An Analysis of Categorical Injury Data using Mixtures of Multinomials

Andrew M. Raim Brandon E. Fleming Nagaraj K. Neerchal

Abstract
Finite mixture models are useful for data that exhibit heterogeneity from unobserved sources. Such models can assign observations into a set of latent classes, and may be helpful in understanding the nature of the heterogeneity. In this paper, the finite mixture of multinomials model is applied to an injury dataset in order to study the probabilities of several injury types common among emergency service providers. Computational techniques from (Raim et al., 2012) are used to determine the number of mixing components, obtain estimates, and compute standard errors and confidence intervals. We find that three classes provides an adequate model for the data, and that the class compositions differ by geography and gender.

Key Words: Multinomial; Finite mixture, Maximum likelihood, Fisher information matrix, Fisher scoring.

1. Introduction

This paper presents an analysis of injury data which was first explored in the Master's thesis of Fleming (2012). The dataset 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 secretaries, mechanics, and administrators. For the rest of this work, we will refer to all workers collectively as emergency service providers. Fleming's analysis focuses on the injury counts of individuals in the data, and the issue of estimating the number of individuals having zero counts (which are not observed). Poisson, truncated Poisson, and related models are considered in carrying out the analysis. In this work, we analyze the counts of several types of injuries commonly suffered by emergency service providers. We consider multinomial and finite mixture of multinomial models using computational techniques discussed in (Raim et al., 2012). Our interest is in the probabilities of several common types of injuries. Finite mixtures are considered to account for heterogeneity in the data.

The remainder of this paper is organized as follows. Section 2 describes the dataset. Section 3 discusses the finite mixture of multinomials model and recalls the computational methods from (Raim et al., 2012). Section 4 presents a model selection for the number of mixing components in the data; an approximation to the Fisher information matrix is used to carry out the study. In section 5, a more accurate approximation is proposed so that standard errors and confidence intervals can be obtained. The completed analysis is given in section 6 and concluding remarks are made in section 7.

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, U.S.A, Email: {araim1, bflemi1, nagaraj}@umbc.edu.

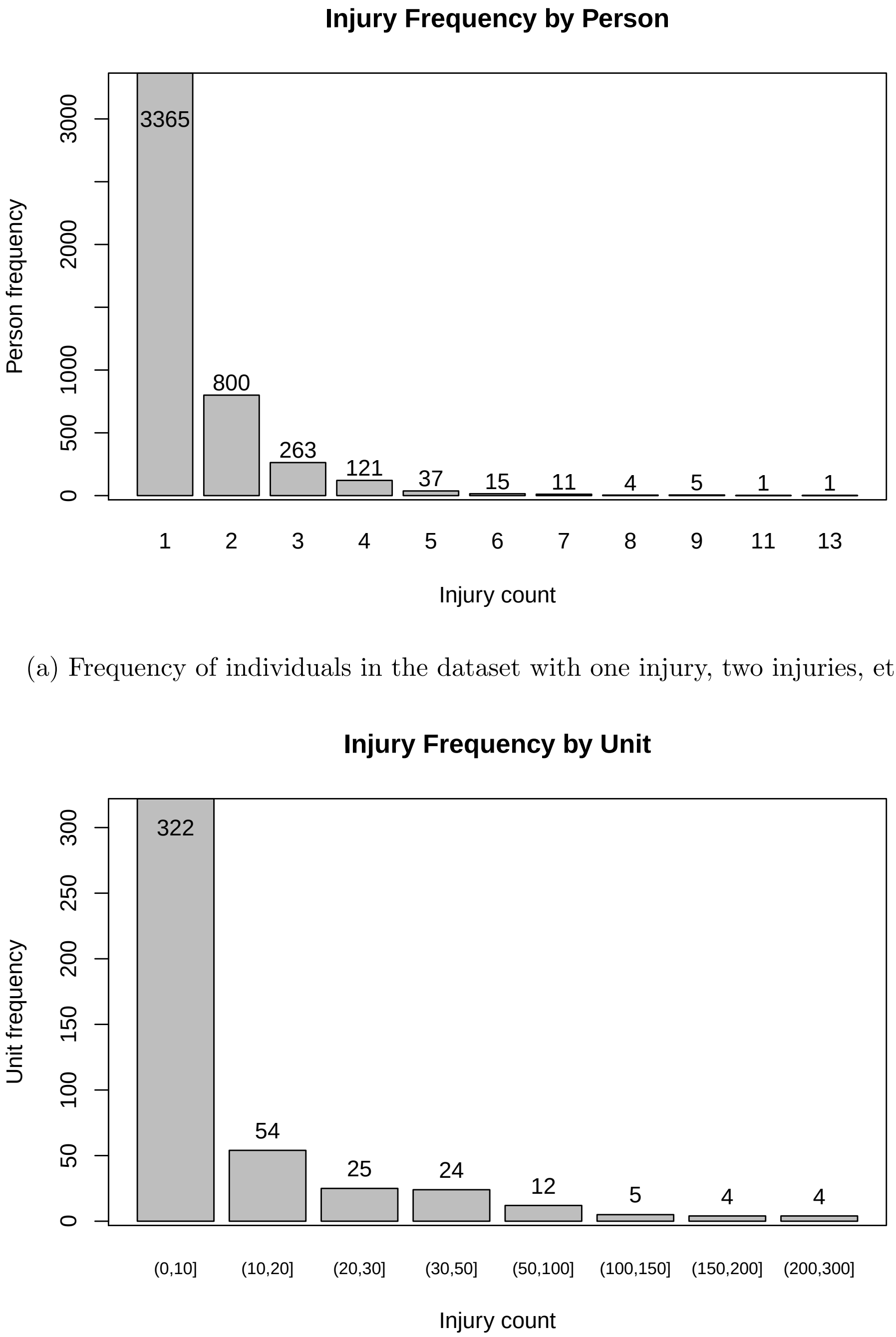
2. Description of the data

In its original form, the data consists of 6,691 individual injuries. Variables include the type of injury, the location of the injury (arm, leg, etc.), the occupation (EMT, firefighter, accountant, etc.), the ambulance unit or station for which the injured person worked, the person's gender, and information about lost wages. Table 1 shows the complete set of variables available. Of the 6,691 reported injuries, there are 4,623 unique people. Figure 1a shows a bar plot of the frequency in which a person reported one, two, etc. injuries. There are 450 distinct ambulance units; Figure 1b shows the frequency in which $(0, 10]$, $(10, 20]$, etc. injuries were reported per unit. There are 55 types of injuries reported, ranging from nausea, to seizure, to death; Table 2 reports the 20 most frequent injury types. There are 600 unique occupations given in the data.

In this study, we are interested in estimating the probabilities of specific injury types, given that an injury occurs. For simplicity, we will collapse the 55 observed injury types into ten: strain, contusion, sprain, puncture, laceration, torn cartilage/ligament/tendon (abbreviated C/L/T from here on), fracture, inflammation, respiratory, and other. Figure 2a plots the relationship between the two most frequent injuries: strain and contusion. We see that a majority of ambulance units have a higher proportion of reported sprains than contusions. Figure 2b shows a plot of the proportion of strains among ambulance units. There appear to be multiple modes in this plot, which is an indication that the standard binomial would not fit well.

It is reasonable to assume that the injury probabilities vary with occupation; for example, a firefighter faces vastly different risks than an accountant. We may wish to imagine the presence of latent factors, associated with each occupation, which influence the probabilities of the injury types. However, there are difficulties in using the `CLMNT_OCCUP` variable directly. The entry of the occupation labels has not been completely systematic, yielding many small variations or misspellings of the same occupation. For example, the values "PARAMEDIC", "PARADEMIC", "PARAMADIC", "PARAMDEIC", and "PARA-MEDIC" are all present in the data. There are also entries such as "EMT FIREFIGHTER", which blurs the distinction between EMT and firefighter. In addition to the variation by occupation, it is reasonable to assume that the probabilities vary between individuals as well, based on factors such as carefulness/carelessness on the job and pre-existing medical conditions. However, such data has not been collected for this study.

Consider the injury count for each ambulance unit, which is the sum of the injury counts from all workers in that unit. This provides a natural clustering of the individual injuries, with `UNITNAME_5` as the variable defining the clusters. Hence the data becomes a multinomial sample of 450 observations, each having a count vector with ten categories. Note that individual injuries sharing the same value of `UNITNAME_5` will be similar geographically; this suggests that they were subject to similar safety standards on the job and similar regional hazards such as crime and weather. Due to variations not modeled explicitly in the analysis (such as by occupation and individual, discussed earlier), we anticipate a heterogeneity in the probabilities of the ten injury types among observations. This motivates the choice of the mixture of multinomials model for our analysis. Several covariates can be made available at the cluster level, such as "proportion of females" and "average age", but we have opted not to use these in our analysis.



(a) Frequency of individuals in the dataset with one injury, two injuries, etc.

(b) Frequency of ambulance units in the dataset with 1–10 injuries, 11–20 injuries, etc.

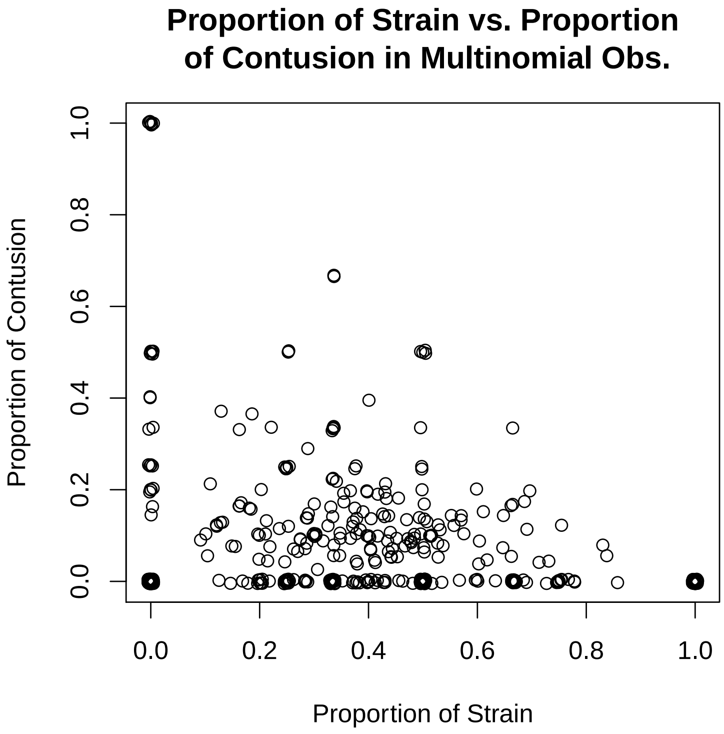
Figure 1: Histograms of injury frequencies.

Table 1: All variables in injury dataset.

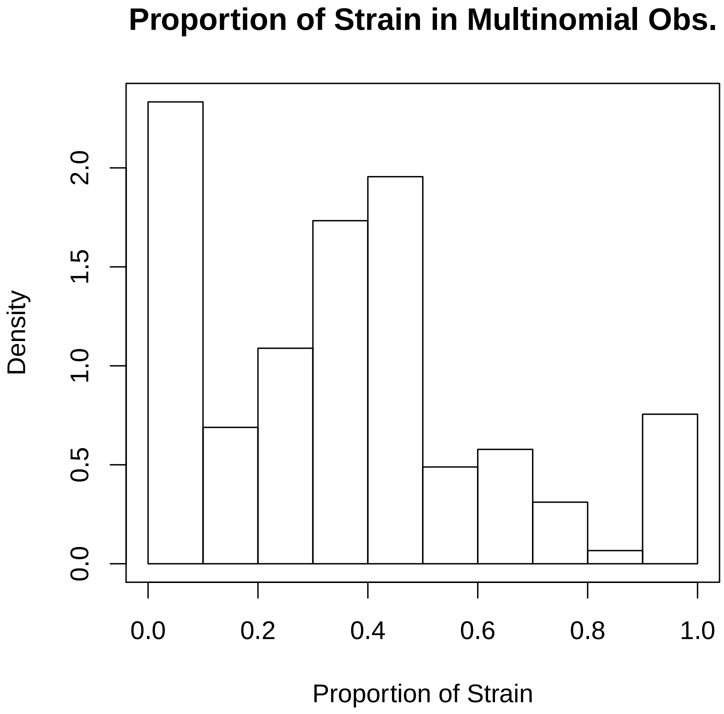
Variable Category	Variable Name	Description
Patient Information	ID	Unique person ID based on SSN, BIRTH_DATE, and CLMNT_SEX
	CLAIM_NUM	Unique ID for the injury
	SSN	Last four digits of Social Security number
	BIRTH_DATE	Date of birth
	CLMNT_SEX	Gender
	CLMNT_OCCUP	Occupation
	UNITNAME_5	Ambulance unit / station for which the person was working
	ACC_STATE	State (in the U.S.A.)
	CLM_STATUS	Claim status
Injury Information	ACC_DATE	Date of injury
	LOSS_DESC	Free-form text description of injury cause
	NL_DESC_BI	A classification of the injury type (strain, burn, etc.)
	POB_DESC	Location of the injury on the body (neck, head, etc.)
	SRCE_DESC	Source of Injury
	TYPE_DESC	Type of injury
Loss Wage Information	TOT_LWD	
	RTW_DATE	
	WC_MED_IND	
	NET_PD_LOS	
	NET_PD_EXP	
	REM_RS_LOS	
	REM_RS_EXP	
	TOT_EXPER	

Table 2: Observed frequencies of top 20 injuries.

Rank	Injury Name	Frequency
1	Strain	2785
2	Contusion, bruise	653
3	Sprain	554
4	Not Otherwise Classified	530
5	Puncture	292
6	Laceration, open wound	279
7	No physical injury	197
8	Multiple physical injuries	170
9	Torn cartilage / ligament / tendon	114
10	Fracture	105
11	Foreign body	104
12	Inflammation / irritation of joint / nerve	103
13	Respiratory disorders	102
14	Herniation, rupture	85
15	Scratch, abrasion	78
16	Communicable Disease	71
17	Occupational health disorder, NOC	49
18	Burn (heat)	47
19	Bite or sting	39
20	Allergic Reaction	31
	ALL OTHERS	303



(a) Scatter plot of proportion of strains vs. proportion of contusions among the $n = 450$ multinomial observations. Some jitter has been added to help distinguish points.



(b) Histogram of strains among the $n = 450$ multinomial observations.

Figure 2: Marginal plots of strain vs. contusion and density of strains.

3. Mixture of Multinomials

Denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})$ as the vector of injury counts for the ith ambulance unit, for $i = 1, \ldots, n$ and $n = 450$. Here, $k = 10$ since there are ten injury types under consideration. Denote $m_i = \sum_{j=1}^{k} Y_{ij}$ as the corresponding cluster size. We will assume the mixture of multinomials model, denoted $\boldsymbol{Y}_i \overset{\text{ind}}{\sim} \text{MultMix}(\boldsymbol{\theta}, m_i)$ for $i = 1, \ldots, n$, which suggests the likelihood

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{\ell=1}^{s} \pi_\ell \left[\frac{m_i!}{y_{i1}! \ldots y_{ik}!} p_{\ell 1}^{y_{i1}} \ldots p_{\ell k}^{y_{ik}} \right] \right\}, \tag{1}$$

with $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s, \boldsymbol{\pi})$. The number of mixing components s corresponds to a number of latent classes, where observations belonging to a class have similar probabilities for the k injury types. The value of s is not known, and must be inferred from the data.

Maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ are desired under the likelihood (1). As is usually the case in finite mixture distributions, closed form expressions for $\hat{\boldsymbol{\theta}}$ are not available, and iterative techniques must be used to maximize the likelihood. Several methods are considered in Raim et al (2012), which we will recall here. Let $\mathcal{I}_m(\boldsymbol{\theta})$ denote the $(sk-1) \times (sk-1)$ Fisher information matrix, with respect to a single multinomial observation with cluster size m, evaluated at $\boldsymbol{\theta}$. The Fisher information matrix will hereafter be referred to as the "FIM" or "exact FIM". Define the $(sk-1) \times (sk-1)$ matrix

$$\begin{aligned}
\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) &:= \text{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right), \text{where} \\
\boldsymbol{F}_\ell &= m \left[\text{diag}(p_{\ell 1}^{-1}, \ldots, p_{\ell,k-1}^{-1}) + p_{\ell k}^{-1} \boldsymbol{1}\boldsymbol{1}^T \right], \quad \ell = 1, \ldots, s, \\
\boldsymbol{F}_\pi &= \text{diag}(\pi_1^{-1}, \ldots, \pi_{s-1}^{-1}) + \pi_s^{-1} \boldsymbol{1}\boldsymbol{1}^T.
\end{aligned} \tag{2}$$

where $\boldsymbol{1}$ denotes a vector of ones of the appropriate dimension. Note that the $\boldsymbol{F}_\ell$ are $(k-1) \times (k-1)$ matrices corresponding to the FIM of the $\text{Mult}(\boldsymbol{p}_\ell, m)$ distribution, and $\boldsymbol{F}_\pi$ is a $(s-1) \times (s-1)$ matrix which is the FIM of $\text{Mult}(\boldsymbol{\pi}_\ell, 1)$. Raim et al (2012) justify $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ as a large cluster approximation (as $m \to \infty$) to $\mathcal{I}_m(\boldsymbol{\theta})$. We will denote the information matrix and its approximation for the entire sample as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{m_1}(\boldsymbol{\theta}) + \cdots + \mathcal{I}_{m_n}(\boldsymbol{\theta}) \quad \text{and} \quad \widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \widetilde{\mathcal{I}}_{m_1}(\boldsymbol{\theta}) + \cdots + \widetilde{\mathcal{I}}_{m_n}(\boldsymbol{\theta})$$

respectively.

Recall that the Fisher scoring algorithm is written as

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)}) S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots, \tag{3}$$

where iterations are repeated, given an initial starting value $\boldsymbol{\theta}^{(0)}$, until some convergence criterion is reached. Here we take the criterion to be

$$\left| \log L(\boldsymbol{\theta}^{(g+1)}) - \log L(\boldsymbol{\theta}^{(g)}) \right| < \varepsilon \tag{4}$$

for some given $\varepsilon > 0$. An approximate Fisher scoring algorithm (AFSA) can be formulated by replacing $\mathcal{I}(\boldsymbol{\theta})$ with $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ in (3) to obtain the iterations

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}^{(g)}) S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots. \tag{5}$$

Raim et al (2012) show that AFSA is approximately equivalent to a standard Expectation-Maximization algorithm used for the finite mixture of multinomials model; similar estimates are obtained at a similar convergence rate. They demonstrate that AFSA and EM are more robust to the choice of initial value $\boldsymbol{\theta}^{(0)}$ than standard Fisher scoring. However, the matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is not recommended to be used in place of $\mathcal{I}(\boldsymbol{\theta})$ for inference (e.g. to compute standard errors), unless the cluster size is large. To address this, a hybrid Fisher scoring algorithm is proposed, where AFSA iterations are used from an initial $\boldsymbol{\theta}^{(0)}$ until some preliminary tolerance ε_0 is reached on (4). Then, exact Fisher scoring iterations are used to reach the desired tolerance ε. This is a more general version of the "one additional step" estimator proposed by Neerchal and Morel (2005), where just a single iteration of exact Fisher scoring is used. Hybrids between EM and Fisher scoring have also been considered; see (McLachlan and Peel, 2000).

4. Selecting the Number of Classes

A complication in fitting finite mixture models is the choice for the number of mixing components s (McLachlan and Peel, 2000). One possibility for selecting s is to consider commonly used information criteria such as the Akaike information criterion (AIC) and Bayesian information criteria (BIC). Consider

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

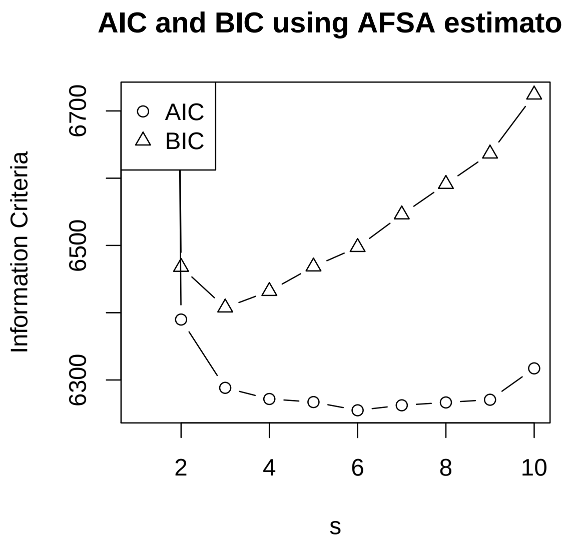
where $q = sk - 1$ is the total number of parameters, and the log-likelihood is evaluated at the AFSA estimate $\tilde{\theta}$. AFSA has been used because of its computational convenience, as opposed to exact Fisher scoring. In section 5, we will more concretely discuss the difficulty in computing $\mathcal{I}(\boldsymbol{\theta})$ under the likelihood (1), and propose a better approximation than $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$. Figure 3 shows the AIC and BIC values using AFSA when $s = 1, \ldots, 10$ mixing components are used. Note that 20 randomly chosen initial values were used for each setting of s, except for $s = 1$ which corresponds to the standard multinomial. Of those, the run with the largest log-likelihood was selected to compute AIC and BIC for that s. Note that not all initial values led to convergence; particularly in the case that $s = 10$, only 6 of 20 runs converged. Under BIC, three mixing components appears to be reasonable. Under AIC, four to eight components appears reasonable. To keep interpretations as simple as possible, we will use the choice $s = 3$ suggested by BIC. Denote b_{jj} as the diagonal elements of $\widetilde{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\theta}})$. Table 3 shows AFSA estimates under this model, along with standard errors given by $\sqrt{b_{jj}}$, and $(1-\alpha)$ confidence intervals computed by $\tilde{\theta}_j \pm z_{\alpha/2}\sqrt{b_{jj}}$ with $\alpha = 0.05$.

5. A Closer Approximation for the Exact Fisher Information Matrix of a Sample with Varying Cluster Sizes

In the previous section, the FIM approximation $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ was used to formulate AFSA iterations so that a preliminary estimate $\tilde{\boldsymbol{\theta}}$ could be computed, and a study of AIC and BIC could easily be carried out. The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ may not provide accurate standard errors for small to moderate cluster sizes, as discussed earlier. However, the matrix $\mathcal{I}(\boldsymbol{\theta})$ is difficult to compute exactly. Note that the basic EM elgorithm does not yield standard errors. Computation of an exact FIM using the EM estimator is one way to obtain standard errors; other methods are discussed in (McLachlan and Peel, 2000). In this section we propose an obvious improvement to the FIM

Table 3: Estimates, standard errors, and confidence intervals using AFSA estimator $\tilde{\boldsymbol{\theta}}$ with $s = 3$.

					95% CI	
Class	Param	Description	Estimate	SE	Lower	Upper
1	π_1	Mixing	0.5347	0.0235	0.4886	0.5808
	p_{11}	Strain	0.4783	0.0084	0.4619	0.4947
	p_{12}	Contusion	0.1025	0.0051	0.0926	0.1124
	p_{13}	Sprain	0.0981	0.0050	0.0884	0.1078
	p_{14}	Puncture	0.0342	0.0030	0.0282	0.0402
	p_{15}	Torn C/L/T	0.0349	0.0031	0.0289	0.0409
	p_{16}	Laceration	0.0144	0.0020	0.0105	0.0183
	p_{17}	Fracture	0.0105	0.0017	0.0072	0.0138
	p_{18}	Inflammation	0.0116	0.0018	0.0081	0.0151
	p_{19}	Respiratory	0.0088	0.0016	0.0057	0.0119
2	π_2	Mixing	0.3961	0.0231	0.3509	0.4413
	p_{21}	Strain	0.2940	0.0088	0.2767	0.3113
	p_{22}	Contusion	0.0743	0.0051	0.0643	0.0843
	p_{23}	Sprain	0.0598	0.0046	0.0508	0.0688
	p_{24}	Puncture	0.0624	0.0047	0.0532	0.0716
	p_{25}	Torn C/L/T	0.0588	0.0046	0.0498	0.0678
	p_{26}	Laceration	0.0178	0.0026	0.0128	0.0228
	p_{27}	Fracture	0.0298	0.0033	0.0233	0.0363
	p_{28}	Inflammation	0.0128	0.0022	0.0085	0.0171
	p_{29}	Respiratory	0.0059	0.0015	0.0030	0.0088
3	p_{31}	Strain	0.3364	0.0220	0.2933	0.3795
	p_{32}	Contusion	0.1311	0.0157	0.1003	0.1619
	p_{33}	Sprain	0.0423	0.0094	0.0240	0.0606
	p_{34}	Puncture	0.0550	0.0106	0.0342	0.0758
	p_{35}	Torn C/L/T	0.0394	0.0090	0.0217	0.0571
	p_{36}	Laceration	0.0333	0.0083	0.0169	0.0497
	p_{37}	Fracture	0.0108	0.0048	0.0014	0.0202
	p_{38}	Inflammation	0.0500	0.0101	0.0301	0.0699
	p_{39}	Respiratory	0.0879	0.0132	0.0621	0.1137



(a)

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

(b)

Figure 3: AIC and BIC for $s = 1, \ldots, 10$ using AFSA estimator.

approximation $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ for the sample, where clusters with smaller m_i are computed with an exact FIM instead. This yields a matrix closer to $\mathcal{I}(\boldsymbol{\theta})$, but which is more easily computed, and will provide more accurate computation of standard errors and confidence intervals than those obtained from $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$.

Suppose the observed cluster sizes are ordered $m_1 \leq \cdots \leq m_n$, without loss of generality. The exact FIM for the sample is computed as

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{I}_{m_i}(\boldsymbol{\theta}) = \sum_{i=1}^{n^*} F_i \, \mathcal{I}_{r_i}(\boldsymbol{\theta}), \tag{6}$$

where $r_1 \leq \ldots \leq r_{n^*}$ are the distinct values of $\{m_1, \ldots, m_n\}$ with $n^* \leq n$, and $F_1, \ldots, F_{n^*}$ are the corresponding freqencies. Using the definition of expectation, the terms in (6) can be computed exactly as

$$\mathcal{I}_m(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \Omega} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}, m) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}, m) \right\}^T f(\boldsymbol{x}; \boldsymbol{\theta}, m) \tag{7}$$

but recall that there are $\binom{m+k-1}{m}$ elements in the multinomial sample space Ω. This number may be quite large, making the naive calculation of the FIM in (7) infeasible in practice. The FIM approximation $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ discussed in section 3 was intended for the scenario for when m is large, which is one way that $\binom{m+k-1}{m}$ may be made large. Denote $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ as the FIM approximation for a single multinomial observation with cluster size m; then the FIM approximation for the entire sample would be

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta}). \tag{8}$$

Raim et al (2012) find that $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ is not an accurate approximation to $\mathcal{I}(\boldsymbol{\theta})$ unless m is large. Therefore, we would expect to obtain a better approximation by identifying the $\widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta})$ in (8) with small cluster sizes, and replacing them with $\mathcal{I}_{m_i}(\boldsymbol{\theta})$. To formalize this idea, let

$$\begin{aligned}\mathcal{I}^*(\boldsymbol{\theta}, C) &= \sum_{i:m_i \leq C} \mathcal{I}_{m_i}(\boldsymbol{\theta}) + \sum_{i:m_i > C} \widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta}) \\ &= \sum_{i:r_i \leq C} F_i\, \mathcal{I}_{r_i}(\boldsymbol{\theta}) + \sum_{i:m_i > C} F_i\, \widetilde{\mathcal{I}}_{r_i}(\boldsymbol{\theta}) \end{aligned} \tag{9}$$

where $C \in [0, m_n]$ is a tuning parameter. Selecting C to be large will ensure that $\mathcal{I}^*(\boldsymbol{\theta}, C) \approx \mathcal{I}(\boldsymbol{\theta})$, but the amount of computation will be close to that of the exact FIM. On the other hand, if C is selected to be small the amount of computation will be dramatically reduced, but $\mathcal{I}^*(\boldsymbol{\theta}, C)$ will be closer to the approximation $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$. Note that the matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ can easily be transformed to another cluster size m^* — recall the expression given in (2), simply transform $\boldsymbol{F}_\ell$ to $(m^*/m)\boldsymbol{F}_\ell$. Hence, computation of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is quite convenient.

We would like to find a value of C such that the Frobenius norms

$$\|\mathcal{I}^*(\boldsymbol{\theta}, C) - \mathcal{I}(\boldsymbol{\theta})\| = \left\| \sum_{i:m_i > C} F_i \left(\widetilde{\mathcal{I}}_{r_i}(\boldsymbol{\theta}) - \mathcal{I}_{r_i}(\boldsymbol{\theta}) \right) \right\|, \quad \text{and} \tag{10}$$

$$\begin{aligned}\|\mathcal{I}^{*-1}(\boldsymbol{\theta}) - \mathcal{I}^{-1}(\boldsymbol{\theta})\| &= \left\| \mathcal{I}^{*-1}(\boldsymbol{\theta}) \Big[\mathcal{I}^*(\boldsymbol{\theta}, C) - \mathcal{I}(\boldsymbol{\theta}) \Big] \mathcal{I}^{-1}(\boldsymbol{\theta}) \right\| \\ &= \left\| \mathcal{I}^{*-1}(\boldsymbol{\theta}) \left[\sum_{i:m_i > C} F_i \left(\widetilde{\mathcal{I}}_{r_i}(\boldsymbol{\theta}) - \mathcal{I}_{r_i}(\boldsymbol{\theta}) \right) \right] \mathcal{I}^{-1}(\boldsymbol{\theta}) \right\| \end{aligned} \tag{11}$$

are small, but the amount of computation for $\mathcal{I}^*(\tilde{\boldsymbol{\theta}}, C)$ is not too intensive. In (11), we have used the identity $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{B}^{-1}$ for non-singular matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Notice that the difference $\mathcal{I}^*(\boldsymbol{\theta}, C) - \mathcal{I}(\boldsymbol{\theta})$ depends only on those clusters with $m_i > C$. To measure the amount of computation needed, let

$$N_{C,k} = \sum_{i:r_i > C} \binom{r_i + k - 1}{r_i} \tag{12}$$

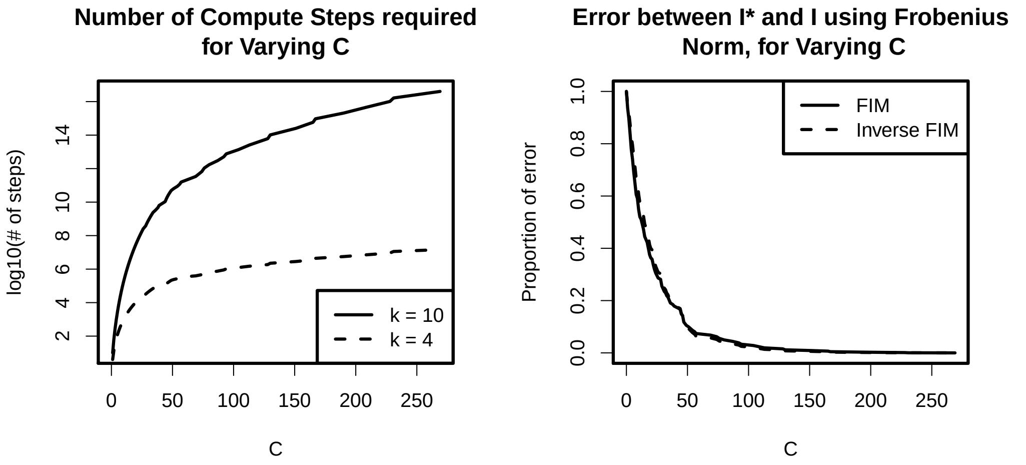
be the number of iterations of (7) needed to compute the exact FIM terms of the "Hybrid FIM" proposed in (9). Note that $N_{C,k}$ does not depend on $\boldsymbol{\theta}$, but only on the sample and the given C.

A small study was carried out to empirically determine a good value of C. We have fixed $\boldsymbol{\theta}$ to $\tilde{\boldsymbol{\theta}}$, the AFSA estimator using $s = 3$, which was found to be an appropriate number of classes for the injury data in section 4. Figure 4a shows the number of compute steps $N_{C,k}$ required to compute $\mathcal{I}^*(\tilde{\boldsymbol{\theta}}, C)$ for $C = 0, m_1, \ldots, m_n$. Both $k = 4$ and $k = 10$ are shown on the log-base-10 scale. Computing the exact FIM for $k = 10$ is intractible using the naive method, requiring a summation over more than 10^{15} terms. Because of this, we evaluate the selection of C under the more modest sample space generated at $k = 4$.

Figure 4b shows the norms given by (10) and (11). Rather than plotting the norms directly, the proportions

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

are plotted for $C = 0, m_1, \ldots, m_n$. Note that $C = 0$ maximizes the distance $\|\mathcal{I}^*(\boldsymbol{\theta}, C) - \mathcal{I}(\boldsymbol{\theta})\|$ over C, and that $\mathcal{I}^*(\boldsymbol{\theta}, C = 0) \equiv \widetilde{\mathcal{I}}(\boldsymbol{\theta})$. From Figure 4b, we



(a) Number of compute steps $N_{C,k}$ for $k = 4$ and $k = 10$ categories.

(b) Proportion of errors p_C and q_C using $k = 4$ categories.

Figure 4: Number of compute steps and proportion of errors as C varies.

see that the sequences p_C and q_C are very similar. For $k = 4$, the choice $C = 50$ appears to give a good balance between computability and accuracy. However, for $k = 10$, the number of compute steps becomes highly impractical after about $C = 20$. Therefore, we proceed with $C = 20$ for our injury data analysis using ten multinomial categories, with some assurance that $\mathcal{I}^*(\boldsymbol{\theta}, C)$ will provide more accurate inference than $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$.

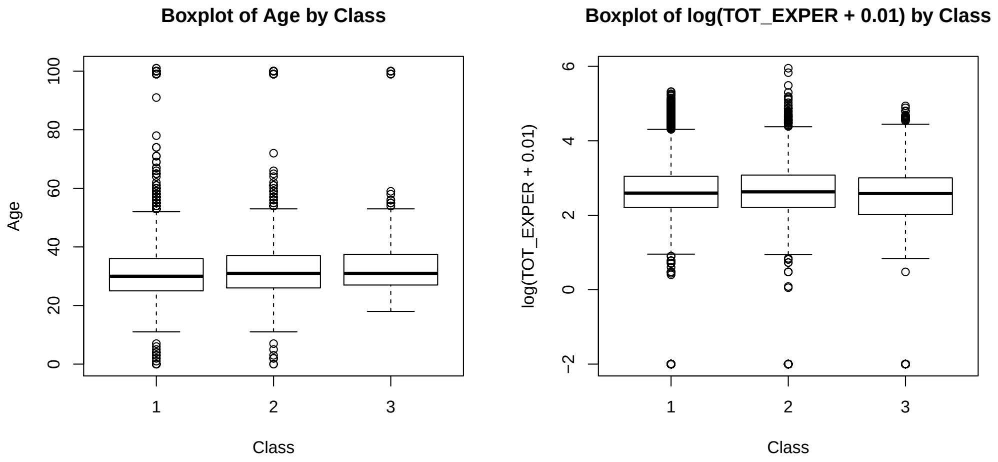
6. Data Analysis using Mixture of Multinomials

Starting with the AFSA estimate $\tilde{\boldsymbol{\theta}}$ from section 4 and using $C = 20$, we apply the "one additional step" iteration

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + \mathcal{I}^{*-1}(\tilde{\boldsymbol{\theta}}, C) S(\tilde{\boldsymbol{\theta}})$$

discussed in section 3 to obtain final estimates for the injury dataset. The diagonal elements of $\mathcal{I}^{*-1}(\hat{\boldsymbol{\theta}}, C)$, denoted d_{jj}, are used to compute asymptotic standard errors via $\sqrt{d_{jj}}$, and level $(1-\alpha)$ confidence intervals are computed by $\hat{\theta}_j \pm z_{\alpha/2}\sqrt{d_{jj}}$ using $\alpha = 0.05$. Table 4 displays these results. We note that the estimates themselves are very close to those in Table 3, hence the additional Fisher scoring iteration did not move $\hat{\boldsymbol{\theta}}$ far from $\tilde{\boldsymbol{\theta}}$. The standard errors in Table 4 are larger than the corresponding values in Table 3, indicating that those derived from the matrix $\widetilde{\mathcal{I}}(\tilde{\boldsymbol{\theta}})$ are too optimistic. However, note that many of the differences are quite small; larger differences can be seen in standard errors with respect to the parameters π_1 and π_2.

There are two large classes in the population — classes 1 and 2 having estimated proportions 53.47% and 39.62% — and one small class — class 3 having estimated proportion 6.91%. Notice that the mixing proportions π_1 and π_2 have the highest standard errors among all parameters, giving the fairly wide 95% confidence intervals $\pi_1 \in [0.4539, 0.6154]$ and $\pi_2 \in [0.3179, 0.4744]$. Observations in class 1 have higher probabilities of strain and sprain than those in classes 2 and 3. On the other hand, observations in class 2 have higher probability of fracture than the



(a) Boxplot of injury ages by class.

(b) Boxplot of wage losses by class.

Figure 5: Boxplots for classified data.

other classes. Finally, observations in class 3 have higher probabilities of contusion, laceration, inflammation, and respiratory problems than the other two classes.

We assign the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into these three classes using the posterior probability rule

$$\text{Class for } i\text{th observation} = \underset{\ell}{\operatorname{argmax}} \frac{\hat{\pi}_\ell f(\boldsymbol{y}_i \mid \hat{\boldsymbol{p}}_\ell, m_i)}{\sum_{a=1}^{s} \hat{\pi}_a f(\boldsymbol{y}_i \mid \hat{\boldsymbol{p}}_a, m_i)}.$$

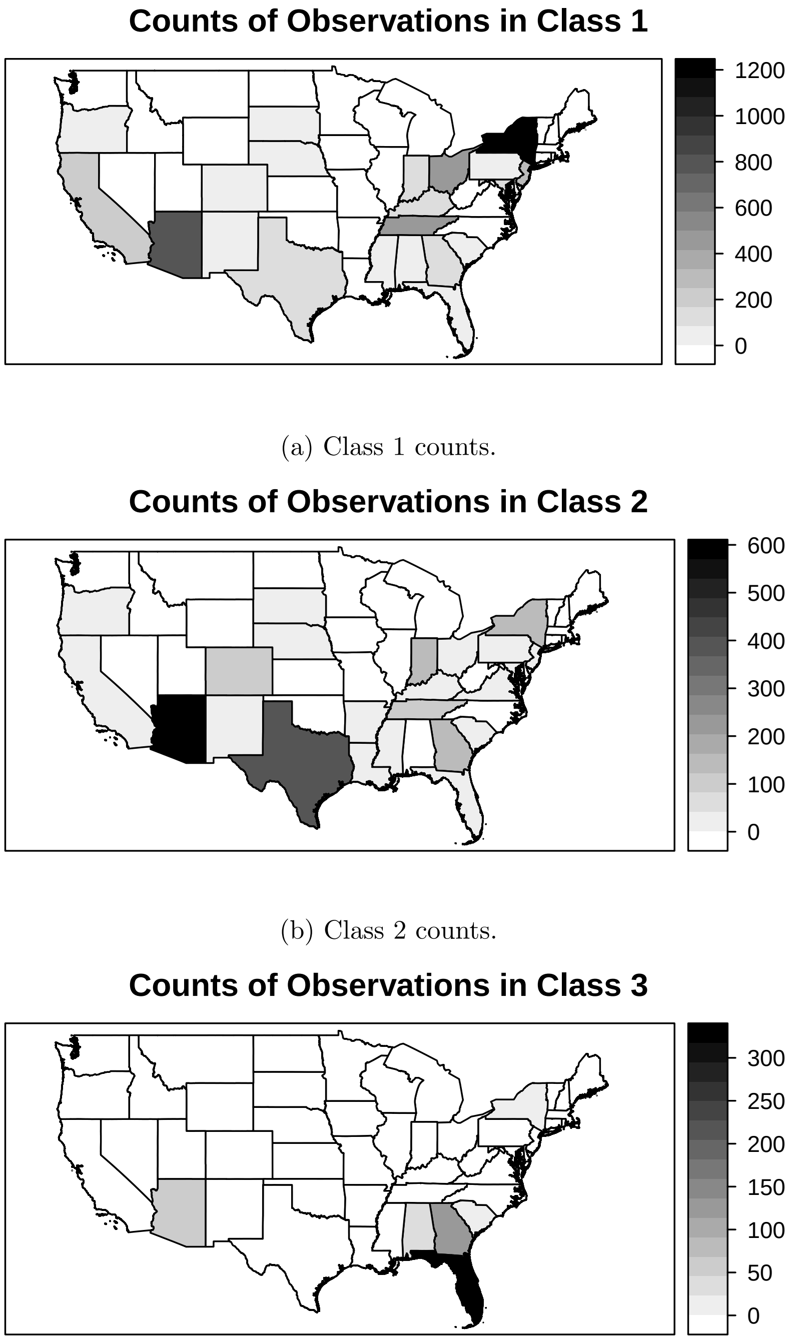
and find that there are 265 observations in class 1, 172 in class 2, and 13 in class 3. Within these multinomial clusters, there are 4,382 individual injuries in class 1, 1,762 in class 2, and 547 in class 3. Using the classified data on individual injuries, some interesting comparisons can be made between classes. The proportion of females is 47.92% in class 1, 48.26% in class 2, but only 30.77% in class 3. Figure 5a shows that the majority of ages are about the same for all three classes, but with less younger individuals in group 3. (Note that ages such as 0 and 100+ are likely data entry issues). Figure 5b compares the value of `TOT_EXPER`, which involves lost wages, between groups at the log-base-10 scale. Before the log is taken, 0.01 is added to prevent taking log of zero. Again the majority of values are about the same, but with class 2 having more large values than classes 1 and 3. Figure 6 shows the distribution of individual injuries across states in the U.S.A. for each class. Class 1 contains significantly more injuries in New York than the other two classes, while class 2 has large counts in Arizona and Texas, and class 3 has many of its injuries in Florida and Georgia. Note that multinomial observations may contain injuries from multiple states.

7. Conclusions

Using the computational methods discussed in (Raim et al., 2012), we have carried out a mixture of multinomials analysis on an injury dataset. AIC and BIC were used in model selection to select an appropriate number of mixing components. The AFSA algorithm was helpful in carrying out this computation efficiently. To obtain standard errors, a "one additional step" estimator was computed. Rather than using the exact FIM, a "better approximation" (than the matrix used in AFSA

Table 4: Final estimates, standard errors, and confidence intervals using "one additional step" estimator $\hat{\boldsymbol{\theta}}$.

					95% CI	
Class	Param	Description	Estimate	SE	Lower	Upper
1	π_1	Mixing	0.5347	0.0412	0.4539	0.6154
	p_{11}	Strain	0.4783	0.0089	0.4608	0.4958
	p_{12}	Contusion	0.1025	0.0053	0.0920	0.1130
	p_{13}	Sprain	0.0981	0.0052	0.0879	0.1083
	p_{14}	Puncture	0.0343	0.0033	0.0278	0.0407
	p_{15}	Torn C/L/T	0.0349	0.0033	0.0285	0.0414
	p_{16}	Laceration	0.0144	0.0021	0.0102	0.0186
	p_{17}	Fracture	0.0105	0.0019	0.0069	0.0142
	p_{18}	Inflammation	0.0115	0.0019	0.0078	0.0153
	p_{19}	Respiratory	0.0088	0.0017	0.0054	0.0121
2	π_2	Mixing	0.3962	0.0399	0.3179	0.4744
	p_{21}	Strain	0.2940	0.0097	0.2749	0.3130
	p_{22}	Contusion	0.0743	0.0055	0.0635	0.0851
	p_{23}	Sprain	0.0598	0.0050	0.0500	0.0696
	p_{24}	Puncture	0.0624	0.0050	0.0526	0.0721
	p_{25}	Torn C/L/T	0.0588	0.0048	0.0493	0.0683
	p_{26}	Laceration	0.0178	0.0027	0.0124	0.0232
	p_{27}	Fracture	0.0298	0.0035	0.0230	0.0366
	p_{28}	Inflammation	0.0128	0.0024	0.0082	0.0174
	p_{29}	Respiratory	0.0059	0.0017	0.0027	0.0091
3	p_{31}	Strain	0.3363	0.0247	0.2879	0.3848
	p_{32}	Contusion	0.1310	0.0174	0.0969	0.1652
	p_{33}	Sprain	0.0422	0.0106	0.0215	0.0630
	p_{34}	Puncture	0.0550	0.0118	0.0319	0.0782
	p_{35}	Torn C/L/T	0.0394	0.0101	0.0196	0.0593
	p_{36}	Laceration	0.0333	0.0092	0.0153	0.0513
	p_{37}	Fracture	0.0109	0.0054	0.0002	0.0215
	p_{38}	Inflammation	0.0500	0.0110	0.0284	0.0716
	p_{39}	Respiratory	0.0879	0.0144	0.0597	0.1161


Counts of Observations in Class 1
1200
1000
800
600
400
200
0
Counts of Observations in Class 2
600
500
400
300
200
100
0
Counts of Observations in Class 3
300
250
200
150
100
50
0

(a) Class 1 counts.

(b) Class 2 counts.

(c) Class 3 counts.

Figure 6: Counts of classified individuals plotted by state.

iterations) was formulated. We conducted a small study to obtain a reasonably accurate approximation to the FIM using a tolerable amount of computation. The "one additional step" estimator provided final estimates and standard errors, as well as a classification of the clustered multinomial observations and the individuals within those clusters. We found that three classes fit well to the data, with each class responding differently to the ten injury types under consideration. Some interesting contrasts were found between classes, such as the proportion of females, and the distribution of injuries across the U.S.A.

References

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

G. McLachlan and D. Peel. *Finite Mixture Models.* Wiley-Interscience, 2000.

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.

Andrew M. Raim, Minglei Liu, Nagaraj K. Neerchal, and Jorge G. Morel. An Approximate Fisher Scoring Algorithm for finite mixtures of multinomials. Technical Report HPCF–2012–14, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2012.