An Analysis of Categorical Injury Data using Mixtures of Multinomials

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

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Summary

- We present an analysis of injury data which was first explored by Fleming (2012). Fleming's analysis focuses on injury counts of individuals, and the issue of estimating the (unobserved) number of individuals having zero counts using truncated Poisson.
- Counts of several common injury types are analyzed in a multinomial setting. Finite mixture of multinomial models are considered to address heterogeneity in the data.
- Computational techniques from Raim et al. (2013) are used to determine the number of mixing components, obtain estimates, and compute standard errors and confidence intervals. We find that three latent classes provides an adequate model.

Injury Dataset

- Data consists of injuries reported to a national database maintained by an ambulance service company.
- Records are associated with emergency service providers such as EMTs, paramedics, and firefighters, along with adjunct workers such as administrators.
- 6,691 total injuries in 4,623 unique people. Individuals are grouped into 450 distinct ambulance units.
- 600 different occupations are listed, many with small subtle differences between them e.g. EMT, PARAMEDIC, EMT FIREFIGHTER, FIREFIGHTER
- 55 types of injuries are reported, from nausea, to seizure, to death. We focus on the $k = 10$ most common: strain, contusion, sprain, puncture, laceration, torn cartilage/ligament/tendon (C/L/T), fracture, inflammation, respiratory, and other

$$\boldsymbol{T}_i = \left.\begin{pmatrix} T_{i1} & \leftarrow \#\text{ strains} \\ T_{i2} & \leftarrow \#\text{ contusions} \\ \vdots & \\ T_{ik} & \leftarrow \#\text{ other injuries} \end{pmatrix}\right\} \text{out of } m_i \text{ injuries for the } i\text{th ambulance unit for } i = 1, \ldots, n = 450.$$

Finite Mixture of Multinomials

- Injuries in same ambulance unit may have similar conditions: e.g. weather and crime.
- But heterogeneity is expected
  - ✓ Between different occupations — e.g firefighter vs. office worker
  - ✓ Between individuals — carefulness, pre-existing medical conditions, etc.
- Therefore, consider finite mixture for analysis: $\boldsymbol{T}_i \overset{\text{ind}}{\sim} \text{MultMix}_k(\boldsymbol{\theta}, m_i)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{\ell=1}^{s} \pi_\ell \left[\frac{m_i!}{t_{i1}! \ldots t_{ik}!} p_{\ell 1}^{t_{i1}} \ldots p_{\ell k}^{t_{ik}} \right] \right\}, \text{ with } \boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s, \boldsymbol{\pi})$$

- Several covariates are available, such as gender and amount of lost wages due to injury, but these are not used in the analysis.

References

Brandon E. Fleming. Estimating risk of occupational injury in the presence of unreported zeros. Master's thesis, University of Maryland, Baltimore County, 2012.

Andrew M. Raim, Brandon E. Fleming, and Nagaraj K. Neerchal. An analysis of categorical injury data using mixtures of multinomials. In *JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association*, pages 2444–2458, 2012.

Andrew M. Raim, Minglei Liu, Nagaraj K. Neerchal, and Jorge G. Morel. On the method of approximate fisher scoring for finite mixtures of multinomials, 2013. (Submitted).

See `www.umbc.edu/~araim1/bib` for preprint of 2nd paper and tech report version of 3rd paper.

Approximate Information Matrix and Scoring

- As usual under finite mixture distributions, closed form expressions for the MLE $\hat{\boldsymbol{\theta}}$ are not available, and iterative techniques must be used to compute estimates.
- A standard iterative estimation method is Fisher scoring

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

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \quad \text{is score vector wrt the sample}$$

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left[S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^T \right] \quad \text{is Fisher information matrix (FIM) wrt the sample}$$

- Let $\mathcal{I}_m(\boldsymbol{\theta})$ be FIM for $\text{MultMix}_k(\boldsymbol{\theta}, m)$. Then $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{m_1}(\boldsymbol{\theta}) + \cdots + \mathcal{I}_{m_n}(\boldsymbol{\theta})$.
- Simple expressions for $\mathcal{I}_m(\boldsymbol{\theta})$ are also not available. We can use definition of expectation

$$\mathcal{I}_m(\boldsymbol{\theta}) = \sum_{\boldsymbol{t} \in \Omega} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{t}; \boldsymbol{\theta}, m) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{t}; \boldsymbol{\theta}, m) \right\}^T f(\boldsymbol{t}; \boldsymbol{\theta}, m) \quad (*)$$

but number of terms $\binom{m+k-1}{m}$ grows quickly with m (or k).

- Raim et al. (2013) justify the following matrix as a large cluster approximation (as $m \to \infty$) to $\mathcal{I}_m(\boldsymbol{\theta})$, as well as its use in Fisher scoring iterations ("approximate Fisher scoring")

$$\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) = \text{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right), \text{ where}$$

$$\boldsymbol{F}_\ell = m \left[\text{diag}(p_{\ell 1}^{-1}, \ldots, p_{\ell, k-1}^{-1}) + p_{\ell k}^{-1} \mathbf{1}\mathbf{1}^T \right], \quad \ell = 1, \ldots, s,$$

$$\boldsymbol{F}_\pi = \text{diag}(\pi_1^{-1}, \ldots, \pi_{s-1}^{-1}) + \pi_s^{-1} \mathbf{1}\mathbf{1}^T.$$

- Can be shown that $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$ and $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$, and that approximate Fisher scoring is "close" to Expectation-Maximization.

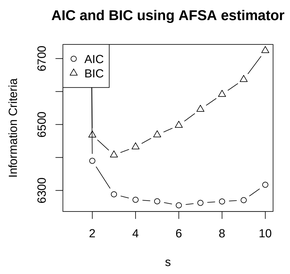
Model Selection

- To select the # of mixture components s supported by the data, consider the information criteria

$$\text{AIC} = -2 \log L(\tilde{\boldsymbol{\theta}}) + 2q \quad \text{and} \quad \text{BIC} = -2 \log L(\tilde{\boldsymbol{\theta}}) + q \log n.$$

where $q = sk - 1$ is the total number of parameters.

- The approximate Fisher scoring estimator $\tilde{\boldsymbol{\theta}}$ is used because exact Fisher scoring (using $(*)$) is intractible for this data.



s	AIC	BIC
1	13977.69	14014.68
2	6389.88	6467.96
3	6288.35	6407.52
4	6271.77	6432.03
5	6267.17	6468.53
6	6254.95	6497.40
7	6262.40	6545.94
8	6266.63	6591.26
9	6270.66	6636.39
10	6317.28	6724.10

- $s = 3$ is selected because it provides a reasonably simple model, and at least minimizes one of the criteria (BIC)

Improvement to FIM Approximation

- Standard errors can be obtained from approximate Fisher scoring using the diagonal elements of $\widetilde{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\theta}})$. But it can be shown that they are systematically too small (i.e. too optimistic).
- It is natural to consider an improved approximation using the exact FIM for smaller m_i and the approximate FIM for larger m_i. Define the "hybrid approximate FIM" as

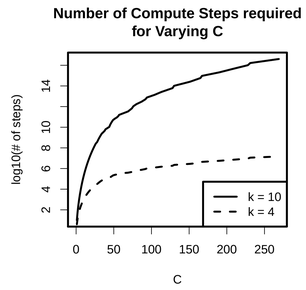
$$\mathcal{I}^*(\boldsymbol{\theta}, C) = \sum_{i: m_i \le C} \mathcal{I}_{m_i}(\boldsymbol{\theta}) + \sum_{i: m_i > C} \widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta}).$$

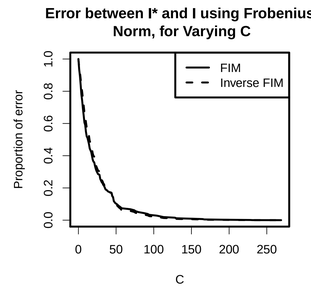
- We would like to find $C \ge 0$ to yield small values for

$$N_{C,k} = \sum_{i: r_i > C} \binom{r_i + k - 1}{r_i}, \quad \text{the \# of terms in the summation } (*)$$

$$p_C = \frac{\|\mathcal{I}^*(\tilde{\boldsymbol{\theta}}, C) - \mathcal{I}(\tilde{\boldsymbol{\theta}})\|}{\|\widetilde{\mathcal{I}}(\tilde{\boldsymbol{\theta}}) - \mathcal{I}(\tilde{\boldsymbol{\theta}})\|} \quad \text{and} \quad q_C = \frac{\|\mathcal{I}^{*-1}(\tilde{\boldsymbol{\theta}}, C) - \mathcal{I}^{-1}(\tilde{\boldsymbol{\theta}})\|}{\|\widetilde{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\theta}}) - \mathcal{I}^{-1}(\tilde{\boldsymbol{\theta}})\|}$$

where $r_1, \ldots, r_{n^*}$ represent unique m_i.



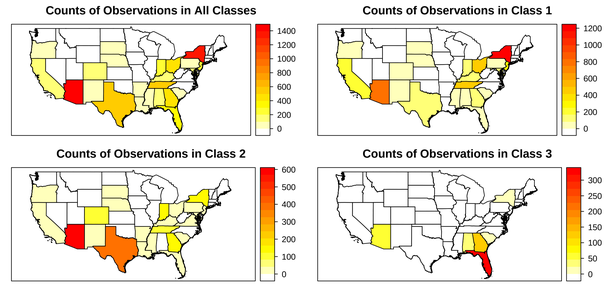


- $C = 50$ (shown above right) is a good choice using only $k = 4$ categories, but too expensive to compute for all $k = 10$ categories. We proceed with $C = 20$ for the final analysis.

Results

- Results after running one additional Fisher scoring step with $\mathcal{I}^*(\tilde{\boldsymbol{\theta}}, C = 20)$
- Class for ith observation assigned according to: $\arg\max_\ell \frac{\hat{\pi}_\ell f(\boldsymbol{t}_i; \hat{\boldsymbol{p}}_\ell, m_i)}{\sum_{a=1}^{s} \hat{\pi}_a f(\boldsymbol{t}_i; \hat{\boldsymbol{p}}_a, m_i)}$.

Estimate (Stderr)	Class 1	Class 2	Class 3
Pr[Mixing]	0.5347 (0.0412)	0.3962 (0.0399)	—
Pr[Strain]	0.4783 (0.0089)	0.2940 (0.0097)	0.3363 (0.0247)
Pr[Contusion]	0.1025 (0.0053)	0.0743 (0.0055)	0.1310 (0.0174)
Pr[Sprain]	0.0981 (0.0052)	0.0598 (0.0050)	0.0422 (0.0106)
Pr[Puncture]	0.0343 (0.0033)	0.0624 (0.0050)	0.0550 (0.0118)
Pr[Torn C/L/T]	0.0349 (0.0033)	0.0588 (0.0048)	0.0394 (0.0101)
Pr[Laceration]	0.0144 (0.0021)	0.0178 (0.0027)	0.0333 (0.0092)
Pr[Fracture]	0.0105 (0.0019)	0.0298 (0.0035)	0.0109 (0.0054)
Pr[Inflammation]	0.0115 (0.0019)	0.0128 (0.0024)	0.0500 (0.0110)
Pr[Respiratory]	0.0088 (0.0017)	0.0059 (0.0017)	0.0879 (0.0144)



- Class 1 had 47.92% females, while class 2 had 48.26% and class 3 had 30.77%.
- See Raim et al. (2012) for the complete study.