12-1-2006

GAMMA SHAPE MIXTURES FOR HEAVY-TAILED DISTRIBUTIONS

Sergio Venturini
Universita Bocconi, Milan Italy, sergio.venturini@unibocconi.it

Francesca Dominici
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Gamma Shape Mixtures for Heavy-tailed Distributions

S. Venturini[1], F. Dominici[2], G. Parmigiani[3]

December 1, 2006

[1] *Università Bocconi, Milan, Italy;* [2] *Johns Hopkins Bloomberg School of Public Health, Baltimore, USA;* [3] *Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins*

Abstract

An important question in health services research is the estimation of the proportion of medical expenditures that exceed a given threshold. Typically, medical expenditures present highly skewed, heavy tailed distributions, for which a) simple variable transformations are insufficient to achieve a tractable low- dimensional parametric form and b) nonparametric methods are not efficient in estimating exceedance probabilities for large thresholds. Motivated by this context, in this paper we propose a general Bayesian approach for the estimation of tail probabilities of heavy-tailed distributions, based on a mixture of gamma distributions in which the mixing occurs over the shape parameter. This family provides a flexible and novel approach for modeling heavy-tailed distributions, it is computationally efficient, and it only requires to specify a prior distribution for a single parameter. By carrying out simulation studies, we compare our approach with commonly used methods, such as the log-normal model and non parametric alternatives. We found that the mixture-gamma model significantly improves predictive performance in estimating tail probabilities, compared to these alternatives. We also applied our method to the Medical Current Beneficiary Survey (MCBS), for which we estimate the probability of exceeding a given hospitalization cost for smoking attributable diseases. The `R` software that implements the method is available from the authors[1].

[1] `sergio.venturini@unibocconi.it.`

1 Introduction

There is an extensive health services research literature on developing models for predicting health costs or health services utilization. These prediction problems are usually complicated by the nature of the distributions being analyzed: high skewness, heaviness of the right tail, and significant fractions of zeros or token amounts are commonly encountered. At present, there is no agreement about the best methods to use (see Mullahy and Manning [29], Kilian et al. [21], Buntin and Zaslavsky [5], Barber and Thompson [2], Manning and Mullahy [34], Powers et al. [35], Dodd et al. [13]; for a recent survey see Willan and Briggs [49]).

In important and still open research question is how to best predict the proportion of (total or single-event related) medical expenditures that will exceed a given threshold (see for example Briggs and Gray [3], Conwell and Cohen [9]). For example, insurance companies and governmental health departments are often interested in predicting how many customers or citizens will ask for a reimbursement above a certain threshold. Similarly, financial institutions are often interested in estimating the probability of the potential loss that could take place in the next day, week or month. In all these situations the parameter of interest is a tail probability of a highly skewed distribution. Thus it is important to develop methods that do not simply smooth the distribution of the data, but that are able to perform well from a predictive point of view.

The development of this work has been motivated by an analysis of medical expenditures from the Medicare Beneficiaries Survey (MCBS). We are interested in modeling the distribution of medical costs paid by the Medicare program for treating smoking attributable diseases, specifically lung cancer (LC) and coronary heart disease (CHD). We need to estimate the probability that the hospitalization cost for a smoking attributable disease exceeds a certain value.

MCBS is a continuous, multipurpose survey of a U.S. nationally representative sample of Medicare beneficiaries (people aged 65 or older, some people under age 65 with disabilities and people with permanent kidney failure requiring dialysis or a kidney transplant). The central goal of MCBS is to determine expenditures and sources of payment for all services used by Medicare beneficiaries. The data set includes medical expenditures for LC or CHD as primary diagnosis for 26,834 hospitalizations of 9,782 individuals for the period 1999-2002. For our analyses, we extract medical expenditures on the first hospitalization for 7,615 individuals.

A typical assumption in health services research is that medical costs are log-normally distributed (Zhou et al. [50], Tu and Zhou [47], Zhou et al. [51], Briggs et al. [4]). In our case, as well as many others, this assumption is not appropriate,

since the distribution of log-transformed expenditures is still far from being symmetric. For this reason, new methods have been recently proposed, especially for estimating the cost mean difference between cases and controls (Johnson et al. [20], Dominici et al. [14], Dominici and Zeger [15]). However few methods have been proposed for modeling the entire distribution and for prediction.

Skewed distributions typically arise in situations where few large values of the quantity under examination are present. It is well known that these observations heavily influence the results of statistical analysis. The remedies proposed in the health research literature are either to transform the data (see Duan [16], Mullahy [33], Manning[28], Mullahy and Manning [29]) or to use robust methods (see Conigliani and Tancredi [8], Cantoni and Ronchetti [6]). A different approach in modeling the medical costs distribution that has not been explored in the literature is to use a mixture distribution. Mixture models are parametric models which are flexible enough to represent a large spectrum of different phenomena. The mixture models literature is extensive (for general overviews, see Titterington et al. [46], Lindsay [24], McLachlan and Peel [32]; for a comprehensive list of applications see Titterington [45]). Particularly relevant for this paper is the well-developed parametric Bayesian literature on mixture distributions (see Diebolt and Robert [12], Robert [37], Roeder and Wasserman [41], Marin et al. [30]).

The purpose of this paper is to propose an innovative Bayesian approach for density estimation of very skewed distributions and for predicting the proportion of medical expenditures that exceed a given threshold. We model the distribution of medical expenditures by use of a mixture of gamma density functions with unknown weights. Using this model, we then estimate the tail probability $\mathbb{P}(Y > k)$, for different values of k. Each gamma distribution in the mixture is indexed by a component-specific shape parameter and a single unknown scale parameter θ. This parametrization allows to create a very parsimonious model with just one parameter for all the gamma components, plus the ordinary set of mixture weights. Moreover it overcomes the well-known identifiability problems that always affect any mixture model estimation because this parametrization automatically provides an ordering of the mixture components (for a recent survey on identifiability problems in Bayesian mixture modeling see Jasra et al. [19]). We assume that the number of mixture components in our model is known. We provide practical advice on how to choose it, as well as the hyperparameters of the prior on the scale parameter. Based on a simulation study we illustrate that our method has a better predictive performance compared to standard approaches.

In Section 2 we introduce the gamma shape mixture model, the estimation approach, and provide guidance on how to choose prior hyperparameters. In Section 3

we illustrate the results of the simulation study and the data analysis. Section 4 contains a discussion and concluding remarks. A final appendix contains technical details about the Gibbs sampler used.

2 The Gamma Shape Mixture Model

In this section we introduce the gamma shape mixture (GSM) model. We start by presenting the likelihood and an overview of its main properties. In particular we show that the GSM model does not suffer from identifiability problems common to mixture distributions. We then introduce the prior distribution and posterior inferences.

2.1 Likelihood and Prior Structure

Let Y be a positive random variable, for example non-zero medical expenditures. The GSM model is defined as

$$f(y|\pi_1,\ldots,\pi_J,\theta) = \sum_{j=1}^{J} \pi_j f_j(y|\theta)\,, \tag{1}$$

where $f_j(y|\theta) = \frac{\theta^j}{\Gamma(j)} y^{j-1} e^{-\theta y}$, the density function of a gamma $\mathcal{G}a(j,\theta)$ random variable. We assume that the number of components J is known and fixed, while $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$ is an unknown vector of mixture weights. Discussion on how to choose J will be provided later. In what follows, we denote (1) as $\mathcal{GSM}(\boldsymbol{\pi},\theta|J)$.

The GSM model has the following nice properties:

1. $\frac{1}{\theta}$ is a scale parameter (Lehmann and Casella [22]) for the whole model, since

$$f(y|\pi_1,\ldots,\pi_J,\theta) = \theta \cdot f(\theta \cdot y|\pi_1,\ldots,\pi_J,1)\,.$$

2. Its moments are convex combinations of the moments of the $Y_j|\theta \sim f_j(y|\theta)$ mixture components, so that the m-th moment is given by

$$\mathbb{E}\left[Y^m|\theta\right] = \sum_{j=1}^{J} \pi_j\, \mathbb{E}\left[Y_j^m|\theta\right] = \sum_{j=1}^{J} \pi_j \frac{\prod_{\ell=1}^{m}(j+\ell-1)}{\theta^m}\,.$$

A further issue related to mixture modeling is *label switching*, that is invariance to permutations of the components' indexes (see Jasra et al. [19]). A typical solution is to impose an *identifiability constraint*, usually an ordering of either the components means or the variances or the mixture weights (see Aitkin and Rubin [1]). A nice

feature of the GSM model (1) is that automatically imposes a constraint on both the means and the variances, since

$$\frac{1}{\theta} < \frac{2}{\theta} < \cdots < \frac{J-1}{\theta} < \frac{J}{\theta}\,,$$
$$\frac{1}{\theta^2} < \frac{2}{\theta^2} < \cdots < \frac{J-1}{\theta^2} < \frac{J}{\theta^2}\,.$$

Therefore the model is always identified and label switching is not a concern.

We assume that θ and $\boldsymbol{\pi}$ are independent a priori and we specify the following conjugate priors, that is

$$\theta \sim \mathcal{G}a(\alpha, \beta),$$
$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J) \sim \mathcal{D}_J\left(\frac{1}{J}, \ldots, \frac{1}{J}\right),$$

We choose the prior hyperparameters of the Dirichlet prior to favor selecting only a small subset of the mixture weights with high prior probability.

Given a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ of iid observations from (1), the likelihood is given by

$$\mathbb{L}\left(\boldsymbol{\pi}, \theta | \boldsymbol{y}\right) = \prod_{i=1}^{n} \sum_{j=1}^{J} \pi_j \, f_j\left(y_i | \theta\right). \tag{2}$$

Unfortunately this expression is untreatable because it includes J^n different terms (Marin et al. [30]). To overcome this hurdle we use the so called *missing data* representation of the mixture (Diebolt and Robert [10], [12]).

Consider a random sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ from model (1). It is possible to associate to each y_i an integer x_i between 1 and J that identifies the component of the mixture generating observation y_i. Thus the variable x_i takes value j with prior probability π_j, $1 \leq j \leq J$. The vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ of component labels is the *missing data* part of the sample since it is not observed. Figure 1 illustrates this for our model, highlighting that $\boldsymbol{y}$ is conditionally independent from the mixture weights $\boldsymbol{\pi}$, given the missing data $\boldsymbol{x}$.

Suppose the missing data $x_1, \ldots, x_n$ were available. Then the model could be written as

$$p(y_1, \ldots, y_n | x_1, \ldots, x_n, \theta) = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} \Gamma(x_i)} \left(\prod_{i=1}^{n} y_i^{x_i - 1}\right) e^{-\theta \sum_{i=1}^{n} y_i}\,. \tag{3}$$

Thus, using (3) and the priors, the posterior distribution is

$$p(\pi_1, \ldots, \pi_J, \theta | y_1, \ldots, y_n, x_1, \ldots, x_n) \propto \left(\prod_{j=1}^{J} \pi_j^{\frac{1}{J} + n_j - 1}\right) \theta^{\,\alpha + \left(\sum_{i=1}^{n} x_i\right) - 1} e^{-\left(\beta + \sum_{i=1}^{n} y_i\right)\theta}, \tag{4}$$

where $n_j = \sum_{i=1}^{n} \mathbb{I}(x_i = j)$, $j = 1, \ldots, J$, and $\mathbb{I}(\cdot)$ is the indicator function. The main consequence of this conditional decomposition is that, for a given missing data vector $x_1, \ldots, x_n$, the conjugacy is preserved and therefore the simulation can be performed conditional on the missing data $x_1, \ldots, x_n$.

2.2 Posterior calculation

We implement two approaches for estimating the unknown parameters of interests. In the first approach we estimate the posterior distribution of $\boldsymbol{\pi}$, $\boldsymbol{x}$ and θ by using a Gibbs sampler (details are reported in Appendix). To increase the efficiency, we also propose a second estimation approach where we integrate out the scale parameter θ analytically. The advantage of this second strategy is both computational, since the chain runs in a smaller space, and theoretical, since generally simulated values are less autocorrelated after partial marginalization (Liu [25], MacEachern [26], MacEachern et al. [27]).

After having integrated out θ, the full conditional distribution of the mixture weights is given by

$$p(\pi_1, \ldots, \pi_J | y_1, \ldots, y_n, x_1, \ldots, x_n) \propto \prod_{j=1}^{J} \pi_j^{\frac{1}{J} + n_j - 1},$$

that is, the $\mathcal{D}_J\left(\frac{1}{J} + n_1, \ldots, \frac{1}{J} + n_J\right)$ Dirichlet distribution. In addition, the full conditional of the i-th missing label is then given by

$$p(x_i | \boldsymbol{y}, \boldsymbol{x}_{(-i)}, \boldsymbol{\pi}) = \sum_{j=1}^{J} \frac{\pi_j \frac{y_i^{j-1}}{\Gamma(j)} \frac{\left(\alpha + \sum_{(-i)} x_r\right)_j}{\left(\beta + \sum_{r=1}^{n} y_r\right)^j}}{\sum_{k=1}^{J} \pi_k \frac{y_i^{k-1}}{\Gamma(k)} \frac{\left(\alpha + \sum_{(-i)} x_r\right)_k}{\left(\beta + \sum_{r=1}^{n} y_r\right)^k}} \mathbb{I}(x_i = j), \tag{5}$$

where $\boldsymbol{x}_{(-i)}$ is the $\boldsymbol{x} = (x_1, \ldots, x_n)$ vector with the i-th element deleted, $\sum_{(-i)} x_r$ denotes the sum of all the component labels except for the i-th one, $(n)_k$ is the Pochhammer symbol. Moreover, α is constrained to be an integer (see Appendix for a justification). Note that, the integration of θ implies that the missing data are no longer independent.

`R` software implementing this approach is available from the authors[2].

2.3 Choice of the Hyperparameters

In this subsection we describe how we choose the values of α and β, the hyperparameters of the prior on θ, and J, the number of components in the GSM.

[2] sergio.venturini@unibocconi.it

Our proposal for choosing the hyperparameters could be described as an informal empirical Bayes approach since we use summary statistics of the data, like the maximum and the sum of the observations, to get reasonable values.

It is useful to note first that the mean of model (1) is

$$\mu = \mathbb{E}[Y|\theta] = \sum_{j=1}^{J} \pi_j \, \frac{j}{\theta}\,, \tag{6}$$

so that we can write

$$\theta = \frac{1}{\mu} \sum_{j=1}^{J} \pi_j j \tag{7}$$

and that the expected value of the full conditional distribution of θ is

$$\begin{aligned} \mathbb{E}[\theta|\boldsymbol{y}, \boldsymbol{x}] &= \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + \sum_{i=1}^{n} y_i} \\ &= \frac{\beta}{\beta + \sum_{i=1}^{n} y_i} \cdot \frac{\alpha}{\beta} + \frac{\sum_{i=1}^{n} y_i}{\beta + \sum_{i=1}^{n} y_i} \cdot \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \\ &= \omega \cdot \frac{\alpha}{\beta} + (1-\omega) \cdot \frac{\bar{x}}{\bar{y}}\,, \end{aligned} \tag{8}$$

Equation (7) indicates that θ can be interpreted as the (normalized) weighted average of the J component labels. Equation (8) indicates that the posterior mean of θ is a weighted average of α/β and $\bar{x}/\bar{y}$ where β is the *prior sample size* and α is the prior mean of θ. When both $\alpha \to 0$ and $\beta \to 0$ the prior becomes improper. Then, for a given value of J, a strategy for choosing α and β is:

1. Calculate the quantity $\widetilde{\theta} = \frac{J}{\max(y_1,\ldots,y_n)}$ and check that $\frac{1}{\widetilde{\theta}} \leq \min(y_1, \ldots, y_n)$; the idea is that on average θ should take values that allow the set of gamma distributions in (1) to completely span the range of observed values (the last gamma distribution should have a mean not smaller than the maximum observation and the first gamma distribution a mean not greater than the minimum observation). $\widetilde{\theta}$ is hence a candidate for the prior mean $\frac{\alpha}{\beta}$.
2. Choose a value for the weight of the prior information ω in (8). Values between 0.2 and 0.5 are usually reasonable choices. Fix β to $\frac{\omega \cdot \sum_{i=1}^{n} y_i}{1-\omega}$.
3. Set α by rounding to the closest integer the quantity $\widetilde{\theta} \cdot \beta$. The rounding is needed because of the assumption used to get (5).

Concerning the choice of J, the goodness of fit of the GSM is the result of the interplay among the grids of m-th order moments

$$\left(\frac{\prod_{\ell=1}^{m} \ell}{\theta^m}, \frac{\prod_{\ell=1}^{m} (\ell+1)}{\theta^m}, \ldots, \frac{\prod_{\ell=1}^{m} (\ell+J-2)}{\theta^m}, \frac{\prod_{\ell=1}^{m} (\ell+J-1)}{\theta^m} \right),$$

and the ordered sequence of observations. These grids should contain sufficient elements to fit the data, therefore J should be calibrated to the specific set of data under examination. Generally speaking, a small value of J can create a severe limitation to the model as the set of densities available in the class being mixed may not be sufficiently rich with elements that have a large mean. On the other had, too large a value does not cause serious difficulties as the fit is often robust when there are several gamma distribution in the class that can serve as building blocks for a particular mixture component. However, large J can cause numerical problems. Sometimes a transformation of the data (like a log or a root) can be useful to handle these numerical issues. In practice, the choice of J may require more than one iteration. Inspection of the predictive density is a practical diagnostic to identify misspecifications of J.

3 Results

In this section we carry out a simulation study to assess the predictive performance of the GSM model in estimating the right tail of the medical expenditures with respect to common alternatives. In addition, we apply our methods to the MCBS data for estimating the risk for persons affected by smoking attributable diseases to exceed a given medical costs threshold in a single hospitalization.

3.1 Simulation design

From the complete MCBS data, we extract expenditure data on hospitalizations in which the first diagnosis has been either CHD or LC (or both), for a total of $7,615$ hospitalizations. Tables 1 and 2 report a brief summary of the dataset. From this population, we drew 500 sub-samples, the *training sets*, of size equal to 10% of the original sample. The remaining 90% constitute the *test sets*.

On each training set we calculate four estimates of the tail probability $\hat{p} = \mathbb{P}(y^* > k|\boldsymbol{y})$ using the following approaches:

- the empirical distribution function (EDF),
- a log-normal distribution (LN),
- a normal mixture distribution (MN),
- a gamma shape mixture distribution (GSM).

For EDF, LN and MN, the estimators of the tail probability are all known. For the MN, we estimate the tail probability by using the `mclust` package in `R`. For the GSM,

the estimator of the tail probability is defined as

$$\mathbb{P}(y^* > k|\boldsymbol{y}) = \int \mathbb{P}\left(y^* > k|\boldsymbol{y}, \theta, \boldsymbol{\pi}\right) f(\theta, \boldsymbol{\pi}|\boldsymbol{y})\, d\theta\, d\boldsymbol{\pi}\,. \tag{9}$$

This predictive probability can be estimated from Gibbs sampling realizations by the rao-blackwellized estimator:

$$\begin{aligned}\widehat{\mathbb{P}}(y^* > k|\boldsymbol{y}) &= \frac{1}{M}\sum_{m=1}^{M} \mathbb{P}\left(y^* > k \,\middle|\, \theta^{(m)}, \boldsymbol{\pi}^{(m)}\right) \\ &= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{J} \pi_j^{(m)} \left[1 - F_j\left(k \,\middle|\, \theta^{(m)}\right)\right],\end{aligned}$$

where $F_j(\cdot|\theta)$ is the distribution function of a $\mathcal{G}a(j, \theta)$ random variable.

On each test set the sample proportion $\#\{i : y_i > k\}/n_{\text{test}}$ is calculated and the entire analysis is repeated for the following values of the medical cost threshold k: \$10,000, \$15,000, \$20,000, \$30,000, \$50,000 and \$80,000 (higher threshold values are too rare to be included in the analysis, see Table 2).

The data are transformed using a cubic root. For any random variable X and strictly monotonic function $g(\cdot)$,

$$\mathbb{P}\left(X > k\right) = \mathbb{P}\left(g(X) > g(k)\right)\,,$$

so no bias is introduced in the analysis, at least for our purposes. Figure 2 reports graphical summaries for both the untransformed and transformed MCBS data.

The parameters chosen for the simulation are J=200, $\alpha = 12,380$, $\beta = 3,420$, $5,000$ iterations ($1,000$ of which for burn-in). These have been chosen following the indications provided in Section 2.3.

3.2 Simulation results

For each estimation method and each k, Figure 3 presents in panel (a) the relative mean squared errors, defined by $\left(\text{mse}_{\text{EDF}} - \text{mse}_T\right)/\text{mse}_{\text{EDF}}$, where mse_T indicates the mean squared error for the tail probability $\mathbb{P}(y^* > k|\boldsymbol{y})$ estimated with T, and in panel (b) the relative bias $\left(\mathbb{E}\left(T\right) - p_{\text{TRUE}}\right)/p_{\text{TRUE}}$, where $p_{\text{TRUE}} = \#\{i : y_i > k\}/n$ for the whole sample, both in percentage. Negative values of the relative mean squared error imply that the EDF estimator is preferred, while positive values are in favor of the compared estimator.

For almost all the medical expenditure thresholds the GSM is more efficient than the estimator based on the empirical distribution function. As expected, we observe a

trend for the efficiency to increase with the threshold, up to a 27% improvement for the highest threshold.

The tail probability estimator based on the log-normal distribution performs poorly. Its mean squared error relative to the EDF estimator is below -100% for all the thresholds, with the worst performance for hospitalization costs above $20,000$\$. This result is important practically since many models in health services research and health economics are based on the assumption that the medical expenditures can be handled using log-normal distributions.

Also the estimator based on the mixture of normal distributions performs worse than the EDF, and the performance worsens as the threshold increases. This finding implies that, to overcome the problems with the log-normal distribution, it does not suffice to use any mixture.

Figure 3(b) reports the relative biases, showing that overall the mixture of gamma distributions is slightly biased, but less so than the other methods. Once again, the estimator based on the log-normal distribution is more biased than the alternatives, and it almost always underestimates the tail probability for the reference population. The bias increases systematically (in absolute value) with the thresholds, indicating that the log-normal distribution is not sufficiently heavy-tailed to mimic the right tail of these data, despite the prior cubic root transformation.

Figure 4 allows a further comparison of the GSM with the other models. Each panel shows the estimated tail probabilities for a combination of cost threshold and estimation method. A dot in these graphs represents a sub-sample of the simulation study. On the horizontal axes we show the absolute value of the difference between the estimated tail probability on the training set using the GSM and the (empirical) tail probability on the corresponding test set, while on the vertical axes we show the same quantity obtained using one of the other methods (as indicated on top of each panel). As an aid in visualization, in each panel the shading shows the density of the points above and below the 45 degrees line separately. These graphs allow to conclude that the GSM performs better than the other estimation methods from a predictive point of view, since in every panel the majority of points are above the diagonal. The labels above the 45 degrees line in each panel indicate the percentage of sub-samples for which the GSM performs better than the compared method, while labels below the 45 degrees line indicate the percentage of sub-samples for which the GSM performs worst. Note that the higher the threshold, the more pronounced the result. Moreover, even if slightly biased, most of the times[3] the GSM method works better than the estimator based on the empirical distribution function.

[3]In Figure 4 only the graphs for some of the available thresholds are shown to avoid cluttering.

3.3 Analysis of MCBS Medical Costs Data

In this section we illustrate a data analysis of the MCBS dataset. The aims of the analysis are to provide estimates of the density function and of the risk of exceeding a given medical cost threshold k in a single hospitalization, with associated probability intervals. As in the simulation study, we restrict the analysis to hospitalizations in which the first diagnosis has been for a smoking attributable disease, CHD or LC. The size of the sample is $n = 7,615$.

The parameters for the Gibbs sampler have been set to $J = 200$, $\alpha = 124,960$, $\beta = 34,520$ and we used $6,000$ sampling iterations ($1,000$ of which as burn-in). The data have been transformed using a cubic root transformation.

Figure 5 shows the fit of the GSM model to the MCBS data. Panel (a) presents the fitted model density together with the data histogram, while Panel (b) reports the QQ-plot of the model cumulative probabilities, evaluated at the posterior mean of the mixture weights and scale parameter, versus the empirical cumulative probabilities $p_i = i/(n+1)$, $i = 1, \ldots, n$. As it is clear from these graphs, the GSM provides a very good representation of the data at hand.

Figure 6 contains additional results that provide insight on how the model works. Even though $J = 200$ components were available, the estimation procedure selects at every iteration just a small subset of them, sufficient to fit the data. In fact, panel (a) shows that, a posteriori, the number of selected components is in between 9 and 20, with the mode equal to 14, while panel (b) shows the posterior means of the mixture weights. This conclusion is reinforced by the observation that only approximately 10 components have a posterior mean weight that is substantially greater than zero. Panels (c) and (d) report the histograms of the model mean and variance evaluated at each Gibbs sampler iteration. The vertical dashed lines indicate the overall sample mean and variance. These plots are useful to qualitatively assess whether the choices for the hyperparameters α, β and for the number of mixture components J are appropriate for the sample at hand.

Figure 7 displays the estimates of the tail probability for different threshold values. In other words, this graph presents the "risk" of exceeding a given medical costs threshold in a single hospitalization for people affected by smoking attributable diseases. The 95% credible intervals for each threshold estimate are also shown.

We performed a set of sensitivity analyses to prior hyperparameters, not reported here in detail, and found that the results are not significantly affected.

4 Discussion

In this paper we introduced a Bayesian mixture model and computation for density estimation of very skewed distributions. Our approach is based on a mixture of gamma distributions over the shape parameter. This family of distributions includes components whose means and variances increase together, offering a parsimonious way of representing populations in which a small fraction of individuals has an outlying behaviour that is difficult to predict.

The development of our work is motivated, by the estimation of the proportion of subjects affected by smoking attributable diseases, specifically CHD and LC, that, in a single hospitalization, have a medical bill exceeding a given threshold. In particular, we used data from MCBS, a multipurpose survey of a U.S. nationally representative sample of Medicare beneficiaries.

Although the model works only for positive variables, we demonstrated that the method we proposed represents an improvement on the standard modeling strategy commonly adopted in the health economics and health services research literature for the distribution of medical costs. Since this distribution is usually very skewed, the typical assumption in that literature is to use a log-normal model. We show that for highly skewed data this choice is not always appropriate and leads to highly biased and inefficient estimates of tail probability, especially for high thresholds. The GSM model presented here (i) does not exhibit any identifiability problem, (ii) represents a density estimation method that works well for skewed distributions and (iii) allows to estimate tail probabilities more efficiently than other common methods used in these kinds of public health applications. Additionally, being a complete probabilistic specification of the data generation process, it can be used for any other inferential purpose.

We also develop a computationally efficient algorithm which is a a modification of the standard Gibbs sampler used in the Bayesian literature on mixture distributions (Robert [37]). In our approach we integrate out the the scale parameter θ, thus making computations more efficient (Liu [25]). The `R` software that implements the method is available from the authors[4].

In our model we fix a priori the number of mixture components J. A possible generalization is to incorporate the model into a reversible jump approach, as proposed by Richardson and Green [36], in which the number of component is random. Due to the particular features of our model, one way to implement it could be to let the algorithm add or delete not simply one component, but rather 5 to 10 components, at every step. The obvious consequence of this modification is a substantial increase in

[4] `sergio.venturini@unibocconi.it`.

the computational time needed for estimation. However, because of the limited loss in specifying a large J, the tool we described is likely to be very useful as currently proposed.

As a future generalization of our basic framework, we suggest that it may be useful to model the scale parameter as a function of covariates of interest. This extension would allow to get a general and robust regression model.

Appendix

Gibbs sampler for mixture estimation with θ integrated out

The posterior distribution of $(\pi_1, \ldots, \pi_J, \theta)$, given the sample $(y_1, \ldots, y_n)$, can be written as

$$p(\pi_1, \ldots, \pi_J, \theta | y_1, \ldots, y_n) \propto \left(\prod_{j=1}^{J} \pi_j^{\frac{1}{J}-1} \right) \theta^{\alpha-1} e^{-\beta\theta} \prod_{i=1}^{n} \left(\sum_{j=1}^{J} \pi_j \frac{\theta^j}{\Gamma(j)} y_i^{j-1} e^{-\theta y_i} \right).$$

The standard algorithm to implement the posterior simulation is reported in the next subsection. However, to increase efficiency, in our estimation approach we integrate out the scale parameter θ (Liu [25], MacEachern [26], MacEachern et al. [27]). Then (3) becomes

$$\begin{aligned} p(y_1, \ldots, y_n | x_1, \ldots, x_n) &= \int_0^\infty \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n \Gamma(x_i)} \left(\prod_{i=1}^{n} y_i^{x_i-1} \right) e^{-\theta \sum_{i=1}^n y_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \, d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\prod_{i=1}^n y_i^{x_i-1}}{\prod_{i=1}^n \Gamma(x_i)} \int_0^\infty \theta^{\alpha+\left(\sum_{i=1}^n x_i\right)-1} e^{-\left(\beta+\sum_{i=1}^n y_i\right)\theta} \, d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\prod_{i=1}^n y_i^{x_i-1}}{\prod_{i=1}^n \Gamma(x_i)} \frac{\Gamma\left(\alpha + \sum_{i=1}^n x_i\right)}{\left(\beta + \sum_{i=1}^n y_i\right)^{\alpha+\left(\sum_{i=1}^n x_i\right)}}. \end{aligned} \tag{10}$$

Note that the observed data, conditionally on the non-observed ones, are no longer independent. The interpretation of this fact is that θ was a parameter shared by all the (y_i, x_i) pairs, $i = 1, \ldots, n$. Removing θ has introduced dependence among the data.

The full conditional of the mixture weights is hence given by

$$p(\pi_1, \ldots, \pi_J | y_1, \ldots, y_n, x_1, \ldots, x_n) \propto \prod_{j=1}^{J} \pi_j^{\frac{1}{J}+n_j-1},$$

while, to get the full conditional of the missing data, we decompose it into the individual full conditionals

$$p(x_i | y_1, \ldots, y_n, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \pi_1, \ldots, \pi_J),$$

$i \in \{1, \ldots, n\}$. Note that

$$\begin{aligned} p(x_i|\boldsymbol{y}, \boldsymbol{x}_{(-i)}, \boldsymbol{\pi}) &= \sum_{j=1}^{J} \frac{p(x_i, \boldsymbol{x}_{(-i)}|\boldsymbol{y}, \boldsymbol{\pi})}{\sum_{k=1}^{J} p(k, \boldsymbol{x}_{(-i)}|\boldsymbol{y}, \boldsymbol{\pi})} \mathbb{I}(x_i = j) \\ &= \sum_{j=1}^{J} \frac{p(\boldsymbol{y}|x_i, \boldsymbol{x}_{(-i)}) \cdot p(x_i, \boldsymbol{x}_{(-i)}|\boldsymbol{\pi})}{\sum_{k=1}^{J} p(\boldsymbol{y}|k, \boldsymbol{x}_{(-i)}) \cdot p(k, \boldsymbol{x}_{(-i)}|\boldsymbol{\pi})} \mathbb{I}(x_i = j)\,, \end{aligned}$$

for $i \in \{1, \ldots, n\}$ and where $\boldsymbol{x}_{(-i)}$ denotes the $\boldsymbol{x} = (x_1, \ldots, x_n)$ vector with the i-th element deleted. The second equality follows from $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\pi}) \cdot p(\boldsymbol{y}|\boldsymbol{\pi}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\pi}) \cdot p(\boldsymbol{x}|\boldsymbol{\pi})$ and from the conditional independence of $\boldsymbol{y}$ from $\boldsymbol{\pi}$, given the missing data $\boldsymbol{x}$. Substituting (10) we obtain

$$p(x_i|\boldsymbol{y}, \boldsymbol{x}_{(-i)}, \boldsymbol{\pi}) = \sum_{j=1}^{J} \frac{\pi_j \frac{y_i^{j-1}}{\Gamma(j)} \frac{\Gamma\left(\alpha + \sum_{(-i)} x_r + j\right)}{\left(\beta + \sum_{r=1}^{n} y_r\right)^j}}{\sum_{k=1}^{J} \pi_k \frac{y_i^{k-1}}{\Gamma(k)} \frac{\Gamma\left(\alpha + \sum_{(-i)} x_r + k\right)}{\left(\beta + \sum_{r=1}^{n} y_r\right)^k}} \mathbb{I}(x_i = j)\,, \tag{11}$$

where the $\sum_{(-i)} x_r$ denotes the sum of all the component labels apart from the i-th one. If one further assumes that $\alpha \in \mathbb{N}$, then (11) can be further simplified[5] to

$$p(x_i|\boldsymbol{y}, \boldsymbol{x}_{(-i)}, \boldsymbol{\pi}) = \sum_{j=1}^{J} \frac{\pi_j \frac{y_i^{j-1}}{\Gamma(j)} \frac{\left(\alpha + \sum_{(-i)} x_r\right)_j}{\left(\beta + \sum_{r=1}^{n} y_r\right)^j}}{\sum_{k=1}^{J} \pi_k \frac{y_i^{k-1}}{\Gamma(k)} \frac{\left(\alpha + \sum_{(-i)} x_r\right)_k}{\left(\beta + \sum_{r=1}^{n} y_r\right)^k}} \mathbb{I}(x_i = j)\,, \tag{12}$$

where $(n)_k$ denotes the Pochhammer symbol. We refer to the whole fraction inside the leftmost summation as κ_{ij}. The steps to implement this simulation algorithm are then summarized below:

- Simulate
$$\boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{x} \sim \mathcal{D}_J\left(\frac{1}{J} + n_1, \ldots, \frac{1}{J} + n_J\right),$$
where $n_j = \sum_{i=1}^{n} \mathbb{I}(x_i = j)$, $j = 1, \ldots, J$.
- Simulate, for every $i = 1, \ldots, n$,
$$p(x_i|y_1, \ldots, y_n, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \pi_1, \ldots, \pi_J) = \sum_{j=1}^{J} \kappa_{ij} \mathbb{I}(x_i = j),$$
with κ_{ij} as defined above, $j = 1, \ldots, J$.
- Update n_j, $j = 1, \ldots, J$.

[5]This simplification helps to avoid overflow errors during the computation.

Standard Gibbs sampler for mixture estimation

The implementation of the standard Gibbs sampling is straightforward and involves the iterative simulation from (4), for the parameters of the model, and from

$$p(x_1, \ldots, x_n | \pi_1, \ldots, \pi_J, \theta, y_1, \ldots, y_n)\,,$$

for the missing data. The steps for the algorithm are (see for example Robert [37]):

- Simulate

$$\begin{aligned} \theta|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\pi} &\sim \mathcal{G}a\left(\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} y_i\right) \\ \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{x}, \theta &\sim \mathcal{D}_J\left(\frac{1}{J} + n_1, \ldots, \frac{1}{J} + n_J\right), \end{aligned}$$

  where $n_j = \sum_{i=1}^{n} \mathbb{I}(x_i = j)$, $j = 1, \ldots, J$.

- Simulate, for every $i = 1, \ldots, n$,

$$p(x_i | y_i, \pi_1, \ldots, \pi_J, \theta) = \sum_{j=1}^{J} \pi_{ij} \mathbb{I}(x_i = j),$$

  where

$$\pi_{ij} = \frac{\pi_j f_j(y_i|\theta)}{\sum_{k=1}^{J} \pi_k f_k(y_i|\theta)}, \qquad j = 1, \ldots, J.$$

- Update n_j, $j = 1, \ldots, J$.

References

[1] M. Aitkin and D. B. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, B*, **47**:67–75, 1985.

[2] J. Barber and S. Thompson. Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research & Policy*, **9**:197–204, 2004.

[3] A. Briggs and A. Gray. The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Economics*, **25**:198–213, 2006.

[4] A. Briggs, R. Nixon, S. Dixon, and S. Thompson. Parametric modelling of cost data: some simulation evidence. *Health Economics*, **14**:421–428, 2005.

[5] M. B. Buntin and A. M. Zaslavsky. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, **23**:525–542, 2004.

[6] E. Cantoni and E. Ronchetti. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Services Research & Policy*, **3**:233–245, 1998.

[7] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton, Second edition, 2000.

[8] C. Conigliani and A. Tancredi. Semi-parametric modelling for costs of health care technologies. *Statistics in Medicine*, **24**:3171–3184, 2005.

[9] L. J. Conwell and J. W. Cohen. Characteristics of Persons with High Medical Expenditures in the U.S. Civilian Noninstitutionalized Population, 2002. Technical Report www.meps.ahrq.gov/papers/st73/stat73.pdf, Agency for Healthcare Research and Quality, 2005.

[10] J. Diebolt and C. P. Robert. Bayesian estimation of finite mixture distributions, Part I: Theoretical aspects. Technical Report 110, Université Paris VI, Paris, 1990a.

[11] J. Diebolt and C. P. Robert. Bayesian estimation of finite mixture distributions, Part II: Sampling implementation. Technical Report 111, Université Paris VI, Paris, 1990b.

[12] J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, B*, **56**:363–375, 1994.

[13] S. Dodd, A. Bassi, K. Bodger, and P. Williamson. A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice*, **9**:197–204, 2004.

[14] F. Dominici, L. Cope, D. Q. Naiman, and S. L. Zeger. Smooth quantile ratio estimation (SQUARE). *Biometrika*, **92**:543–557, 2005.

[15] F. Dominici and S. L. Zeger. Smooth quantile ratio estimation with regression: estimating medical expenditures for smoking attributable diseases. *Biostatistics*, 2006. (to appear).

[16] N. Duan. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, **78**:605–610, 1983.

[17] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[18] P. J. Huber. Robust Statistical Procedures. In *CBMS-NSF Regional Conference Series in Applied Mathematics, Number 68*. Soc. Industr. Appl. Math., Philadelphia, Pennsylvania, Second edition, 1996.

[19] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**:50–67, 2005.

[20] E. Johnson, F. Dominici, M. Griswold, and S. L. Zeger. Disease cases and their medical costs attribuitable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics*, **112**:135–151, 2003.

[21] R. Kilian, H. Matschinger, W. Loeffler, C. Roick, and M. C. Angermeyer. A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *The Journal of Mental Health Policy and Economics*, **5**:21–31, 2002.

[22] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York, Second edition, 1998.

[23] B. G. Lindsay. The geometry of likelihoods: a general theory. *Annals of Statistics*, **11**:86–94, 1983.

[24] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics, Hayward, California, 1995.

[25] J. S. Liu. The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association*, **89**:958–966, 1994.

[26] S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, **23**:727–741, 1994.

[27] S. N. MacEachern, M. Clyde, and J. S. Liu. Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, **27**:251–267, 1999.

[28] W. G. Manning. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, **17**:283–295, 1998.

[29] W. G. Manning and J. Mullahy. Estimating log models: to transform or not to transform? *Journal of Health Economics*, **20**:461–494, 2001.

[30] J. M. Marin, K. Mengersen, and C. P. Robert. Bayesian Modelling and Inference on Mixtures of Distributions. In D. Dey and C. R. Rao, editors, *Handbook of Statistics 25*. Elsevier-Sciences, 2005. (to appear).

[31] R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.

[32] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[33] J. Mullahy. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, **17**:247–281, 1998.

[34] J. Mullahy and W. G. Manning. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, **24**:465–488, 2005.

[35] C. A. Powers, C. M. Meyer, M. C. Roebuck, and B. Vaziri. Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Medical Care*, **43**:1065–1072, 2005.

[36] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *Journal of the Royal Statistical Society, B*, **59**:731–792, 1997.

[37] C. P. Robert. Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 441–464. Chapman & Hall/CRC, New York, 1996.

[38] C. P. Robert. *The Bayesian Choice*. Springer, New York, Second edition, 2001.

[39] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, Second edition, 2004.

[40] K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**:617–624, 1990.

[41] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**:894–902, 1997.

[42] S. J. Sheather. Density estimation. *Statistical Science*, **19**:588–597, 2004.

[43] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, London, 1986.

[44] M. Stephens. Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *Annals of Statistics*, **28**:40–74, 2000.

[45] D. M. Titterington. Mixture distributions. In *Encyclopedia of Statistical Sciences*, pages 399–407. Wiley, New York, 1997.

[46] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester, 1985.

[47] W. Tu and X-H. Zhou. A Wald test comparing medical cost based on log-normal distributions with zero valued costs. *Statistics in Medicine*, **18**:2749–2761, 1999.

[48] L. Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2006.

[49] A. R. Willan and A. H. Briggs. *Statistical Analysis of Cost-Effectiveness Data*. Wiley, New York, 2006.

[50] X-H. Zhou, S. Gao, and S. L. Hui. Methods for comparing the means of two independent log-normal samples. *Biometrics*, **53**:1129–1135, 1997.

[51] X-H. Zhou, C. Li, S. Gao, and W. M. Tierney. Methods for testing equality of means of health care costs in a paired design study. *Statistics in Medicine*, **20**:1703–1720, 2001.

Table 1: *Summary of the MCBS dataset: high-order quantiles of medical expenditures of first hospitalization for LC, CHD or both.*

Quantile order	75	90	95	97.5	99	99.9
Quantile ($)	8,187.5	15,457.2	22,485.6	29,009.4	40,955.9	115,060.6

Table 2: *Summary of the MCBS dataset: number of hospitalizations with a cost above a specified threshold.*

Threshold ($)	$10,000$	$15,000$	$20,000$	$30,000$	$50,000$	$80,000$	$100,000$
Count	1,468	799	503	179	48	24	9
Proportion	0.1928	0.1049	0.0661	0.0235	0.0063	0.0032	0.0012

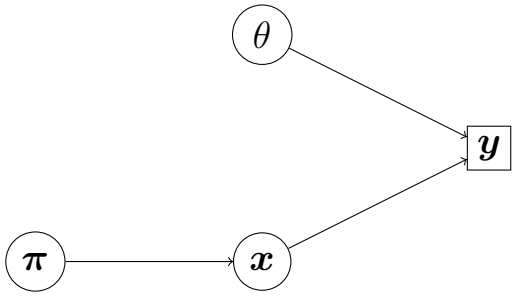


Figure 1: *Directed acyclic graph (DAG) for the missing data representation of the* $\mathcal{GSM}(\boldsymbol{\pi}, \theta | J)$ *model.*

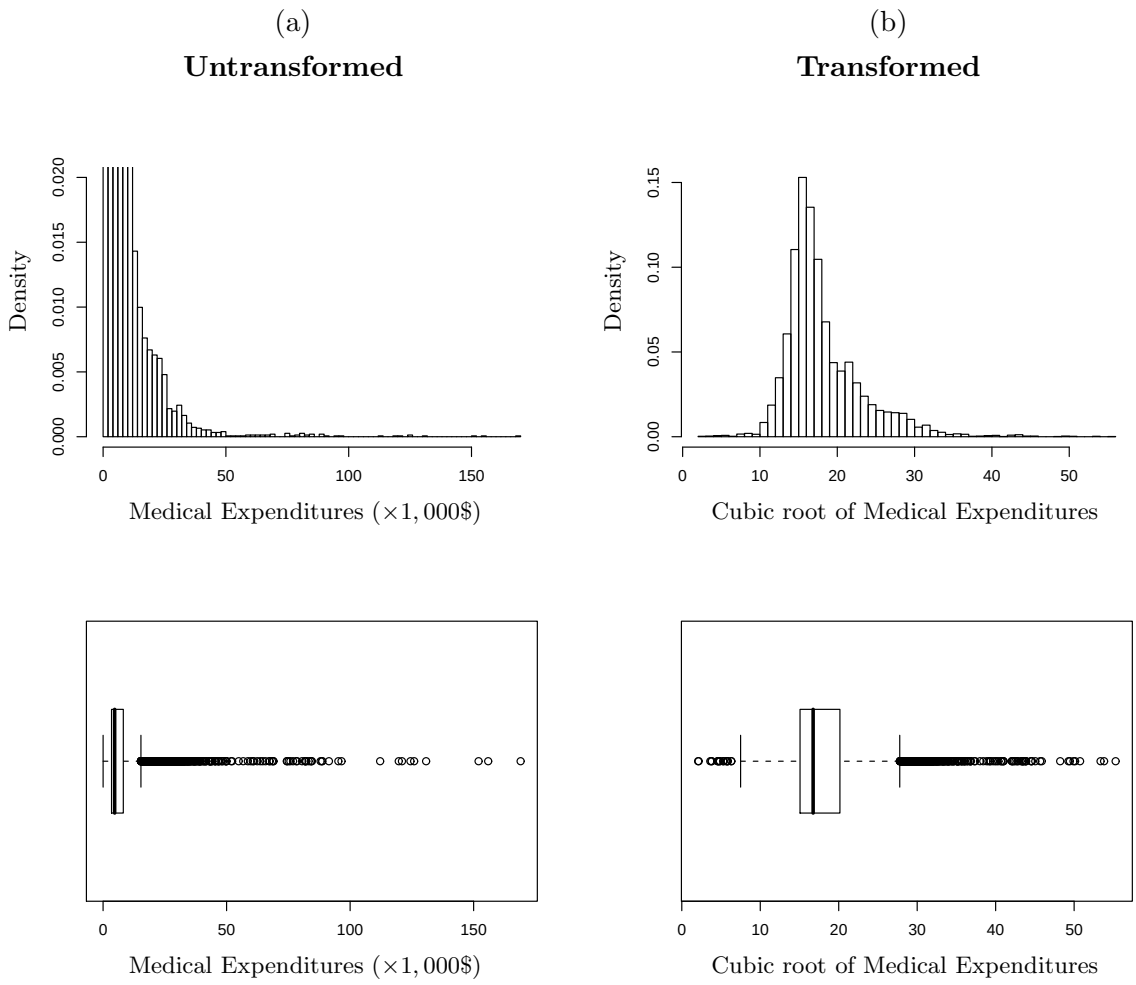
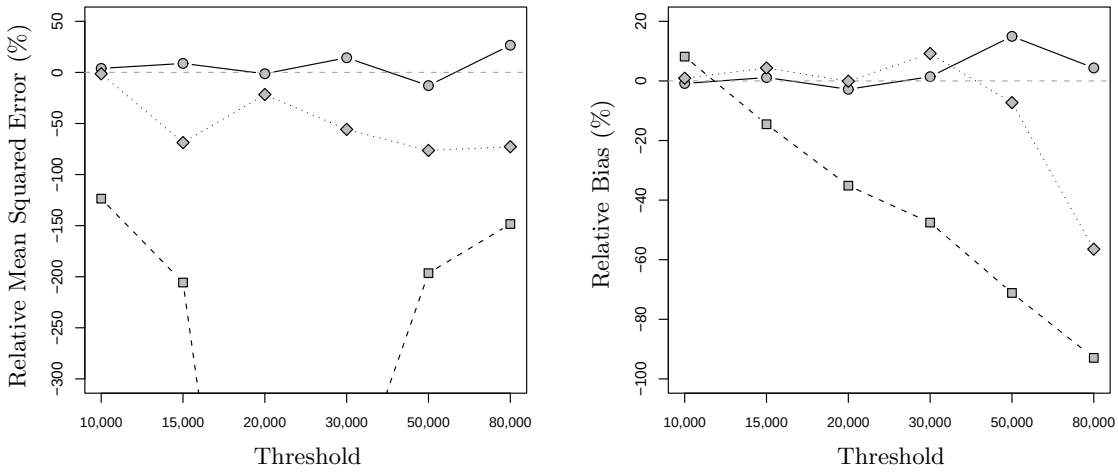


Figure 2: *Histograms and boxplots of positive Medicare expenditures for hospitalizations regarding smoking attributable diseases (lung cancer and coronary heart disease) from the 1999-2002 Medicare Current Beneficiaries Survey (for clarity of exposition, the histogram of the original expenditures has been truncated at the top.)*



(a) (b)

Figure 3: *Simulation study. (a) For different threshold values, the percentage mean squared error relative to the EDF estimator defined by* $[(mse_{EDF} - mse_{\hat{p}})/mse_{EDF}] \times 100$ *is shown. Circles indicate the gamma shape mixture tail probability estimator, diamonds the mixture of normals estimator and squares that based on the log-normal distribution. (b) For different threshold values* k*, the percentage bias relative to the sample proportion for the entire sample* $p_{TRUE} = \#\{i : y_i > k\}/500$ *defined by* $[(\mathbb{E}(\hat{p}) - p_{TRUE})/p_{TRUE}] \times 100$ *is shown. Circles indicate the gamma shape mixture tail probability estimator, diamonds the mixture of normals estimator and squares that based on the log-normal distribution.*

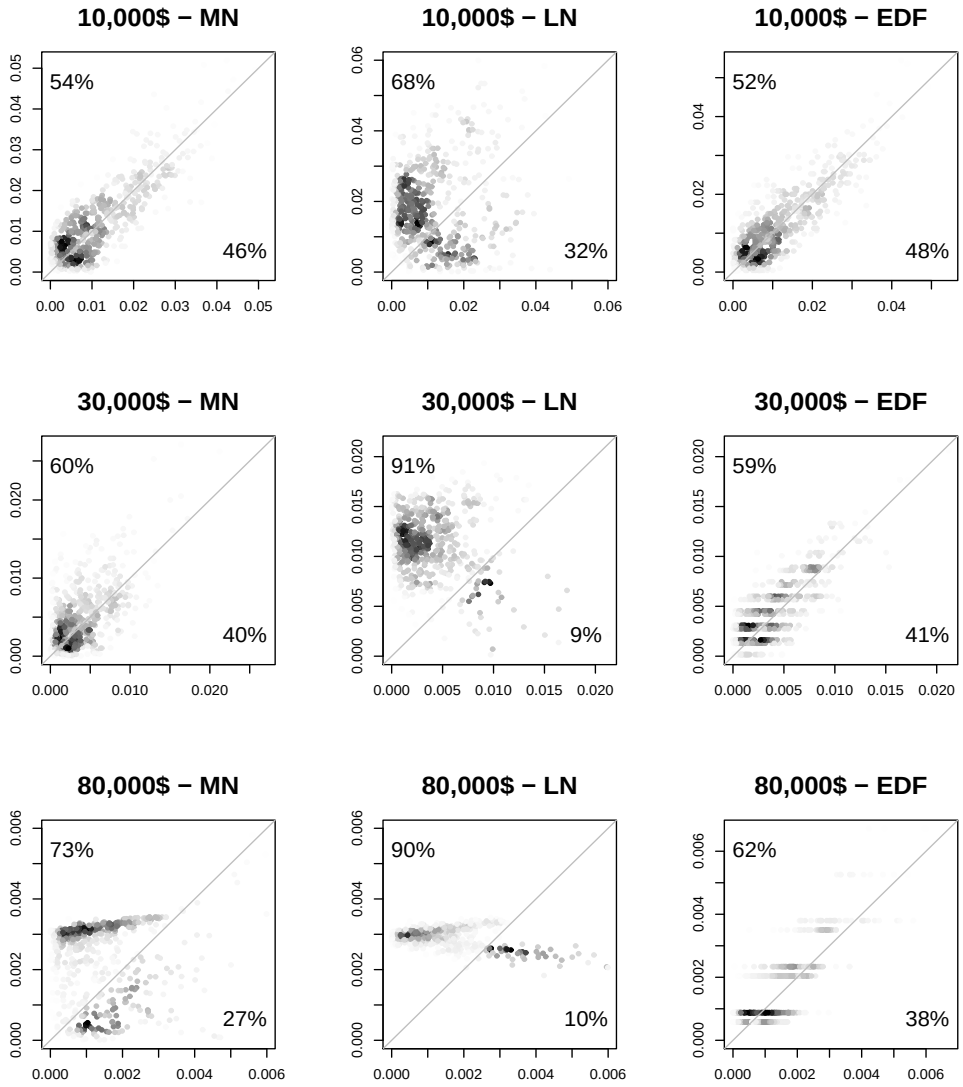
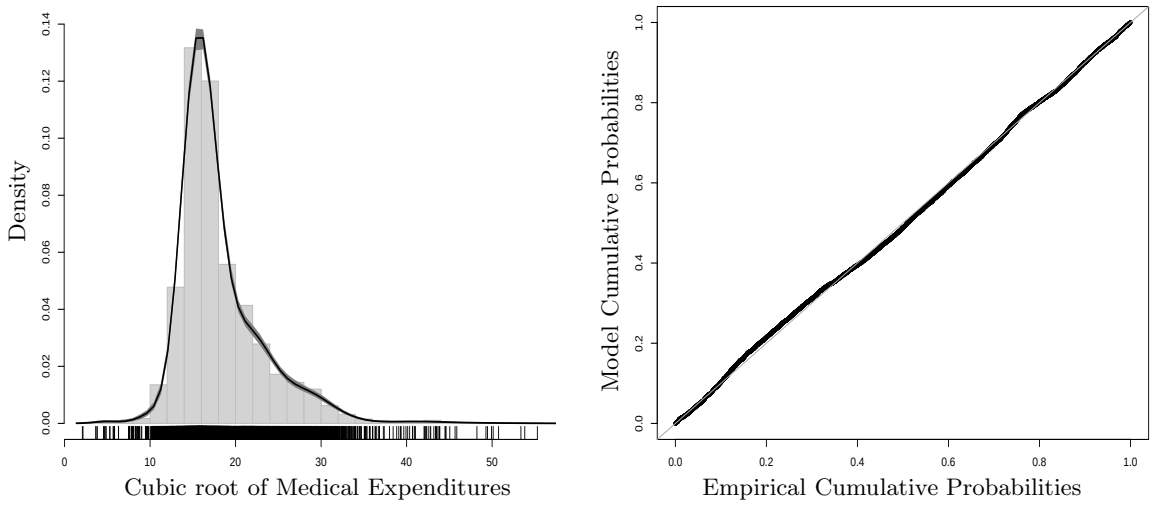


Figure 4: *Simulation study: pairwise comparison of predictive performances. Each point represents a sub-sample of the simulation study. The coordinates of each point are given by the absolute value of the difference between the estimated tail probability on the training and test set for the* $\mathcal{GSM}(\boldsymbol{\pi}, \theta|J)$ *model (on the horizontal axes) and for an alternative model (on the vertical axes), as indicated on top of each plot. Graphs on different rows refer to different medical expenditure thresholds.*



(a) (b)

Figure 5: *MCBS data analysis. Fit of the* $\mathcal{GSM}(\boldsymbol{\pi}, \theta|J)$ *model to the MCBS medical costs related to smoking attributable diseases* ($n = 7,615$ *hospitalizations). (a) The solid line is the posterior mean, while the shaded area is the correspondent 95% credible interval (on the horizontal axis the cubic root transformed data are reported). (b) QQ-plot of the model cumulative probabilities versus the empirical ones.*

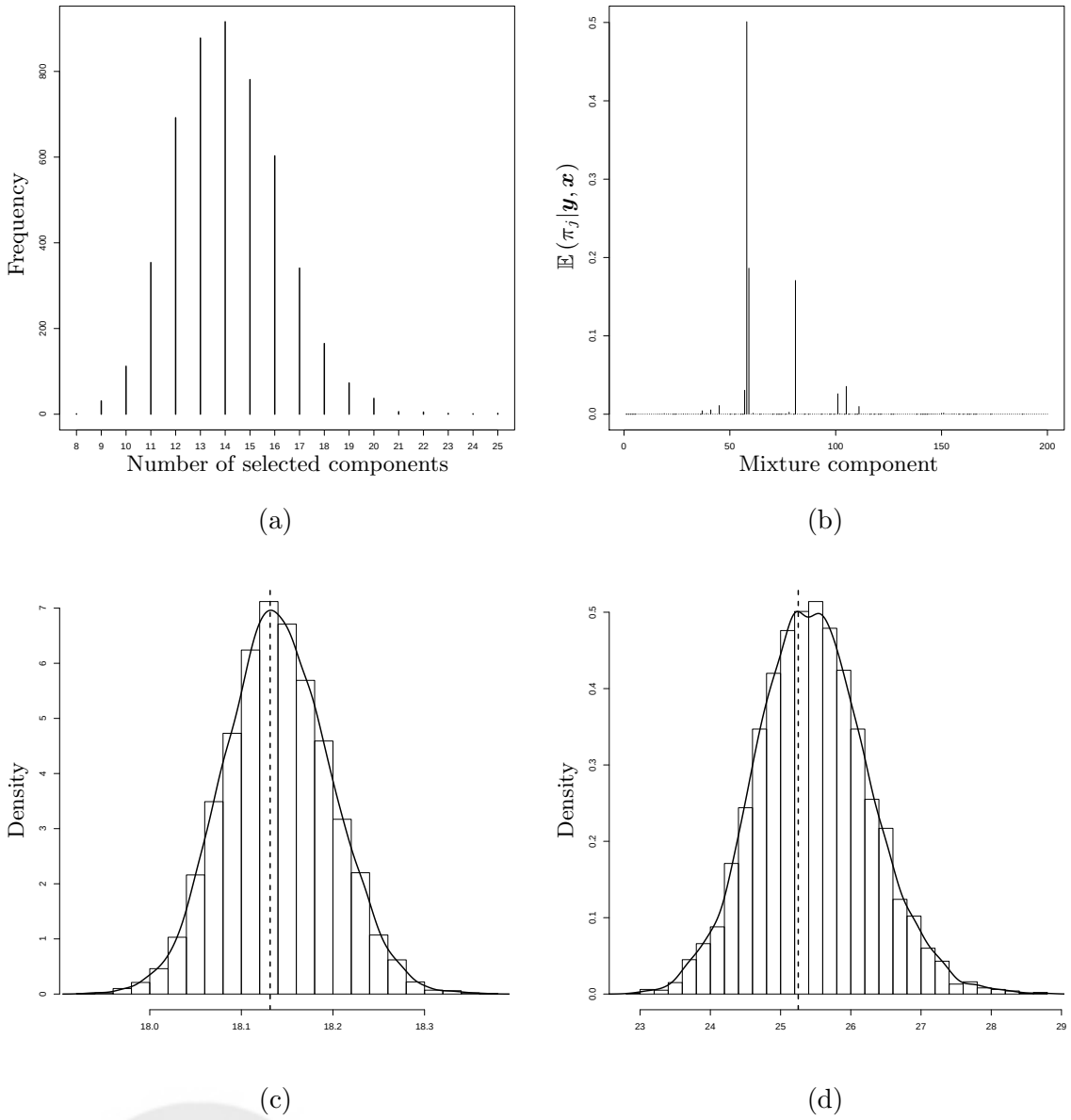


Figure 6: *MCBS data analysis. (a) Posterior mean of the mixture weights. (b) Number of selected mixture components. (c) Posterior distribution of the model mean. The vertical dashed line represents the data sample mean and the over-imposed solid line is the kernel density estimator. (d) Posterior distribution of the model variance. The vertical dashed line indicates the data sample variance, while the over-imposed solid line is the kernel density estimator.*

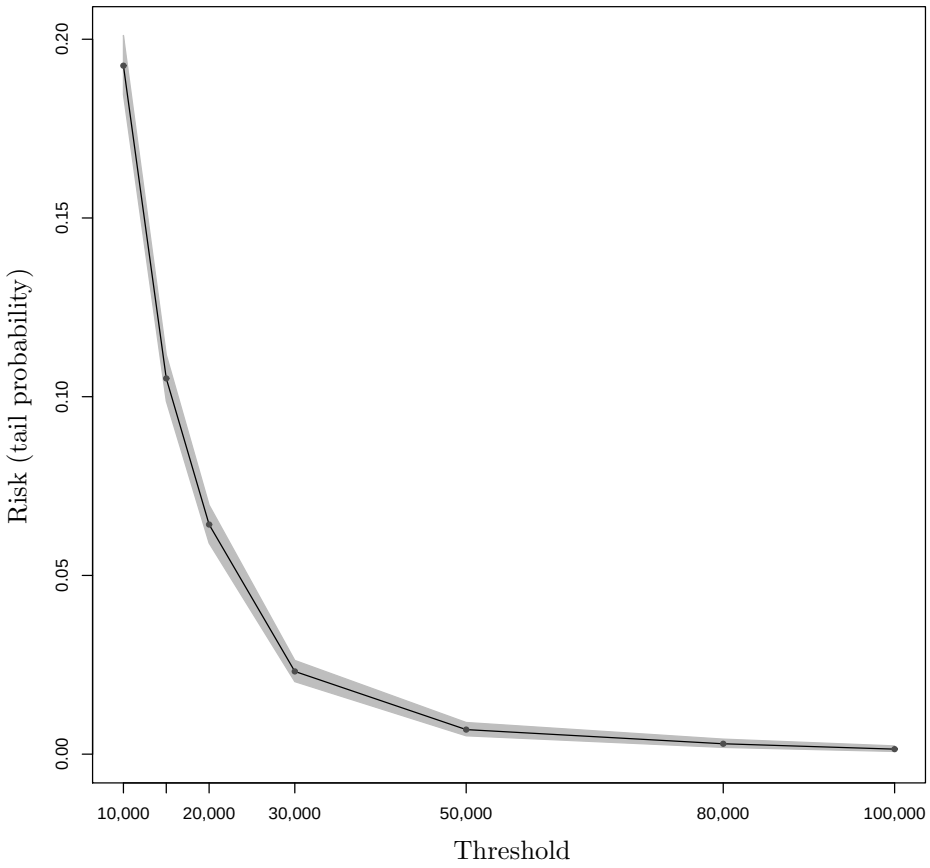


Figure 7: *MCBS data analysis. Risk to exceed a given medical costs threshold in a single hospitalization. Each point corresponds to the estimate of the predictive posterior probability* $\widehat{\mathbb{P}}(y^* > k|\boldsymbol{y})$ *obtained with the* $\mathcal{GSM}(\boldsymbol{\pi}, \theta|J)$ *model on the MCBS data. Shading represents the 95% credible intervals.*