# The Approximate Fisher Information Matrix for Multinomial Mixture Models

Andrew M. Raim and Nagaraj K. Neerchal
Department of Mathematics and Statistics,
University of Maryland, Baltimore County

UMBC
AN HONORS UNIVERSITY IN MARYLAND

## Summary


- Mixture distributions are useful in many problems, but can also be difficult to work with
- Minglei Liu (2005, PhD Thesis) studied estimation in multinomial mixture models. Related work has been done by Morel & Neerchal (1993, 1998, 2005)
- Here we present one of the key results, a large cluster approximation to the FIM
- The FIM approximation for the general multinomial mixture was shown to be useful in the Fisher Scoring algorithm. Here we consider its direct usage in inference


## Mixture multinomial model

- Suppose we have $s$ multinomial populations

$$f(\boldsymbol{x} \mid \boldsymbol{p}_1, m), \ldots, f(\boldsymbol{x} \mid \boldsymbol{p}_s, m), \qquad \boldsymbol{p}_\ell = (p_{\ell 1}, \ldots, p_{\ell k})$$

- If population $\ell$ occurs with proportion $\pi_\ell$, and we draw $X$ from the mixed population

$$\boldsymbol{X} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\ell=1}^{s} \pi_\ell f(\boldsymbol{x} \mid \boldsymbol{p}_\ell, m), \qquad \boldsymbol{x} \in \mathcal{X}, \qquad \boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s, \boldsymbol{\pi})$$

- Mixture distributions are a natural way to deal with mixed populations. A housing satisfaction survey from J. R. Wilson (1989) is an example featuring multinomials

| Non-metropolitan area | | | | Metropolitan area | | | |
|---|---|---|---|---|---|---|---|
| Neighborhood | US | S | VS | Neighborhood | US | S | VS |
| 1 | 3 | 2 | 0 | 1 | 0 | 4 | 1 |
| 2 | 3 | 2 | 0 | 2 | 0 | 5 | 1 |
| 3 | 0 | 5 | 0 | 3 | 0 | 3 | 2 |
| | | | | ... | | | |
| 17 | 4 | 1 | 0 | 17 | 4 | 1 | 0 |
| 18 | 5 | 0 | 0 | | | | |

- Some related problems (✓ means mixtures are often applied here)
  - Classification: Given samples $(\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_{n_\ell}^{(\ell)})$ from each population $\ell = 1, \ldots, s$, classify a new observation $\boldsymbol{x}$
  - Discriminant Analysis: Given samples $(\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_{n_\ell}^{(\ell)})$, find a rule to best distinguish between the $s$ groups
  - ✓ Clustering: Given a sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ from the mixed population, try to determine which observations belong to the same groups. If $s$ is not known the problem is harder
  - ✓ Modeling overdispersion: In usual inference problems such as point estimation, confidence intervals, and hypothesis testing, model a mixed population with a mixture model of simpler distributions to capture the differences between groups

## References


[1] M. Liu, *Estimation for Finite Mixture Multinomial Models*, PhD Thesis, University of Maryland, Baltimore County, Department of Mathematics and Statistics, 2005.

[2] J.G. Morel and N.K. Nagaraj, *A Finite Mixture Distribution for Modelling Multinomial Extra Variation*, Biometrika 80 (1993), pp. 363–371.

[3] N.K. Neerchal and J.G. Morel, *Large Cluster Results for Two Parametric Multinomial Extra Variation Models*, Journal of the American Statistical Association 93 (1998), pp. 1078–1087.

[4] 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 (2005), pp. 33–43.


The computational resources used for this work were provided by the UMBC 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.

## FIM Approximation

- The Fisher Information Matrix (FIM, "outer product" form)

$$\mathcal{I}(\boldsymbol{\theta}) := \mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X)\right)^T\right]$$

  is extremely useful for inference and model selection
- But a closed form expression cannot be obtained for most mixture models
- For the mixture of multinomials, the expectation can be computed exactly by summing over the sample space (which is finite but grows quickly with $k$ and $m$)

$$\mathcal{I}(\boldsymbol{\theta}) = n \sum_{\boldsymbol{x} \in \mathcal{X}} \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right\}^T f_{\boldsymbol{\theta}}(\boldsymbol{x})$$

- Liu and Morel & Nagaraj found an approximation for $\mathcal{I}(\boldsymbol{\theta})$

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) := \begin{pmatrix} \pi_1 \boldsymbol{F}_1 & & & 0 \\ & \ddots & & \\ & & \pi_s \boldsymbol{F}_s & \\ 0 & & & \boldsymbol{F}_\pi \end{pmatrix} \qquad (sk-1) \times (sk-1)$$

$$\boldsymbol{F}_\ell = m\left[\mathrm{diag}(p_{\ell 1}^{-1}, \ldots, p_{\ell, k-1}^{-1}) - p_{\ell k}^{-1} \boldsymbol{1}\boldsymbol{1}^T\right] \qquad (k-1) \times (k-1)$$

$$\boldsymbol{F}_\pi = \mathrm{diag}(\pi_1^{-1}, \ldots, \pi_{s-1}^{-1}) - \pi_s^{-1} \boldsymbol{1}\boldsymbol{1}^T \qquad (s-1) \times (s-1)$$

- This has a simple closed form that requires little computation to construct. There are also simple forms for the inverse approximate FIM and determinant
- Has been shown that $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta}) \to 0$ as $m \to \infty$, for multinomial mixtures
- The approximation can also be used for more complicated mixtures such as Random-Clumped Multinomial and Dirichlet Multinomial. These multinomial mixtures feature parameters with functional dependencies on each other. See Neerchal & Morel (1998)
- In some cases the approximation was shown to work very well even for moderate $m$

## Relationship to complete data FIM

- $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ turns out to be equivalent to the complete data FIM obtained by considering latent class variables
- This technique is also used in Expectation Maximization (EM)
- Suppose we observe an iid sample $(\boldsymbol{X}_i, Z_i)$, $i = 1, \ldots n$, where

$$Z_i = \begin{cases} 1 & \text{wp } \pi_1 \\ \ldots & \\ s & \text{wp } \pi_s, \end{cases} \qquad \boldsymbol{X}_i \mid Z_i = \ell \quad \sim \mathrm{Mult}(\boldsymbol{p}_\ell, m)$$

- Then the complete data likelihood is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \prod_{\ell=1}^{s} \left[\pi_\ell f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell, m)\right]^{I(z_i = \ell)}$$

- Computing the FIM ("Hessian" form) with respect to this likelihood, we obtain that

$$\mathrm{E}\left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\right] \equiv \widetilde{\mathcal{I}}(\boldsymbol{\theta})$$

- The classes $(Z_1, \ldots, Z_n)$ aren't observable, but they are only used inside the expectation hence we don't need to observe them

## Simulation Study

The Wald statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$T_n(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \qquad T_n(\hat{\boldsymbol{\theta}}) \xrightarrow{\mathcal{L}} \chi_q^2, \qquad q = sk - 1.$$

This $T$ can be used to construct an approximate $1 - \alpha$ level Wald-type confidence region

$$R(\hat{\boldsymbol{\theta}}) = \left\{\boldsymbol{\theta}_0 : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \le \chi_{q,\alpha}^2\right\},$$

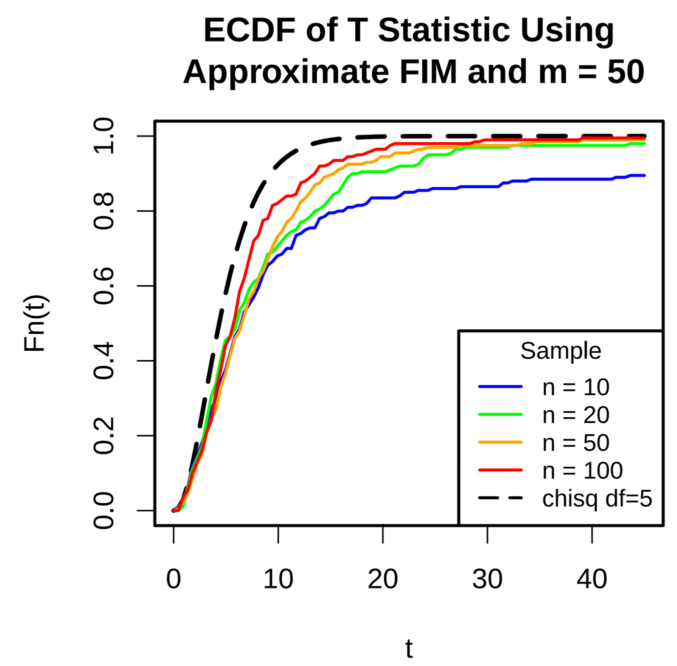
an ellipsoid in $\mathbb{R}^q$ centered at the MLE $\hat{\boldsymbol{\theta}}$, with shape determined by the FIM. Consider replacing $\mathcal{I}(\hat{\boldsymbol{\theta}})$ with the approximate FIM, which is much easier to compute

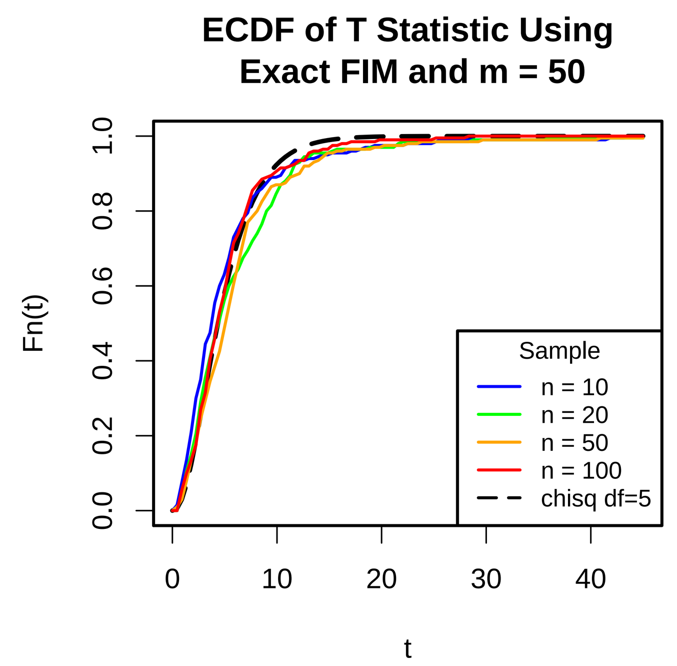
$$\widetilde{T}_n(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

We compare $T$ and $\widetilde{T}$ with a simulation, choosing the parameters $\boldsymbol{\theta}$ as

$$(\boldsymbol{p}_1 \quad \boldsymbol{p}_2 \quad \boldsymbol{p}_3) \propto \begin{pmatrix} 1 & 1 & 2 \\ 6 & 2 & 1 \end{pmatrix}, \qquad \pi \propto \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Samples were drawn from this binomial mixture 200 times for several $m$ and $n$. For each sample we compute $T$ and $\widetilde{T}$, obtaining their empirical distributions under $H_0$.

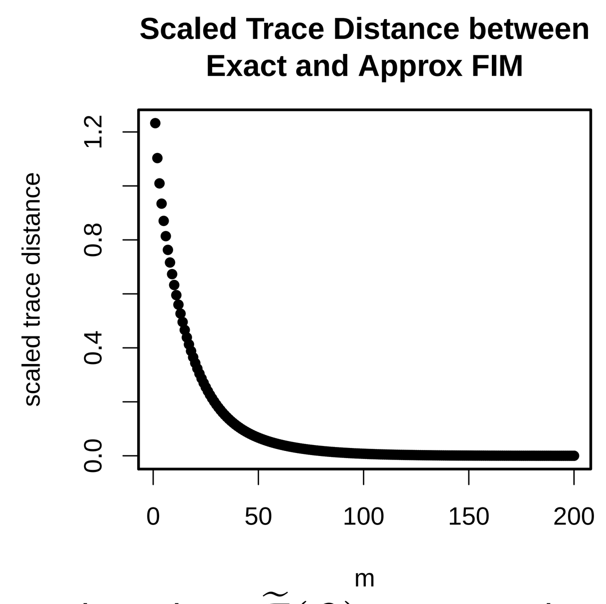


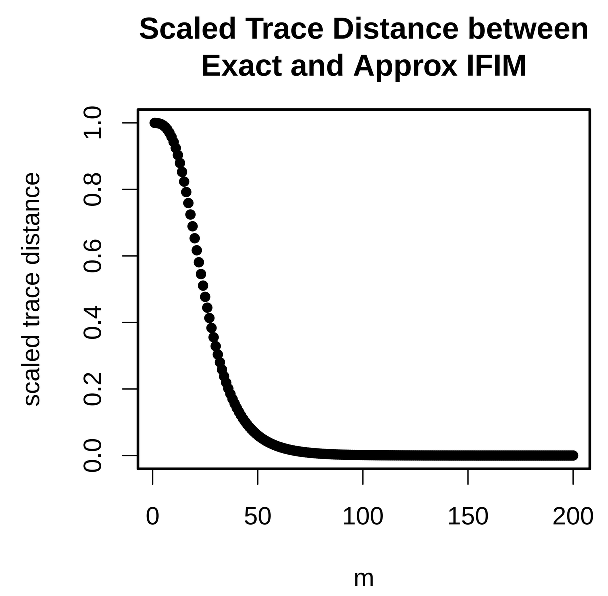




$\widetilde{T}$ seems to be lagging behind $T$ in terms of the large sample $\chi_q^2$, even for $m = 50$ which may be considered a fairly large cluster size. To see why this is happening, we compare the two FIMs directly. Consider the following criteria based on the trace distance

$$d(A, B) = \frac{\mathrm{tr}(A - B)^T (A - B)}{\mathrm{tr}\, B^T B} = \frac{\sum_i \sum_j (a_{ij} - b_{ij})^2}{\sum_i \sum_j b_{ij}^2}.$$

For $\boldsymbol{\theta}$ given above, we compute distances for varying $m$. The left plot shows $d\left(\widetilde{\mathcal{I}}(\boldsymbol{\theta}), \mathcal{I}(\boldsymbol{\theta})\right)$ and the right plot shows $d\left(\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}), \mathcal{I}^{-1}(\boldsymbol{\theta})\right)$ which corresponds to the asymptotic covariance matrix.




- Inference based on $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ may not be correct for small to moderate $m$ in the general mixture considered here
- Interesting that $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ works well in estimation procedures like Fisher Scoring, even when $m$ is not large. But it may be too far from $\mathcal{I}(\boldsymbol{\theta})$ to work well in inference
- $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ may continue to be useful as a computational aid. Consider the approach of Neerchal and Morel (2005), where the approximation is used in Fisher Scoring iterations until convergence, and then one additional iteration is performed with the exact FIM to produce a final result