An Approximate Fisher Scoring Algorithm for Finite Mixtures of Multinomials

Andrew M. Raim

Department of Mathematics and Statistics
University of Maryland, Baltimore County
Baltimore, MD, USA

6th Annual Probability and Statistics Day at UMBC
Spring 2012

Joint work with Nagaraj K. Neerchal (UMBC), Minglei Liu (Medtronic), Jorge G. Morel (Procter & Gamble)

Background

- Morel and Neerchal (1991, 1993, 1998, 2005) studied estimation in their multinomial model for overdispersion: "Random Clumped Multinomial".
- They obtained a large cluster approximation to the Fisher Information Matrix (FIM), and used it to formulate an Approximate Fisher Scoring Algorithm (AFSA).
- Liu (2005, PhD Thesis) extended the idea to general mixtures of multinomials, and found some interesting connections between AFSA and Expectation Maximization (EM).
- This work extends Liu (2005), further investigating the quality of the FIM approximation and the connection between AFSA and EM.

Mixture of Multinomials Example

Example: Housing satisfaction survey

Non-metropolitan area				Metropolitan area			
Neighborhood	US	S	VS	Neighborhood	US	S	VS
1	3	2	0	19	0	4	1
2	3	2	0	20	0	5	1
3	0	5	0	21	0	3	2
⋮				⋮			
17	4	1	0	35	4	1	0
18	5	0	0				

With labels, a reasonable likelihood is product of two multinomials

$$L(\boldsymbol{\theta}) = \left[\prod_{i=1}^{18} f(\mathbf{x}_i \mid \mathbf{p}_1, m)\right] \left[\prod_{i=19}^{35} f(\mathbf{x}_i \mid \mathbf{p}_2, m)\right], \qquad m = 5.$$

J. R. Wilson, Chi-Square Tests for Overdispersion with Multiparameter Estimates. Journal of the Royal Statistical Society (Series C), 38(3):441–453, 1989.

Mixture of Multinomials Example

Example: Housing satisfaction survey

???				???			
Neighborhood	US	S	VS	Neighborhood	US	S	VS
1	3	2	0	19	0	4	1
2	3	2	0	20	0	5	1
3	0	5	0	21	0	3	2
⋮				⋮			
17	4	1	0	35	4	1	0
18	5	0	0				

Without labels, a reasonable likelihood is mixture of two multinomials

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{35} \left\{ \pi f(\mathbf{x}_i \mid \mathbf{p}_1, m) + (1-\pi) f(\mathbf{x}_i \mid \mathbf{p}_2, m) \right\}, \qquad m = 5.$$

J. R. Wilson, Chi-Square Tests for Overdispersion with Multiparameter Estimates. Journal of the Royal Statistical Society (Series C), 38(3):441–453, 1989.

Mixture of Multinomials

- Suppose we have s multinomial populations

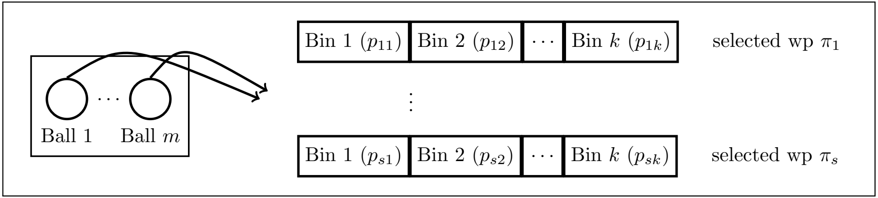
$$f(\mathbf{x} \mid \mathbf{p}_\ell, m) = \frac{m!}{x_1! \dots x_k!} p_{\ell 1}^{x_1} \dots p_{\ell k}^{x_k} \cdot I(\mathbf{x} \in \Omega), \qquad \ell = 1, \dots, s$$

  which occur in the total population with probabilities $\pi_1, \dots, \pi_s$.

- If we draw $\mathbf{T}$ from the mixed population,

$$\mathbf{T} \sim f(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell f(\mathbf{x} \mid \mathbf{p}_\ell, m), \qquad \boldsymbol{\theta} = (\mathbf{p}_1, \dots, \mathbf{p}_s, \boldsymbol{\pi})$$

  We'll write $\mathbf{T} \sim \text{MultMix}_k(\boldsymbol{\theta}, m)$.



Estimation Problem

- Suppose our sample is $\mathbf{X}_i \stackrel{\text{ind}}{\sim} \text{MultMix}_k(\boldsymbol{\theta}, m_i), \qquad i = 1, \ldots, n$

- Likelihood

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{\ell=1}^{s} \pi_\ell \left[\frac{m_i!}{x_{i1}! \ldots x_{ik}!} p_{\ell 1}^{x_{i1}} \ldots p_{\ell k}^{x_{ik}} \cdot I(\mathbf{x}_i \in \Omega) \right] \right\}$$

- To find MLE $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_s, \hat{\boldsymbol{\pi}})$, which maximizes the (log) likelihood

- Some options
  - No nice closed form
  - Newton-Raphson, **Fisher Scoring**, Quasi-Newton methods

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} - \alpha \mathbf{H}^{-1} S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots$$

  - Expectation Maximization (EM)

Score: $S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})$

FIM: $\mathcal{I}(\boldsymbol{\theta}) = \text{E}\left\{ -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta}) \right\}$

Fisher Scoring Algorithm

- The iterations become

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)}) S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots,$$

  but $\mathcal{I}(\boldsymbol{\theta})$ may not be easy to compute.

- Naive summation works when sample space Ω is small

$$\mathcal{I}(\boldsymbol{\theta}) := \sum_{\mathbf{x} \in \Omega} \left\{ -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\mathbf{x} \mid \boldsymbol{\theta}) \right\} f(\mathbf{x} \mid \boldsymbol{\theta}).$$

- Monte Carlo approximation

- For large clusters ($m \uparrow$), Morel & Nagaraj (1991) and Liu (2005, PhD thesis) propose an approximation (shown for $\mathbf{X}_1 \sim \text{MultMix}_k(\boldsymbol{\theta}, m)$)

$$\begin{aligned} \widetilde{\mathcal{I}}(\boldsymbol{\theta}) &:= \text{Blockdiag}\left(\pi_1 \mathbf{F}_1, \ldots, \pi_s \mathbf{F}_s, \mathbf{F}_\pi\right), \\ \mathbf{F}_\ell &= m \left[\text{Diag}(p_{\ell 1}^{-1}, \ldots, p_{\ell, k-1}^{-1}) + p_{\ell k}^{-1} \mathbf{1}\mathbf{1}^T \right] \\ \mathbf{F}_\pi &= \text{Diag}(\pi_\ell^{-1}, \ldots, \pi_{s-1}^{-1}) + \pi_s^{-1} \mathbf{1}\mathbf{1}^T \end{aligned}$$

- Result: $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$.

Approximate FIM Properties I

- $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is a block diagonal matrix of Multinomial FIMs.
  - Simple forms for inverse, trace, and determinant

- Result: $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is "complete data" FIM of $(\mathbf{X}, Z)$

$$Z = \begin{cases} 1 & \text{wp } \pi_1 \\ & \vdots \\ s & \text{wp } \pi_s, \end{cases} \quad \text{and} \quad (\mathbf{X} \mid Z = \ell) \sim \text{Mult}_k(\mathbf{p}_\ell, m).$$

So that we have

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{E}\left\{ -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\mathbf{x}, z \mid \boldsymbol{\theta}) \right\}$$

- Note that EM is based on maximizing

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \text{E}_{\boldsymbol{\theta}'}\left[\log f(\mathbf{x}, z \mid \boldsymbol{\theta}) \mid \mathbf{x} \right].$$

Approximate FIM Properties II

- Can also show that the inverses converge

$$\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \to \mathbf{0} \quad \text{as } m \to \infty.$$

- For any non-singular $\mathbf{A}, \mathbf{B}$, and sub-multiplicative matrix norm

$$\begin{aligned} \mathbf{B}^{-1} - \mathbf{A}^{-1} &= \mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\mathbf{B}^{-1} \\ &\implies \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| \leq \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A} - \mathbf{B}\| \cdot \|\mathbf{B}^{-1}\|. \end{aligned}$$

- Taking $\mathbf{A} = \mathcal{I}(\boldsymbol{\theta})$ and $\mathbf{B} = \widetilde{\mathcal{I}}(\boldsymbol{\theta})$

$$\|\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\| \leq \|\mathcal{I}^{-1}(\boldsymbol{\theta})\| \cdot \|\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})\| \cdot \|\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\|$$

  which can be shown to converge to 0.

- $\mathcal{I}(\boldsymbol{\theta})$ may be singular if identifiability fails to hold on the model.
  - See Rothenberg (1971) about the connection.

Approximate FIM Properties III

- Large cluster size (m) needed for

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) \approx \mathcal{I}(\boldsymbol{\theta}) \quad \text{and} \quad \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \approx \mathcal{I}^{-1}(\boldsymbol{\theta})$$

  (with inverses apparently converging faster).

- Approximate FIM and inverse are not recommended for general inference.

- But useful as a tool for estimation, as we will see.

Approximate Fisher Scoring Algorithm

- Using the approximate FIM in place of the true FIM gives AFSA

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}^{(g)}) S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots$$

until $\left| \log L(\boldsymbol{\theta}^{(g+1)}) - \log L(\boldsymbol{\theta}^{(g)}) \right| < \varepsilon$.

- Liu (2005, PhD Thesis) derives explicit iterations for each parameter in $\boldsymbol{\theta}$ for both EM and AFSA.

- Under $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} \text{MultMix}_k(\boldsymbol{\theta}, m)$, EM and AFSA iterations are "equivalent", given the same starting place $\boldsymbol{\theta}^{(g)}$

$$\tilde{\pi}_\ell^{(g+1)} = \hat{\pi}_\ell^{(g+1)}, \qquad \tilde{p}_{\ell j}^{(g+1)} = \left(\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} \right) \hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} \right) p_{\ell j}^{(g)}.$$

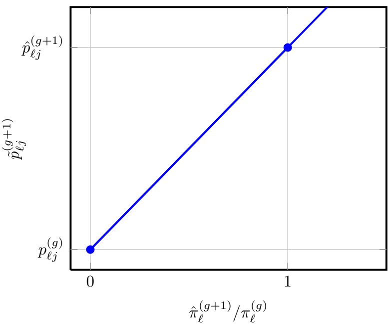
- Doesn't hold under the "independent but not iid" case.

Equivalence of AFSA and EM

AFSA steps are linear combinations of the next EM step and the previous iterate

$$\tilde{\pi}_\ell^{(g+1)} = \hat{\pi}_\ell^{(g+1)}, \qquad \tilde{p}_{\ell j}^{(g+1)} = \left(\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} \right) \hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} \right) p_{\ell j}^{(g)}.$$

AFSA step compared to previous iterate and EM step



When EM is close to convergence, we will have $\tilde{p}_{\ell j}^{(g+1)} \approx \hat{p}_{\ell j}^{(g+1)}$.

Equivalence of AFSA and EM II

- A more general connection is known between EM and iterations of the form
$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}_c^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \qquad g = 1, 2, \ldots.$$

- Titterington (1984) shows the two are approximately equivalent (under regularity conditions)

- And the equivalence is exact when the complete data likelihood is a regular exponential family
$$L(\boldsymbol{\mu}) = \exp\left\{ b(\mathbf{x}) + \boldsymbol{\eta}^T \mathbf{t} + a(\boldsymbol{\eta}) \right\},$$
$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu}) : \text{natural parameter},$$
$$\mathbf{t} = \mathbf{t}(\mathbf{x}) : \text{sufficient statistic},$$
$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{t}(\mathbf{X})) : \text{the parameter of interest}.$$

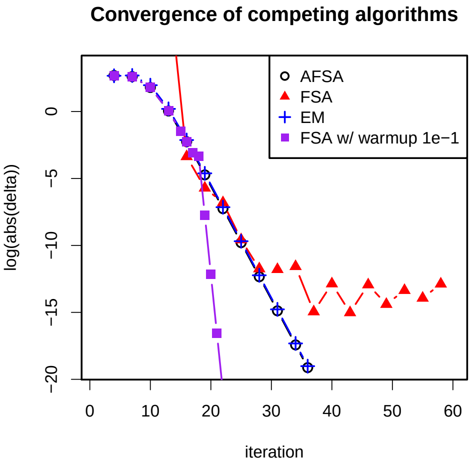
- For MultMix problem, equivalance is approximate not exact.
  - Justification for AFSA originally came from $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and Blischke (1964).
  - But this result justifies AFSA for finite mixtures other than multinomial.

Comparison between algorithms

Consider the mixture of two trinomials

$$\mathbf{X}_i \overset{\text{iid}}{\sim} \text{MultMix}_3(\boldsymbol{\theta}, m = 20), \qquad i = 1, \ldots, n = 500$$

$$\begin{pmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}, \qquad \begin{pmatrix} \pi \\ 1-\pi \end{pmatrix} = \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}.$$



method	ε_0	tol	iter
AFSA	—	4.94×10^{-09}	36
EM	—	5.50×10^{-09}	36
FSA	∞	-1.26×10^{-07}	100
FSA	10	4.46×10^{-10}	16

Monte Carlo Comparison of EM and AFSA

Consider a scenario with varying cluster sizes

$$\mathbf{Y}_i \overset{\text{ind}}{\sim} \text{MultMix}_k(\boldsymbol{\theta}, m_i), \qquad i = 1, \ldots, n = 500, \qquad \boldsymbol{\pi} = (0.75, 0.25)$$
$$W_1, \ldots, W_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta), \qquad m_i = \lceil W_i \rceil.$$

Ran 1000 reps of nine scenarios and looked at the quantity

$$\frac{1}{1000} \sum_{r=1}^{1000} \left\{ \bigvee_{j=1}^{q} \left| \frac{\tilde{\theta}_j^{(r)} - \hat{\theta}_j^{(r)}}{\tilde{\theta}_j^{(r)}} \right| \right\}.$$

(kth probability not shown) $\mathbf{p}_1$	$\mathbf{p}_2$	m_i equal $m_i = 20$	$\alpha = 100$ $\text{Var}(m_i) \approx 4.083$	$\alpha = 25$ $\text{Var}(m_i) \approx 16.083$
(0.1)	(0.5)	2.178×10^{-6}	2.019×10^{-6}	2.080×10^{-6}
(0.3)	(0.5)	4.073×10^{-5}	3.501×10^{-5}	3.890×10^{-5}
(0.35)	(0.5)	8.683×10^{-4}	2.625×10^{-4}	2.738×10^{-4}
(0.4)	(0.5)	9.954×10^{-3}	6.206×10^{-2}	6.563×10^{-2}
$(0.1, 0.3)$	$(1/3, 1/3)$	1.342×10^{-3}	1.009×10^{-3}	1.878×10^{-3}
$(0.1, 0.5)$	$(1/3, 1/3)$	1.408×10^{-6}	1.338×10^{-6}	1.334×10^{-6}
$(0.3, 0.5)$	$(1/3, 1/3)$	3.884×10^{-6}	3.943×10^{-6}	3.885×10^{-6}
$(0.1, 0.1, 0.3)$	$(0.25, 0.25, 0.25)$	8.389×10^{-7}	8.251×10^{-7}	8.440×10^{-7}
$(0.1, 0.2, 0.3)$	$(0.25, 0.25, 0.25)$	1.523×10^{-6}	1.472×10^{-6}	1.408×10^{-6}

Conclusions

AFSA is obtained as a Newton-type algorithm using an approximate FIM.

- Nearly equivalent to EM iterations — similar solutions are obtained at similar rates of convergence
- (EM advantange) M-step can be formulated so it won't wander outside parameter space.
- (AFSA advantange) May be easier to formulate when missing data structure is complicated.
  E.g. Random-Clumped Multinomial (Morel & Neerchal 1993).

Result of Titterington (1984) suggests AFSA approach is reasonable for finite mixtures in general.

Both EM and AFSA suffer from a slow convergence rate.

- Hybrid is recommended for fast convergence and robustness.
- . . . if true FIM is feasible to compute.

References I

[1] W. R. Blischke. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59(306):510–528, 1964.

[2] M. Liu. *Estimation for Finite Mixture Multinomial Models*. Phd thesis, University of Maryland, Baltimore County, Department of Mathematics and Statistics, 2005.

[3] J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modeling multinomial extra variation. Technical Report Research report 91–03, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1991.

[4] J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.

[5] N. K. Neerchal and J. G. Morel. Large cluster results for two parametric multinomial extra variation models. *Journal of the American Statistical Association*, 93(443):1078–1087, 1998.

References II

[6] N. K. Neerchal and J. G. Morel. An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis*, 49(1):33–43, 2005.

[7] T. J. Rothenberg. Identification in parametric models. *Econometrica*, 39:577–591, 1971.

[8] D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B*, 46:257–267, 1984.

Acknowledgement: The computational resources used for this work were provided by the High Performance Computing Facility at the University of Maryland, Baltimore County (UMBC). See `www.umbc.edu/hpcf` for information on the facility and its uses. Andrew additionally thanks the facility for financial support as an RA.

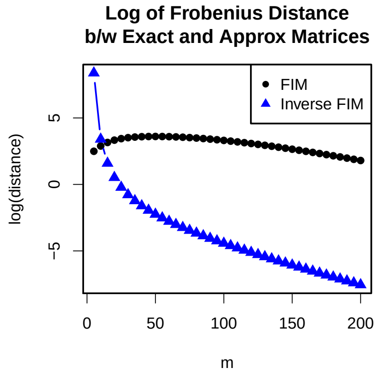
How good is the FIM approximation?

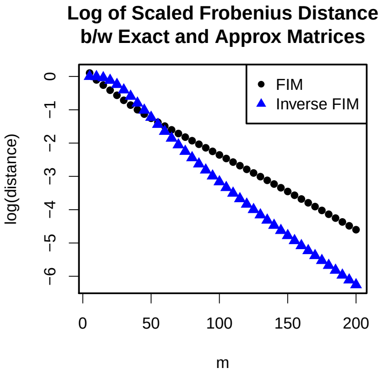
Consider a mixture $\text{MultMix}_2(\boldsymbol{\theta}, m)$ of three binomials, with parameters

$$\begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} = \begin{pmatrix} 1/7 & 1/3 & 2/3 \end{pmatrix}, \qquad \boldsymbol{\pi} = \begin{pmatrix} 1/6 & 2/6 & 3/6 \end{pmatrix},$$

and two matrix distances

$$d(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_\text{F} \qquad\qquad d(\mathbf{A}, \mathbf{B}) = \frac{\|\mathbf{A} - \mathbf{B}\|_\text{F}}{\|\mathbf{B}\|_\text{F}}$$





Large m is needed for a good approximation. Inverses are converging faster.