider the case when θ is one-dim, and an iid sample is given.
- We first obtain the "MLE" $\hat{\Psi}_0$ under the null hypothesis (maximizing $p\ell_n$).
- Let $\hat{\theta}_{j0}$, $j = 1, 2, \ldots, m_0$ be estimated value of sub-population parameters.
- Let I_j's be the interval that contain $\hat{\theta}_{j0}$ and partition Θ evenly.

Define EM-test for $H_0 : m = m_0$, II

- We define a specific class of order-$2m_0$ mixture models

$$\Omega_{2m_0} = \{\sum_{j=1}^{m_0} \{\beta_j f(x;\theta_{j1}) + (1-\beta_j) f(x;\theta_{j2})\} : \theta_j \in I_j\}.$$

  where $\beta_j \in \{0.1, 0.3, 0.5\}$.
- Next, we find a $\hat{\Psi}^{(0)} \in \Omega_{2m_0}$ that maximizes $\ell_n(\Psi)$.
- Last, use EM-iteration to improve the fit of $\hat{\Psi}^{(k)}$.
- Multiple initial β_j will be used.

Define EM-test for $H_0 : m = m_0$, III

- After a pre-chosen iterations $k = K$, the EM-statistic is

$$M_n^{(K)} = 2\{\ell_n(\Psi^{(K)}) - \ell_n(\hat{\Psi}_0)\}$$

(take the largest out of multiple initial β).

- The EM-test rejects $H_0 : m = m_0$ in favour of $m > m_0$ if $M_n^{(K)}$ exceeds some threshold value.

“Tricks” in this EM-test

- We confined the initial alternative to Ω_{2m_0}.
  - It prevents wild models from being fitted.
- For each sub-population fitted under null model, we examine its possibility to be split into two sub-subpopulations.
  - We have a sub-homogeneity test within each initially fitted sub-population.
  - If these initial subpopulations spread out far away from each other, the limiting distribution would be a convolution of m_0 $0.5\chi_0^2 + 0.5\chi_1^2$.

EM-test: limiting distribution (1)

Theorem 2

Under some regularity conditions on $f(x;\theta)$ and penalty function $p(\beta)$, and assume $0.5 \in \mathrm{B}$ (set of initial values),

$$EM_n^{(K)} \to \sup_{\mathbf{v}\geq 0}(2\mathbf{v}^\tau \mathbf{w} - \mathbf{v}^\tau \Omega \mathbf{v}) = \sum_{h=0}^{m_0} a_h \chi_h^2$$

for some $a_h \geq 0$ and $\sum_{h=0}^{m_0} a_h = 1$, under Ψ_0 and fixed K.

- $\mathbf{w} = (w_1, \ldots, w_{m_0})^\tau$: a 0-mean multivariate normal random vector with correlation matrix $\Omega = (\omega_{ij})$.
- $\mathbf{v} = (v_1, \ldots, v_{m_0})^\tau$ and $\{\mathbf{v} \geq 0\} = \{v_1 \geq 0, \ldots, v_{m_0} \geq 0\}$.
- The weights $(a_0, \ldots, a_{m_0})$ depend on Ω.
- Ω can be calculated based on Ψ_0 or $\hat{\Psi}_0$.

EM-test: limiting distribution (2)

Theorem 2 (continued)

In particular,

1. when $m_0 = 1$, $a_0 = a_1 = 0.5$;
2. when $m_0 = 2$, $a_0 = (\pi - \arccos \omega_{12})/(2\pi)$, $a_1 = 0.5$, and $a_0 + a_2 = 0.5$;
3. when $m_0 = 3$, $a_0 + a_2 = a_1 + a_3 = 0.5$ and

$$\begin{aligned} a_0 &= (2\pi - \arccos \omega_{12} - \arccos \omega_{13} - \arccos \omega_{23})/(4\pi), \\ a_1 &= (3\pi - \arccos \omega_{12:3} - \arccos \omega_{13:2} - \arccos \omega_{23:1})/(4\pi), \end{aligned}$$

where

$$\omega_{ij:k} = \frac{(\omega_{ij} - \omega_{ik}\omega_{jk})}{\sqrt{(1-\omega_{ik}^2)(1-\omega_{jk}^2)}}.$$

Further progress is desired

- The previous result of Li and Chen (2010, JASA) succeeded at testing hypothesis of $H_0 : m = m_0$ against $H_a : m > m_0$.
- Yet the result is only applicable for one-dim Θ.
- The suggested model for SLC data is a finite normal mixture. Its $\theta = (\mu, \sigma^2)$ is 2-dimensional.
- Keep working!

Further progress is desired

- The previous result of Li and Chen (2010, JASA) succeeded at testing hypothesis of $H_0 : m = m_0$ against $H_a : m > m_0$.
- Yet the result is only applicable for one-dim Θ.
- The suggested model for SLC data is a finite normal mixture. Its $\theta = (\mu, \sigma^2)$ is 2-dimensional.
- Keep working!

EM-test for normal mixture model

- While the result of Li and Chen (2010, JASA) is not applicable, the EM-test principle is.
- Chen and Li (2009, AOS) worked out EM-test for homogeneity under finite normal mixture models.
- Surprisingly, the limiting distributions of $EM_n^{(k)}$ (defined similarly) are very simple and beautiful.

EM-test for homogeneity with equal-variance assumption

Theorem 3

Suppose the penalty function $p(\cdot)$ introduced in $p\ell_n$ satisfies some conditions.
The initial set of value B contains 0.5.
The alternative H_a is under equal-variance assumption.
Then under the homogeneous null distribution $N(\theta_0, \sigma_0^2)$ and for any finite K, as $n \to \infty$,

$$\Pr(EM_n^{(K)} \leq x) \to F(x - \Delta)\{0.5 + 0.5F(x)\},$$

where $F(x)$ is the cumulative density function (cdf) of the χ_1^2 and

$$\Delta = 2 \max_{\alpha_j \neq 0.5} \{p(\alpha_j) - p(0.5)\}.$$

EM-test for homogeneity without equal-variance assumption

Theorem 4

Suppose the penalty function $p(\cdot)$ introduced in $p\ell_n$ satisfies some conditions.

The initial set of value B contains 0.5.

The alternative H_a is any two component normal mixture.

Under the homogeneous null distribution $N(\theta_0, \sigma_0^2)$ and for any finite K, as $n \to \infty$,

$$EM_n^{(K)} \to \chi_2^2.$$

SLC DATA EXAMPLE AGAIN

- The results in Chen and Li (2009) is designed for finite normal mixture models. Hence model-wise, the method is applicable.
- A simple application shows the homogeneity assumption is rejected soundly.
- We are more interested in checking whether $H_0 : m = 2$ will be rejected in favour of $H_a : m > 2$.
- Charge forward further!

SLC DATA EXAMPLE AGAIN

- The results in Chen and Li (2009) is designed for finite normal mixture models. Hence model-wise, the method is applicable.
- A simple application shows the homogeneity assumption is rejected soundly.
- We are more interested in checking whether $H_0 : m = 2$ will be rejected in favour of $H_a : m > 2$.
- Charge forward further!

SLC DATA EXAMPLE AGAIN

- The results in Chen and Li (2009) is designed for finite normal mixture models. Hence model-wise, the method is applicable.
- A simple application shows the homogeneity assumption is rejected soundly.
- We are more interested in checking whether $H_0 : m = 2$ will be rejected in favour of $H_a : m > 2$.
- Charge forward further!

EM-test on the order of finite normal mixture model

Theorem 5 (Chen, Li and Fu, submitted)

Assume the same conditions on penalty functions placed in $p\ell_n$.
The initial set of value B contains 0.5.
Under the null distribution $f(x; \Psi_0)$ of order m_0, and for any fixed finite K, as $n \rightarrow \infty$,

$$EM_n^{(K)} \rightarrow \chi^2_{2m_0}.$$

- We have not worked on the case when σ_j are equal;
- The statistic is defined similarly but needed special care on $p\ell_n$.
- The method is fully applicable to the SLC data analysis.

Back to SLC data, null-fit

- We test the hypothesis of $H_0 : m = 2$ against $H_a : m = 3$.
- The best null model divides the population into two sub-populations with proportions: 65.4% and 34.6%.
- The fitted means and variances of two sub-populations are:

	mean	variance	proportion
Comp 1	2.194	0.557	65.4%
Comp 2	3.457	1.081	34.6%

Back to SLC data, conclusion

- Whether or not we reject $H_0 : m = 2$ in favor of $H_a : m = 3$ depends on how much better higher order models can fit the data.
- This question of "how much better" is answered through EM-statistics: we find

$$EM_n^{(1)} = 4.597, \ EM_n^{(2)} = 4.639, \ EM_n^{(3)} = 4.659.$$

- So when H_0 is true, EM-statistic can attain or exceed the above level with probability 33%.
- That is, such better fits as measured by EM-statistic can be easily explained by random fluctuation. Hence, H_0 is not rejected.

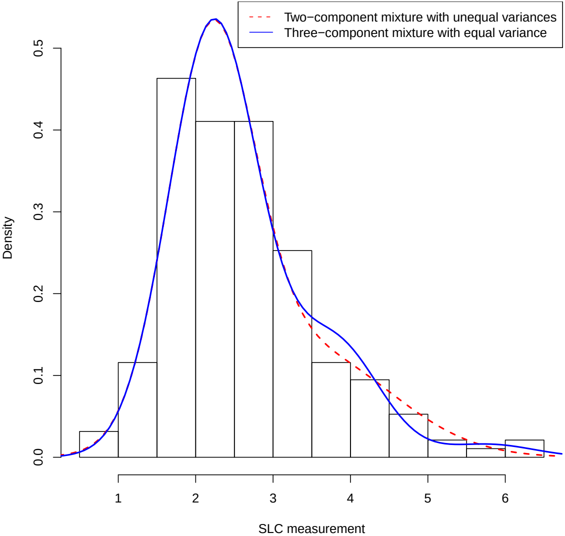
Roeder's conclusion

- Roeder (1994) uses diagnostic tool and finds a 3-component model is favoured.
- The diagnostic tool requires equal-component-variance assumption which is unfortunate.

  A formal test can be easily deviced to show that the equal-variance assumption is not plausible.
- Her conclusion can be read as: if component variances must be equal, then one needs a 3-component model to describe the data properly.
- We believe that the EM-test is superior when applied to this and many other real data examples.

SLC data again

Figure: SLC and 2/3-component normal mixture models again.



Key references

- Hartigan, J. A. (1985) A failure of likelihood asymptotics for normal mixtures, in *Proc. Berkeley Conf. in Honor of J. Neyman and Kiefer, Volume 2*, eds L. LeCam and R. A. Olshen, 807-810.
- Chernoff, H. and Lander, E. (1995) Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, **43**, 19-40.
- Chen, H., Chen, J. and Kalbfleisch, J.D. (2001). "A modified likelihood ratio test for homogeneity in finite mixture models". *Journal of the Royal Statistical Society, B.*, **63**, 19-29.
- Chen, H., Chen, J., and Kalbfleisch, J. D. (2004) Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society, Series B*, **66**, 95-115.

Key references

- Liu, X. and Shao, Y. (2004) Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, **123**, 61-81.
- Chen, J. and Li, P. (2009) Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*. **37**, 2523-2542.
- Li, P., Chen, J., and Marriott, P. (2009) Non-finite Fisher information and homogeneity: The EM approach. *Biometrika*, **96**, 411-426.
- Li, P. and Chen, J. (2010) "Testing the order of a finite mixture". *the Journal of American Statistical Association*. **105**, 1084-1092

Thank you

Questions are welcome