Scandinavian ACTUARIAL Journal

Scandinavian Actuarial Journal

Taylor & Francis
Taylor & Francis Group

ISSN: 0346-1238 (Print) 1651-2030 (Online) Journal homepage: http://www.tandfonline.com/loi/sact20

Dirichlet process mixture models for insurance loss data

Liang Hong & Ryan Martin

To cite this article: Liang Hong & Ryan Martin (2017): Dirichlet process mixture models for insurance loss data, Scandinavian Actuarial Journal, DOI: 10.1080/03461238.2017.1402086

To link to this article: https://doi.org/10.1080/03461238.2017.1402086

Published online: 20 Nov 2017.

SCANDINAVIAN ACTUARIAL JOURNAL, 2017
https://doi.org/10.1080/03461238.2017.1402086

Taylor & Francis
Taylor & Francis Group

Dirichlet process mixture models for insurance loss data

Liang Hong[a] and Ryan Martin[b]

[a]Department of Mathematics, Robert Morris University, Moon, PA, USA; [b]Department of Statistics, North Carolina State University, Raleigh, NC, USA

ABSTRACT

In the recent insurance literature, a variety of finite-dimensional parametric models have been proposed for analyzing the hump-shaped, heavy-tailed, and highly skewed loss data often encountered in applications. These parametric models are relatively simple, but they lack flexibility in the sense that an actuary analyzing a new data-set cannot be sure that any one of these parametric models will be appropriate. As a consequence, the actuary must make a non-trivial choice among a collection of candidate models, putting him/herself at risk for various model misspecification biases. In this paper, we argue that, at least in cases where prediction of future insurance losses is the ultimate goal, there is reason to consider a single but more flexible nonparametric model. We focus here on Dirichlet process mixture models, and we reanalyze several of the standard insurance data-sets to support our claim that model misspecification biases can be avoided by taking a nonparametric approach, with little to no cost, compared to existing parametric approaches.

ARTICLE HISTORY
Received 23 March 2017
Accepted 2 November 2017

KEYWORDS
Danish fire losses data; general insurance; model misspecification; nonparametric Bayes; Norwegian fire losses data; property and casualty insurance; US allocated loss adjustment expenses data

1. Introduction

Prediction of future insurance losses is arguably one of the most important duties for actuaries. This task requires actuaries to adequately model insurance loss data which are often found to be hump-shaped, highly skewed, and heavy-tailed. Several standard parametric models, such as log-normal distribution, log-gamma distribution, Pareto distribution, generalized Pareto distribution, and Weibull distribution (e.g. Embrechts et al. 1997, McNeil 1997, Resnick 1997, Beirlant et al. 2004), have been used to model insurance loss data. While these models are relatively standard (e.g. Klugman et al. 2008), they have some inherent drawbacks (e.g. Cooray & Ananda 2005, Calderín-Ojeda & Kwok 2016). For example, a log-normal distribution is able to capture the fact that loss data are typically hump-shaped and highly skewed, but the tail of log-normal distribution is too thin. On the other hand, Pareto distribution has a heavy tail but it does not fit hump-shaped data very well. This motivated several authors to propose a variety of new parametric models; see, for example, Cooray & Ananda (2005), Scollnik (2007), Pigeon & Denuit (2011), Brazauskas & Kleefeld (2011, 2014, 2016), Eling (2012), Nadarajah & Bakar (2014), Cooray & Cheng (2015), and Calderín-Ojeda & Kwok (2016).

Each of these models is valuable for analyzing data of the appropriate form. However, an actuary charged with analyzing a new set of insurance loss data will not know which of these parametric models is most appropriate. Therefore, he/she must entertain a collection of candidate models and make a non-trivial choice among them. The challenges of defining and selecting a 'most appropriate' model, and the corresponding risk of model misspecification, are fairly well-understood. In addition,

CONTACT Liang Hong hong@rmu.edu

there are yet-to-be-understood challenges in *post-selection inference and/or prediction* that are only recently being explored in the statistics literature (e.g. Taylor & Tibshirani 2015). In light of these serious issues, especially if the ultimate goal of the analysis is prediction of future insurance losses, there is sufficient reason to abandon classical model selection – either via exploratory data analysis or automatic selection tools – in favor of a single but more flexible nonparametric model that can adapt to the important structural features present in the data itself.

In this paper, following the suggestion of Nadarajah & Bakar (2014), we consider a Bayesian nonparametric approach. In particular, we focus on the class of *Dirichlet process mixture models*, which have been widely used, especially for density estimation applications, in the statistics literature. Benefits of this approach include:

- flexibility of the model, via the approximation-theoretic properties of mixtures;
- statistical efficiency, via the nearly optimal posterior rate of convergence;
- and ease of computation, via Markov chain Monte Carlo.

We review each of these points in detail in Section 3. The take-away message is that there really are no theoretical or computational reasons not to consider a Bayesian nonparametric approach. Moreover, the practical benefit of the flexibility of the Dirichlet process mixture model is revealed in the real-data examples in Section 4, where this single nonparametric approach provides a better, or at least no worse, fit compared to the various parametric methods which were motivated by their performance on these very same data-sets.

The remainder of this paper is organized as follows. First, in Section 2, we discuss the general benefits of nonparametric methods for modeling insurance losses. In Section 3, we give a brief review of the Dirichlet process mixtures, their basic properties, and computational methods. In Section 4, we demonstrate the performance of this model by fitting it to several widely studied insurance data-sets to highlight its flexibility. Then some concluding remarks are given in Section 5.

2. Parametric vs. nonparametric

Most would agree that parametric models are simpler in their mathematical form, that fitting these models is a relatively straightforward computational problem, and that (asymptotic) properties of the resulting estimators, confidence intervals, etc., can usually be derived from standard statistical theory. However, this simplicity does not come for free: the actuary must actually take a serious risk in specifying a parametric model. The reality is that the actuary never knows whether their choice of parametric model is appropriate, and making a ‘wrong’ choice may result in some severe model misspecification biases. In particular, since parameters only have an interpretation relative to the model, point estimates, confidence intervals, etc, under a misspecified model are potentially meaningless. For example, if the specified model is gamma, but the true distribution is log-normal, then a confidence interval for the gamma shape parameter has no interpretation because the log-normal has no shape parameter. For prediction problems, the issues are more subtle, but the point is that a ‘predictive distribution’ from a misspecified model will have an incorrect shape, and the derived point and interval predictions are potentially misleading. Of course, a competent actuary would not blindly take a particular parametric model without comparing the quality of fit against that of several other candidate models. But despite the fact that every applied statistics textbook describes various approaches – formal tests, informal graphical procedures, etc – for model comparison and selection, there are potentially serious dangers lurking behind the scenes, only recently discussed in the statistics literature. The specific concern is the so-called *selection effect*, i.e. that using the data itself to select a model may have a non-negligible and elusive effect on the distribution theory used to derive confidence intervals, prediction intervals, etc, making these invalid. A nice general-audience overview of the effects of selection is Taylor & Tibshirani (2015); a discussion of the selection effect in prediction, with an emphasis on insurance applications, is in Hong et al. (2016).

To be clear, the above remarks should not be taken as criticism of parametric models. In fact, we believe that, in most cases, there exists a good parametric model that would lead to an effective statistical analysis. However, identifying that 'good' model may not be so straightforward, and the actuary risks making invalid inference/prediction if the chosen model turns out to be not so good. Moreover, even the standard practice of using the data to help choose a model from a set of candidates can introduce biases that affect the validity of the conclusions. Therefore, it would be advantageous, at least in some cases, if actuaries could avoid these risks associated with model specification.

The only available strategy to avoid the issues arising from the need to select a suitable parametric model is to start with a single sufficiently flexible, all-encompassing model that, roughly speaking, contains all the parametric models the actuary might entertain as special cases. In the context of this paper, this boils down to treating the loss distribution itself as the unknown parameter. Such models are called *nonparametric*,[1] although perhaps a better name is infinite-dimensional parametric. That is, the distribution will be described by a density function p and, without any specified form, identifying p requires knowledge of the value $p(x)$ for each $x \in \mathbb{R}$, and there are infinitely many such values. Of course, if there is no structure imposed on p, then this setup would cover all possible models and, therefore, no risk of model misspecification. But it is desirable to have some structure in p because it effectively reduces the dimension of the problem, thereby improving the actuary's inferential or prediction efficiency based on necessarily finite loss data. Therefore, the actuary must balance their model misspecification risk with statistical and computational efficacy, i.e. specify a nonparametric model which is just wide enough to capture the relevant features of the data, but no wider. Fortunately, there are nonparametric models available which are known to be sufficiently wide and relatively easy to compute; see Section 3.

In what follows, since our primary goal is prediction of future insurance losses, we take a Bayesian approach for the way it naturally incorporates uncertainty for prediction. Therefore, our *nonparametric Bayes* setup is to treat the density function p as the unknown parameter, introduce a prior distribution π for p supported on a space $\mathbb{P}$ of density functions, and combine with (conditionally iid) loss data $X_1, \ldots, X_n$ via Bayes's formula to obtain a posterior distribution $\pi_n(\cdot) = \pi(\cdot \mid X_1, \ldots, X_n)$, given by

$$\pi_n(B) = \frac{\int_B \prod_{i=1}^n p(X_i)\,\pi(\mathrm{d}p)}{\int_{\mathbb{P}} \prod_{i=1}^n p(X_i)\,\pi(\mathrm{d}p)}, \quad B \subseteq \mathbb{P}.$$

To perform the above calculation, one must choose a prior on the space of positive densities. As Ferguson (1973) suggested, the prior should have a sufficiently large support to ensure that the model is flexible enough but, on the other hand, the prior needs to be sufficiently tractable so that posterior computations can be carried out. There are several nonparametric priors available in the literature, such as Dirichlet process prior and Polya tree prior (e.g. Müller & Quintana 2004, Hjort et al. 2010). In this paper, we will focus on the Dirichlet process prior, commonly used for density estimation; see Section 3.

Given the chosen prior π and the loss data $X_1, \ldots, X_n$, the actuary's uncertainty about the loss distribution density function p is encoded in the posterior π_n. This posterior cannot be computed in closed-form, so numerical methods are needed, and visualization of this distribution is a challenge because it is defined on an infinite-dimensional space. However, finite-dimensional marginal posterior distributions can be derived (numerically) for any functional of p, including the mean, various quantiles, and the density function evaluated at a finite collection points. Since our primary focus is prediction, we are interested in the *predictive density*

$$\hat{p}_n(x) = \int p(x)\,\pi_n(\mathrm{d}p), \quad x \in \mathbb{R}. \tag{1}$$

[1] Note that this is different from the classical notion of 'nonparametric' as in Lehmann (2006), where the goal is inference on a finite-dimensional quantity, such as a distribution quantile, under minimal conditions on the distribution itself. Here we are interested in the distribution itself, an infinite-dimensional object.

Formally, this is the (conditional) density function for X_{n+1}, given $X_1, \ldots, X_n$, under the specified Bayesian model and, therefore, is the appropriate distribution to use when considering prediction of a future insurance loss. A decision-theoretic justification for this claim follows from the fact that $\hat{p}_n$ is the Bayes estimator of p (relative to Kullback–Leibler loss). The key point here is that the flexibility inherent in the nonparametric model propagates to the predictive distribution, meaning that $\hat{p}_n$ can take virtually any form, providing quality fit to loss data of various shapes, and accurate predictions; see Section 4.

3. Dirichlet process mixture models

As described above, a nonparametric Bayesian analysis requires a prior distribution for the density function p. This prior π is a probability measure defined on the space $\mathbb{P}$ of density functions, a very complicated object. Here we focus on a particular prior distribution, namely, *Dirichlet process mixtures*. This presentation is designed for readability, so we make no attempt to be mathematically rigorous.

As mentioned in Section 2, the goal is to define a model which balances flexibility with mathematical and computational tractability. Motivated by this, Ferguson (1973) proposed the *Dirichlet process* as a prior distribution on the space of probability measures; see Ghosal (2010) for a detailed review. The Dirichlet process, denoted by $\mathsf{DP}(\alpha, G_0)$, is a distribution characterized by two parameters: a base probability measure G_0 and a precision $\alpha > 0$. These two parameters are to be interpreted as follows: if B is a measurable subset of the support of G_0 and $G \sim \mathsf{DP}(\alpha, G_0)$ is a random probability distribution, then $G(B)$ is a random variable and its expected value and variance are given by

$$\mathsf{E}\{G(B)\} = G_0(B) \quad \text{and} \quad \mathsf{V}\{G(B)\} = \frac{G_0(B)\{1 - G_0(B)\}}{\alpha + 1}.$$

Therefore, G_0 is the mean of $\mathsf{DP}(\alpha, G_0)$ and α controls the concentration of G around the mean G_0 in the sense that larger α means tighter concentration around the mean. The name ‘Dirichlet process’ derives from a close connection with the finite-dimensional Dirichlet distribution (see Ferguson 1973), but a more informative description can be seen from the stick-breaking representation of Sethuraman (1994). Indeed, a random probability measure G having a Dirichlet process distribution, i.e. $G \sim \mathsf{DP}(\alpha, G_0)$, can be expressed as

$$G = \sum_{j=1}^{\infty} w_j \delta_{Z_j},$$

where δ_z denotes a degenerate point-mass measure at z, the random locations $Z_1, Z_2, \ldots$ are iid samples from G_0, and the random weights $w_1, w_2, \ldots$, independent of locations and satisfying $\sum_j w_j = 1$ with probability 1, are obtained by the ‘stick-breaking’ rule

$$w_1 = v_1, \quad w_j = v_j \prod_{k<j} (1 - v_k), \quad \text{where} \quad v_1, v_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Beta}(1, \alpha).$$

This representation reveals that, despite being distribution on distributions, there is actually a lot of structure in a random $G \sim \mathsf{DP}(\alpha, G_0)$. In particular, Sethuraman’s result implies that G is discrete with probability 1. Therefore, a Dirichlet process would not be a suitable prior distribution for the situations we have in mind with continuous loss data. But this does not make the Dirichlet process useless for our purposes.

Mixture models are widely used for density estimation and classification problems (e.g. McLachlan & Peel 2000). Motivated by this, a natural idea is to create a prior for continuous densities via a mixture, where the mixing distribution is given a Dirichlet process prior. To be concrete, consider a normal mixture model

$$p(x) = p_G(x) = \int \mathsf{N}(x \mid \mu, \sigma^2)\, G(d\mu, d\sigma),$$

where $\mathsf{N}(x \mid \mu, \sigma^2)$ is the $\mathsf{N}(\mu, \sigma^2)$ density function, and G is a mixing distribution defined on $\mathbb{R} \times \mathbb{R}^+$. Now construct a prior for the continuous density $p = p_G$ by taking $G \sim \mathsf{DP}(\alpha, G_0)$; the corresponding prior for $p = p_G$ is called a *Dirichlet process mixture of normals* (e.g. Lo 1984, Escobar 1988, MacEachern 1994). This model has been extensively studied and, below, we give a far-from-comprehensive summary of the available results.

Computation: Though the Dirichlet process mixture model is infinite-dimensional, it follows from Sethuraman's representation that it is at most only countably infinite. Intuitively, one may expect that data of size n will not admit a posterior that is overly complex. A famous result of Blackwell & MacQueen (1973) confirms this intuition by showing, among other things, that samples p from the posterior π_n are mixtures with at most n mixture components. Therefore, the posterior under a Dirichlet process mixture of normal models is effectively finite-dimensional, though the dimension is adaptive, determined by data, as opposed to fixed like in the parametric models. This adaptive dimensionality is what gives the model its flexibility and its effective finite-dimensionality is what makes posterior computation possible. The various summaries of π_n that actuaries might be interested in, e.g. probabilities, expectations, etc., all involve integrals, and since integrals can be approximated by averages, the available computational methods (e.g. Escobar & West 1995, MacEachern and Müller 1998, Neal 2000, Kalli et al. 2011) all focus on producing (approximate) samples $p \sim \pi_n$ via Markov chain Monte Carlo. In particular, given a sample $\{p^{(m)} : m = 1, \ldots, M\}$ of size M from the posterior π_n, the predictive density in (1) is approximated by

$$\hat{p}_n(x) \approx \frac{1}{M} \sum_{m=1}^{M} p^{(m)}(x), \quad x \in \mathbb{R}^+.$$

Theory: Recall that virtually any continuous density function on $\mathbb{R}$ can be well-approximated by a finite location–scale mixture of normal densities (e.g. DasGupta 2008, p. 572). Then the fact that the posterior π_n corresponding to the Dirichlet process mixture of normals prior is supported on finite mixtures with an adaptive number of components suggests that it is flexible enough to identify a continuous density of virtually any form, given a sufficient amount of data. Indeed, Tokdar (2006) shows, among other things, that if $X_1, \ldots, X_n$ are iid with true density $p^\star$,

$$\lim_{n\to\infty} \int |\hat{p}_n(x) - p^\star(x)|\, dx = 0 \quad \text{almost surely (with respect to } p^\star),$$

provided that $p^\star$ does not have tails heavier than Cauchy; see, also, Wu & Ghosal (2008). The point is that the true density need not be a mixture of normals for it to be accurately estimated by the predictive $\hat{p}_n$ for large n, and this is what reduces the actuary's model misspecification risk. A more precise picture of the statistical efficiency can be seen by looking at the rate of convergence, i.e. the most rapidly decaying sequence ε_n such that

$$\lim_{n\to\infty} \frac{1}{\varepsilon_n} \int |\hat{p}_n(x) - p^\star(x)|\, dx = 0 \quad \text{in probability (with respect to } p^\star).$$

The rate of convergence depends on features of $p^\star$, such as smoothness, and several cases are investigated in, e.g. Ghosal & van der Vaart (2001, 2007). Without getting into unnecessary technical details here, the general conclusion is that the rate ε_n for the Dirichlet process mixture of normals predictive is roughly equal to the optimal (minimax) rate over various classes of true densities $p^\star$. For example, if $p^\star$ is twice continuously differentiable, then the optimal rate is $n^{-2/5}$ and Ghosal and van der Vaart (2007) show that a particular choice of Dirichlet process mixture model attains the rate $\varepsilon_n = (\log n)^{4/5} n^{-2/5}$, so the two rates are of the same order, up to the negligible logarithmic term. This theory implies that the actuary cannot do appreciably better in terms of asymptotic accuracy using a method other than the Dirichlet process mixture of normals model described above.

Insurance losses are typically non-negative, so the loss distribution should be supported on $\mathbb{R}^+$. Therefore, a mixture of normals, which is supported on $\mathbb{R}$, may not be appropriate. A simple fix, suggested by Hong & Martin (2017), is to fit the Dirichlet process mixture of normal model to the log-loss variables, i.e. to $Y_i = \log X_i$, $i = 1, \ldots, n$. This turns out to be equivalent to a Dirichlet process mixture of log-normal model, where the continuous density p supported on $\mathbb{R}^+$ has a prior distribution π described by the formula

$$p(x) = \int \mathsf{logN}(x \mid \mu, \sigma^2)\, G(d\mu, d\sigma), \quad G \sim \mathsf{DP}(\alpha, G_0),$$

where $\mathsf{logN}(x \mid \mu, \sigma^2) = x^{-1}\mathsf{N}(\log x \mid \mu, \sigma^2)$ is the density function of a log-normal distribution. To complete specification of the prior π for p, the actuary must specify the base measure G_0 and the precision parameter α. For computational simplicity, it is common to take the base measure G_0 to be the joint distribution of independent μ and σ, where $\mu \sim \mathsf{N}(m, s^2)$ and $(\sigma^2)^{-1} \sim \mathsf{Gamma}(a, b)$; here $\mathsf{Gamma}(a, b)$ denotes the gamma distribution with shape and rate parameters a and b, respectively. The choice of hyperparameters (m, s^2, a, b) is up to the actuary, and interpretation of G_0 as the prior guess can be helpful in eliciting values for these parameters; it is also possible to build a deeper hierarchical model by introducing prior distributions for these hyperparameters. In cases where no genuine prior information is available about these hyperparameters, some 'default' choices can be recommended. For example, Hong & Martin (2017) set (m, s^2, a, b) according to the rule of thumb presented in Green & Richardson (2001) and, for the precision parameter, they take $\alpha = 1$.

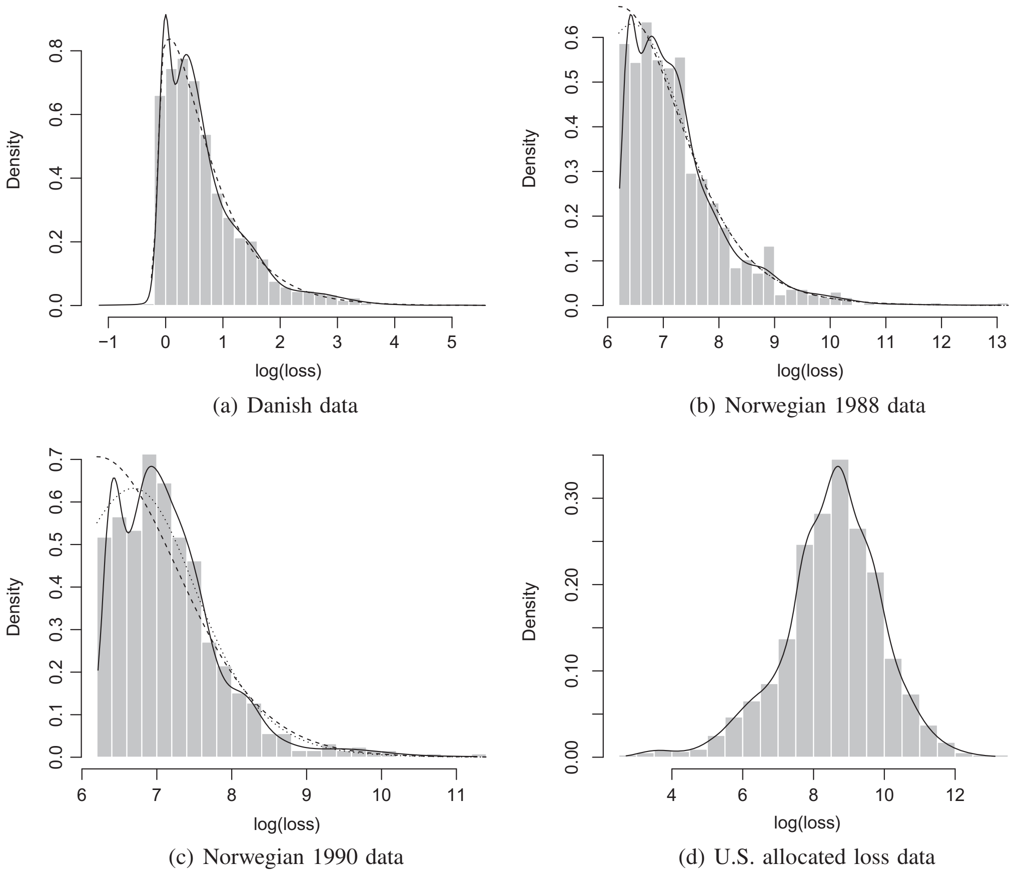
Given that this mixture of log-normal model can be viewed as a mixture of normals applied to the log-loss data, the Markov chain Monte Carlo techniques discussed above can be readily applied to evaluate the posterior distribution π_n for the density p and, in particular, to obtain the predictive density $\hat{p}_n$. Details for sampling $p^{(m)}$ from the Dirichlet process mixture of log-normal model, via the slice-sampling algorithm in Kalli et al. (2011), can be found in Hong & Martin (2017), and R codes are available in supplementary materials.

Again, since the mixture of log-normal model is closely tied to the more classical mixture of normals model described above, much of the theory for the posterior and predictive carries over word for word. Indeed, Hong & Martin (2017) establish conditions under which the predictive is consistent in the sense described above and, in particular, consistency holds even for cases where the true density $p^\star$ is not a mixture of log-normals. Details about the rate of convergence have yet to be worked out, but our conjecture is that the conclusions are the same as those coming from the references given above for normal mixtures.

4. Numerical examples

There are several loss data-sets that have been regularly analyzed in the insurance literature, and here we compare the results of fitting the Dirichlet process mixture of log-normal model described above, using the 'default' prior specification, and $M = 5000$ posterior samples, with the most cutting-edge method applied to these data, as of writing this paper. These existing methods are based on various data transformations and, here, to provide a unified comparison, we fit the models based on the transformations proposed in the respective papers, and then transform the fitted densities to the log-loss scale for display.

(a) *Danish fire insurance loss data.* The Danish data have been analyzed many authors such as Cooray & Ananda (2005), Scollnik (2007), Eling (2012), Nadarajah & Bakar (2014), Bakar



(a) Danish data

(b) Norwegian 1988 data

(c) Norwegian 1990 data

(d) U.S. allocated loss data

Figure 1. Histograms of the four data-sets described above, along with the various fitted densities. In each panel, the solid line corresponds to the Dirichlet process mixture of normals model fit to the log-loss data.

et al. (2015), Cooray & Cheng (2015), Calderín-Ojeda & Kwok (2016). It consists of $n = 2,492$ fire insurance losses (in Danish krones) in Copenhagen from 1980 to 1990, adjusted for inflation relative to the 1985 values. Calderín-Ojeda & Kwok (2016) recently proposed a few composite Stoppa models and compared their method to several others in the literature, concluding that their Weibull–Stoppa model is superior. The fitted density for this method, along with that for the Dirichlet process mixture of normals are shown in Figure 1(a).

(b) *1988 Norwegian fire claims data.* The Norwegian fire claims data have also been studied by quite a few authors such as Brazauskas (2009), Brazauskas & Kleefeld (2011, 2014, 2016), Nadarajah & Bakar (2015), and Scollnik (2014). This set is comprised of $n = 827$ fire loss claims exceeding 500 thousand Norwegian krones during the year 1988. Plots of the two fitted densities based on the folded-t (dashed) and truncated folded-t (dotted) models in Brazauskas and Kleefeld (2011, 2014, 2016), along with the Dirichlet process mixture of normals, are shown in Figure 1(b).

(c) *1990 Norwegian fire claims data.* The 1990 Norwegian fire claims data consist of $n = 628$ fire loss claims exceeding 500 thousand Norwegian krones during the year 1990. Figure 1(c) compares the Dirichlet process mixture with the two fitted densities from same two folded-t models in Example (b).

(d) *US allocated loss adjustment expenses data.* Eling (2012) and Bakar et al. (2015) recently analyzed the US allocated loss adjustment expenses data. This set has $n = 1,500$ general liability claims recorded in US dollars. Each claim is comprised of an indemnity payment and an allocated loss adjustment expense, both in US dollars. Here, we follow Eling (2012) and Bakar et al. (2015) to focus on the latter part only. Figure 1(d) shows the data histogram (on the log-loss scale) with the Dirichlet process mixture of normals predictive density overlaid.

The overall conclusion we draw from these four analyses is that the same nonparametric approach, in this case, based on the Dirichlet process mixture of normals model is able to capture the shape of all four data-sets very well, whereas different parametric approaches are needed for each data-set and their respective fits are arguably no better overall. The quality fit of the nonparametric models is important, but the fact that it does not require the actuary to consider a collection of parametric models that may or may not be appropriate suggests that the various model misspecification biases can be avoided. Therefore, if the actuary's goal is to fit a model for the purpose of prediction, and there is a concern about the choice of a satisfactory parametric model, then we argue that there is no reason not to consider a nonparametric approach such as the one illustrated here, which can avoid potential model misspecification biases.

5. Concluding remarks

Recent insurance literature saw a variety of parametric models being motivated by some specific sets of insurance data such as Danish fire data and Norwegian fire data. In this paper, we demonstrated that a Dirichlet process mixture of log-normal model does as good a job as these models in capturing the shape of the true densities uniformly cross these insurance data-sets. To reiterate the point made in Section 1, we are not claiming that this particular nonparametric approach is 'better' than any of these parametric models, only that this nonparametric approach is (by definition) more flexible, has theoretical support, and performs competitively in the standard test cases.

Our motivation here was to address the potential biases that can result from the consideration of multiple parametric models, either through the choice of a 'wrong' model or from the fact that data would be used to help choose a particular model, the so-called selection effect. Switching from a set of candidate parametric models to a larger nonparametric model helps to alleviate these biases, but there is no free lunch. Certainly, theory and computation is more difficult in nonparametric models than in the more familiar parametric ones, but, fortunately, there is already a large statistical literature on this. Beyond these technical matters, there are other nonparametric approaches besides the one here can be entertained, either based on different hyperparameter settings or an altogether different prior for p. Since all these nonparametric approaches would be sufficiently flexible, it is reasonable to expect that there is less risk of bias in the choice among these, but nevertheless there is a choice to be made. We make no claims that the model employed in this paper is universally 'best' among these nonparametric methods, only that it is relatively simple, with strong theoretical support, and good empirical performance in a range of problems. So, actuaries concerned with the potential biases resulting from selecting a parametric model have at least one justifiable and relatively simple nonparametric method at hand to turn to.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Bakar, S. A. A., Hamzah, N. A., Maghsoudi, M. & Nadarajah, S. (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics* **61**, 146–154.

Beirlant, J., Joossens, E. & Segers, J. (2004). Discussion of Generalized Pareto fit to the society of actuaries' large claims database. *North American Actuarial Journal* **8**(2), 108–110.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics* **1**, 353–355.

Brazauskas, V. (2009). Robust and efficient fitting of loss models: diagnostic tools and insights. *North American Actuarial Journal* **13**(3), 356–369.

Brazauskas, V. & Kleefeld, A. (2011). Folded and log-folded-*t* distributions as models for insurance loss data. *Scandinavian Actuarial Journal* **1**, 59–74.

Brazauskas, V. & Kleefeld, A. (2014). Author's reply to letter to the editor: regarding folded models and the paper by Brazauskas and Kleefeld (2011) by Scollnik. *Scandinavian Actuarial Journal* **1**, 59–74.

Brazauskas, Y. & Kleefeld, A. (2016). Modeling severity and measuring tail risk of Norwegian fire claims. *North American Actuarial Journal* **20**(1), 1–16.

Calderín-Ojeda, E. & Kwok, C. F. (2016). Modeling claims data with composite Stoppa models. *Scandinavian Actuarial Journal* **9**, 817–836.

Cooray, K. & Ananda, M. A. M. (2005). Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal* **5**, 321–334.

Cooray, K. & Cheng, C. I. (2015). Bayesian estimators of the lognormal-Pareto composite distribution. *Scandinavian Actuarial Journal* **6**, 500–515.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. New York: Springer.

Eling, M. (2012). Fitting insurance claims to skewed distributions: are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics* **51**, 239–248.

Embrechts, P., Kluüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. New York: Springer-Verlag.

Escobar, M. D. (1988). *Estimating the means of several normal populations by estimating the distribution of the means*. Ph.D. dissertation. Department of Statistics, Yale University.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T. S. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ghosal, S. (2010). *The Dirichlet process, related priors and posterior asymptotics*, Bayesian nonparametrics pp. 35–79. Cambridge: Cambridge University Press.

Ghosal, S. & van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics* **29**, 1233–1263.

Ghosal, S. & van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics* **35**, 697–723.

Green, P. J. & Richardson, S. (2001). Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Hjort, N., Holmes, C., Müller, P. & Walker, S. G. (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hong, L., Kuffner, T. & Martin, R. (2016). On prediction of future insurance claims when the model is uncertain. http://ssrn.com/abstract=2883574

Hong, L. & Martin, R. (2017). A flexible Bayesian nonparametrics model for predicting future insurance claims. *North American Actuarial Journal* **21**, 228–241.

Kalli, M., Griffin, J. E. & Walker, S. G. (2011). Slice sampling mixture models. Statistical. *Computing* **21**, 93–105.

Klugman, S. A., Panjer, H. H. & Willmot, G. E. (2008). *Loss models: from data to decisions*, 3rd ed. Hoboken: Wiley.

Lehmann, E. L. (2006). *Nonparametrics: statistical methods based on ranks*, revised ed. Springer: New York.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I. *Density estimates. Annals of Statistics* **12**, 351–357.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation & Computation* **23**, 727–741.

MacEachern, S. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. Hoboken, NJ: Wiley.

McNeil, A. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin* **27**, 117–137.

Müller, P. & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

Nadarajah, S. & Bakar, S. A. A. (2014). New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal* **2**, 180–187.

Nadarajah, S. & Bakar, S. A. A. (2015). New folded models for the log-transformed Norwegian fire claim data. *Communications in Statistics-Theory and Methods* **44**, 4408–4440.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Pigeon, M. & Denuit, M. (2011). Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal* **3**, 177–192.

Resnick, S. I. (1997). Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin* **27**, 139–151.

Scollnik, D. P. M. (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal* **1**, 20–33.

Scollnik, D. P. M. (2014). Letter to editor: regarding folded models and the paper by Brazauskas and Kleefeld (2011). *Scandinavian Actuarial Journal* **2014**(3), 278–281.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Taylor, J. & Tibshirani, R. J. (2015). Statistical learning and selective inference. *PNAS* **112**, 7629–7634.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā* **67**(4), 90–110.

Wu, Y. & Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics* **2**, 298–331.