Modeling Hidden Exposures in Claim Severity via the EM Algorithm

Grzegorz A. Rempala and Richard A. Derrig

Modeling Hidden Exposures in Claim Severity via the EM Algorithm

Grzegorz A. Rempala
Department of Mathematics, University of Louisville,
Richard A. Derrig
Automobile Insurers Bureau of Massachusetts

Abstract

We consider the issue of modeling the so-called hidden severity exposure occurring through either incomplete data or an unobserved underlying risk factor. We use the celebrated EM algorithm as a convenient tool in detecting latent (unobserved) risks in finite mixture models of claim severity and in problems where data imputation is needed. We provide examples of applicability of the methodology based on real-life auto injury claim data and compare, when possible, the accuracy of our methods with that of standard techniques.

1 Introduction

Actuarial analysis can be viewed as the process of studying profitability and solvency of an insurance firm under a realistic and integrated model of key input random variables such as loss frequency and severity, expenses, reinsurance, interest and inflation rates, and asset defaults. In a modern analysis of financial models of property-casualty companies, these input variables typically can be classified into financial market variables and underwriting variables (cf. e.g., D'Arcy et al. 1997). The financial variables generally refer to asset-side generated cash flows of the business, and the underwriting variables relate to the cash flows of the liabilities side. The process of developing any actuarial model begins with the creation of probability distributions of these input variables, including the establishment of the proper range of values of input parameters. The use of parameters is generally determined by the use of the parametric families of distributions, although the non-parametric techniques have a role to play as well (see, e.g., Derrig, et al. 2001). In this article we consider an issue of hidden or "lurking" risk factors or parameters and point out the possible use of the celebrated EM algorithm to uncover those factors. We begin by addressing the most basic questions concerning hidden loss distributions. To keep things in focus we will be concerned here only with two applications to modeling the severity of loss, but the methods discussed may be easily applied to other problems like loss frequencies, asset returns, asset defaults, and combining those into models of Risk Based Capital, Value at Risk, and general Dynamic

Financial Analysis, including Cash Flow Testing and Asset Adequacy Analysis. Our applications will illustrate the use of the EM algorithm (i) to impute missing values in an asset portfolio and (ii) to screen medical bills for possible fraud or abusive practices.

1.1 Hidden Exposures in Loss Severity Distributions

In many instances one would be interested in modeling hidden risk exposures as additional dimension(s) of the loss severity distribution. This in turn in many cases leads to considering mixtures of probability distributions as the model of choice for losses affected by hidden exposures; some parameters of the mixtures will be considered missing (i.e., unobservable in practice). During the last 20 years or so there has been a considerable advancement in statistical methodologies dealing with partially hidden or incomplete data models. Empirical data imputation has become more sophisticated and the availability of ever faster computing power have made it increasingly possible to solve these problems via iterative algorithms.

In our paper we shall illustrate a possible approach to two types of problems arising often in practical situations of modeling the severity of losses: (i) imputation of partially *missing* multivariate observations and (ii) identification of *latent* risks via fitting finite mixtures models.

The common feature of both of these issues is, generally speaking, the unavailability of complete information on the variables or parameters of interest. The statistical methodology which is especially well-suited for this type of circumstances is the so-called EM algorithm.

1.2 The EM Algorithm

In their seminal paper Dempster, Laird and Rubin (1977) have proposed the methodology which they have called the Expectation-Maximization (EM) algorithm as an iterative way of finding maximum likelihood estimates.[1] They demonstrated that the method was especially appropriate for finding the parameters of an underlying distribution from a given data set where the data was *incomplete* or had missing values. At present there are two basic applications of the EM methodology considered in the statistical literature. The first occurs when the data indeed has missing values, due to problems with or limitations of the data collection process. The second occurs when the original likelihood estimation problem is altered by assuming the existence of the *hidden* parameters or factors. It turns out that both these circumstances can be, at least initially, described in the following statistical setting. Let us consider a density function (possibly multivariate) $p(\cdot|\Theta)$ that is indexed by the set of parameters Θ. As a simple example we may take p to be a univariate Gaussian density and $\Theta = \{(\mu, \sigma)| -\infty < \mu < \infty, \sigma > 0\}$. Additionally, we have an observed data set $\mathcal{X}$ of size n, drawn from the distribution p. More precisely, we assume that the points of $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are the realizations of some independent random variables distributed according to $p(\cdot|\Theta)$. We shall call $\mathcal{X}$ the *incomplete* data. In addition to $\mathcal{X}$, we also

[1] A full explanation of the role of the EM algorithm in missing data problems can be found in Little and Rubin, (1987) or in a somewhat more mathematically advanced monograph by McLachlan and Krishnan (1997).

consider a *complete* data set $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and specify the joint density

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\mathbf{x}, \Theta)\, p(\mathbf{x}|\Theta). \tag{1}$$

As we can see from the last equality, this joint density $p(\mathbf{z}|\Theta)$ arises from considering the marginal density $p(\mathbf{x}|\Theta)$ and the specific assumptions on the relation between hidden (or missing) variables $\mathcal{Y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$ and the observed incomplete data $\mathcal{X}$. Associated with the joint density is the joint likelihood function

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i|\Theta)$$

which is often called the *complete* likelihood. For the sake of computational simplicity it is often more convenient to consider the logarithm of the complete likelihood

$$l(\Theta|\mathcal{Z}) = \log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{y}_i|\Theta). \tag{2}$$

Note that the function above may be thought of as a random variable since it depends on the unknown or missing information $\mathcal{Y}$ which by assumption is governed by an underlying probability distribution. Note also that in accordance with the likelihood principle, we now regard $\mathcal{X}$ as constant.

The EM algorithm as described in Dempster, Laird and Rubin (1977) consists of two steps repeated iteratively. In its *expectation* step or the E-step, the algorithm first finds the expected value of the complete log-likelihood function $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$ with respect to the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current parameter estimates. That is, instead of the complete log-likelihood (2) we consider the following

$$Q(\Theta, \Theta^{(i-1)}) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\Theta)|\mathcal{X}, \Theta^{(i-1)}\right]. \tag{3}$$

Note the presence of the second argument in the function $Q(\Theta, \Theta^{(i-1)})$. Here $\Theta^{(i-1)}$ stands for the current value of the parameter Θ at the iteration $(i-1)$, that is, the value which is used to evaluate the conditional expectation.

After the completion of the E-step, the second step of the algorithm is to maximize the expectation computed in the first step. This is called the *maximization* or the M-step, at which time the value of Θ is updated by taking

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta, \Theta^{(i-1)}) \tag{4}$$

The steps are repeated until convergence. It can be shown (via the relation (1) and Jensen's

inequality) that if Θ^* maximizes $Q(\Theta, \Theta^{(i-1)})$ with respect to Θ for fixed $\Theta^{(i-1)}$ then

$$l(\Theta^*|\mathcal{Z}) - l(\Theta^{(i-1)}|\mathcal{Z}) \geq Q(\Theta^*, \Theta^{(i-1)}) - Q(\Theta^{(i-1)}, \Theta^{(i-1)}) \geq 0$$

and each iteration of the procedure indeed increases the value of complete log-likelihood (2). Let us note that from the above argument it follows that a full maximization in the M-step is not necessary: it suffices to find any value of $\Theta^{(i)}$ such that $Q(\Theta^{(i)}, \Theta^{(i-1)}) > Q(\Theta^{(i-1)}, \Theta^{(i-1)})$. Such procedures are called GEM (*generalized EM*) algorithms. For a complete set of references see, for instance, the monograph by McLachlan and Krishnan (1997) where also the issues of convergence rates for the EM and GEM algorithms are thoroughly discussed. For some additional references and examples see also Wu (1983) or the monographs by Little and Rubin (1987) and Hastie,Tibshirani, and Friedman (2001).

2 Modeling Hidden Risks via the EM Algorithm

As indicated in the previous section the primary application of the EM algorithm is in fitting the maximum likelihood models. Since this is accomplished by the M-step of the algorithm, the role of the E-step is, therefore, secondary – it is needed to facilitate the performance of the M-step in the presence of the missing or incomplete data. However, as in this paper we shall focus on the usefulness of the EM procedure in modeling hidden risks or variables, in our setup we shall be in fact more interested in the E-step of the algorithm, as it will provide us with the way to estimate or impute missing data and uncover hidden factors and variables. In our examples below we shall consider two types of hidden (latent) variables. The first one will arise when, due to some problems with the data collection, parts of the observations are missing from the observed dataset. We consider this problem via the EM method in the particular context of multivariate (loss) models.

2.1 Multivariate Severity Distributions. Data Imputation with EM

Although insurance has been traditionally build on the assumption of independence and the law of large numbers has governed the determination of premiums, the increasing complexity of insurance and reinsurance products has lead over past decade to increased actuarial interest in the modeling of dependent risks (see, e.g., Wang 1998 or Embrechts et al. 2000). Multivariate loss and risk models (and especially those based on elliptically contoured distributions) have been hence of interest in such areas as Capital Asset Pricing Model and the Arbitrage Pricing Theory (cf. e.g., Campbell, Lo, and MacKinlay 1996).

In some circumstances, however, parts of the observed multivariate data may be missing. Claim reporting systems depend heavily on the front-line adjusters to provide data elements beyond the simple payment amounts. In the absence of, or even in the presence of, system edits, daily work load pressures and the lack of interest in the coded data provide a deadly combination of disincentives for accurate and complete coding. Actuaries are quite familiar with missing data fields, which when

Table 1: 10 fictitious observed gains and losses from two risk portfolios in thousands.

0.914	2.088	2.644	0.477	-1.940	-0.245	0.362	1.147	?	?
3.855	4.025	2.092	3.400	1.520	2.626	?	?	5.473	6.235

essential to the analysis most often results in throwing the record out, thereby creating unknown 'hidden' biases. Likewise, financial time series data may be interrupted, unavailable, or simply lost for securities or portfolios that are not widely tracked.

As an illustration of an application of the EM algorithm in this setting let us consider a hypothetical example of 10 losses/gains from a two-dimensional vector of risk portfolios, which we have generated using a bivariate normal distribution. The data is presented in Table 1 (in thousands of dollars). As we can see parts of the last four observations are missing from the table. In fact, for the purpose of our example, they have been removed from the generated data. We shall illustrate the usefulness of the EM algorithm in estimating these missing values.

If we denote by $\mathcal{X}$ the observed (incomplete) data listed in Table 1 then following our notation from previous section we have the complete data vector $\mathcal{Z}$ given by

$$\mathcal{Z} = (\mathbf{z}_1 \dots \mathbf{z}_n) = \left(\mathbf{x}_1 \dots, \mathbf{x}_6, (x_{1,7}, y_{2,7})^T, (x_{1,8}, y_{2,8})^T, (y_{1,9}, x_{2,9})^T, (y_{1,10}, x_{2,10})^T\right)$$

where $\mathbf{x}_j = (x_{1,j}, x_{2,j})^T$ for $j = 1 \dots, 6$ is the set of pairwise complete observations. The missing data (corresponding to ? marks in Table 1) is, therefore,

$$\mathcal{Y} = (y_{2,7}, y_{2,8}, y_{1,9}, y_{1,10}). \tag{5}$$

Let us note that under our assumption of normality, the equation (2) now becomes

$$l(\Theta|\mathcal{Z}) = -n \log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(\mathbf{z}_j - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{z}_j - \boldsymbol{\mu})$$

where $n = 10$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is a vector of means and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

is covariance matrix. The vector of unknown parameters, therefore, can be represented as

$$\Theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}). \tag{6}$$

In order to describe the EM algorithm in this setting we need to find the particular form of $Q(\Theta, \Theta^{(i-1)})$ defined by (3). Due to the independence of the $\mathbf{z}_i$'s this is equivalent, in effect, to evaluating

$$E_{\Theta^{(i-1)}}(Y|\mathcal{X}) \qquad \text{and} \qquad E_{\Theta^{(i-1)}}(Y^2|\mathcal{X})$$

where Y is the underlying random variable for $\mathcal{Y}$, assumed to be normal. From the general formulae for conditional moments of a bivariate normal variable $X = (X_1, X_2)$ with the set of parameters Θ as above, we have that

$$\begin{aligned} E(X_2|X_1 = x_1) &= \mu_2 + \sigma_{12}/\sigma_{11}(x_1 - \mu_1) \\ Var(X_2|X_1 = x_1) &= \sigma_{22.1} = \sigma_{22}(1 - \rho^2) \end{aligned} \tag{7}$$

where ρ stands for the correlation coefficient. Interchanging the subscripts 1 and 2 in (7) gives the formulae for the conditional mean and variance of the distribution $X_1|X_2 = x_2$. Using the relations (7) and the usual formulae for ML estimators of the normal mean vector μ and the covariance matrix Σ, we may now state the EM algorithm for imputing missing data in Table 1 as follows.

Algorithm 1 *(EM version of Buck's algorithm)*

1. *Define the initial value* $\Theta^{(0)}$ *of the set of parameters* (5). *Typically, it can be obtained on the basis of the set of complete pairs of observations (i.e.,* $\mathbf{x}_1 \ldots, \mathbf{x}_6$ *in Table 1).*

2. *The E-step: given the value of* $\Theta^{(i)}$ *calculate via* (7) *the vector* $\mathcal{Y}^{(i)}$ *of the imputations of the missing data* $\mathcal{Y}$ *given by* (6).

$$y_{2k}^{(i)} = \mu_2^{(i)} + \frac{\sigma_{12}^{(i)}}{\sigma_{11}^{(i)}}\left(x_{1k} - \mu_1^{(i)}\right) \quad \text{and} \quad {y_{2k}^2}^{(i)} = \left(y_{2k}^{(i)}\right)^2 + \sigma_{22.1}^{(i)} \quad \text{for } k = 7, 8$$

$$y_{1k}^{(i)} = \mu_1^{(i)} + \frac{\sigma_{12}^{(i)}}{\sigma_{22}^{(i)}}\left(x_{2k} - \mu_2^{(i)}\right) \quad \text{and} \quad {y_{1k}^2}^{(i)} = \left(y_{1k}^{(i)}\right)^2 + \sigma_{11.2}^{(i)} \quad \text{for } k = 9, 10$$

3. *The M-step: given the current value of the imputed complete data vector* $\mathcal{Z}^{(i)} = (\mathcal{X}, \mathcal{Y}^{(i)})$ *set* $M_k = \sum_{j=1}^n z_{k,j}^{(i)}/n$ *and* $M_{kl} = \sum_{j=1}^n z_{k,j}^{(i)} z_{l,j}^{(i)}/n$ *for* $k, l = 1, 2$, *and calculate* $\Theta^{(i+1)}$ *as*

$$\begin{aligned} \left(\mu_1^{(i+1)}, \mu_2^{(i+1)}\right) &= (M_1, M_2) \\ \sigma_{kl}^{(i+1)} &= M_{kl} - M_k M_l \qquad \text{for } k, l = 1, 2 \end{aligned}$$

4. *Repeat steps 2 and 3 until the relative difference of the subsequent values of* $l(\Theta^{(i+1)}|\mathcal{Z}^{(i)})$ *is sufficiently small.*

The above algorithm in its non-iterative version was first introduced by Buck (1960) who used the method of imputation via linear regression with subsequent covariance correction to estimate means and covariance matrices of p dimensional random vectors in case when some parts of the vector components were missing. For more details about Buck's imputation procedure, we refer to his original paper (Buck 1960) or to Chapter 3 of Little and Rubin (1987) or Chapter 2 of McLachlan and Krishnan (1997).

The numerical illustration of the algorithm is presented in Table 2. As we can see from the

Table 2: Selected iterations of the EM algorithm for data in Table 1.

Iteration	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	$y_{2,7}$	$y_{2,8}$	$y_{1,9}$	$y_{1,10}$	$-2Q$
1	0.6764	3.5068	1.8170	0.3868	2.0671	3.4399	3.6069	1.0443	1.1867	65.7704
5	0.8779	3.6433	1.8618	0.8671	2.2030	3.4030	3.7685	1.5982	1.8978	64.7568
10	0.9279	3.6327	1.9463	0.9837	2.1724	3.3466	3.7433	1.7614	2.1061	64.5587
20	0.9426	3.6293	1.9757	1.0181	2.1639	3.3301	3.7345	1.8102	2.1683	64.5079
30	0.9435	3.6291	1.9775	1.0202	2.1634	3.3291	3.7339	1.8132	2.1722	64.5048
35	0.9436	3.6291	1.9776	1.0203	2.1633	3.3290	3.7339	1.8134	2.1724	64.5047
40	0.9436	3.6291	1.9777	1.0204	2.1633	3.3290	3.7339	1.8134	2.1724	64.5046
45	0.9436	3.6291	1.9777	1.0204	2.1633	3.3290	3.7339	1.8134	2.1725	64.5046

table with the accuracy of up to three significant digits, the algorithm seems to converge after about 30 steps or so and the estimated or imputed values of (5) are given by

$$\mathcal{Y}^{(em)} = (3.329, 3.734, 1.813, 2.173).$$

Let us note, for the sake of comparison, that if we were to employ the standard, "naive" linear or polynomial regression model based on 6 complete observations in order to fit the missing values in Table 1 we would have obtained in this case

$$\mathcal{Y}^{(reg)} = (2.834, 3.063, 2.700, 3.269).$$

Both $\mathcal{Y}^{(em)}$ and $\mathcal{Y}^{(reg)}$ can be now compared with the actual values removed from Table 1 which were

$$\mathcal{Y} = (3.362, 3.657, 1.484, 3.410).$$

As we can see, in our example the EM method did reasonably well in recovering the missing values.

2.2 Massachusetts Auto Bodily Injury Liability Data. Fraud and Build-up Screening via Mixture Models

By now it is fairly well known that fraud and build-up, exaggerated injuries and/or excessive treatment, are key components of the auto injury loss distributions (Derrig et al. 1994, Cummins and Tennyson 1996, Abrahamse and Carroll 1999). Indeed, injury loss distributions are prime candidates for mixture modeling, for at least the differing of payment patterns by injury type. Even within an injury type as predominant as strain and sprain,[2] there can be substantial differences in subpopulations arising from fraud and build-up. One common method of identifying these claims has been to gather additional features of the claim, the so-called fraud indicators, and to build models to identify those bogus claims (Brockett, et al. 1998). The acquisition of reliable indicators some of which may be highly subjective, is costly, and may not be efficient in uncovering abusive patterns in injury claims (Crocker and Tennyson 1999). The use of more flexible methods such as the fuzzy logic (see more below) may overcome the lack of this precision in subjective features in an economically efficient manner by running a background algorithm on adjusters' electronic files (see, for example, Derrig and Ostaszewski 1995, 1999).

Another approach to uncovering fraud and build up, perhaps grounded more in practical considerations, is to construct a filter, or screening algorithm, for medical provider bills (Derrig 2002). Routinely, excessive medical bills can be reduced to "reasonable and customary" levels by computer algorithms that compare incoming bills to right censored billing distributions with "excessive" being operationally defined to be above the censoring point. Less routine is the implementation of systematic analysis of the patterns of *a provider's billing practices* (Major and Riedinger 1992). Our second application of the EM algorithm is to build a first level screening device to uncover potential abusive billing practices and the appropriate set of claims to review. We perform the pattern analysis by uncovering abusive-like distributions within mixture models parametrized by the estimates obtained via the EM algorithm. An illustration of the method follows.

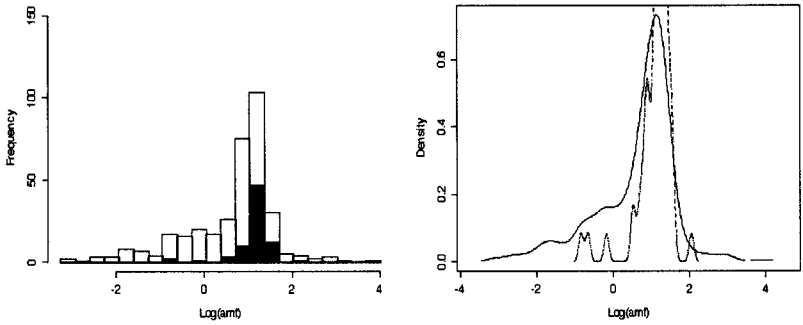
In the table provided in Appendix B we present a set of outpatient medical provider's total billings on the set of 348 auto bodily injury liability claims closed in Massachusetts during 2001. For illustration purposes, 76 claims with one "outlier" provider ("A") were chosen based on a pattern typical of abusive practice; namely, an empirical kurtosis more than five times the overall average. The "outlier" was then combined with medical bills in claims from a random sample of providers. The losses are recorded in thousands and are presented in column two. Column 4 identifies each medical billing amount as provider "A" or "other". We will use the EM algorithm applied to a normal (log) mixture model attempting to uncover provider A.

The relatively large volume of provider A's claims is clearly visible in the left panel of Figure 1, where it is presented as a portion of the overall claims

Whereas the volume of claims by itself never constitutes a basis for the suspicion of fraud or build-up, it certainly might warrant a closer look at the data at hand, especially via some type of

[2]Currently, Massachusetts insured bodily injury claims are upwards of 80 percent strain and sprain claims as the most costly part of the medical treatment. Of course, that may have a dependency on the $2,000 dollar threshold to file a tort claim.

Figure 1: Overall distribution of the 348 BI medical bill amounts from Appendix B compared with that submitted by provider A. Left panel: frequency histograms (provider A's histogram in filled bars). Right panel:density estimators (provider A's density in dashed line).



homogeneity analysis, since the second panel in Figure 1 clearly indicates the difference between the overall claims distribution and that of the provider A. Hence in this problem we shall be looking for a hidden exposure which could manifest itself as a non-homogenous component of the data, albeit we shall not be assuming that this component is necessarily due to provider A. In fact, as the initial inspection of the *overall* data distribution does not immediately indicate non-homogeneity we shall not make any prior assumptions about the nature or source of the possible non-homogeneity.

Since the standard analysis of the data by fitting a kernel density estimator (see the solid curve in the right panel of Figure 1) appears to give no definite indication of multimodality, it seems, that some more sophisticated methods are needed in order to identify any foreign components of the claims. Whereas many different approaches to this difficult problem are possible, we have chosen one that shall illustrate the applicability of the EM methodology in our setting. Namely, we shall attempt to fit a *log-mixture-normal* distribution to the data, that is, we shall model the logarithm of the claim outpatient medical billing distribution as a mixture of several normal variables. The use of normal distributions here is mostly due to convenience of the EM implementation and in more complicated real life problems can be inappropriate. However, the principle that we shall attempt to describe here is, in general, applicable to any mixture of distributions, even including non-parametric ones.[3]

[3]The notion of fitting non-parametric distributions via likelihood methods, which at first may seem contradiction in terms, has become very popular in statistics over the last decade. This is due to intensive research into the so called empirical likelihood methods (see for instance a recent monograph by Owen 2001 and references therein). In

In order to describe our method in the context of the EM algorithm we shall again relate the problem at hand to our EM methodology introduced in Section 1. In our current setting we shall consider the set of logarithms of the BI claim medical bills as the incomplete data $\mathcal{X}$. According to our model assumption we identify the underlying random variable X, of which $\mathcal{X}$ is a realization, as a mixture of several (say, $m \geq 2$) normal distributions[4]

$$X_i \sim N(\mu_j, \sigma_j) \quad \text{for} \quad j = 1, \ldots, m$$
$$X = \sum_{j=1}^{m} Y_j \cdot X_j, \tag{8}$$

where $Y_j \in \{0, 1\}$ with $P(Y_j = 1) = \pi_j$ such that $\sum \pi_j = 1$ and the joint distribution of the vector $(Y_1, \ldots, Y_m)$ is multinomial with one trial, (i.e., $\sum Y_j = 1$). The right hand side of (8) is sometimes known as *generative* representation of a mixture. Indeed, if we generate a multinomial variable $(Y_1, \ldots, Y_m)$ with probabilities of $Y_j = 1$ equal to π_j, and depending on the index j for which outcome is a unity, deliver X_j, then it can be shown that the density of X is

$$\sum_{j=1}^{m} \pi_j \, p(x|\Theta_j) \tag{9}$$

where $p(\cdot|\Theta_j)$ is a normal density with the parameter

$$\Theta_j = (\mu_j, \sigma_j) \quad \text{for} \quad j = 1, 2 \ldots, m$$

Hence X is indeed a mixture of the X_j's. The density given by (9) is less helpful in our approach as it doesn't explicitly involve the variables Y_j's. Moreover, fitting the set of parameters[5]

$$\Theta = (\Theta_1, \ldots, \Theta_m, \pi_1, \ldots, \pi_{m-1}). \tag{10}$$

by considering log-likelihood of (9) is known to be numerically difficult as it involves evaluation of the sums under the logarithm. In contrast, the representation (8) provides for a simpler approach, which also suits better our purpose of illustrating the use of the EM methodology. In the spirit of the search for hidden exposure, we consider the (unobserved) realizations of random vector $(Y_1, \ldots, Y_m)$ in (8) as the missing data $\mathcal{Y}$. Let us note that unlike in the example discussed in Section 2 here we have in some sense artificially created the set $\mathcal{Y}$. In this setting the complete set of data is now $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ or $\mathbf{z}_j = (x_j, y_{jk})$ for $j = 1, \ldots, n$, and $k = 1 \ldots, m$. Here $n = 348$ is the number of obervations, m is the number of components in the mixture, unspecified for now, x_j is (logarithm of) the observed medical bill value, and $y_{jk} \in \{0, 1\}$ is the auxiliary indicator variable

principle, with some modifications, the mixture approach discussed in this section and the associated EM algorithm can be applied to the empirical likelihood as well.

[4]Note that in our notation σ denotes the variance, not standard deviation.

[5]Note that we only need to estimate $m - 1$ proportions since $\sum \pi_i = 1$.

indicating whether or not x_j arrives from the distribution of X_k. In this setting the complete log-likelihood function (2) takes the form

$$l(\Theta|\mathcal{Z}) = \sum_{j=1}^{n}\sum_{k=1}^{m} y_{jk} \log p(x_j|\Theta_k), \tag{11}$$

and the conditional expectation (3) is given by

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{j=1}^{n}\sum_{k=1}^{m} \delta_{jk} \log p(x_j|\Theta_k), \tag{12}$$

where

$$\delta_{jk} = E(Y_{jk}|\Theta^{(i-1)}, \mathcal{Z}) = P(Y_{jk} = 1|\Theta^{(i-1)}, \mathcal{X}) \quad \text{for} \quad j = 1 \ldots, n; \quad k = 1, \ldots, m. \tag{13}$$

As we can see from the above formulae, in this particular case $Q(\Theta, \Theta^{(i-1)})$ is obtained from the complete data likelihood by substituting for the unknown y_{jk}'s their conditional expectations δ_{jk}'s calculated under the current value of the estimates of Θ.[6] The quantity δ_{jk} is often referred to as the *responsibility* of the component X_k for the observation j. This terminology reflects the fact that we may think about final δ_{jk} as the conditional (posterior) probability of the j-th observation arriving from the distribution of X_k.

Once we have replaced the y_{jk}'s in (11) by the δ_{jk}'s, the maximization step of the EM algorithm is straightforward and applied to (12) gives the usual weighted ML estimates of the normal means, variances, and the mixing proportions (see below for the formulae). However, in order to proceed with the EM procedure we still need to construct the initial guesses for the set of parameters (10). A good way to do so (for a discussion, see, for instance, Chapter 8 of Hastie et al. 2001 or Xu and Jordan 1996) is to simply choose at random m of the observed claim values as the initial estimates of the means, and set all the estimates of the variances to the overall sample variance. The mixing proportion can be set uniformly over all components. This way of initiating the parameters ensures the relative robustness of the final estimates obtained via EM against any particular initial conditions. In fact, in our BI data example we have randomly selected several initial sets of values for the means and in all case have obtained convergence to the same set of estimates. Below we present the detailed EM algorithm we have used to analyze the Massachusetts auto BI data. In order to identify the number m of the mixture components in the model we have used the EM method to obtain the estimates of the complete log-likelihood function (as the final values of (12)) for $m = 2, 3, 4$ (we have had determined earlier that for $m > 4$ the BI mixture model becomes too cumbersome). The results are presented in Table 3. As can be seen from the last row of the table, $m = 3$ is the number of components minimizing the negative of the estimated log-likelihood (12). Henceforth we shall, therefore, take $m = 3$ for the BI mixture model.

[6] It may happen that some of the values y_{jk} are in fact available. In such cases, we would take $\delta_{jk} = y_{jk}$.

Table 3: Comparison of the mixture fit for the different values of m for the BI data

Parameter	$m=2$	$m=3$	$m=4$
μ_1	0.071	0.107	-0.01
μ_2	1.110	0.874	0.218
μ_3	-	1.248	0.911
μ_4	-	-	1.258
$\sigma_1^{1/2}$	1.265	1.271	1.201
$\sigma_2^{1/2}$	0.252	0.178	1.349
$\sigma_3^{1/2}$	-	0.146	0.214
$\sigma_4^{1/2}$	-	-	0.144
π_1	0.470	0.481	0.250
π_2	0.530	0.205	0.224
π_3	-	0.314	0.247
π_4	-	-	0.279
$-2Q$	819.909	811.381	811.655

Table 4: Selected iterations of the EM algorithm for the BI data with $m=3$.

Iteration	μ_1	μ_2	μ_3	$\sigma_1^{1/2}$	$\sigma_2^{1/2}$	$\sigma_3^{1/2}$	π_1	π_2	π_3	$-2Q$
1	0.229	0.785	0.885	1.172	0.89	0.843	0.35	0.329	0.321	973.115
5	-0.129	0.946	1.054	1.374	0.525	0.356	0.337	0.301	0.361	854.456
6	-0.131	0.953	1.083	1.357	0.499	0.300	0.349	0.281	0.370	839.384
10	-0.041	0.917	1.137	1.324	0.456	0.223	0.396	0.217	0.387	820.903
20	0.042	0.875	1.166	1.302	0.364	0.207	0.438	0.177	0.385	817.363
30	0.064	0.876	1.184	1.29	0.301	0.200	0.453	0.176	0.372	816.143
40	0.074	0.871	1.204	1.285	0.259	0.188	0.460	0.186	0.354	814.957
50	0.084	0.868	1.226	1.281	0.222	0.17	0.467	0.197	0.336	813.367
60	0.099	0.871	1.243	1.275	0.190	0.153	0.476	0.204	0.320	811.838
64	0.105	0.873	1.247	1.272	0.180	0.147	0.48	0.205	0.315	811.454
65	0.107	0.874	1.248	1.271	0.178	0.146	0.481	0.205	0.314	811.381

Algorithm 2 *(The EM algorithm for fitting m-component normal mixture)*

1. *Define the initial estimate $\Theta^{(0)}$ of the set of parameters* (10) *(see discussion above).*
2. *The E-step: given the current value of $\Theta^{(i)}$ compute the responsibilities δ_j as*

$$\delta_{jk} = \frac{\pi_k^{(i)}\, p(x_j|\Theta_k^{(i)})}{\sum_{l=1}^{m} \pi_l^{(i)}\, p(x_j|\Theta_l^{(i)})} \qquad j = 1, \ldots, n \quad \textit{and} \quad k = 1, \ldots, m.$$

3. *The M-step: compute the ML estimators of* (12) *as*

$$\pi_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk}}{n} \quad \textit{for} \quad k = 1, \ldots, m-1,$$

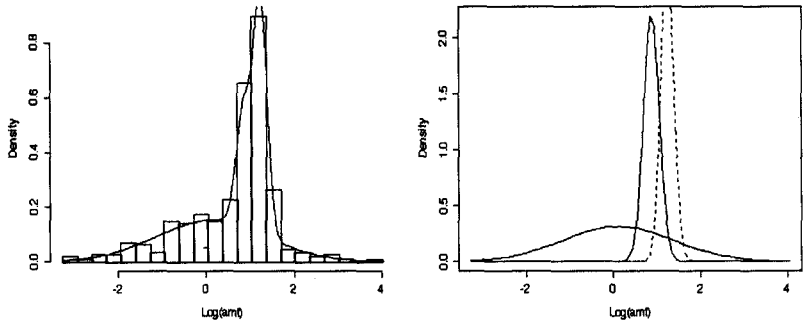
and

$$\mu_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk}\, x_j}{\sum_{j=1}^{n} \delta_{jk}}$$

$$\sigma_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk} \left(x_j - \mu_k^{(i+1)}\right)^2}{\sum_{j=1}^{n} \delta_{jk}} \quad \textit{for} \quad k = 1, \ldots, m.$$

4. *Repeat steps 2 and 3 until the relative difference of the subsequent values of* (12) *is sufficiently small.*

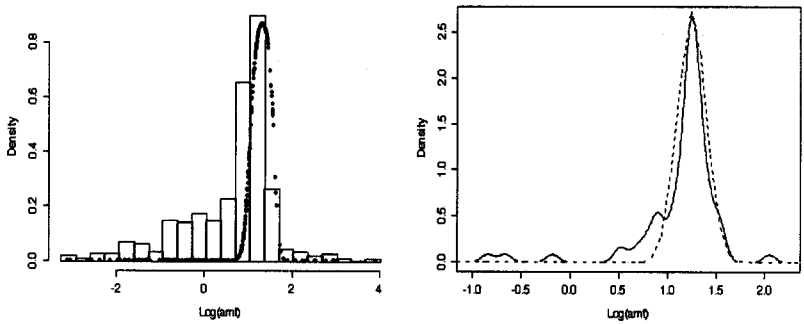
Figure 2: EM Fit. Left panel: mixture of normal distributions fitted via the EM algorithm to BI data. Right panel: Three normal components of the mixture. The values of all the parameters are given in the last row of Table 4.



In Figures 2 and 3 we present graphically the results of the analysis of the BI data via the

mixture model with $m = 3$ using the EM algorithm as described above. Some selected iterations of the EM algorithm for the three component normal mixture are presented in Table 4. In the left panel of Figure 2 we show the fit of the normal mixture fitted to the data using Algorithm 2 (with parameters values given by the last row of Table 4) . As we can see the fit looks reasonable and the fitted mixture distribution looks similar to the standard density estimator (solid curve in the right panel of Figure 1). The mixture components identified by the EM method are presented in the right panel of Figure 2 and clearly indicate non-homogeneity of the data which seems to consist of two (in fact, three) different types of claims. This is, obviously, related to a high volume of claims in the interval around 1.8–4.5 thousands (corresponding to the values .6–1.5 on the log scale). This feature of the data is modeled by the two tall and thin (i.e., with small dispersion) components of the mixture (corresponding in our notation to X_2 and X_3, marked as solid and dashed curves, respectively). Let us also note the very pronounced difference (over seven-fold) in the spread between the first and the two last components.

Figure 3: Latent risk in BI data modeled by the EM algorithm with $m = 3$. Left panel: set of the responsibilities δ_{j3}. Right panel: the third component of the normal mixture compared with the distribution of provider A's claims ("A" claims density estimator is a solid curve).



In the left panel of Figure 3 we present the set of responsibilities (δ_{j3}) of the model (or component) X_3 as calculated by the EM algorithm superimposed on the histogram of the BI data. The numerical values of the responsibilities for each data point are also listed in the last column of the table in Appendix B. The relationship between the set of responsibilities obtained via the EM procedure and the apparent lack of homogeneity of the data, demonstrated by Figure 2, is easy to see. The high responsibilities are clustered around the claim values within two standard deviations of the estimated mean (1.25) of the tallest distribution X_3. Hence the plot of responsibilities

superimposed on the data distribution again uncovers the non-homogeneity or the risk factor which was initially hidden. As we can see from the right panel in Figure 3 the observed non-homogeneity may be attributed largely, as initially expected (and as the illustration intended), to the high kurtosis of "A" claims. Indeed, the superimposing of the distribution of "A" claims (solid curve) on the component X_3 (dashed curve) in the right panel of Figure 3 reveals a reasonably close match in the interval (.8, 1.7) or so. Outside this interval the normal approximation to the provider A's claims fails, mostly due to the fact that the normal mixture model employed is not sufficiently "fine tuned" in its tails to look for this particular type of distribution. The deficiency could be perhaps rectified in this particular case by incorporating some different (non-normal) components into the mixture model. However, our main task in this analysis was to merely uncover hidden factors (if any) and not necessarily to model them precisely, which should be done afterwards using some different, more sophisticated modeling approach depending on the type of problem at hand. See, for instance, Bilmes, (1998) who presents the extension of our Algorithm 2 to the so-called general hidden Markov model (HMM). For a full review of some possible approaches to fitting the finite mixtures models and the use of the EM methodology in this context, readers are referred to the recent monograph by McLachlan and Peel (2000) which also contains some descriptions of the currently available software for fitting a variety of non-normal mixtures via the EM method.

2.3 The EM Algorithm Output and Fuzzy Set Membership Function

As we have seen above, each run of the EM algorithm estimating an m-mixture model will produce responsibilities for each claim and for each one of the m mixture distributions. As mentioned earlier, they can be interpreted as the (posterior) probability that the claim "arises" from each of the components of the mixing distributions. They also can be interpreted as the membership functions for the fuzzy sets of "arising from the i-th mixture component". If for any claim the responsibility (membership) of a particular model component equals one, we say that the claim arises from that model component. When the responsibility is less than one, the claim arises partially from that component, and if the responsibility equals zero, we can say the claim does not arise from that component. In that context, every claim "belongs to" each of the mixing component with measurement value equal to the responsibility. Putting the EM algorithm within the fuzzy set context provides us with the well-known tools of fuzzy arithmetic to help interpret the EM output in a way that matches real-life actuarial choices (c.f., e.g., Derrig and Ostaszewski 1999).

Another advantage of portraying the responsibility probabilities as fuzzy sets relations is that the defuzzification operator known as the α-cut[7], can be used to illustrate the type I and II errors when the α-cut criterion is used to classify the claim as belonging to one of the mixture distributions. The α-cut classification table is presented in Table 5 below and shows the portions of "A" claims contained in each α-cut from 0.1 to 0.9 for each mixture component distribution. In particular, the α-cut analysis confirms our previous findings that "A" claims belong predominantly to the third

[7]For α equal to a number between zero and one, the α-cut of a fuzzy set consists of the (crisp) set of all elements that have a membership value greater than or equal to α (see, Derrig and Ostaszewski 1995).

mixing distribution (i.e., distribution of X_3). Indeed, the α-cut at about 0.5 provides us with a good indication that "A" arises from the third mixing distribution (corresponding to the value 75% in the table) but not from the first one (corresponding to 8% value only). These findings are consistent with those illustrated by Figure 3. In contrast, the second mixing distribution (distribution of X_2) does not allow us to classify correctly "A" and "other" in our three-mixture model. The low proportion of "A" claims assigned to the model X_2 indicates that they are generally unlikely to arrive from X_2 which may be an indication of some further non-homogeneity among claims, even after adjusting for the type "A". The X_2 component could be, therefore, the manifestation of some additional hidden factors, which again confirms the findings summarized in the previous section.[8]

Table 5: Fuzzy membership via responsibility probabilities

α	Resp. X_1		Resp. X_2		Resp. X_3	
	A	Other	A	Other	A	Other
0.9	0.05	0.42	0.00	0.00	0.00	0.00
0.8	0.05	0.42	0.00	0.00	0.54	0.13
0.7	0.07	0.45	0.09	0.11	0.62	0.19
0.6	0.08	0.46	0.13	0.20	0.70	0.22
0.5	**0.08**	**0.48**	**0.13**	**0.24**	**0.75**	**0.24**
0.4	0.11	0.49	0.16	0.28	0.78	0.26
0.3	0.16	0.54	0.21	0.30	0.79	0.30
0.2	0.22	0.64	0.24	0.35	0.79	0.34
0.1	0.95	1.00	0.33	0.41	0.82	0.38

2.4 Accuracy Assessment for the EM Output via Parametric Bootstrap

In our analysis of the BI data conducted in the previous sections we have used the numeric values of the estimated parameters (10) and the responsibilities (13). Since these values were estimated from the data via the EM algorithm, it is important to learn about their accuracy. In general, for the set of parameters (10) the usual approach to assessing accuracy based on the asymptotic normality of the maximum likelihood estimators can be applied here, as soon as we calculate the information matrix for Θ. This is slightly more complicated for the set of responsibilities (13) as they are the functions of Θ and hence require the appropriate transformation of the information matrix. However, a simpler method of obtaining, for instance, confidence intervals for the set of responsibilities and the model parameters can be also used, based on the so-called *parametric*

[8] An analysis of the mixture model applied only to 272 "other" claims shows that X_2 has a more pronounced representation (high α-cut proportions) of (i) chiropractic and physical therapy treatment, (ii) special investigations and independent medical examinations, and (iii) extended treatment delays.

bootstrap method outlined in Algorithm 3 below. The method can be shown to be asymptotically equivalent to the normal approximation approach and is known to be often more reliable for smaller sample sizes or for the heavily biased estimators (which will often be the case for the responsibilities (13)). The algorithm below describes how to obtain confidence intervals for the parameters given by (10) and (13) using bootstrap. For some more examples and further discussion see, for instance, McLachlan and Peel (2000) or the forthcoming paper by Rempala and Szatzschneider (2002) where also the issue of the hypothesis testing for the number of mixture components via the parametric bootstrap method is discussed.

Algorithm 3 *(Bootstrap confidence intervals)*

1 *Using the values of the model parameters* (10) *obtained from the EM algorithm generate the set of pseudo-data* $\mathcal{X}^*$ *(typically of the same length as the original data* $\mathcal{X}$*).*

2 *With* $\mathcal{X}^*$ *at hand, use Algorithm 2 in order to obtain a set of pseudo-values* Θ^**.*

3 *Using the set of the original data values* $\mathcal{X}$ *and* Θ^* *from step 2 above, calculate the pseudo-responsibilities* δ^*_{jk} *as in Algorithm 2 step 2.*

4 *Repeat the steps 1–3 a large number of times, say,B.* [9]

5 *Use the empirical quantiles of the distributions of pseudo-values* Θ^* *and* δ^*_{jk} *to obtain confidence bounds for* Θ *and* δ_{jk}*.*

For illustration purpose we present the set of confidence intervals for the three-mixture-normal model parameters and the responsibilities (of X_3) obtained via the above algorithm for the BI data in Tables 6 and 7 below. The term "bootstrap estimate" in the tables refers to the average value of the B bootstrap pseudo-values obtained in steps 2 or 3.

3 Summary and Conclusion

This paper has introduced the statistical methodology for inference in the presence of missing data, known as the EM algorithm, into the actuarial settings. We have shown that this methodology is particularly appropriate for those practical situations which require consideration of the missing or incomplete data, the "lurking" variables, or the hidden factors. We believe that due to its conceptual simplicity, the EM method could become a standard tool of actuarial analysis in the future. Herein we have given only some example of its usefulness in modeling loss severity. Specifically, in modeling claim severities, the EM algorithm was used to impute missing values in a more sophisticated and statistically less biased way than simple regression methods as well as to uncover (hidden) patterns in the claim severity data. Actual auto bodily injury liability claims closed in Massachusetts in 2001 were used to illustrate a first stage screen for abusive medical providers, and

[9] In our setting B needs to be fairly large, typically at least a thousand. For a discussion see, for instance, McLachlan and Peel (2001).

Table 6: Accuracy of the parameter estimates for the BI data with B=1000

Parameter	Value	Bootstrap Estimate	95% CI
μ_1	0.107	0.104	(-0.115, 0.298)
μ_2	0.874	0.871	(0.809, 0.924)
μ_3	1.248	1.249	(1.216, 1.284)
$\sigma_1^{1/2}$	1.271	1.269	(1.132, 1.389)
$\sigma_2^{1/2}$	0.178	0.175	(0.125, 0.222)
$\sigma_3^{1/2}$	0.146	0.144	(0.117, 0.174)
π_2	0.205	0.207	(0.157, 0.253)
π_3	0.314	0.317	(0.268, 0.375)

Table 7: Accuracy of the selected responsibilities δ_{j3}

No (j)	Log Claim Value	δ_{j3} Value	Bootstrap Estimate	95% CI
100	0.380	0.000	0.000	(3.90e-12, 2.04e-06)
200	1.031	0.410	0.396	(0.243, 0.531)
300	1.353	0.854	0.863	(0.802 0.912)

their abusive claims, utilizing the EM algorithm. The usefulness of the EM output for classification purpose and its connections with fuzzy logic techniques were discussed. Namely, the EM algorithm output of posterior probabilities called responsibilities were reinterpreted as fuzzy set membership function in order to bring the machery of fuzzy logic to bear in the classification problem. The Monte-Carlo based method of assessing the accuracy of the model parameters fitted via the EM algorithm, known as the parametric bootstrap was also presented and the appropriate algorithm for its implementation was developed. The set of functions written in the statistical language R, implementing the EM algorithms discussed in the paper, have been included in Appendix A to allow readers to try different actuarial situations where missing data and hidden components might be found. A large variety of actuarial and financial applications of the presented methodology are possible, including its incorporation into models of Risk Based Capital, Value at Risk, and general Dynamic Financial Analysis. We hope that this paper shall promote enough interest in the EM methodology for further exploration of those opportunities.

Acknowledgement. The authors wish to acknowledge Dr Krzysztof M. Ostaszewski, FSA for his encouragements and helpful comments at the initial stages of this project.

References

Abrahamse, Alan F. and Stephan J. Carroll (1999). The Frequency of Excess Claims for Automobile Personal Injuries, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation,* Dionne, Georges and Claire Laberge-Nadeau, Eds., Kluwer Academic Publishers, pp. 131-150.

Bilmes, Jeff A. (1998). *Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, International Computer Science Institute. UC Berkeley.

Brockett, Patrick L., Xiaohua Xia and Richard A. Derrig (1998). Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud, *Journal of Risk and Insurance*, June, Volume 65, No. 2.

Campbell, John W., Andrew Y. Lo, and Archie Craig MacKinlay (1996). *The Econometrics of Financial Markets.* Princeton University Press.

Crocker, Keith J., and Sharon Tennyson (1999). Costly State Falsification or Verification? Theory and Evidence from Bodily Injury Liability Claims, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation,* Dionne, Georges and Claire Laberge-Nadeau, Eds., Kluwer Academic Publishers, pp.119-131.

Cummins, J. David, and Sharon Tennyson (1996). Controlling Automobile Insurance Costs, *Journal of Economic Perspectives*, Spring, Volume 6, No. 2. pp. 95–115.

D'Arcy, S. P.,Gorvett, R. W., Herbers, J. A., and Hettinger, T. E. (1997). Building a Dynamic Financial Analysis Model that Flies. *Contingencies* November/December, 40–45.

Dempster, Allan P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data using EM algorithm, *Journal of Royal Statistical Society Series B*, Volume 39, No. 1. pp. 1–38.

Derrig, Richard A. (2002). Insurance Fraud. *Journal of Risk and Insurace*, Volume 69, No. 3. pp. 271–287.

Derrig, Richard A. and K.M. Ostaszewski (1999). Fuzzy Sets Methodologies in Actuarial Science, *Practical Applications of Fuzzy Technologies*, Hans-Jurgen Zimmerman Eds, Kluwer Academic Publishers, Boston, (November).

Derrig, Richard A., K.M. Ostaszewski, and G.A. Rempala (2001). Applications of Resampling Methods in Actuarial Practice, *Proceedings of the Casualty Actuarial Society*, Volume LXXXVII, pp. 322–364.

Derrig, Richard A., and K.M. Ostaszewski (1995). Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification, *Journal of Risk and Insurance*, Volume 62, No. 3. pp. 447–482.

Derrig, Richard A., Herbert I. Weisberg and Xiu Chen (1994). Behavioral Factors and Lotteries Under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims", *Journal of Risk and Insurance*, June, Volume 61, No. 2. pp. 245–275.

Embrechts, Paul, Alexander McNeil and Daniel Straumann (2000). Correlation and Dependence in Risk Management: Properties and Pitfalls, in *Extremes and Integrated Risk Management*, Paul Embrechts, Ed. pp 71–76. Risk Books. London.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2001). *The Elements of Statistical Learning*, Springer-Verlag New York, Inc., New York.

Little, Roderick J.A., and Donald B. Rubin (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., Canada.

McLachlan, Geoffrey and David Peel (2000). *Finite mixture models*, Wiley-Interscience, New York.

McLachlan, Geoffrey and Thriyambakam Krishnan (1997). *The EM algorithm and extensions*, Wiley, New York.

Major, John A., and Dan R. Riedinger (1992). EFD: A Hybrid Knowledge/Statistical -Based System for the Detection of Fraud, International Journal of Intelligent Systems. Volume 7, pp. 687–703 (reprinted Journal of Risk and Insurance, Vol 69, No.3, pp 309–324 September, 2002).

Owen, Art, B. (2001). *Empirical Likelihood*. Chapman and Hall. New York.

Rempala, Grzegorz A. and Konrad Szatzschneider (2002). Bootstrapping Parametric Models of Mortality. Manuscript, to appear in *Scandinavian Actuarial Journal*.

Wang, Shaun (1998). Aggregation of Correlated Risk Portfolios: Models and Algorithms *Proceedings of the Casualty Actuarial Society*, Volume LXXXIII, pp 848–939

Wu, Jeff A (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, Volume 11, No. 1. pp. 95–103.

Xu, Li and Mike I Jordan (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural computation*, Volume 8, pp. 129–151.

Appendix A. R Functions

We present here the implementation of Algorithms 1 and 2 in statistical software R which is a freeware version of the award winning statistical software S+ and is available from http://www.r-project.org. The functions below were used in the numerical examples discussed in the text.

#Algorithm 1: EM version of Buck's imputation procedure####################

```
#auxiliary function
inv<-function(m,...)  solve(m,diag(rep(1,length(m[,1]))),...);
#defining matrix inverse (for compatibility with older versions of R)
#input parameters
# d -dataframe of two columns containing complete observations
# d1-list of observations with missing second coordinate
# d2-list of observations with missing first coordinate
# B-maximal number of iterations (default value 500)
# eps- convergence criterion (default value .0001)
#########################################################################

em.buck <-function(d,d1,d2,B=500,eps=.0001) {
n<-length(d[,1]);
n1<-length(d1);
n2<-length(d2);
m<-apply(d,2,mean);
R<-cov(d);
rho<-cor(d)[1,2]; nLL.old<-eps; nLL.new<-100;
w<-rbind(d,cbind(d1,rep(m[2],n1)),cbind(rep(m[1],n2),d2));
i<-1; #mainloop#
while (abs(nLL.new-nLL.old)/nLL.old>eps && i<=B){
T1<-sum(w[,1]); T2<-sum(w[,2]); T12<-sum(w[,1]*w[,2]);
T11<-sum(w[(n+n1+1):(n+n1+n2),1]^2
+R[1,1]*(1-rho^2))+sum(w[-((n+n1+1):(n+n1+n2)),1]^2);
T22<-sum(w[(n+1):(n+n1),2]^2+R[2,2]*(1-rho^2))+sum(w[-((n+1):(n+n1)),2]^2);
R<-array(c(T11-T1^2/(n+n1+n2),T12-T1*T2/(n+n1+n2),T12-T1*T2/(n+n1+n2),
T22-T2^2/(n+n1+n2))/(n+n1+n2), c(2,2));
m<-c(T1/(n+n1+n2),T2/(n+n1+n2));
rho<-R[1,2]/sqrt(R[1,1]*R[2,2]);
w[(n+1):(n+n1),2]<-m[2]+R[1,2]*(w[(n+1):(n+n1),1]-m[1])/R[1,1];
w[(n+n1+1):(n+n1+n2),1]<-m[1]+R[1,2]*(w[(n+n1+1):(n+n1+n2),2]-m[2])/R[2,2];
nLL.old<-nLL.new;
s<-0; for (k in 1:(n+n1+n2)) s<-(w[k,]-m)%*%inv(R)%*%(w[k,]-m)+s;
nLL.new<-2*(n+n1+n2)*log(2*pi)+s+(n+n1+n2)*log(abs(det(R)));
i<-i+1; }; #end mainloop#
print(paste("n=", n, n1, n2, "Theta estimates=", m[1],m[2], R[1,1],
R[1,2], R[2,2],"iter=",i-1,"-2LL=",nLL.new,"rho=",rho))
return(list(m=m, R=R,iter=i-1,LL=nLL.new,w=w)); }
#output parameters:  list of objects (m,R,iter,LL,w)
```

```
# m -vector of estimated means
# R -estimated covariance matrix
# iter -number of iterations until convergence
# w-concatenated dataframe of d,d1,d2 along with imputed missing values

# Algorithm 2: EM for normal mixtures ##############################
#auxiliary function
lsum<-function(a,p.new,m,s){ k<-length(m); ss<-0;
for (i in 1:k) ss<-ss+p.new[[i]]*dnorm(a,m[[i]],s[[i]]);
return(ss)}
# facilitates calculation of LL in the main procedure below

# input parameters:
# a -any list of numeric data
# pi -initial estimate of mixing proportions (default value:  uniform over three components)
# eps -desired convergence accuracy (default value .0001)
# B -maximal number of iterations allowed (default value 100)
# m -initial values of means estimates (default value:  random selection from a)
#####################################################################
em.multnorm<-function(a, pi=c(1/3,1/3,1/3),eps=.0001,B=100,m=sort(sample(a,3)))
{n<-length(a); k<-length(m); s<-rep(sd(a),k);
i<-1; p.new<-pi;
m0<-m;
logl.old<-1;
logl.new<-sum(log(lsum(a,p.new,m,s)));
#mainloop#
while (abs((logl.new-logl.old)/logl.old)>eps && i<=B)
{g<-NULL;
for (t in 1:k) g<-rbind(g, p.new[[t]]*dnorm(a,m[[t]],s[[t]])/lsum(a,p.new,m,s));
m<-g%*%a/g%*%rep(1,n);
s<-sqrt(g%*%a^2/g%*%rep(1,n)-m^2);
p.old<-p.new; p.new<-g%*%rep(1,n)/n; i<-i+1;
logl.old<-logl.new;
logl.new<-sum(log(lsum(a,p.new,m,s)));
};
#end mainloop#
print(paste("Theta estimates",m,s,"pi=",p.new,"iter=",i-1,"-2LL=",-2*logl.new));
return(list(m=m,s=s,pi=p.new,iter=i-1,start=m0,logl=-2*logl.new,resp=t(g),data=a)) }

# output parameters:  list of objects (m,s,pi,iter,logl,resp)
# m - vector of estimated values of means
# s - vector of estimated values of standard deviations
# pi -vector of estimated value of mixing proportions
# iter- number of iterations until convergence
# logl- final value of -2Q
# resp- matrix of responsibilities (columns correspond to mixture components)
```

Appendix B. Massachusetts Auto Insurance Bodily Injury Liability Data

Below we present the set of Auto Insurance Data discussed in the paper. Medical bill claim amounts are given in thousands. Responsibilities δ_{j3} are calculated according to Algorithm 2.

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
1	0.045	-3.101	Other	0.00	2	0.047	-3.058	Other	0.00
3	0.07	-2.659	Other	0.00	4	0.075	-2.590	Other	0.00
5	0.077	-2.564	Other	0.00	6	0.092	-2.386	Other	0.00
7	0.117	-2.146	Other	0.00	8	0.117	-2.146	Other	0.00
9	0.14	-1.966	Other	0.00	10	0.145	-1.931	Other	0.00
11	0.149	-1.904	Other	0.00	12	0.165	-1.802	Other	0.00
13	0.167	-1.790	Other	0.00	14	0.169	-1.778	Other	0.00
15	0.18	-1.715	Other	0.00	16	0.18	-1.715	Other	0.00
17	0.199	-1.614	Other	0.00	18	0.202	-1.599	Other	0.00
19	0.212	-1.551	Other	0.00	20	0.225	-1.492	Other	0.00
21	0.23	-1.470	Other	0.00	22	0.242	-1.419	Other	0.00
23	0.264	-1.332	Other	0.00	24	0.275	-1.291	Other	0.00
25	0.285	-1.255	Other	0.00	26	0.29	-1.238	Other	0.00
27	0.363	-1.013	Other	0.00	28	0.384	-0.957	Other	0.00
29	0.4	-0.916	Other	0.00	30	0.4	-0.916	Other	0.00
31	0.413	-0.884	Other	0.00	32	0.414	-0.882	Other	0.00
33	0.416	-0.877	Other	0.00	34	0.425	-0.856	Other	0.00
35	0.425	-0.856	Other	0.00	36	0.43	-0.844	Other	0.00
37	0.43	-0.844	A	0.00	38	0.431	-0.842	Other	0.00
39	0.45	-0.799	Other	0.00	40	0.46	-0.777	Other	0.00
41	0.486	-0.722	Other	0.00	42	0.5	-0.693	Other	0.00
43	0.5	-0.693	Other	0.00	44	0.514	-0.666	A	0.00
45	0.531	-0.633	Other	0.00	46	0.54	-0.616	Other	0.00
47	0.556	-0.587	Other	0.00	48	0.564	-0.573	Other	0.00
49	0.6	-0.511	Other	0.00	50	0.605	-0.503	Other	0.00
51	0.605	-0.503	Other	0.00	52	0.65	-0.431	Other	0.00
53	0.66	-0.416	Other	0.00	54	0.66	-0.416	Other	0.00
55	0.685	-0.378	Other	0.00	56	0.69	-0.371	Other	0.00
57	0.698	-0.360	Other	0.00	58	0.7	-0.357	Other	0.00
59	0.705	-0.350	Other	0.00	60	0.725	-0.322	Other	0.00
61	0.74	-0.301	Other	0.00	62	0.75	-0.288	Other	0.00
63	0.78	-0.248	Other	0.00	64	0.785	-0.242	Other	0.00
65	0.785	-0.242	Other	0.00	66	0.806	-0.216	Other	0.00
67	0.825	-0.192	Other	0.00	68	0.825	-0.192	Other	0.00
69	0.83	-0.186	Other	0.00	70	0.836	-0.179	A	0.00
71	0.87	-0.139	Other	0.00	72	0.9	-0.105	Other	0.00
73	0.934	-0.068	Other	0.00	74	0.95	-0.051	Other	0.00
75	0.954	-0.047	Other	0.00	76	0.956	-0.045	Other	0.00
77	0.962	-0.039	Other	0.00	78	0.97	-0.030	Other	0.00
79	0.975	-0.025	Other	0.00	80	0.988	-0.012	Other	0.00

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
81	1.015	0.015	Other	0.00	82	1.053	0.052	Other	0.00
83	1.058	0.056	Other	0.00	84	1.08	0.077	Other	0.00
85	1.161	0.149	Other	0.00	86	1.167	0.154	Other	0.00
87	1.195	0.178	Other	0.00	88	1.215	0.195	Other	0.00
89	1.242	0.217	Other	0.00	90	1.26	0.231	Other	0.00
91	1.295	0.259	Other	0.00	92	1.31	0.270	Other	0.00
93	1.319	0.277	Other	0.00	94	1.33	0.285	Other	0.00
95	1.34	0.293	Other	0.00	96	1.355	0.304	Other	0.00
97	1.39	0.329	Other	0.00	98	1.444	0.367	Other	0.00
99	1.455	0.375	Other	0.00	100	1.463	0.380	Other	0.00
101	1.49	0.399	Other	0.00	102	1.5	0.405	Other	0.00
103	1.542	0.433	Other	0.00	104	1.598	0.469	Other	0.00
105	1.616	0.480	Other	0.00	106	1.623	0.484	Other	0.00
107	1.64	0.495	Other	0.00	108	1.645	0.498	A	0.00
109	1.65	0.501	Other	0.00	110	1.66	0.507	Other	0.00
111	1.68	0.519	Other	0.00	112	1.695	0.528	Other	0.00
113	1.7	0.531	A	0.00	114	1.758	0.564	Other	0.00
115	1.759	0.565	Other	0.00	116	1.76	0.565	Other	0.00
117	1.896	0.640	Other	0.00	118	1.92	0.652	Other	0.00
119	1.923	0.654	Other	0.00	120	1.941	0.663	A	0.00
121	1.96	0.673	Other	0.00	122	1.972	0.679	Other	0.00
123	1.99	0.688	Other	0.00	124	2.005	0.696	Other	0.00
125	2.018	0.702	Other	0.00	126	2.02	0.703	Other	0.00
127	2.02	0.703	Other	0.00	128	2.03	0.708	Other	0.00
129	2.042	0.714	A	0.00	130	2.062	0.724	Other	0.00
131	2.063	0.724	Other	0.00	132	2.08	0.732	Other	0.00
133	2.087	0.736	Other	0.00	134	2.089	0.737	Other	0.00
135	2.1	0.742	Other	0.00	136	2.115	0.749	Other	0.01
137	2.12	0.751	Other	0.01	138	2.155	0.768	Other	0.01
139	2.159	0.770	Other	0.01	140	2.161	0.771	Other	0.01
141	2.184	0.781	A	0.01	142	2.188	0.783	Other	0.01
143	2.191	0.784	Other	0.01	144	2.196	0.787	Other	0.01
145	2.224	0.799	Other	0.02	146	2.237	0.805	Other	0.02
147	2.251	0.811	A	0.02	148	2.253	0.812	Other	0.02
149	2.288	0.828	Other	0.03	150	2.295	0.831	Other	0.03
151	2.318	0.841	Other	0.03	152	2.325	0.844	Other	0.03
153	2.325	0.844	A	0.03	154	2.335	0.848	Other	0.04
155	2.341	0.851	Other	0.04	156	2.35	0.854	Other	0.04
157	2.374	0.865	Other	0.05	158	2.39	0.871	Other	0.05
159	2.406	0.878	Other	0.06	160	2.434	0.890	Other	0.07
161	2.45	0.896	Other	0.08	162	2.453	0.897	A	0.08
163	2.468	0.903	Other	0.09	164	2.468	0.903	A	0.09
165	2.48	0.908	Other	0.10	166	2.48	0.908	A	0.10
167	2.49	0.912	A	0.10	168	2.498	0.915	Other	0.11
169	2.5	0.916	Other	0.11	170	2.5	0.916	Other	0.11
171	2.5	0.916	A	0.11	172	2.51	0.920	Other	0.11
173	2.532	0.929	Other	0.13	174	2.54	0.932	Other	0.14

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
175	2.543	0.933	Other	0.14	176	2.559	0.940	Other	0.15
177	2.572	0.945	Other	0.16	178	2.593	0.953	Other	0.18
179	2.601	0.956	Other	0.19	180	2.616	0.962	Other	0.20
181	2.619	0.963	Other	0.20	182	2.63	0.967	Other	0.21
183	2.635	0.969	Other	0.22	184	2.635	0.969	Other	0.22
185	2.653	0.976	Other	0.24	186	2.655	0.976	Other	0.24
187	2.675	0.984	Other	0.26	188	2.679	0.985	Other	0.26
189	2.697	0.992	Other	0.28	190	2.718	1.000	Other	0.31
191	2.73	1.004	Other	0.32	192	2.734	1.006	Other	0.32
193	2.755	1.013	Other	0.35	194	2.758	1.015	Other	0.35
195	2.773	1.020	Other	0.37	196	2.775	1.021	Other	0.37
197	2.78	1.022	Other	0.38	198	2.785	1.024	A	0.38
199	2.795	1.028	Other	0.40	200	2.805	1.031	Other	0.41
201	2.805	1.031	Other	0.41	202	2.808	1.032	A	0.41
203	2.88	1.058	Other	0.49	204	2.881	1.058	Other	0.50
205	2.881	1.058	A	0.50	206	2.924	1.073	A	0.54
207	2.93	1.075	Other	0.55	208	2.934	1.076	A	0.55
209	2.94	1.078	Other	0.56	210	2.972	1.089	Other	0.59
211	2.975	1.090	Other	0.59	212	3	1.099	Other	0.62
213	3	1.099	A	0.62	214	3.025	1.107	Other	0.64
215	3.058	1.118	Other	0.67	216	3.082	1.126	A	0.68
217	3.085	1.127	Other	0.69	218	3.095	1.130	Other	0.69
219	3.1	1.131	Other	0.70	220	3.102	1.132	A	0.70
221	3.106	1.133	Other	0.70	222	3.135	1.143	Other	0.72
223	3.17	1.154	Other	0.74	224	3.187	1.159	Other	0.75
225	3.192	1.161	A	0.75	226	3.193	1.161	Other	0.75
227	3.2	1.163	Other	0.76	228	3.21	1.166	Other	0.76
229	3.23	1.172	Other	0.77	230	3.23	1.172	Other	0.77
231	3.23	1.172	A	0.77	232	3.232	1.173	Other	0.77
233	3.235	1.174	Other	0.78	234	3.243	1.176	A	0.78
235	3.248	1.178	A	0.78	236	3.249	1.178	Other	0.78
237	3.26	1.182	Other	0.79	238	3.261	1.182	Other	0.79
239	3.272	1.185	A	0.79	240	3.29	1.191	Other	0.80
241	3.295	1.192	Other	0.80	242	3.304	1.195	Other	0.80
243	3.332	1.204	A	0.81	244	3.333	1.204	Other	0.81
245	3.338	1.205	Other	0.81	246	3.34	1.206	Other	0.82
247	3.341	1.206	A	0.82	248	3.349	1.209	A	0.82
249	3.349	1.209	A	0.82	250	3.349	1.209	A	0.82
251	3.353	1.210	A	0.82	252	3.36	1.212	Other	0.82
253	3.378	1.217	A	0.83	254	3.385	1.219	A	0.83
255	3.387	1.220	A	0.83	256	3.416	1.228	Other	0.84
257	3.429	1.232	A	0.84	258	3.438	1.235	A	0.84
259	3.444	1.237	A	0.84	260	3.469	1.244	A	0.85
261	3.473	1.245	A	0.85	262	3.473	1.245	A	0.85
263	3.475	1.246	A	0.85	264	3.477	1.246	A	0.85
265	3.505	1.254	Other	0.85	266	3.517	1.258	A	0.85
267	3.518	1.258	Other	0.85	268	3.527	1.260	A	0.85

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
269	3.535	1.263	A	0.86	270	3.547	1.266	A	0.86
271	3.55	1.267	Other	0.86	272	3.552	1.268	Other	0.86
273	3.567	1.272	A	0.86	274	3.57	1.273	Other	0.86
275	3.575	1.274	Other	0.86	276	3.58	1.275	Other	0.86
277	3.583	1.276	A	0.86	278	3.59	1.278	A	0.86
279	3.603	1.282	A	0.86	280	3.615	1.285	A	0.86
281	3.623	1.287	A	0.86	282	3.647	1.294	A	0.86
283	3.655	1.296	Other	0.86	284	3.655	1.296	A	0.86
285	3.658	1.297	Other	0.87	286	3.675	1.302	Other	0.87
287	3.675	1.302	Other	0.87	288	3.687	1.305	A	0.87
289	3.72	1.314	Other	0.87	290	3.72	1.314	Other	0.87
291	3.742	1.320	Other	0.87	292	3.757	1.324	A	0.87
293	3.765	1.326	Other	0.87	294	3.8	1.335	A	0.87
295	3.809	1.337	Other	0.86	296	3.848	1.348	A	0.86
297	3.857	1.350	A	0.86	298	3.867	1.352	Other	0.86
299	3.867	1.352	A	0.86	300	3.87	1.353	Other	0.86
301	3.883	1.357	Other	0.86	302	3.89	1.358	Other	0.86
303	3.905	1.362	A	0.86	304	3.907	1.363	A	0.86
305	4	1.386	Other	0.85	306	4.011	1.389	Other	0.85
307	4.039	1.396	A	0.84	308	4.065	1.402	A	0.84
309	4.095	1.410	Other	0.83	310	4.134	1.419	Other	0.82
311	4.147	1.422	Other	0.82	312	4.155	1.424	A	0.82
313	4.17	1.428	Other	0.81	314	4.179	1.430	A	0.81
315	4.2	1.435	Other	0.81	316	4.215	1.439	Other	0.80
317	4.257	1.449	A	0.79	318	4.3	1.459	Other	0.78
319	4.489	1.502	A	0.70	320	4.593	1.525	A	0.64
321	4.595	1.525	Other	0.64	322	4.63	1.533	A	0.62
323	4.653	1.538	Other	0.60	324	4.7	1.548	A	0.57
325	4.731	1.554	Other	0.55	326	4.741	1.556	A	0.55
327	4.75	1.558	Other	0.54	328	4.761	1.560	Other	0.53
329	4.81	1.571	Other	0.50	330	5.072	1.624	Other	0.31
331	5.161	1.641	Other	0.25	332	5.24	1.656	Other	0.20
333	5.64	1.730	Other	0.06	334	5.779	1.754	Other	0.03
335	6.166	1.819	Other	0.01	336	6.406	1.857	Other	0.00
337	6.725	1.906	Other	0.00	338	7.717	2.043	A	0.00
339	8	2.079	Other	0.00	340	9.5	2.251	Other	0.00
341	10.295	2.332	Other	0.00	342	12.533	2.528	Other	0.00
343	12.688	2.541	Other	0.00	344	16.043	2.775	Other	0.00
345	18.847	2.936	Other	0.00	346	19.5	2.970	Other	0.00
347	20.827	3.036	Other	0.00	348	50	3.912	Other	0.00