Insurance: Mathematics and Economics 108 (2023) 177–189

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Nonparametric density estimation and risk quantification from tabulated sample moments

Philippe Lambert [a,b,*]

[a] Institut de Mathématique, Université de Liège, Grande Traverse 12 (B37), B-4000 Liège, Belgium
[b] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received July 2021
Received in revised form October 2022
Accepted 16 December 2022
Available online 22 December 2022

JEL classification:
C14
C10
C18

Keywords:
Nonparametric density estimation
Grouped data
Tabulated sample moments
Value-at-Risk
P-splines

ABSTRACT

Continuous data such as losses are often summarized by means of histograms or displayed in tabular format: the range of data is partitioned into consecutive interval classes and the number of observations falling within each class is provided to the analyst. This paper investigates how the additional report of sample moments within each class can be integrated to obtain a smooth nonparametric estimate of the density and credible intervals for the loss quantiles. Extensive simulations confirm the merits of the proposed methodology with correctly estimated densities based on tabulated sample moments of increasing orders and effective coverages of credible intervals close to their nominal values, even when the number of classes is small. An application on motor insurance data further illustrates the usefulness of the method with an estimation of the loss density and of Value-at-Risk.

1. Introduction and motivation

In risk analysis, monetary losses are generally modelled as non-negative random variables entering aggregate definition of the underlying risks, see e.g. Bolancé et al. (2003) and Klugman et al. (2019) for an in-depth treatment in an actuarial framework. Analysts often need easy-to-compute approximations of quantities relating to the risks they consider. Several methods based on two to four moments of the underlying loss distribution have been proposed in the literature: it includes the classical Central Limit theorem, the Normal Power approximation (Pentikäinen, 1987) based on Edgeworth expansion, the shifted-Gamma approximation (Hardy, 2006), the maximum entropy principle (Brockett et al., 1995), mixed Erlang distributions (Cossette et al., 2016) or generalized Gamma convolutions (Laverny et al., 2021; Furman et al., 2021) with the references therein, to cite a few. Numerous moment bounds have also been developed in probability and actuarial science, originating in the analytic problem of moments. Actuaries and quantitative risk managers indeed sometimes act in a conservative way by basing their decisions on the least attractive risk that is consistent with the incomplete available information (here, the range and the first moments). Since Markov fundamental inequality, a number of improvements have been obtained under additional assumptions on the underlying distribution function. Bernard et al. (2018) recently provided a new derivation of moment bounds on distribution functions and Value-at-Risk (VaR) measures, revisiting previous contributions to the literature. Besides distribution functions and VaR measures, bounds have also been derived on stop-loss premiums and Tail-VaR, for instance. Hürlimann (2008) provides a useful review of the available results. Notice that these bounds are also related to local moment matching methods, see Courtois and Denuit (2007).

Density estimation from grouped data has been studied in multiple publications with applications in e.g. actuarial science, biology, demography, economy, environmental science or epidemiology. Most of them use the bin limits and the associated data frequencies to produce an estimate of the underlying density. It is relatively straightforward using likelihood-based techniques when the latter is assumed to belong to a known parametric class, see e.g. Hasselblad et al. (1980) with the Lognormal distribution and the expectation-maximization (EM) algorithm (Dempster et al., 1977) in an epidemiological context. Local likelihood methods based on kernels with coefficients of the

* Correspondence to: Institut de Mathématique, Université de Liège, Grande Traverse 12 (B37), B-4000 Liège, Belgium.
E-mail address: p.lambert@uliege.be.

https://doi.org/10.1016/j.insmatheco.2022.12.004
0167-6687/

polynomials selected using the EM-algorithm have been studied by Braun et al. (2005) to estimate a density from interval-censored or aggregated data. When the density is just assumed to be smooth without any additional restrictive parametric assumption, the combination of P-splines (Eilers and Marx, 1996) and of the composite link model (Thompson and Baker, 1981) was successfully used in a frequentist (Eilers, 2007; Rizzi et al., 2015) or in a Bayesian framework (Lambert and Eilers, 2009; Jaspers et al., 2016). Papkov and Scott (2010) directly smooth a histogram using splines with penalized least-squares and account for unconditional bin moments.

In the present paper, we propose an efficient nonparametric estimation procedure for the loss density based on histograms or grouped data, with additional information about class-specific sample moments. Specifically, the analyst has access to a set of observed losses, grouped into consecutive classes. Graphically, this corresponds to an histogram. In addition to these grouped data, the average value of the observations in each class is provided, as well as the corresponding local variance, skewness and kurtosis. This format is often encountered in practice. For instance, in banking and insurance contexts, operational risk loss data in the ORX annual report (published by the operational risk management association) are tranched and the total number of loss events as well as the total gross loss falling within loss size boundaries are provided. Reinsurers also often display the information about insurance losses in this way. Confidentiality issues may sometimes justify this grouping procedure. We show how to obtain a smooth, nonparametric density estimate based on this information, while taking into account the observed moments and their joint sampling distribution. The simulation study reported in Section 4 suggests that the additional information contained in class-specific moments greatly improves the accuracy of the estimation. Of course, the proposed method can also be applied to individual data. It suffices to group them in an arbitrary number of classes and to compute the corresponding sample moments.

The remainder of this paper is organized as follows. Section 2 formally describes the problem under investigation. In Section 3, we explain how to get a smooth estimate of the density based on tabulated summary data including frequencies and up to four class-specific sample moments. As intermediate statistical goals, we also aim to quantify uncertainty for the density estimate and derived quantities and to evaluate the contribution of the different descriptive measures on the density estimate (to issue recommendations for future reporting). Section 4 is devoted to a simulation study assessing the performances of the proposed approach. In Section 5, we analyze a set of insurance losses and we illustrate the added value of our new method. The final Section 6 discusses the results and proposes some research perspectives.

2. Problem under investigation

Consider a univariate random variable X with values in a bounded interval $\mathcal{X}$, distribution function (c.d.f.) F and continuous density function f. Our starting point is a set of n independent observations $x_1, x_2, \ldots, x_n$ of X summarized in a tabular form. More specifically, $\mathcal{X} = [a, b]$ has been partitioned into consecutive class intervals $\mathcal{C}_j = (a_{j-1}, a_j]$, $j = 1, \ldots, J$, where the cut points $a_0, a_1, \ldots, a_J$ satisfy

$$-\infty < a = a_0 < \min x_i < a_1 < \ldots < a_{J-1} < \max x_i < a_J = b < +\infty.$$

In addition to the number n_j of observations belonging to $\mathcal{C}_j$, we also have summary statistics about observations in each class. Specifically, we assume that we know the class-specific means

$$\overline{x}_j = \frac{1}{n_j} \sum_{x_i \in \mathcal{C}_j} x_i, \quad j = 1, \ldots, J,$$

as well as sample centered moments m_{kj}, $k = 2, 3, 4$, defined as

$$m_{kj} = \frac{1}{n_j} \sum_{x_i \in \mathcal{C}_j} \left(x_i - \overline{x}_j\right)^k, \quad j = 1, \ldots, J.$$

Here, we consider the cases where variances, $s_j^2 = m_{2j}$, skewness coefficients, $g_{1j} = m_{3j}/m_{2j}^{3/2}$, and kurtosis, $g_{2j} = m_{4j}/m_{2j}^{4/2} - 3$, are available in addition to the means $\overline{x}_j$.

Nonparametric computations are often carried out using the empirical distribution function, assuming a uniform distribution of the class relative frequencies n_j/n over $\mathcal{C}_j$, see e.g. Klugman et al. (2019, Chap. 14). This allows the risk analyst, for a given real function g integrable on the domain of X, to estimate $\mathrm{E}[g(X)]$ by

$$\sum_{j=1}^{J} \frac{n_j}{n(a_j - a_{j-1})} \int_{a_{j-1}}^{a_j} g(t)dt.$$

This standard approach does not use any information about the structure of the observed data inside each class. Arbitrarily assuming a uniform distribution of the losses in each risk class $\mathcal{C}_j$ is contradicted by data for instance if $\overline{x}_j$ significantly differs from $(a_{j-1} + a_j)/2$. The approach proposed in the present paper integrates the information about class-specific sample moments of order 1 to k ($k \leq 4$) in the smooth density estimate under the assumption that X has finite moments of order 1 to $2k$ in any subset of the partition of $\mathcal{X}$ defined by $\mathcal{C}_j$, $j = 1, \ldots, J$. Using that estimate, expectations of functions of the real random variable X can be computed, as well as risk measures defined from the quantile function (Embrechts and Hofert, 2013), $Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ for $0 \leq p \leq 1$, such as the Value-at-Risk (Jorion, 2006) at probability level p, $\mathrm{VaR}_p(X) = Q(p)$.

3. Methodology

3.1. Density estimation for given class frequencies using the EM algorithm

The starting point of the proposed approach, summarized in this subsection, is based on Lambert and Eilers (2009), but with the EM algorithm substituted to the modified Langevin-Hastings algorithm for parameter estimation. Estimation could also be performed using the

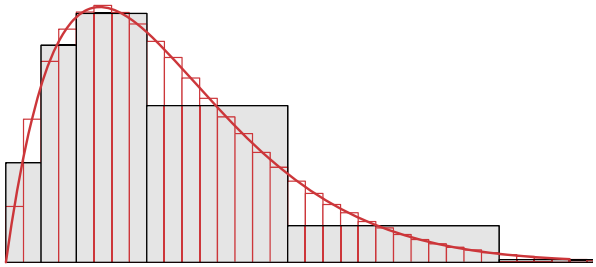

Fig. 1. Density f (continuous, red curve), latent distribution $\boldsymbol{\pi}$ (thin rectangles) and observed histogram (gray blocks). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

strategy described in Eilers (2007) based on the composite link model (Thompson and Baker, 1981), but the complete likelihood coupled with the EM algorithm turns out to be ideally suited to integrate information based on tabulated sample moments, see Section 3.2.

Assume that the support $\mathcal{X}$ of the density f is partitioned into I narrow bins $\mathcal{I}_i = (b_{i-1}, b_i]$ with midpoint u_i $(i = 1, \ldots, I)$ and with equal width Δ taken small enough to give an accurate description of f for computing quantiles or other indices accurately. The relationship between class $\mathcal{C}_j$ and the narrow bins $\mathcal{I}_i$ is quantified by means of the $J \times I$ matrix $\mathbf{C} = (c_{ji})$ where c_{ji} is the proportion of $\mathcal{I}_i$ included in $\mathcal{C}_j$. In particular, it implies that $c_{ji} = 1$ if $\mathcal{I}_i \subset \mathcal{C}_j$, while $c_{ji} = 0$ if $\mathcal{I}_i \cap \mathcal{C}_j = \emptyset$. Let $\pi_i = \int_{b_{i-1}}^{b_i} f(t)dt = f(u_i)\Delta + \mathcal{O}(\Delta^2)$ and $\boldsymbol{\pi} = (\pi_i)_{i=1}^I$ denote the probability masses associated to the partition of the support. Fig. 1 illustrates the construction (for a value of Δ much larger than what we use in practice). Consider a cubic B-spline basis $\{b_k(\cdot)\}_{k=1}^K$ associated to a large number of equidistant knots on $\mathcal{X}$. Given the multinomial sampling distribution of the frequencies, we model the probabilities in $\boldsymbol{\pi}$ using multinomial logistic regression (also called polytomous logistic regression) (see e.g. Hosmer and Lemeshow, 2000),

$$\pi_i = \pi_i(\boldsymbol{\theta}) = \frac{\exp(\eta_i)}{\exp(\eta_1) + \ldots + \exp(\eta_I)} \;\; ; \;\; \eta_i = \eta_i(\boldsymbol{\theta}) = \sum_{k=1}^{K} b_k(u_i)\theta_k = [\mathbf{B}\boldsymbol{\theta}]_i \,, \tag{1}$$

with each column of the matrix $[\mathbf{B}]_{ik} = b_k(u_i)$ containing one of the K B-splines in the basis evaluated at the small bin midpoints. In this setting, one does not observe empirical counterparts to $\boldsymbol{\pi}$ itself, but only sums over the $J < I$ classes $\mathcal{C}_j$, $j = 1, \ldots, J$. The probability mass associated to $\mathcal{C}_j$ is $\gamma_j = \sum_{i=1}^I c_{ji}\pi_i$, or in matrix form, $\boldsymbol{\gamma} = \mathbf{C}\boldsymbol{\pi}$.

We propose to use the EM algorithm to estimate the probabilities $\boldsymbol{\pi}$ from the latent (unobserved) small bin frequencies $\mathbf{k} = (k_1, \ldots, k_I)$ for given observed class frequencies $\mathcal{D}_0 = \{\mathbf{n}\}$ where $\mathbf{n} = (n_1, \ldots, n_J)$. Given that $X_1, \ldots, X_n$ are independent and identically distributed, the small bin frequencies $k_1, \ldots, k_I$ can be handled as the realization of a Multinomial random vector with the sample size n as exponent and probability vector $\boldsymbol{\pi}$. While the observed likelihood is $\ell(\boldsymbol{\theta}|\mathcal{D}_0) = \sum_{j=1}^J n_j \log \gamma_j(\boldsymbol{\theta})$, the complete log-likelihood is given by $\ell^c(\boldsymbol{\theta}|\mathcal{D}_0^c) = \sum_{i=1}^I k_i \log \pi_i(\boldsymbol{\theta})$ where $\mathcal{D}_0^c = \{\mathbf{k}\}$. A discrete roughness penalty is put on rth differences of consecutive B-splines coefficients to force smoothness on the density estimate (Eilers and Marx, 1996). Then, the penalized log-likelihood is

$$\ell_p(\boldsymbol{\theta}|\mathcal{D}_0, \lambda) = \ell(\boldsymbol{\theta}|\mathcal{D}_0) - \frac{\lambda}{2}\|\mathbf{D}\boldsymbol{\theta}\|^2 \,,$$

when using the observed class frequencies and

$$\ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c, \lambda) = \ell^c(\boldsymbol{\theta}|\mathcal{D}_0^c) - \frac{\lambda}{2}\|\mathbf{D}\boldsymbol{\theta}\|^2 \,,$$

when using the complete frequency data, where $\lambda > 0$ is a penalty parameter tuning the smoothness of the density estimate and $\mathbf{D}$ the rth order differencing matrix of size $(K - r) \times K$. For instance, with 2nd order differences ($r = 2$), we have

$$\mathbf{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

We propose to use Algorithm 1 to perform estimation in the described context.

Algorithm 1 *(Density estimation for given class frequencies).* The following EM algorithm alternates the update of the estimates for the latent frequencies $\mathbf{k}$, $\boldsymbol{\theta}$ and, possibly, λ using the following steps till convergence:

1. E-step: $k_i \leftarrow \mathrm{E}(k_i|\boldsymbol{\theta}, \mathcal{D}_0) = \sum_{j=1}^J n_j c_{ji}\pi_i(\boldsymbol{\theta})/\gamma_j(\boldsymbol{\theta})$.
2. M-step: $\boldsymbol{\theta} = \arg\max_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c, \lambda)$. This can be done using penalized iteratively weighted least squares (P-IWLS) or a Newton-Raphson (N-R) algorithm. Let us detail the last algorithm. Based on the following explicit forms for the gradient $\nabla_\theta \ell_p^c$ and the Hessian matrix $\mathbf{H}_p^c = \nabla_\theta^2 \ell_p^c$,

$$\nabla_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c, \lambda) = \mathbf{B}^\top(\mathbf{k} - n\boldsymbol{\pi}) - \lambda\mathbf{P}\boldsymbol{\theta} \;\; ; \;\; -\mathbf{H}_p^c = -\nabla_\theta^2 \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c, \lambda) = \mathbf{B}^\top\mathbf{W}\mathbf{B} + \lambda\mathbf{P} \,, \tag{2}$$

where $\mathbf{P} = \mathbf{D}^\top \mathbf{D}$ is the $K \times K$ penalty matrix and $(\mathbf{W})_{ii'} = n(\pi_i \delta_{ii'} - \pi_i \pi_{i'})$, the N-R algorithm repeats, till convergence, the following substitution: $\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \left(-\mathbf{H}_p^c + \epsilon\, \mathrm{I}_K\right)^{-1} \nabla_\theta \ell_p^c$. The addition of a small multiple of the identity matrix to the Hessian before inversion in the N-R step is a ridge penalty that conveniently handles the identification problem in (1), as $\pi_i(\boldsymbol{\theta}) = \pi_i(\boldsymbol{\theta} + c)$ for any constant c.
3. Penalty update: $\lambda \leftarrow (\mathrm{edf} - r)/\|\mathbf{D}\boldsymbol{\theta}\|^2$ where the effective number of spline parameters is given by the trace of a matrix, $\mathrm{edf}(\lambda) = \mathrm{Tr}\left((-\mathbf{H}_p^c + \epsilon\, \mathrm{I}_K)^{-1}(-\mathbf{H}_p^c - \lambda \mathbf{P})\right)$.

The last step is optional as one might prefer to perform the estimation procedure for a given value of the penalty parameter and rely on an external ad-hoc strategy to select λ. At convergence, one obtains the penalized MLE $\hat{\boldsymbol{\theta}}_\lambda$ (given the selected value for the penalty parameter λ). □

3.2. Moment-based extensions

3.2.1. Density estimation for given class-specific sample means and frequencies

Let us now further assume that, together with the frequencies n_j, the sample means $\bar{x}_j$ within $\mathcal{C}_j$ $(j = 1, \ldots, J)$ are also reported. Then, the likelihood based on the observed data, $\mathcal{D}_1 = \mathcal{D}_0 \cup \{\bar{x}_j : j = 1, \ldots, J\}$, becomes

$$L(\boldsymbol{\theta}|\mathcal{D}_1) = \Pr\left(N_1 = n_1, \ldots, N_J = n_J\right) \prod_{j=1}^{J} f_{\overline{X}_j}(\bar{x}_j|n_j) \propto \prod_{j=1}^{J} \left(\gamma_j^{n_j} f_{\overline{X}_j}(\bar{x}_j|n_j)\right),$$

where $f_{\overline{X}_j}(\cdot|n_j)$ is the conditional density of $\overline{X}_j$ given the class frequency $N_j = n_j$. The central limit theorem provides the asymptotic distribution of the sampling distribution of $\bar{X}_j$ that will be used here to approximate $f_{\overline{X}_j}(\cdot|n_j)$ for given finite class frequency n_j. Formally, denote the class-specific population moments as

$$\mu_{1j} = \frac{1}{\gamma_j} \int_{a_{j-1}}^{a_j} x f(x)\, dx \text{ and } \sigma_j^2 = \frac{1}{\gamma_j} \int_{a_{j-1}}^{a_j} (x - \mu_{1j})^2 f(x) dx.$$

Then,

$$f_{\overline{X}_j}(\bar{x}_j|n_j) \approx \sqrt{\frac{n_j}{2\pi\sigma_j^2}} \exp\left(-\frac{n_j}{2\sigma_j^2}(\bar{x}_j - \mu_{1j})^2\right).$$

Given the preceding spline approximation to the (log-)density using polytomous logistic regression for the probabilities to be in the small bins partitioning the support, see (1), one has

$$\begin{aligned}
\mu_{1j}(\boldsymbol{\theta}) &= \mathrm{E}[X|X \in \mathcal{C}_j] = \sum_{i=1}^{I} u_i \frac{c_{ji}\pi_i(\boldsymbol{\theta})}{\gamma_j(\boldsymbol{\theta})} + \mathcal{O}(\Delta^2), \\
\sigma_j^2(\boldsymbol{\theta}) &= \mathrm{Var}[X|X \in \mathcal{C}_j] = \sum_{i=1}^{I} \left(u_i - \mu_{1j}(\boldsymbol{\theta})\right)^2 \frac{c_{ji}\pi_i(\boldsymbol{\theta})}{\gamma_j(\boldsymbol{\theta})} + \mathcal{O}(\Delta^2).
\end{aligned} \tag{3}$$

Then, for given roughness penalty parameter λ, the penalized log-likelihood based on $\mathcal{D}_1$ becomes

$$\ell_p(\boldsymbol{\theta}|\mathcal{D}_1, \lambda) = \ell_p(\boldsymbol{\theta}|\mathcal{D}_0, \lambda) - \frac{1}{2}\sum_{j=1}^{J}\left(\log\sigma_j^2(\boldsymbol{\theta}) + \frac{n_j}{\sigma_j^2(\boldsymbol{\theta})}\left(\bar{x}_j - \mu_{1j}(\boldsymbol{\theta})\right)^2\right),$$

while its counterpart based on latent small bins frequencies, $\mathcal{D}_1^c = \mathcal{D}_0^c \cup \{\bar{x}_j : j = 1, \ldots, J\}$, is

$$\ell_p^c(\boldsymbol{\theta}|\mathcal{D}_1^c, \lambda) = \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c, \lambda) - \frac{1}{2}\sum_{j=1}^{J}\left(\log\sigma_j^2(\boldsymbol{\theta}) + \frac{n_j}{\sigma_j^2(\boldsymbol{\theta})}\left(\bar{x}_j - \mu_{1j}(\boldsymbol{\theta})\right)^2\right). \tag{4}$$

It leads to Algorithm 2 for an estimation of the density using the EM algorithm.

Algorithm 2 *(Density estimation given class-specific sample means and frequencies).* The following EM algorithm alternates the update of the estimates for the latent frequencies $\mathbf{k}$, $\boldsymbol{\theta}$ and, possibly, λ. Denote by $\tilde{\mathbf{B}}$ the $I \times K$ matrix such that $(\tilde{\mathbf{B}})_{ik} = \tilde{b}_{ik} = b_k(u_i) - \sum_{t=1}^{I} b_k(u_t)\pi_t$.

Repeat the following steps till convergence:

1. E-step: $k_i \leftarrow \mathrm{E}\left(k_i|\boldsymbol{\theta}, \mathcal{D}_0\right) = \sum_{j=1}^{J} n_j c_{ji}\pi_i(\boldsymbol{\theta})/\gamma_j(\boldsymbol{\theta})$.
2. M-step: compute $\boldsymbol{\theta} = \arg\max_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_1^c, \lambda)$. This can be done using a Newton-Raphson (N-R) algorithm with the iterative substitution, $\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \left(-\mathbf{H}_p^c + \epsilon\, \mathrm{I}_K\right)^{-1} \nabla_\theta \ell_p^c$, till convergence. Explicit forms for the gradient and the Hessian matrix are available after conditioning on the value $\tilde{\sigma}_j^2$ of $\sigma_j^2(\boldsymbol{\theta})$ in (3) at the start of the iteration:

$$\begin{aligned}\nabla_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_1^c,\lambda) &= \nabla_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c,\lambda) + \sum_{j=1}^{J} \frac{n_j}{\tilde{\sigma}_j^2}(\overline{x}_j - \mu_{1j})\frac{\partial \mu_{1j}}{\partial \boldsymbol{\theta}} \ ; \\ -\mathbf{H}_p^c = -\nabla_\theta^2 \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_1^c,\lambda) &\approx -\nabla_\theta^2 \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c,\lambda) + \sum_{j=1}^{J} \frac{n_j}{\tilde{\sigma}_j^2}\frac{\partial \mu_{1j}}{\partial \boldsymbol{\theta}}\frac{\partial \mu_{1j}}{\partial \boldsymbol{\theta}^\top},\end{aligned} \tag{5}$$

where $\partial \mu_{1j}/\partial \theta_k = \frac{1}{\gamma_j}\sum_{i=1}^{I} c_{ji}\pi_i(u_i - \mu_{1j})b_{ik}$, with the approximation to $\mathbf{H}_p^c$ coming from the neglect of zero expectation terms.

3. Penalty update: $\lambda \leftarrow (\text{edf} - r)/\|\mathbf{D}\boldsymbol{\theta}\|^2$ where the effective number of spline parameters is given by the trace of a matrix, $\text{edf}(\lambda) = \text{Tr}\left((-\mathbf{H}_p^c + \epsilon\, \mathrm{I}_K)^{-1}(-\mathbf{H}_p^c - \lambda\mathbf{P})\right)$.

At convergence, one obtains the penalized MLE $\hat{\boldsymbol{\theta}}_\lambda$ (given the selected value for λ). □

3.2.2. Density estimation given class-specific sample central moments and frequencies

Assume now that, together with the frequencies, the sample mean $\overline{X}_j$, variance S_j^2, skewness G_{1j} and kurtosis G_{2j} within $\mathcal{C}_j$ are reported. If M_{rj} denotes the rth sample central moment in $\mathcal{C}_j$, we have

$$\overline{X}_j = M_{1j} \ ; \ S_j^2 = M_{2j} \ ; \ G_{1j} = M_{3j}/M_{2j}^{3/2} \ ; \ G_{2j} = M_{4j}/M_{2j}^{4/2} - 3.$$

Denote by $\mathbf{M}_j = (M_{1j}, M_{2j}, M_{3j}, M_{4j})$ the random vector of sample central moments in $\mathcal{C}_j$ and by $\mathbf{m}_j$ its observed counterpart. Assume that moments of order p ($p \leq 8$) for X exist when X is restricted to take values in any of the classes $\mathcal{C}_j$ ($j = 1, \ldots, J$). The central limit theorem provides the asymptotic distribution of $\mathbf{M}_j$, $\mathbf{M}_j \xrightarrow{d} \mathcal{N}_4\left(\boldsymbol{\mu}_j, \Sigma_j\right)$, where $\boldsymbol{\mu}_j = (\mu_{1j}, \ldots, \mu_{4j})$ with $\mu_{1j} = \mu_j$ and

$$\begin{aligned}\mu_{rj} = \mu_{rj}(\boldsymbol{\theta}) &= \mathrm{E}\left[(X - \mu_{1j}(\boldsymbol{\theta}))^r | X \in \mathcal{C}_j\right], \quad r = 2, 3, \ldots \\ &= \frac{1}{\gamma_j(\boldsymbol{\theta})}\sum_{j=1}^{J} c_{ji}\pi_i(\boldsymbol{\theta})\left(u_i - \mu_{1j}(\boldsymbol{\theta})\right)^r + \mathcal{O}(\Delta^2).\end{aligned}$$

It is used here as an approximation to the sampling distribution of $\mathbf{M}_j$ for finite n_j. Using the Generalized Method of Moments (GMM) (Hansen, 1982), one can show that

$$\Sigma_j = \frac{1}{n_j}\begin{pmatrix} \mu_{2j} & \mu_{3j} & \mu_{4j} & \mu_{5j} \\ \mu_{3j} & \mu_{4j} - \mu_{2j}^2 & \mu_{5j} - \mu_{2j}\mu_{3j} & \mu_{6j} - \mu_{2j}\mu_{4j} \\ \mu_{4j} & \mu_{5j} - \mu_{2j}\mu_{3j} & \mu_{6j} - \mu_{3j}^2 & \mu_{7j} - \mu_{3j}\mu_{4j} \\ \mu_{5j} & \mu_{6j} - \mu_{2j}\mu_{4j} & \mu_{7j} - \mu_{3j}\mu_{4j} & \mu_{8j} - \mu_{4j}^2 \end{pmatrix}, \tag{6}$$

see Arellano-Valle et al. (2021) for a similar argumentation. Based on the observed data $\mathcal{D} = \mathcal{D}_0 \cup \{\mathbf{m}_j : j = 1, \ldots, J\}$ and for a given roughness penalty parameter λ, the penalized log-likelihood becomes

$$\ell_p(\boldsymbol{\theta}|\mathcal{D},\lambda) = \ell_p(\boldsymbol{\theta}|\mathcal{D}_0,\lambda) - \frac{1}{2}\sum_{j=1}^{J}\left\{\log|\Sigma_j| + (\mathbf{m}_j - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{m}_j - \boldsymbol{\mu}_j)\right\}, \tag{7}$$

comparable to (4) when only the tabulated sample means were available. When, in addition, the latent small bins frequencies are given, $\mathcal{D}^c = \mathcal{D}_0^c \cup \{\mathbf{m}_j : j = 1, \ldots, J\}$, inference is based on the penalized complete log-likelihood,

$$\ell_p^c(\boldsymbol{\theta}|\mathcal{D}^c,\lambda) = \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c,\lambda) - \frac{1}{2}\sum_{j=1}^{J}\left\{\log|\Sigma_j| + (\mathbf{m}_j - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{m}_j - \boldsymbol{\mu}_j)\right\}. \tag{8}$$

The maximization of the penalized log-likelihood (8) and the selection of λ can be made using Algorithm 3. A Laplace approximation to the joint posterior of $\boldsymbol{\theta}$ for the selected λ value can be obtained using the Hessian calculated from the observed penalized log-likelihood (7) evaluated at convergence, see Section 3.3 and the precision matrix obtained by combining (11) and (12).

Algorithm 3 *(Density estimation given class-specific sample central moments and frequencies).* The following EM algorithm alternates the update of the estimates for the latent frequencies $\mathbf{k}$, $\boldsymbol{\theta}$ and, possibly, λ using the following steps till convergence:

1. E-step: $k_i \leftarrow \mathrm{E}\left(k_i|\boldsymbol{\theta},\mathcal{D}_0\right) = \sum_{j=1}^{J} n_j c_{ji}\pi_i(\boldsymbol{\theta})/\gamma_j(\boldsymbol{\theta})$.
2. M-step: compute $\boldsymbol{\theta} = \arg\max_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}^c,\lambda)$. This can be done using a Newton-Raphson (N-R) algorithm with the iterative substitution, $\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \left(-\mathbf{H}_p^c + \epsilon\, \mathrm{I}_K\right)^{-1}\nabla_\theta \ell_p^c$, till convergence. Explicit forms for the gradient and the Hessian matrix are available after conditioning on the value $\tilde{\Sigma}_j$ of $\Sigma_j(\boldsymbol{\theta})$ in (8) at the start of the iteration:

$$\begin{aligned}\big(\nabla_\theta \ell_p^c\big)_k &= \big(\nabla_\theta \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c,\lambda)\big)_k + \sum_{j=1}^{J}(\mathbf{m}_j-\boldsymbol{\mu}_j)^\top \tilde{\Sigma}_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial\theta_k}\ ;\\ -\big(\mathbf{H}_p^c\big)_{ks} &\approx -\big(\nabla_\theta^2 \ell_p^c(\boldsymbol{\theta}|\mathcal{D}_0^c,\lambda)\big)_{ks} + \sum_{j=1}^{J}\frac{\partial \boldsymbol{\mu}_j^\top}{\partial\theta_s}\tilde{\Sigma}_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial\theta_k}\end{aligned} \tag{9}$$

with

$$\begin{aligned}\frac{\partial \mu_{rj}}{\partial\theta_k} &= \frac{1}{\gamma_j}\sum_{i=1}^{I} c_{ji}\pi_i b_{ik}\big\{(u_i-\mu_{1j})^r-\mu_{rj}\big\} - r\mu_{r-1,j}\frac{\partial\mu_{1j}}{\partial\theta_k}\\ &= \frac{1}{\gamma_j}\sum_{i=1}^{I} c_{ji}\pi_i b_{ik}\big\{(u_i-\mu_{1j})^r-\mu_{rj}-r\mu_{r-1,j}(u_i-\mu_{1j})\big\}\end{aligned} \tag{10}$$

for $1\le k,s\le K$ and $2\le r\le 4$.

3. Penalty update: $\lambda \leftarrow (\text{edf}-r)/\|\mathbf{D}\boldsymbol{\theta}\|^2$ where the effective number of spline parameters is given by the trace of a matrix, $\text{edf}(\lambda) = \text{Tr}\big((-\mathbf{H}_p^c+\epsilon\,\mathrm{I}_K)^{-1}(-\mathbf{H}_p^c-\lambda\mathbf{P})\big)$.

At convergence, one obtains the penalized MLE $\hat{\boldsymbol{\theta}}_\lambda$ (given the selected value for λ). □

3.3. Quantile estimation

Consider the following shorthand notation for the conditional posterior mode of the vector of spline parameters underlying the density estimate, $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_\lambda - \max\{\hat{\theta}_k\}_{k=1}^K$, with a subtraction of the largest estimated component to handle the identification issue following from $\boldsymbol{\pi}(\boldsymbol{\theta}+c)=\boldsymbol{\pi}(\boldsymbol{\theta})$ for any real number c. Denote by $\hat{k}$ the component for which $(\hat{\boldsymbol{\theta}})_{\hat{k}}=0$ and by $\boldsymbol{\theta}_{-\hat{k}}$ the vector of spline parameters where the $\hat{k}$th component of $\boldsymbol{\theta}$ is omitted. Quantile estimates can be derived using the fitted density estimate,

$$f(x|\hat{\boldsymbol{\theta}}) = \exp\{\eta(x|\hat{\boldsymbol{\theta}})\}/\int_{\mathcal{X}}\exp\{\eta(t|\hat{\boldsymbol{\theta}})\}\,dt\,,$$

where $\eta(x|\boldsymbol{\theta})=\sum_{k=1}^K b_k(x)\theta_k$ and $\mathcal{X}=(a_0,a_J)$ denotes the support of the density. Indeed, as the associated estimate for the c.d.f., $\hat{F}(x)=F(x|\hat{\boldsymbol{\theta}})=\int_{a_0}^x f(t|\hat{\boldsymbol{\theta}})\,dt$, is monotone, it can be inverted to provide an estimate of the quantile function,

$$\hat{Q}(p) = Q(p|\hat{\boldsymbol{\theta}}) = \inf\{x\in\mathbb{R}: F(x|\hat{\boldsymbol{\theta}})\ge p\}.$$

Practically, starting for the fitted probability, $\hat{\pi}_i=\Pr(X\in\mathcal{I}_i|\hat{\boldsymbol{\theta}})$, to have an observation in the small bin $\mathcal{I}_i=(b_{i-1},b_i]$ $(i=1,\ldots,I)$, see Section 3.1, and with $\hat{F}(b_0)=0$, $\hat{F}(b_i)=\hat{\pi}_1+\ldots+\hat{\pi}_i$, a first guess for $\hat{Q}(p)$ is given by

$$x_0 = \max_{0\le i\le I}\big\{b_i:\hat{F}(b_i)\le p\big\}.$$

This first approximation can be improved in an iterative way with, at iteration t,

$$x_t \longleftarrow x_{t-1} + \big(p-F(x_{t-1}|\hat{\boldsymbol{\theta}})\big)/f(x_{t-1}|\hat{\boldsymbol{\theta}})\,,$$

yielding at convergence $Q(p|\hat{\boldsymbol{\theta}})=x_\infty$. The uncertainty in that estimation directly follows from the uncertainty in the choice of $\boldsymbol{\theta}$. The latter is quantified by the conditional posterior distribution of $\boldsymbol{\theta}_{-\hat{k}}$ with Laplace approximation $\big(\boldsymbol{\theta}_{-\hat{k}}|\mathcal{D},\lambda\big)\dot{\sim}\mathcal{N}\big(\hat{\boldsymbol{\theta}}_{-\hat{k}},(\mathcal{J}_{-\hat{k},-\hat{k}})^{-1}\big)$ where

$$\mathcal{J}_{ks} = -\big(\nabla_\theta^2\ell_p(\boldsymbol{\theta}|\mathcal{D}_0,\lambda)\big)_{ks} + \sum_{j=1}^{J}\frac{\partial\boldsymbol{\mu}_j^\top}{\partial\theta_s}\tilde{\Sigma}_j^{-1}\frac{\partial\boldsymbol{\mu}_j}{\partial\theta_k}\,, \tag{11}$$

with the partial derivatives of the theoretical central moments $\boldsymbol{\mu}_j$ in the jth class given in (10). Thanks to the Gaussian Markov field (GMRF) prior (Rue and Held, 2005) induced by the penalty on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\lambda)\propto\exp\{-\frac{\lambda}{2}\|\mathbf{D}\boldsymbol{\theta}\|^2\}$, the preceding Laplace approximation to the conditional posterior is usually excellent, see Rue et al. (2009) for the same argument in latent Gaussian models. The second derivatives of the log-likelihood based on the observed class frequencies $\mathcal{D}_0$ in the first term of (11) have an explicit form given by

$$\begin{aligned}-\big(\nabla_\theta^2\ell_p(\boldsymbol{\theta}|\mathcal{D}_0,\lambda)\big)_{ks} &= n\sum_{i=1}^{I} b_{ik}\pi_i\tilde{b}_{is} + \sum_{j=1}^{J}\frac{n_j}{\gamma_j^2}\sum_{\ell=1}^{I}c_{j\ell}\pi_\ell\tilde{b}_{\ell s}\sum_{i=1}^{I}c_{ji}\pi_i\tilde{b}_{ik} - \sum_{j=1}^{J}\frac{n_j}{\gamma_j}\sum_{i=1}^{I}c_{ji}\pi_i\tilde{b}_{is}\tilde{b}_{ik}\\ &= (\mathbf{B}^\top\mathbf{W}\mathbf{B})_{ks} - \sum_{j=1}^{J}\frac{n_j}{\gamma_j}\sum_{i=1}^{I}c_{ji}\pi_i b_{ik}\left(b_{is}-\frac{1}{\gamma_j}\sum_{\ell=1}^{I}c_{j\ell}\pi_\ell b_{\ell s}\right).\end{aligned} \tag{12}$$

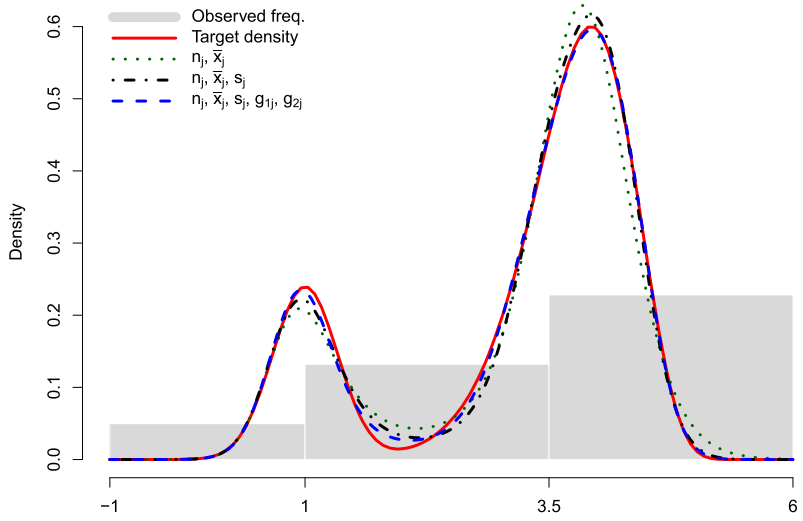

Fig. 2. Simulation study ($n = 1000$, $J = 3$) – Density estimates from tabulated summary statistics $\mathcal{D}_1$ (dotted), $\mathcal{D}_2$ (dot-dashed) and $\mathcal{D}_4$ (dashed) averaged over the $S = 500$ replicates, with the 'true' underlying density (solid curve).

The first term in the last expression, where $(\mathbf{W})_{ii'} = n(\pi_i \delta_{ii'} - \pi_i \pi_{i'})$, corresponds to the information available on $\boldsymbol{\theta}$ based on data frequencies in the absence of class tabulation. The information reduction due to tabulation is quantified by the second term. Based on the following first-order expansion,

$$Q(p|\boldsymbol{\theta}) \approx Q(p|\hat{\boldsymbol{\theta}}) + \frac{\partial Q(p|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^\top_{-\hat{k}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_{-\hat{k}},$$

the conditional posterior distribution of $Q(p|\boldsymbol{\theta})$ can be approximated by

$$\left(Q(p|\boldsymbol{\theta})|\mathcal{D}, \lambda\right) \dot{\sim} \mathcal{N}\left(Q(p|\hat{\boldsymbol{\theta}}), s^2_Q(p)\right), \tag{13}$$

with

$$\frac{\partial Q(p|\hat{\boldsymbol{\theta}})}{\partial \theta_k} = \frac{1}{f\left(Q(p|\hat{\boldsymbol{\theta}})|\hat{\boldsymbol{\theta}}\right)} \left\{ \int_{-\infty}^{Q(p|\hat{\boldsymbol{\theta}})} b_k(x) f(x|\hat{\boldsymbol{\theta}})\, dx - p \int_{\mathbb{R}} b_k(x) f(x|\hat{\boldsymbol{\theta}})\, dx \right\},$$

$$s^2_Q(p) = \frac{\partial Q(p|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^\top_{-\hat{k}}} \left(\mathcal{J}_{-\hat{k},-\hat{k}}\right)^{-1} \frac{\partial Q(p|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{-\hat{k}}}.$$

Therefore, an approximate $100(1-\alpha)\%$ credible interval for $Q(p)$ is given by

$$Q(p|\hat{\boldsymbol{\theta}}) \pm \Phi^{-1}(1-\alpha/2)\, s_Q(p), \tag{14}$$

with $\Phi^{-1}(\cdot)$ denoting the quantile function of the standard Normal distribution.

4. Simulation study

In this section, we evaluate the performances of the estimation method proposed in Section 3 by means of extensive simulations. Independent and identically distributed data were generated from a mixture density,

$$f(x) = w_1 f_1(x) + w_2 f_2(5.6 - x)$$

where $f_1(\cdot)$ corresponds to a Normal density with mean 1.0 and variance 9.0^{-1}, $f_2(\cdot)$ to a Gamma density with mean $11/6$ and variance $11/6^2$, weighted respectively by $w_1 = .20$ and $w_2 = .80$. It corresponds to the solid red curve in Figs. 2 and 3 that could be viewed as the underlying distribution of log transformed positive data. Datasets of size $n = 250$, 1 000 or 3 000 were generated $S = 500$ times and grouped into either $J = 3$ or $J = 5$ partitioning classes $\{C_j : j = 1, \ldots, J\}$ with interval extremities given by $\{-1.0, 1.0, 3.5, 6.0\}$ and $\{-1.0, 1.0, 2.2, 3.5, 4.8, 6.0\}$, respectively. Tabulated frequencies $(n_j)_{j=1}^J$ and the associated local central moments $(m_{rj})_{j=1}^J$ of order $r = 1$ up to 4 were computed and used with the methodology of Section 3 to produce an estimate $\hat{f}(\cdot|\mathcal{D}_r)$ of the underlying density on $(-1.0, 6.0)$, with $\mathcal{D}_r = \bigcup_{j=1}^J \{n_j, m_{1j}, \ldots, m_{rj}\}$ denoting the available data. Selected quantile estimates $\hat{Q}(p|\mathcal{D}_r)$ were computed using that density and compared to the 'true' quantile values associated to $f(\cdot)$. Biases, standard deviations (SD), root mean squared errors (RMSE) and effective coverages of 95% and 90% credible intervals are given in Tables 1, 2 and 3 for different samples sizes and number of classes. As expected, biases for the point estimator of a given quantile tend to decrease with the sample size and the number of classes for which tabulated summary statistics are observed. They are already very small when $n = 250$ with $r = 4$ moments reported in only $J = 3$ classes,

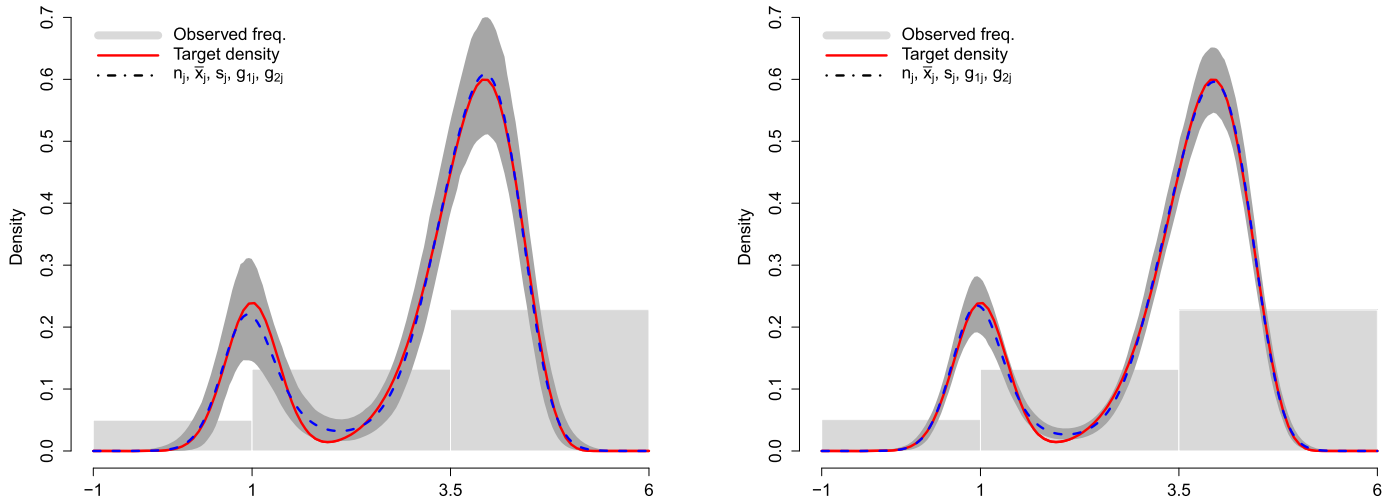

Fig. 3. Simulation study – Connected pointwise intervals containing 95% of the estimated densities over the $S = 500$ datasets and obtained using tabulated summary statistics $\mathcal{D}_4$ over $J = 3$ classes. Left panel: $n = 250$; Right panel: $n = 1\,000$.

Table 1
Simulation study ($n = 250$) – Selected p-quantile estimates using tabulated summary statistics in $\mathcal{D}_r$ with $r = 1, 2, 4$: bias, standard deviation, root mean squared error (RMSE) and effective coverages of 95% and 90% credible intervals (based on $S = 500$ replicates).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$Q(p)$	1.000	1.793	3.122	3.430	3.643	3.822	3.989	4.163	4.375
$J = 3$ (classes)									
$\hat{Q}(p\|\mathcal{D}_1)$	1.084	2.071	3.081	3.421	3.630	3.801	3.965	4.144	4.383
Bias	0.084	0.278	-0.041	-0.009	-0.014	-0.021	-0.024	-0.019	0.007
SD	0.121	0.401	0.154	0.074	0.054	0.045	0.040	0.038	0.048
RMSE	0.147	0.488	0.160	0.075	0.056	0.050	0.046	0.043	0.049
95% CI	0.950	0.694	0.930	0.964	0.958	0.964	0.968	0.934	0.926
90% CI	0.904	0.622	0.880	0.922	0.924	0.926	0.920	0.870	0.866
$\hat{Q}(p\|\mathcal{D}_2)$	1.016	2.003	3.166	3.466	3.663	3.830	3.988	4.155	4.363
Bias	0.016	0.210	0.045	0.036	0.020	0.008	-0.001	-0.009	-0.013
SD	0.082	0.493	0.124	0.064	0.050	0.044	0.040	0.039	0.040
RMSE	0.084	0.536	0.132	0.073	0.054	0.044	0.040	0.040	0.042
95% CI	0.950	0.656	0.860	0.886	0.908	0.932	0.934	0.896	0.840
90% CI	0.906	0.586	0.778	0.830	0.842	0.878	0.876	0.848	0.762
$\hat{Q}(p\|\mathcal{D}_4)$	1.007	2.004	3.121	3.435	3.646	3.823	3.988	4.160	4.369
Bias	0.007	0.211	-0.001	0.005	0.003	0.001	-0.001	-0.004	-0.006
SD	0.081	0.480	0.121	0.068	0.053	0.046	0.041	0.040	0.041
RMSE	0.082	0.525	0.121	0.068	0.054	0.046	0.041	0.040	0.041
95% CI	0.948	0.740	0.946	0.954	0.950	0.948	0.962	0.950	0.948
90% CI	0.910	0.694	0.918	0.912	0.908	0.912	0.924	0.904	0.880
$J = 5$ (classes)									
$\hat{Q}(p\|\mathcal{D}_1)$	0.986	1.920	3.104	3.429	3.648	3.827	3.992	4.159	4.364
Bias	-0.014	0.127	-0.018	-0.001	0.005	0.006	0.002	-0.004	-0.011
SD	0.081	0.532	0.129	0.074	0.057	0.048	0.042	0.041	0.044
RMSE	0.082	0.547	0.131	0.074	0.058	0.048	0.043	0.042	0.046
95% CI	0.910	0.650	0.938	0.930	0.936	0.932	0.912	0.882	0.898
90% CI	0.866	0.598	0.900	0.894	0.892	0.892	0.864	0.818	0.838
$\hat{Q}(p\|\mathcal{D}_2)$	0.998	1.914	3.100	3.423	3.640	3.820	3.988	4.162	4.373
Bias	-0.002	0.121	-0.022	-0.007	-0.003	-0.002	-0.001	-0.002	-0.002
SD	0.081	0.532	0.132	0.073	0.056	0.048	0.043	0.043	0.047
RMSE	0.081	0.545	0.134	0.073	0.056	0.048	0.043	0.043	0.047
95% CI	0.920	0.608	0.948	0.934	0.934	0.924	0.910	0.842	0.732
90% CI	0.878	0.548	0.904	0.880	0.884	0.878	0.860	0.792	0.654
$\hat{Q}(p\|\mathcal{D}_4)$	1.012	2.001	3.120	3.435	3.646	3.822	3.987	4.157	4.368
Bias	0.012	0.208	-0.002	0.005	0.002	0.000	-0.003	-0.006	-0.008
SD	0.091	0.480	0.120	0.068	0.054	0.046	0.041	0.039	0.040
RMSE	0.092	0.523	0.120	0.068	0.054	0.046	0.041	0.040	0.041
95% CI	0.952	0.728	0.946	0.952	0.950	0.948	0.960	0.952	0.940
90% CI	0.914	0.680	0.928	0.906	0.910	0.906	0.922	0.910	0.874

see Fig. 3 for a graphical illustration at the density level when $n = 250$ and $n = 1000$. An exception concerns the 20% quantile ($= 1.793$) that is not so well estimated whatever the simulation setting: it corresponds to the region surrounding the local minimum of the mixture density between the two modes. Increasing the number of reported central moments tends to improve the estimation of density and

Table 2
Simulation study ($n = 1000$) – Selected p-quantile estimates using tabulated summary statistics in $\mathcal{D}_r$ with $r = 1, 2, 4$: bias, standard deviation, root mean squared error (RMSE) and effective coverages of 95% and 90% credible intervals (based on $S = 500$ replicates).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
$Q(p)$	1.000	1.793	3.122	3.430	3.643	3.822	3.989	4.163	4.375	4.530
$J = 3$ (classes)										
$\hat{Q}(p\|\mathcal{D}_1)$	1.026	1.940	3.121	3.436	3.634	3.800	3.960	4.138	4.381	4.586
Bias	0.027	0.147	-0.001	0.005	-0.009	-0.022	-0.029	-0.025	0.006	0.056
SD	0.047	0.257	0.069	0.037	0.028	0.024	0.021	0.020	0.025	0.033
RMSE	0.054	0.297	0.069	0.037	0.030	0.032	0.036	0.032	0.025	0.065
95% CI	0.944	0.836	0.954	0.930	0.954	0.972	0.980	0.948	0.960	0.998
90% CI	0.878	0.786	0.910	0.894	0.916	0.920	0.920	0.892	0.918	0.978
$\hat{Q}(p\|\mathcal{D}_2)$	1.005	1.867	3.159	3.452	3.653	3.826	3.989	4.159	4.369	4.527
Bias	0.006	0.074	0.037	0.022	0.010	0.004	0.000	-0.005	-0.007	-0.003
SD	0.040	0.302	0.058	0.035	0.028	0.025	0.023	0.021	0.021	0.023
RMSE	0.040	0.311	0.068	0.042	0.030	0.025	0.023	0.022	0.022	0.023
95% CI	0.960	0.752	0.834	0.880	0.918	0.916	0.910	0.886	0.874	0.816
90% CI	0.902	0.698	0.768	0.812	0.844	0.866	0.844	0.820	0.788	0.728
$\hat{Q}(p\|\mathcal{D}_4)$	0.993	1.904	3.125	3.429	3.642	3.822	3.991	4.165	4.376	4.530
Bias	-0.006	0.111	0.003	-0.001	-0.001	0.000	0.002	0.002	0.001	0.001
SD	0.037	0.331	0.058	0.037	0.029	0.025	0.023	0.021	0.021	0.023
RMSE	0.037	0.349	0.058	0.037	0.029	0.025	0.023	0.022	0.021	0.023
95% CI	0.950	0.816	0.938	0.942	0.940	0.946	0.948	0.956	0.952	0.958
90% CI	0.904	0.776	0.892	0.898	0.890	0.904	0.898	0.896	0.904	0.908
$J = 5$ (classes)										
$\hat{Q}(p\|\mathcal{D}_1)$	0.989	1.864	3.108	3.427	3.648	3.828	3.992	4.161	4.367	4.522
Bias	-0.010	0.071	-0.014	-0.003	0.005	0.006	0.003	-0.003	-0.009	-0.008
SD	0.034	0.370	0.061	0.039	0.032	0.026	0.023	0.022	0.023	0.025
RMSE	0.036	0.377	0.063	0.040	0.032	0.027	0.023	0.022	0.024	0.027
95% CI	0.950	0.734	0.920	0.928	0.936	0.930	0.916	0.858	0.934	0.966
90% CI	0.900	0.686	0.848	0.894	0.884	0.880	0.830	0.786	0.882	0.928
$\hat{Q}(p\|\mathcal{D}_2)$	0.992	1.836	3.106	3.426	3.642	3.821	3.989	4.164	4.378	4.536
Bias	-0.008	0.043	-0.016	-0.004	-0.001	-0.001	-0.001	0.001	0.003	0.006
SD	0.036	0.357	0.064	0.038	0.030	0.026	0.024	0.023	0.024	0.026
RMSE	0.037	0.360	0.066	0.038	0.030	0.026	0.024	0.023	0.024	0.026
95% CI	0.942	0.730	0.912	0.938	0.932	0.924	0.890	0.816	0.692	0.808
90% CI	0.874	0.690	0.842	0.890	0.892	0.866	0.820	0.744	0.592	0.734
$\hat{Q}(p\|\mathcal{D}_4)$	0.995	1.914	3.121	3.429	3.643	3.823	3.990	4.164	4.374	4.530
Bias	-0.005	0.121	-0.001	-0.001	0.000	0.001	0.001	0.000	-0.002	0.000
SD	0.037	0.342	0.059	0.037	0.029	0.025	0.023	0.022	0.022	0.023
RMSE	0.037	0.363	0.059	0.037	0.029	0.025	0.023	0.022	0.022	0.023
95% CI	0.954	0.804	0.936	0.942	0.940	0.940	0.946	0.956	0.944	0.956
90% CI	0.910	0.754	0.890	0.900	0.896	0.904	0.900	0.880	0.892	0.914

quantiles, with 4 moments being preferable, see Fig. 2 for an evolution of the averaged density estimates (over the $S = 500$ replicates) starting with $\mathcal{D}_1$ (dotted curve) to $\mathcal{D}_4$ (dashed curve) when $n = 1\,000$. This is a remarkable improvement over the estimate that would be obtained using only observed frequencies. Whatever the sample size, the effective coverages of 90% and 95% credible intervals for the reported quantiles (except the 20% one) are in agreement with their nominal values when 4 central moments ($\mathcal{D}_4$) are reported. Moderate undercoverages can be observed with $\mathcal{D}_2$, while the effective coverages can be larger than expected when just the means are reported in addition to frequencies ($\mathcal{D}_1$). Global metrics were also calculated to compare the true and estimated quantile functions,

$$\ell_1(Q, \hat{Q}) = \int_0^1 \left| Q(p|\hat{\boldsymbol{\theta}}) - Q(p) \right| dp$$

as well the true and estimated densities,

$$\text{RIMSE}(f, \hat{f}) = \int_{\mathbb{R}} \left(f(x|\hat{\boldsymbol{\theta}}) - f(x) \right)^2 f(x)\, dx \;\; ; \;\; \text{KL}(f, \hat{f}) = \int_{\mathbb{R}} f(x) \log \left(\frac{f(x)}{\hat{f}(x)} \right) dx \,,$$

see Table 4 for their median values over the $S = 500$ simulated datasets with tabulated summary statistics $\mathcal{D}_r$ ($r = 1, 2, 4$) in 3 or 5 classes. This suggests that the extra information provided by additional central moments for quantile or density estimation is even more valuable when the number of classes is small.

In summary, this simulation study confirms the added value of central moments over isolated frequencies for density estimation. With only 15 numbers (the frequency and the 4 central moments in each of the $J = 3$ classes), an accurate and precise density estimate could be obtained from summary statistics in just 3 classes, see Fig. 3 for a graphical representation. The simulation study also suggests that this method can be used to estimate quantiles and, consequently, values at risk (VaR) in a reliable way.

Table 3
Simulation study ($n = 3000$) – Selected p-quantile estimates using tabulated summary statistics in $\mathcal{D}_r$ with $r = 1, 2, 4$: bias, standard deviation, root mean squared error (RMSE) and effective coverages of 95% and 90% credible intervals (based on $S = 500$ replicates).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
$Q(p)$	1.000	1.793	3.122	3.430	3.643	3.822	3.989	4.163	4.375	4.530	4.778
$J = 3$ (classes)											
$\hat{Q}(p\|\mathcal{D}_1)$	1.013	1.897	3.127	3.438	3.634	3.796	3.954	4.131	4.376	4.591	5.052
Bias	0.013	0.104	0.005	0.008	-0.010	-0.025	-0.035	-0.033	0.000	0.061	0.274
SD	0.025	0.145	0.038	0.021	0.016	0.013	0.011	0.010	0.013	0.020	0.045
RMSE	0.029	0.179	0.038	0.023	0.019	0.029	0.037	0.034	0.013	0.064	0.278
95% CI	0.948	0.934	0.990	0.928	0.960	0.992	0.996	0.994	0.996	1.000	1.000
90% CI	0.886	0.874	0.972	0.866	0.916	0.944	0.974	0.968	0.996	0.996	1.000
$\hat{Q}(p\|\mathcal{D}_2)$	1.004	1.810	3.153	3.444	3.649	3.825	3.990	4.160	4.368	4.524	4.797
Bias	0.005	0.017	0.032	0.014	0.005	0.003	0.001	-0.003	-0.007	-0.006	0.019
SD	0.023	0.170	0.032	0.021	0.017	0.014	0.013	0.012	0.011	0.012	0.021
RMSE	0.023	0.171	0.045	0.025	0.018	0.014	0.013	0.013	0.014	0.013	0.028
95% CI	0.954	0.828	0.762	0.868	0.936	0.942	0.904	0.932	0.886	0.818	0.942
90% CI	0.910	0.762	0.686	0.794	0.866	0.876	0.840	0.832	0.810	0.750	0.866
$\hat{Q}(p\|\mathcal{D}_4)$	0.994	1.868	3.126	3.429	3.643	3.822	3.991	4.165	4.376	4.528	4.782
Bias	-0.006	0.075	0.004	-0.001	-0.000	0.001	0.001	0.002	0.000	-0.002	0.004
SD	0.021	0.223	0.033	0.021	0.017	0.014	0.013	0.012	0.012	0.012	0.018
RMSE	0.022	0.235	0.033	0.022	0.017	0.014	0.013	0.012	0.012	0.013	0.018
95% CI	0.948	0.904	0.946	0.948	0.956	0.956	0.954	0.946	0.966	0.964	0.968
90% CI	0.902	0.850	0.894	0.880	0.894	0.914	0.906	0.910	0.914	0.918	0.918
$J = 5$ (classes)											
$\hat{Q}(p\|\mathcal{D}_1)$	0.999	1.864	3.115	3.430	3.644	3.820	3.986	4.161	4.376	4.530	4.771
Bias	-0.001	0.071	-0.007	0.000	0.001	-0.002	-0.004	-0.002	0.001	0.000	-0.007
SD	0.020	0.258	0.034	0.023	0.018	0.015	0.014	0.013	0.017	0.019	0.019
RMSE	0.020	0.268	0.035	0.023	0.018	0.015	0.014	0.013	0.017	0.019	0.020
95% CI	0.958	0.846	0.944	0.952	0.976	0.982	0.938	0.908	0.974	0.976	0.984
90% CI	0.920	0.772	0.882	0.880	0.922	0.946	0.858	0.826	0.926	0.932	0.934
$\hat{Q}(p\|\mathcal{D}_2)$	0.997	1.831	3.114	3.433	3.646	3.822	3.987	4.161	4.375	4.530	4.777
Bias	-0.003	0.038	-0.008	0.003	0.003	0.000	-0.002	-0.002	-0.001	0.000	-0.001
SD	0.021	0.252	0.035	0.022	0.017	0.014	0.013	0.013	0.013	0.013	0.018
RMSE	0.022	0.255	0.036	0.022	0.017	0.014	0.014	0.013	0.013	0.013	0.018
95% CI	0.942	0.814	0.934	0.932	0.938	0.910	0.896	0.852	0.746	0.894	0.972
90% CI	0.902	0.750	0.860	0.882	0.876	0.858	0.800	0.788	0.668	0.824	0.924
$\hat{Q}(p\|\mathcal{D}_4)$	0.996	1.875	3.123	3.430	3.644	3.823	3.990	4.164	4.375	4.527	4.778
Bias	-0.004	0.082	0.001	0.000	0.000	0.001	0.001	0.001	-0.001	-0.003	0.000
SD	0.021	0.241	0.033	0.021	0.017	0.014	0.013	0.012	0.012	0.013	0.016
RMSE	0.021	0.255	0.033	0.021	0.017	0.014	0.013	0.012	0.012	0.013	0.016
95% CI	0.958	0.860	0.958	0.946	0.954	0.960	0.948	0.942	0.962	0.958	0.982
90% CI	0.902	0.820	0.896	0.884	0.892	0.910	0.904	0.904	0.904	0.928	0.952

Table 4
Simulation study – L_1-distance, $\ell_1(Q, \hat{Q})$, between the true and the estimated quantile functions; Root integrated mean squared error (RIMSE) and Kullback-Liebler divergence (K-L) comparing the true and estimated density functions. Median values of these metrics (over $S = 500$ simulated datasets of size n) are reported with estimation performed from tabulated summary statistics $\mathcal{D}_r$ (with $r = 1, 2, 4$) in J classes.

		$J = 3$ (classes)			$J = 5$ (classes)		
n	Metric	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_4$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_4$
250	$\ell_1(Q, \hat{Q})$	0.087	0.072	0.067	0.072	0.071	0.068
	RIMSE	0.048	0.044	0.037	0.043	0.043	0.037
	K-L	0.034	0.022	0.016	0.018	0.019	0.016
1 000	$\ell_1(Q, \hat{Q})$	0.051	0.038	0.034	0.037	0.036	0.034
	RIMSE	0.041	0.025	0.020	0.025	0.024	0.020
	K-L	0.025	0.011	0.006	0.006	0.006	0.005
3 000	$\ell_1(Q, \hat{Q})$	0.044	0.025	0.021	0.023	0.021	0.021
	RIMSE	0.044	0.018	0.014	0.018	0.016	0.013
	K-L	0.026	0.008	0.003	0.003	0.002	0.002

5. Application

Table 5 provides summary statistics on insurance claim amount data (in euros). For confidentiality reasons, the 3 518 data were rescaled and gathered in $J = 3$ classes of increasing width. Besides the class frequencies n_j, the sample mean, standard deviation, skewness and kurtosis of the $\log_{10}$ transformed claims within each class are also provided. Fig. 4 displays the histogram corresponding to the grouped data frequencies. The thick solid (red) line corresponds to the 'target' density estimated from the precise individual ($\log_{10}$) claims shared with us in confidentiality by the insurance company. Estimation of the density using the grouped summary statistics in Table 5 was performed using the methods described in Sections 3.1 and 3.2 with $K = 25$ B-splines associated to equidistant knots

Table 5
Car insurance data: summary statistics for $n = 3518$ grouped claim data (in euros).

		$\log_{10}$ (Claim)				
Claim Interval	Freq. n_j	Class $\mathcal{C}_j$	Mean $\overline{x}_j$	Std. dev. s_j	Skewness g_{1j}	Kurtosis g_{2j}
(1; 1 000]	1168	(0.00; 3.00]	2.462	0.580	-1.793	2.401
(1 000; 20 000]	2234	(3.00; 4.30]	3.529	0.336	0.375	-0.836
(20 000; 1 500 000]	116	(4.30; 6.18]	4.556	0.275	2.603	9.416

Table 6
Car insurance data: numerical comparison of density and Value-at-Risk estimates obtained using tabulated sample moments of increasing orders.

				$\text{VaR}_{95\%}$		$\text{VaR}_{99\%}$	
Data $\mathcal{D}$	edf	RIMSE	K-L	Est.	95% cred. int.	Est.	95% cred. int.
$\mathcal{D}_0$	6.2	0.069	0.042	16 250	(14 795, 17 848)	34 764	(29 724, 40 658)
$\mathcal{D}_1$	6.7	0.030	0.029	15 885	(14 617, 17 263)	41 502	(37 064, 46 472)
$\mathcal{D}_2$	9.0	0.027	0.019	16 641	(15 355, 17 647)	40 766	(35 261, 47 131)
$\mathcal{D}_4$	11.7	0.012	0.001	16 106	(14 896, 17 413)	38 988	(33 504, 45 371)

Table 7
Car insurance data: observed and fitted central moments within classes using a model based on $\mathcal{D}_4$.

		Central moments for $\log_{10}$ (Claim)							
Class	Freq.	M_1		M_2		M_3		M_4	
$\mathcal{C}_j$	n_j	Obs.	Fitted	Obs.	Fitted	Obs.	Fitted	Obs.	Fitted
(0.00; 3.00]	1168	2.462	2.472	0.336	0.336	-0.350	-0.351	0.611	0.619
(3.00; 4.30]	2234	3.529	3.532	0.113	0.111	0.014	0.013	0.028	0.026
(4.30; 6.18]	116	4.556	4.549	0.075	0.073	0.054	0.051	0.071	0.064

on $(0.0, 6.18)$. Computation was performed in less than one second using the R-package `degross` (Density Estimation from GRouped Summary Statistics) developed and maintained by the author. The top graph in the figure compares the 'target' density with the density estimates obtained from the grouped data frequencies ($\mathcal{D}_0$, dotted line) and from the addition of the grouped sample means ($\mathcal{D}_1$, dashed line). The bottom graph further considers the cumulative addition of the grouped sample standard deviations ($\mathcal{D}_2$, dotted line), skewness and kurtosis ($\mathcal{D}_4$, dashed line) to perform density estimation. An important improvement is observed with the addition of the standard deviations in the dataset. This is confirmed numerically by inspecting the evolution of the root integrated mean squared error (RIMSE) and of the Kullback-Leibler (K-L) divergence between the 'target' density and the estimate obtained using sample moments of increasing orders, see Table 6. When all the tabulated sample moments are used (with $\mathcal{D}_4$), one can see (from the dashed curve in the bottom of Fig. 4 and from the K-L divergence in Table 6) that the target density is nearly perfectly reconstructed. The table also provides information on the effective number of spline parameters $\text{edf}(\hat{\lambda})$ for the roughness penalty parameter $\hat{\lambda}$ selected using Algorithm 3 and quantifying the complexity of the density estimate. The fitted central moments can be compared to their observed counterparts within each of the 3 classes, see Table 7. The observed differences are within the sampling tolerances tuned by the variance-covariance matrix Σ_j of the moment estimators, see (6), with a larger tolerance for class $\mathcal{C}_3$ as it is associated to the smallest frequency, $n_3 = 116$. Value-at-Risk measures corresponding to 95% and 99% quantile estimates were also computed using the theory of Section 3.3 with 95% credible intervals first evaluated on the $\log_{10}$-scale using (14) and transformed back to the original scale (in euros) for reporting purposes. These can be compared with the actual values that were calculated from the confidential raw data, $\text{VaR}_{95\%} = 16\,125$ and $\text{VaR}_{99\%} = 38\,099$ euros, respectively.

6. Discussion

We have shown how to combine tabulated summary statistics involving moments of order one to four with the observed frequencies to obtain a smooth estimate of a density with compact support. The proposed inference strategy, implemented in the R-package `degross`, relies on an EM algorithm with uncertainty measures computed in a final step from the observed penalized log-likelihood. The penalty not only encourages smoothing of the resulting density estimate, but also ensures agreement up to sampling errors between the underlying theoretical moments and their observed values in each class. Simple parametric alternatives might be considered for the density model in specific settings. The nonparametric estimation studied here could then be used to validate or select such proposals, or to point out their possible shortcomings.

Although the transmission of data using tabulated summary statistics may not be fully compliant with the European General Data Protection Regulation (GDPR, EU 2016/679) guidelines, it enables to mask data details by summarizing them with a couple of technical numbers besides the class frequencies, see e.g. Table 5. This is a convenient method to communicate in a fairly accurate and compact way on the distribution of the underlying raw data with a limited loss of information.

That methodology might be combined with regression models where information on the distribution of the response is provided in such a summarized way conditionally on a selected and limited number of subject characteristics (such as the age category in the car insurance example). At the individual level, besides covariates values, the reported loss would take the form of a class indicator. The challenge would be to make inference on the regression model components from such imprecise information on the response data. Flexible forms for the error distribution and for the quantification of covariate effect on the response conditional distribution should be compatible with the available information at the aggregate level.

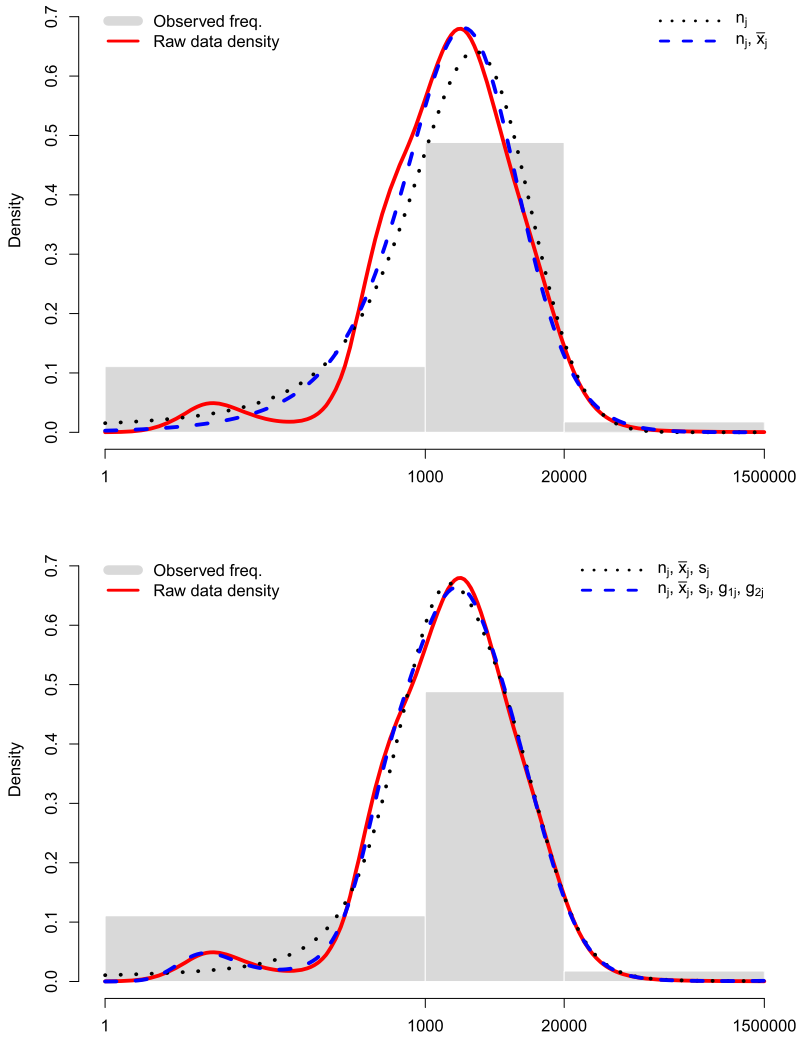

Fig. 4. Car insurance data: observed frequencies and density estimates (on the $\log_{10}$-scale) using tabulated sample moments.

An extension of the methodology to estimate a density with a non-compact support could also be studied. A splicing model (Reynkens et al., 2017) could e.g. be used to combine an estimate of the density of X, obtained using the proposed strategy based on the observed tabulated summary statistics below some threshold, with an extreme value distribution fitted to the right tail.

Declaration of competing interest

No competing interest.

Data availability

The data that has been used is confidential.

Acknowledgements

The author would like to thank Dr. Bernard Lejeune (ULiege, Belgium) for useful discussions about the Generalized Method of Moments and Prof. Michel Denuit (UCLouvain, Belgium) for motivating this project and sharing the data used in the application. Philippe Lambert also acknowledges the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie Universitaire Louvain.

References

Arellano-Valle, R.B., Harnik, S.B., Genton, M.G., 2021. On the asymptotic joint distribution of multivariate sample moments. In: Ghosh, N.B.I., Ng, H. (Eds.), Advances in Statistics – Theory and Applications Honoring the Contributions of Barry C. Arnold. Springer Nature Switzerland AG 2021, pp. 181–206.
Bernard, C., Denuit, M., Vanduffel, S., 2018. Measuring portfolio risk under partial dependence information. The Journal of Risk and Insurance 85, 843–863.
Bolancé, C., Guillen, M., Nielsen, J.P., 2003. Kernel density estimation of actuarial loss functions. Insurance. Mathematics & Economics 32, 19–36.
Braun, J., Duchesne, T., Stafford, J.E., 2005. Local likelihood density estimation for interval censored data. Canadian Journal of Statistics 33, 39–60.
Brockett, P., Cox, S., Golany, B., Phillips, F., Song, Y., 1995. Actuarial usage of grouped data: an approach to incorporating secondary data. Transactions—Society of Actuaries 47, 89–113.

Cossette, H., Landriault, D., Marceau, E., Moutanabbir, K., 2016. Moment-based approximation with mixed Erlang distributions. Variance 10, 161–182.
Courtois, C., Denuit, M., 2007. Local moment matching and s-convex extrema. ASTIN Bulletin 37, 387–404.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, Methodological 39, 1–22.
Eilers, P.H.C., 2007. Ill-posed problems with counts, the composite link model and penalized likelihood. Statistical Modelling 7, 239–254.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–102.
Embrechts, P., Hofert, M., 2013. A note on generalized inverses. Mathematical Methods of Operations Research 77, 423–432.
Furman, E., Kuznetsov, A., Miles, J., 2021. Risk aggregation: a general approach via the class of generalized gamma convolutions. Variance 13, 233–249.
Hansen, L., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.
Hardy, M.R., 2006. Approximating the Aggregate Claims Distribution. John Wiley & Sons, Ltd.
Hasselblad, V., Stead, A.G., Galke, W., 1980. Analysis of coarsely grouped data from the lognormal distribution. Journal of the American Statistical Association 75, 771–778.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. John Wiley and Sons.
Hürlimann, W., 2008. Extremal moment methods and stochastic orders. Boletin de la Associacion Matematica Venezolana 15, 153–301.
Jaspers, S., Lambert, P., Aerts, M., 2016. A Bayesian approach to the semiparametric estimation of a minimum inhibitory concentration distribution. Annals of Applied Statistics 10, 906–924.
Jorion, P., 2006. Value at Risk, 3rd ed. McGraw-Hill.
Klugman, S.A., Panjer, H.H., Wilmot, G.E., 2019. Loss Models. From Data to Decisions, 5th ed. John Wiley & Sons, Ltd.
Lambert, P., Eilers, P.H.C., 2009. Bayesian density estimation from grouped continuous data. Computational Statistics & Data Analysis 53, 1388–1399.
Laverny, O., Masiello, E., Maume-Deschamps, V., Rullière, D., 2021. Estimation of multivariate generalized gamma convolutions through Laguerre expansions. Electronic Journal of Statistics 15, 5158–5202.
Papkov, G.I., Scott, D.W., 2010. Local-moment nonparametric density estimation of pre-binned data. Computational Statistics & Data Analysis 54, 3421–3429.
Pentikäinen, T., 1987. Approximative evaluation of the distribution function of aggregate claims. ASTIN Bulletin 17, 15–39.
Reynkens, T., Verbelen, R., Beirlant, J., Antonio, K., 2017. Modelling censored losses using splicing: a global fit strategy with mixed Erlang and extreme value distributions. Insurance. Mathematics & Economics 77, 65–77.
Rizzi, S., Gampe, J., Eilers, P.H., 2015. Efficient estimation of smooth distributions from coarsely grouped data. American Journal of Epidemiology 182, 138–147.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press, Boca Raton.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society, Series B, Statistical Methodology 71 (2), 319–392.
Thompson, R., Baker, R., 1981. Composite link functions in generalized linear models. Journal of the Royal Statistical Society. Series C. Applied Statistics 30, 125–131.