A new look at the inverse Gaussian distribution with applications to insurance and economic data

Antonio Punzo

Department of Economics and Business, University of Catania, Catania, Italy

ABSTRACT

Insurance and economic data are often positive, and we need to take into account this peculiarity in choosing a statistical model for their distribution. An example is the inverse Gaussian (IG), which is one of the most famous and considered distributions with positive support. With the aim of increasing the use of the IG distribution on insurance and economic data, we propose a convenient mode-based parameterization yielding the reparametrized IG (rIG) distribution; it allows/simplifies the use of the IG distribution in various branches of statistics, and we give some examples. In nonparametric statistics, we define a smoother based on rIG kernels. By construction, the estimator is well-defined and does not allocate probability mass to unrealistic negative values. We adopt likelihood cross-validation to select the smoothing parameter. In robust statistics, we propose the contaminated IG distribution, a heavy-tailed generalization of the rIG distribution to accommodate mild outliers. Finally, for model-based clustering and semiparametric density estimation, we present finite mixtures of rIG distributions. We use the EM algorithm to obtain maximum likelihood estimates of the parameters of the mixture and contaminated models. We use insurance data about bodily injury claims, and economic data about incomes of Italian households, to illustrate the models.

ARTICLE HISTORY

Received 8 March 2018
Accepted 22 October 2018

KEYWORDS

Mode; positive support; insurance losses; heavy-tailed distributions; mixture models

1. Introduction

Insurance and economic data are often positive and their distribution is usually right-skewed. One of the most famous parametric models with these peculiarities is the inverse Gaussian (IG), a two-parameter family of distributions with probability density function (PDF) typically expressed as

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp \left\{ -\frac{\lambda(x - \mu)^2}{2\mu^2x} \right\}, \quad 0 < x < \infty,$$

where $\mu > 0$ is the mean and $\lambda > 0$ is the shape parameter, inversely related to the distribution variability. As well-known (see, e.g. [46, Chapter 15]), the PDF in (1), which is a
memberof the exponential family, is unimodal, with mode located at
\[ \mu \left( \sqrt{1 + \frac{9\mu^2}{4\lambda^2}} - \frac{3\mu}{2\lambda} \right), \tag{2} \]
and positively skewed, with skewness
\[ 3 \sqrt{\frac{\mu}{\lambda}}. \tag{3} \]

For the many attractive properties of this distribution, making it one of the most famous and considered distributions with positive support, see Tweedie [87] and the review paper by Folks and Chhikara [34]. Seshadri [80] provides a detailed list of fields where the IG distribution has been applied with success; see also [46, Chapter 15, 15, Chapter 2].

A problem of the parameterization in (1) is that, for skewed distributions like the IG, the mean \( \mu \) may be far from the mode, but the mode may be what we intuitively want as the ‘descriptive handle’ on the distribution. Moreover, the mode is easily understood, not affected by extremely large or small values, and can be located graphically (see, e.g. [64]). So, to further increase the applicability of the IG distribution, in Section 2, we propose a convenient parameterization based on the mode \( \theta \) and on a parameter \( \gamma \) which is closely related to the distribution variability. We refer to the resulting distribution as reparametrized IG (rIG).

The adopted parameterization simplifies/allows the use of the IG distribution in various branches of statistics, and we give some examples in Section 3. In detail, in Section 3.1, we define finite mixtures of rIG distributions for semiparametric density estimation and clustering of positive data; see also Section 3.1.1. To understand the importance of having mixture components parametrized in terms of the mode, just consider that mixtures have often been praised for their ability to model multimodal distributions, where each mode represents a certain entity [11]. The underlying basic idea is that modes can be associated with important structure in an empirical distribution. For example, as highlighted by Carreira-Perpiñán [11], in visual modeling or object detection, a probabilistic model of a visual scene should account for multimodal distributions so that multiple objects can be represented [62]. In missing data reconstruction and inverse problems, multivalued mappings can be derived from the modes of the conditional distribution of the missing variables given the present ones [12]. In Section 3.1.2, we illustrate an expectation–maximization (EM) algorithm to obtain maximum likelihood (ML) estimates of the mixture parameters. In Section 3.2, we propose a kernel smooth estimator specifically conceived for nonparametric density estimation of positive data. Kernel functions are chosen from the family of rIG distributions (Section 3.2.1); since their support matches the support of the data, the resulting nonparametric density curve does not allocate probability mass to unrealistic negative values. We adopt likelihood cross-validation to select the smoothing parameter (Section 3.2.2). In Section 3.3, we introduce the contaminated IG distribution, a four-parameter heavy-tailed generalization of the rIG distribution to handle the possible presence of mild outliers. In addition to the parameters of the rIG distribution, the contaminated IG distribution has one parameter controlling the proportion of mild outliers and one specifying the degree of contamination (Section 3.3.1). We describe an EM algorithm to obtain ML estimates of the parameters (Section 3.3.2). Advantageously with respect to
the rIG distribution, mild outliers are automatically down-weighted in the estimation of \( \theta \) and \( \gamma \), so providing a robust method of parameter estimation and, once the model is fitted, mild outliers can be directly identified via maximum \textit{a posteriori} probabilities. In Section 4, we use real insurance (Section 4.1) and economic (Section 4.2) data to appreciate the usefulness of the new parameterization and of the proposed models. At last, in Section 5, we summarize the key aspects of the proposal, along with future possible extensions.

2. Reparameterized inverse Gaussian distribution

In this section, we present our parameterization of the IG distribution (Section 2.1) and we give some details about the weighted log-likelihood function (Section 2.2) which can be seen as a generalization of the classical log-likelihood function to be used when sample weights are available.

2.1. The model

The reparametrized IG (rIG) distribution we propose has PDF

\[
f(x; \theta, \gamma) = \sqrt{\frac{\theta (3\gamma + \theta)}{2\pi \gamma x^3}} \exp \left\{ -\frac{(x - \sqrt{\theta (3\gamma + \theta)})^2}{2\gamma x} \right\}, \quad 0 < x < \infty,
\]

where \( \theta, \gamma > 0 \). The link between the parameterizations in (1) and (4) is

\[
\begin{align*}
\mu &= \sqrt{\theta (3\gamma + \theta)} \\
\lambda &= \frac{\theta (3\gamma + \theta)}{\gamma}
\end{align*} \quad \Leftrightarrow \quad \begin{align*}
\theta &= \mu \left( \sqrt{1 + \frac{9\mu^2}{4\lambda^2} - \frac{3\mu}{2\lambda}} \right) \\
\gamma &= \frac{\mu^2}{\lambda}
\end{align*}.
\]

Now, focus on the system on the right-hand side of (5). Recalling (2), the equation on the top guarantees that \( \theta \) is the mode of \( X \); the effect of varying \( \theta \), with \( \gamma \) kept fixed, is illustrated in Figure 1. The equation on the bottom is chosen so that \( \gamma \) is related to the variability of \( X \) without making the PDF formulation more complex. To have an idea, if we chose \( \gamma \) to

![Figure 1. Reparameterized inverse Gaussian PDF in (4) with \( \gamma = 1 \).](image-url)
Figure 2. Reparameterized inverse Gaussian PDF in (4) with $\theta = 1$.

be directly the variance, i.e. $\gamma = \mu^3/\lambda$, then we would get

$$
\begin{align*}
\mu &= \frac{23^{1/3}\theta^2 + 2^{1/3}\left(27\gamma\theta + \sqrt{729\gamma^2\theta^2 - 12\theta^6}\right)^{2/3}}{6^{2/3}\left(27\gamma\theta + \sqrt{729\gamma^2\theta^2 - 12\theta^6}\right)^{1/3}}, \\
\lambda &= \frac{1}{\gamma} \left[3\gamma\theta + \frac{\left(\frac{2}{3}\right)^{1/3}\theta^4}{\left(27\gamma\theta + \sqrt{729\gamma^2\theta^2 - 12\theta^6}\right)^{1/3}} + \frac{\theta^2\left(27\gamma\theta + \sqrt{729\gamma^2\theta^2 - 12\theta^6}\right)^{1/3}}{21^{1/3}3^{2/3}}\right],
\end{align*}
$$

which is more complex if compared with the system on the left-hand side of (5).

We now try to clarify the role of $\gamma$. From the standard theory on the IG distribution with PDF given in (1), the variance is $\mu^3/\lambda$ (see, e.g. [46, Equation (15.6)]); thus, thanks to (5), the variance of the random variable $X$ with PDF (4) is

$$
\gamma \sqrt{\theta} \sqrt{3\gamma + \theta}.
$$

The last expression, analyzed as a function of $\gamma$, is monotone increasing; consequently, fixed $\theta$, the variability increases in line with the value of $\gamma$, confirming that $\gamma$ governs the spread of $X$. The effect of varying $\gamma$, the mode $\theta$ kept fixed, is illustrated in Figure 2.

### 2.2. Maximum weighted likelihood estimation

Several estimation methods of the parameters of the rIG distribution may be considered. Among them, we focus on ML because, for the IG distribution, the usual regularity conditions are satisfied and the method consequently yields a consistent estimator with the usual asymptotic normality properties (cf. [14]). Note also that Barbasi et al. [8] address the problem of ML estimation and random variate generation for the class of tempered positive stable distributions, for which the IG distribution is a special case. In particular, in view of the analysis of Section 4.2 on income data, we first describe the more general maximum weighted likelihood (MWL) estimation – useful in complex surveys where units are drawn from a population with unequal probabilities – and then we make explicit the
closed-form ML estimates of \(\mu\) and \(\lambda\). Note that, for the parametric rIG-based models, we will introduce in Section 3, we only discuss the classical ML approach.

Given a sample \(x_1,\ldots,x_n\) from the PDF in (4), the weighted log-likelihood function related to the rIG distribution is

\[
l(\theta, \gamma) = \sum_{i=1}^{n} w_i \ln[f(x_i; \theta, \gamma)],
\]

where \(w_i \geq 0, i = 1,\ldots,n\), is a given weight. If \(w_1 = \cdots = w_n = 1\), then the classical log-likelihood function is obtained.

The first-order partial derivatives of (6) with respect to \((\theta, \gamma)'\) are

\[
l'(\theta, \gamma) = \sum_{i=1}^{n} w_i u(x_i; \theta, \gamma),
\]

where

\[
u(x_i; \theta, \gamma) = \frac{\partial}{\partial (\theta, \gamma)} \ln[f(x_i; \theta, \gamma)]
\]

is the \(2 \times 1\) vector of the first partial derivatives of \(\ln[f(x_i; \theta, \gamma)]\) with respect to \((\theta, \gamma)'\), for a given observation \(x_i\). Details about the two elements of \(u(x_i; \theta, \gamma)\) are given in Appendix 1. Similarly, the second-order partial derivatives of (6) with respect to \((\theta, \gamma)'\) are

\[
l''(\theta, \gamma) = \sum_{i=1}^{n} w_i v(x_i; \theta, \gamma),
\]

where

\[
v(x_i; \theta, \gamma) = \frac{\partial^2}{\partial (\theta, \gamma) \partial (\theta, \gamma)} \ln[f(x_i; \theta, \gamma)]
\]

is a symmetric \(2 \times 2\) matrix of the second partial derivatives of \(\ln[f(x_i; \theta, \gamma)]\) with respect to \((\theta, \gamma)'\), for a given observation \(x_i\). Details about the three different elements of \(v(x_i; \theta, \gamma)\) are given in Appendix 1.

The values of \(\hat{\theta}\) and \(\hat{\gamma}\) that maximize \(l(\theta, \gamma)\) are the MWL estimates \(\hat{\theta}\) and \(\hat{\gamma}\) and satisfy the condition

\[
l'(\hat{\theta}, \hat{\gamma}) = 0.
\]

Operationally, we obtain maximization of (6), with respect to \(\theta\) and \(\gamma\), by the general-purpose optimizer \texttt{optim()} for R [75], included in the \texttt{stats} package. The BFGS algorithm, passed to \texttt{optim()} via the argument \texttt{method}, is used for maximization. It is an iterative unconstrained variable-metric/quasi-Newton method solving an approximate version of the Newton equations (to seek a zero of the gradient) by using an approximation of the inverse Hessian. A line search is applied to the resulting search direction,
and a new trial solution found. The inverse Hessian is updated at each iteration by the Broyden–Fletcher–Goldfarb–Shanno formula, giving the method its acronym (see [63, for further details]).

Under the special case of uniform weights, i.e. for ML, the estimates of \( \theta \) and \( \gamma \) can be got from the ML estimates of \( \mu \) and \( \lambda \) as follows

\[
\begin{align*}
\hat{\mu} &= \bar{x} \\
\hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} \left( \frac{1}{x_i} - \frac{1}{\bar{x}} \right)} \\
\hat{\theta} &= \bar{x} \left\{ \sqrt{1 + \frac{9x^2}{4n^2} \left[ \sum_{i=1}^{n} \left( \frac{1}{x_i} - \frac{1}{\bar{x}} \right) \right]^2} - \frac{3\bar{x}}{2n} \sum_{i=1}^{n} \left( \frac{1}{x_i} - \frac{1}{\bar{x}} \right) \right\} \\
\hat{\gamma} &= \frac{x^2}{n} \sum_{i=1}^{n} \left( \frac{1}{x_i} - \frac{1}{\bar{x}} \right),
\end{align*}
\]

where \( \bar{x} = \sum_{i=1}^{n} x_i/n \) is the sample mean. The result in (7) is allowed by (5) and by the invariance property of the ML method (see, e.g. [34, Section 3]).

3. Possible uses in statistics

In this section, we show how our parameterization allows/simplifies the use of the IG distribution in various branches of statistics. We define a finite mixture of \( r \) IG distributions for clustering/classification and semiparametric density estimation (Section 3.1), a smoother based on \( r \) IG kernels for nonparametric density estimation (Section 3.2), and a contaminated IG distribution for robustness in the presence of mild outliers (Section 3.3).

3.1. Model-based clustering and semiparametric density estimation

Finite mixtures of distributions are commonly employed for two different purposes ([86, pp. 2–3]). In indirect applications, they are used as semiparametric competitors of nonparametric density estimation techniques, while in direct applications, they are considered a powerful device for clustering/classification by often assuming that each mixture component represents a group (or cluster) in the original data.

Most of the work published is concerned with mixtures of Gaussian distributions. Although using Gaussian components is in principle appropriate when the theoretical support is \( R \), it is not adequate if the support is \((0, \infty)\) due to the allocation of probability mass to unrealistic negative values. This problem can be eliminated by considering component PDFs defined on \((0, \infty)\). Motivated by this consideration, we suggest using \( r \) IG components.

3.1.1. Mixtures of \( r \) IG distributions

Taking advantage of our parameterization in (4), we rewrite the PDF of a finite mixture of \( k \) IG distributions (see [4], in the case \( k = 2 \), and [61]) as

\[
p(x; \mathbf{\pi}, \mathbf{\theta}, \gamma) = \sum_{j=1}^{k} \pi_j f(x; \theta_j, \gamma_j), \quad 0 < x < \infty.
\]
In (8)

- $f(x; \theta_j, \gamma_j)$ is the rIG component PDF with parameters $\theta_j$ and $\gamma_j$;
- $\pi = (\pi_1, \ldots, \pi_k)'$ is the vector of mixture weights, with $\pi_j \in (0, 1]$ and $\sum_{j=1}^{k} \pi_j = 1$;
- $\theta = (\theta_1, \ldots, \theta_k)'$ is the vector of component modes $\theta_j$;
- $\gamma = (\gamma_1, \ldots, \gamma_k)'$ is the vector of component parameters $\gamma_j$.

Thus, there are $3k-1$ unknown parameters to be estimated. Of course, as also underlined by Izenman [44, p. 103] and Bagnato and Punzo [5], there is no guarantee that $p$ will produce a multimodal density with the same number of modes as there are densities in the mixture; similarly, there is no guarantee that those individual modes $\theta_j$ will remain at the same locations in (8). Indeed, the shape of the mixture distribution depends upon both the spacings of the modes and the relative shapes of the component distributions. Nevertheless, we retain that for well-separated components, the values of $\theta_j$ should accurately approximate the location of the mixture modes.

### 3.1.2. ML estimation: The EM algorithm

To find ML estimates of the parameters for model (8) with a fixed number of mixture components $k$, we use the EM algorithm, which is a natural approach for ML estimation when data are incomplete. For model (8), the source of incompleteness, the classical one in the use of mixture models, arises from the fact that for each observation we do not know its component membership; this source, which is especially related to a direct application of the model, is governed by an indicator vector $z_i = (z_{i1}, \ldots, z_{ik})$, where $z_{ij} = 1$ if $x_i$ comes from component $j$ and $z_{ij} = 0$ otherwise. The complete-data likelihood can be written as

$$L_c(\pi, \theta, \gamma) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[ \pi_j f \left( x_i; \theta_j, \gamma_j \right) \right]^{z_{ij}}.$$  

Therefore, the complete-data log-likelihood becomes

$$l_c(\pi, \theta, \gamma) = l_{1c}(\pi) + l_{2c}(\theta, \gamma),$$  \hspace{1cm} (9)

where

$$l_{1c}(\pi) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln \pi_j,$$  \hspace{1cm} (10)

$$l_{2c}(\theta, \gamma) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln \left[ f \left( x_i; \theta_j, \gamma_j \right) \right].$$  \hspace{1cm} (11)

The EM algorithm iterates between two steps, one E-step and one M-step, until convergence. These steps are described below.

#### 3.1.2.1. E-step

The E-step, on the $(r+1)$th iteration of the EM algorithm, requires the calculation of $Q(\pi, \theta, \gamma)$, the current conditional expectation of $l_c(\pi, \theta, \gamma)$. To do this, we
need to calculate

\[ E \left( Z_{ij} | x_i; \pi^{(r)}, \theta^{(r)}, \gamma^{(r)} \right) = \frac{\pi_j^{(r)} f \left( x_i; \theta_j^{(r)}, \gamma_j^{(r)} \right)}{p \left( x_i; \pi^{(r)}, \theta^{(r)}, \gamma^{(r)} \right)} =: z_{ij}^{(r)}. \]

Then, by substituting \( z_{ij} \) with \( z_{ij}^{(r)} \) in (9), and based on (10) and (11), we obtain

\[ Q(\pi, \theta, \gamma) = Q_1(\pi) + Q_2(\theta, \gamma). \]  

3.1.2.2. **M-step.** The M-step on the \((r + 1)\)th iteration of the EM algorithm requires the calculation of \( \pi^{(r+1)}, \theta^{(r+1)}, \) and \( \gamma^{(r+1)} \) as the values of \( \pi, \theta, \) and \( \gamma \) that maximize \( Q(\pi, \theta, \gamma) \). As the two terms on the right-hand side of (12) have zero cross-derivatives, they can be maximized separately. Maximizing \( Q_1(\pi) \) with respect to \( \pi \), subject to the constraints on these parameters, yields

\[ \pi_j^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(r)}, \quad j = 1, \ldots, k. \]

Maximizing \( Q_2(\theta, \gamma) \) with respect to \( \theta \) and \( \gamma \) (subject to the constraints on these parameters), is equivalent to independently maximizing each of the \( k \) expressions

\[ Q_{2j}(\theta_j, \gamma_j) = \sum_{i=1}^{n} z_{ij}^{(r)} \ln \left( f \left( x_i; \theta_j, \gamma_j \right) \right), \quad j = 1, \ldots, k. \]

\( Q_{2j}(\theta_j, \gamma_j) \) is a weighted log-likelihood, with weights \( z_{ij}^{(r)}, i = 1, \ldots, n \), whose maximization has been discussed in Section 2.2.

3.1.3. **ML estimation: issues**

Although the ML estimator is the most popular for mixtures, a number of analytical and computational issues arise; we analyze them in the context of the \( r \)IG mixture in (8).

A first problem is that the (observed-data) likelihood function is unbounded when \( k > 1 \): it tends to infinity when one of the component modes, say \( \theta_j \), coincides with a sample observation \( x_i \) and the corresponding \( \gamma_J \) parameter tends to zero. As well-known, this problem also affects the mixture of \( k \) Gaussian distributions with means \( \theta_1, \ldots, \theta_k \) and standard deviations \( \gamma_1, \ldots, \gamma_k \) [26]. For Gaussian mixtures, there are a number of solutions to this problem. The majority of the existing approaches conveniently constrain the standard deviations. The works by Dennis [27], Hathaway [39,40], DeSarbo and Cron [28], and Tanaka and Takemura [85] go in this direction. Other approaches to deal with the unboundedness of the likelihood function for Gaussian mixtures exist like the penalized ML estimator proposed by Ciuperca et al. [19]. All these approaches may be easily re-adapted to the \( r \)IG mixture, but this is beyond the scope of this work.

A second problem is that the log-likelihood function has several local maxima – including those related to the lack of identifiability due to label switching – and the EM algorithm could be entrapped in a spurious local maximum [93]. This problem is typically circumvented by running the algorithm for a set of different starting values, and then choosing
the EM solution which gives the largest log-likelihood value. Seidel et al. [79] shows that different starting values can lead to quite different estimates in the context of mixtures of exponential distributions, to which we recall the rIG mixture belongs. To mitigate this problem, in the real data analyses of Section 4, we initialize the EM algorithm (for all the competing mixture models) by providing the initial quantities $z_i^{(0)}$, $i = 1, \ldots, n$, to the first M-step: nine times using a random initialization and once with a k-means initialization (as implemented by the kmeans () function for R). The solution maximizing the observed-data log-likelihood among these 10 runs is then selected; see [6,23].

A third problem is that the log-likelihood function can also have flat regions where the log-likelihood has very tiny variations. If the EM algorithm, as often happens, is stopped when two successive iterations differ by less than a small value, then it is likely that it stops at some point which might not even be a local maximum. To reduce this problem, in the fashion of Böhning et al. [10], we use the Aitken acceleration procedure [1] to estimate the limiting value of the log-likelihood at each iteration of the EM algorithm. Based on this estimate, we can decide whether or not the algorithm has reached convergence, i.e. whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration $r+1$ is given by

$$a^{(r+1)} = \frac{l^{(r+2)} - l^{(r+1)}}{l^{(r+1)} - l^{(r)}}$$

where $l^{(r)}$ is the observed-data log-likelihood value from iteration $r$. The estimate of the limit $l_\infty$ of the sequence of log-likelihood values is

$$l_\infty^{(r+2)} = l^{(r+1)} + \frac{1}{1 - a^{(r+1)}} \left( l^{(r+2)} - l^{(r+1)} \right).$$

The EM algorithm can be considered to have converged when $|l_\infty^{(r+2)} - l_\infty^{(r+1)}| < \epsilon$, with $\epsilon > 0$ being the desired tolerance (see [10,82,83] for further details). In the analyses of Section 4, we fix $\epsilon = 0.0001$.

### 3.2. Nonparametric density estimation

Due to their conceptual simplicity and practical and theoretical properties, kernel smoothers are one of the most popular statistical methods for nonparametric density estimation. Given the random sample $X_1, \ldots, X_n$, these estimators are merely a sum of $n$ (usually symmetric) 'bumps' (the so-called kernels), with equal weights $1/n$, placed over each observation. A classical example, the default of various softwares/packages implementing this approach, are the Gaussian kernels. However, as stressed by Chen [13], Gaussian kernels (defined on the whole real line) produce an estimated density function which allocates probability mass outside the theoretical support $(0, \infty)$ of the data. To circumvent this issue, Chen [13] proposes a nice and natural remedy consisting in using kernels defined on the same support of the unknown density $f$, so that no probability mass is assigned outside.

Motivated by these considerations, in Section 3.2.1, rIG distributions are adopted to propose a well-defined kernel smoother – that is an estimator satisfying all the fundamental properties of a PDF – with support $(0, \infty)$. An automatic method to select the smoothing parameter of the proposed density estimator is illustrated in Section 3.2.2.
3.2.1. Reparametrized IG kernel density estimation

We consider the following kernel density smoother

\[
\hat{f}(x; \gamma) = \frac{1}{n} \sum_{i=1}^{n} f(x; \theta = X_i, \gamma) = \frac{1}{n} \sum_{i=1}^{n} k_\gamma(x; X_i), \quad 0 < x < \infty, \tag{13}
\]

where \(k_\gamma(x; X_i)\) and \(\gamma\) are the rIG kernel and the smoothing parameter, respectively. As we can note, an rIG density is placed over each single observation \(X_i\) by fixing \(\theta = X_i\) in (4); this is in line with Punzo [66], Punzo and Zini [72] and Mazza and Punzo [55–59]. By construction, model (13) defines a density function. It is interesting to note that the rIG kernel density estimator in (13) is a special case of the rIG mixture in (8) when \(k = n, \pi_j = 1/n, \theta_j = X_j, \) and \(\gamma_j = \gamma_j, j = i = 1, \ldots, n.\)

Two quantities characterize the nonparametric estimator (13): the smoothing parameter \(\gamma\) and the rIG kernels \(k_\gamma(x; X_i)\). The former can be considered as a smoothing parameter for the following considerations: according to the results of Section 2, if \(\gamma\) is chosen too large, then all details, such as modes, may be obscured by \(\hat{f}(x; \gamma)\). Vice versa, as \(\gamma\) becomes small, spurious fine structure becomes visible. The limit as \(\gamma \to 0^+\) is a sum of \(n\) Dirac delta functions (spikes) over the observations; consequently, \(\hat{f}(x; \gamma)\) converges to the empirical frequency distribution. As regards the rIG kernels, they obey the fundamental graphical properties of a kernel function. In detail, they are non-negative, integrate to one, assume their maximum value when \(x = X_i\) and are smoothly non-increasing as the point \(x\) departs from \(X_i\). The only unconventional property is their skewness: indeed, fixed \(\gamma\), the kernel shape changes naturally according to the position where the observation \(X_i\) falls (see Figure 1). In particular, thanks to (5) and recalling (3), the skewness of the density (4) is

\[
3 \sqrt{\frac{\gamma}{\sqrt{\theta (3\gamma + \theta)}}}; \tag{14}
\]

fixing \(\gamma\) in (14), the skewness is a decreasing function of \(\theta\). This peculiarity of the rIG kernels is shared by other asymmetric kernels used in the literature, such as the gamma kernels [13].

3.2.2. The choice of the smoothing parameter \(\gamma\)

To select the smoothing parameter \(\gamma\), we use likelihood cross-validation (LCV). According to this method, \(\gamma\) is chosen by minimizing the score function

\[
\text{LCV}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \ln \left[\hat{f}_{-i}(x = X_i; \gamma)\right]
\]

over the possible values of \(\gamma\), where \(\hat{f}_{-i}\) is the density estimate in (13) without the data point \(X_i\). The value of \(\gamma\) that minimizes \(\text{CV}(\gamma)\) is referred to as the LCV smoothing parameter, \(\hat{\gamma}_{\text{LCV}}\). We perform minimization via the \texttt{nlm()} function, of the \texttt{stats} package for \texttt{R}, which carries out a non-linear minimization of \(\text{LCV}(\gamma)\) using a Newton-type algorithm.
3.3. Robustness against mild outliers

Although the IG is one of the most considered distributions with support \((0, \infty)\), real data are often ‘contaminated’ by outliers – at one or both ends of the support – that can affect the estimation of the parameters. Thus, the detection of outliers, and the development of robust methods of parameter estimation insensitive to their presence, is an important problem.

Outliers can be roughly distinguished into two types (cf. [77, pp. 79–80]): mild (also referred to as bad points herein, in analogy with [2]) and gross. Mild outliers, on which we focus on, are observations sampled from some population different or even far from the assumed model. Such outliers document mainly the difficulty of the specification problem. In their presence, the statistician is recommended to choose a model flexible enough to accommodate all data points, including the outliers. The classical choice is to consider heavy-tailed distributions; endowed with heavy tails, they offer the flexibility needed for achieving mild outliers robustness. Heavy tails are typically obtained by embedding the reference distribution (the IG in our case) in a larger model with one or more additional parameters denoting deviation from the reference distribution due to mild outliers; for a discussion about the concept of reference distribution, see Davies and Gather [25] and Hennig [41].

By choosing the rIG as reference distribution, in Section 3.3.1, we propose a simple four-parameter contaminated model in order to accommodate all the available data points. The proposed model is a two-component mixture in which one of the components, with a large prior probability, represents the good points (reference distribution), and the other, with a small prior probability, the same mode, and an inflated parameter \(\gamma\), represents the bad points. This is a simple theoretical model for the occurrence of bad points and the two additional parameters, with respect to the parameters of the reference rIG distribution, have a direct interpretation in terms of proportion of good points and degree of contamination (a sort of measure of how different bad points are from the bulk of the good points). Advantageously, the proposed model also allows for automatic detection of bad points via a simple and natural procedure based on maximum a posteriori probabilities. Note that, as we will detail in Section 3.3.1, the parameterization of the IG distribution given in (4) is fundamental for the definition of the contaminated model. We discuss ML estimation of the parameters for the contaminated IG distribution in Section 3.3.2 via the adoption of the EM algorithm.

3.3.1. The contaminated IG distribution

The PDF of the contaminated IG model is given by

\[
p(x; \theta, \gamma, \alpha, \eta) = \alpha f(x; \theta, \gamma) + (1 - \alpha) f(x; \theta, \eta \gamma), \quad 0 < x < \infty.
\]  

(15)

In (15):

- \(f(x; \theta, \gamma)\) is the PDF of the rIG, given in (4), chosen as reference distribution.
- \(\alpha \in (0.5, 1)\) can be seen as the proportion of good points. Note that \(\alpha\) is constrained to be greater than 0.5 because, in robust statistics, it is usually assumed that at least half of the observations are good (cf. [41, p. 250]).
η > 1 denotes the degree of contamination and, because of the assumption η > 1, it can be interpreted as the increase in variability due to the bad points with respect to the reference distribution f(x; θ, γ); hence, it is an inflation parameter.

Of course, p(x; θ, γ, α, η) produces a unimodal density, with mode θ, because both the reference and contaminant PDFs have their maximum in θ. As a limiting case of (15), when α → 1− and η → 1+, the reference distribution f(x; θ, γ) is obtained. Moreover, the contaminated IG model can be seen as a special case of the rIG mixture in (8) when k = 2, π1 = α, with α ∈ (0.5, 1), θ1 = θ2 = θ, γ1 = γ, and γ2 = ηγ.

An advantage of model (15) is that, given θ, γ, α, and η, we can establish whether a generic data point x is either good or bad via the a posteriori probability

\[
P(x \text{ is good } | \theta, \gamma, \alpha, \eta) = \frac{\alpha f(x; \theta, \gamma)}{p(x; \theta, \gamma, \alpha, \eta)}. \quad (16)
\]

Based on (16), x will be considered good if P(x is good | θ, γ, α, η) > 1/2, while it will be considered bad otherwise. Such a decision rule can be equivalently defined in terms of the discriminant functions

\[
D_{\text{good}}(x; \theta, \gamma, \alpha) = \alpha f(x; \theta, \gamma)
\]

and

\[
D_{\text{bad}}(x; \theta, \gamma, \alpha, \eta) = (1 - \alpha) f(x; \theta, \eta \gamma),
\]

so that x will be classified as good if

\[
D_{\text{good}}(x; \theta, \gamma, \alpha) > D_{\text{bad}}(x; \theta, \gamma, \alpha, \eta), \quad (17)
\]

and bad otherwise. More generally, by solving (17) as a function of x, the positive real line is partitioned in two regions, say R_{\text{good}} and R_{\text{bad}}, of good and bad data, respectively. By using the (monotonically increasing) logarithmic transformation of the discriminant functions, as in Ingrassia and Punzo [43], the inequality in (17) becomes the following quadratic inequality

\[
a(\eta)x^2 + b(\theta, \gamma, \alpha, \eta)x + c(\theta, \eta) < 0,
\]

where \(a(\eta) = \eta - 1 > 0\), \(c(\theta, \eta) = (\eta - 1)\theta^2 > 0\), and

\[
b(\theta, \gamma, \alpha, \eta) = -2\gamma \left\{ \ln(\alpha) - \ln(1 - \alpha) \right\} - \eta \gamma \left\{ \ln(\eta(3\gamma + \theta)) - \ln(3\eta\gamma + \theta) \right\} - 2\sqrt{\theta} \left( \eta\sqrt{3\gamma + \theta} - \sqrt{3\eta\gamma + \theta} \right). \quad (18)
\]

Due to the domain of the parameters θ, γ, α, and η, it is straightforward to show that three addends, on the right-hand side of (18), are negative. Consequently, \(b(\theta, \gamma, \alpha, \eta) < 0\). If

\[
\Delta(\theta, \gamma, \alpha, \eta) = b(\theta, \gamma, \alpha, \eta)^2 - 4\theta^2(\eta - 1)^2 > 0,
\]

then x will be considered good when it lies in the interval delimited by the roots

\[
\frac{1}{2(\eta - 1)} \left[ -b(\theta, \gamma, \alpha, \eta) \mp \sqrt{\Delta(\theta, \gamma, \alpha, \eta)} \right]. \quad (19)
\]

Since \(4\theta^2(\eta - 1)^2 > 0\), we have that \(0 < \sqrt{\Delta(\theta, \gamma, \alpha, \eta)} < -b(\theta, \gamma, \alpha, \eta)\) and, consequently, the two extremes of the interval defined in (19) are positive. This means that we
have an interval around \(-b(\theta, \gamma, \alpha, \eta)\) composed by good data, and the rest of the positive real data composed by bad points.

Finally, it is worth to be mentioned that the contaminated IG model in (15) assumes that the contaminant distribution is rIG itself, like the reference distribution. Although this assumption may seem to be restrictive, it allows reference and contaminant distributions to be comparable not only in terms of the mode \(\theta\), but also in terms of the variability parameter \(\gamma\), so that \(\eta > 1\) effectively acts as an inflation parameter.

3.3.2. ML estimation: An EM algorithm

In analogy with Section 2.2, estimates of the parameters \(\theta, \gamma, \alpha, \eta\) can be determined by the maximization of the weighted log-likelihood function if sample weights \(w_1, \ldots, w_n\) are available in addition to the sample \(x_1, \ldots, x_n\) from model (15). Details about the four first-order partial derivatives of \(\ln[p(x; \theta, \gamma, \alpha, \eta)]\) are given in Appendix 2 for the reader interested in this approach.

To find classical ML estimates of the parameters, as for the mixture of rIG distributions, we illustrate the use of the EM algorithm. In this case, the source of incompleteness arises from the fact that we do not know whether the generic data point \(x_i, i = 1, \ldots, n\), is good or bad. To denote this source of incompleteness, we use the indicator variables \(v_1, \ldots, v_i, \ldots, v_n\), where \(v_i = 1\) if \(x_i\) is good and \(v_i = 0\) otherwise, \(i = 1, \ldots, n\). Therefore, the complete-data are given by \((x_1, v_1), \ldots, (x_i, v_i), \ldots, (x_n, v_n)\) and the complete-data likelihood, on which the algorithm works on, can be written as

\[
L_c(\theta, \gamma, \alpha, \eta) = \prod_{i=1}^{n} \left[ \alpha f(x_i; \theta, \gamma) \right]^{v_i} \left[ (1 - \alpha) f(x_i; \theta, \eta \gamma) \right]^{1 - v_i}.
\]

Simple algebra yields the following complete-data log-likelihood

\[
l_c(\theta, \gamma, \alpha, \eta) = l_1(\alpha) + l_2(\theta, \gamma, \eta),
\]

where

\[
l_1(\alpha) = \sum_{i=1}^{n} \left[ v_i \ln \alpha + (1 - v_i) \ln (1 - \alpha) \right]
\]

and

\[
l_2(\theta, \gamma, \eta) = \sum_{i=1}^{n} \left[ v_i \ln f(x_i; \theta, \gamma) + (1 - v_i) \ln f(x_i; \theta, \eta \gamma) \right].
\]

E-step and M-step are described below.

3.3.2.1. E-step. The E-step, on the \((r + 1)\)th iteration of the EM algorithm, requires the calculation of \(Q(\theta, \gamma, \alpha, \eta)\), the current conditional expectation of \(l_c(\theta, \gamma, \alpha, \eta)\). To do this, we need to calculate \(E(V_i|x_i; \theta^{(r)}, \gamma^{(r)}, \alpha^{(r)}, \eta^{(r)})\), where \(V_i\) is the random variable related
to \( v_i, i = 1, \ldots, n \); this expectation is given by

\[
E \left( V_i | x_i; \theta^{(r)}, \gamma^{(r)}, \alpha^{(r)}, \eta^{(r)} \right) = \frac{\alpha^{(r)} f \left( x_i; \theta^{(r)}, \gamma^{(r)} \right)}{p \left( x_i; \theta^{(r)}, \gamma^{(r)}, \alpha^{(r)}, \eta^{(r)} \right)} =: v_i^{(r)},
\]

which is the posterior probability that \( x_i \) is a good point; compare with (16). Then, by substituting \( v_i \) with \( v_i^{(r)} \) in (20), and based on (21) and (22), we obtain

\[
Q(\theta, \gamma, \alpha, \eta) = Q_1(\alpha) + Q_2(\theta, \gamma, \eta).
\]

3.3.2.2. M-step. The M-step on the \((r+1)\)th iteration of the EM algorithm requires the calculation of \( \theta^{(r+1)}, \gamma^{(r+1)}, \alpha^{(r+1)} \), and \( \eta^{(r+1)} \) as the values of \( \theta, \gamma, \alpha, \eta \) that maximize \( Q(\theta, \gamma, \alpha, \eta) \).

The update for \( \alpha \) is calculated independently by maximizing

\[
Q_1(\alpha) = \sum_{i=1}^{n} \left[ v_i^{(r)} \ln \alpha + \left( 1 - v_i^{(r)} \right) \ln \left( 1 - \alpha \right) \right] \tag{23}
\]

with respect to \( \alpha \), subject to the constraint on this parameter. Some simple algebra, detailed in Appendix 3, yields

\[
\alpha^{(r+1)} = \max \left\{ 0.5, \frac{n_{\text{good}}^{(r)}}{n} \right\}, \tag{24}
\]

where \( n_{\text{good}}^{(r)} = \sum_{i=1}^{n} v_i^{(r)} \) is the expected proportion of good data. The updates of \( \theta, \gamma, \) and \( \eta \) are obtained by the maximization of the function

\[
Q_2(\theta, \gamma, \eta) = \sum_{i=1}^{n} \left[ v_i^{(r)} \ln f \left( x_i; \theta, \gamma \right) + \left( 1 - v_i^{(r)} \right) \ln f \left( x_i; \theta, \eta \gamma \right) \right]. \tag{25}
\]

For R users, the \texttt{optim()} function, in the \texttt{stats} package, can be used to perform a numerical search of the maximum \((\theta^{(r+1)}, \gamma^{(r+1)}, \eta^{(r+1)})\)' of the function (25).

4. Real data analysis

In this section, we will show how the rIG-based models, introduced in Section 3, act on real insurance and economic data. The whole analysis is conducted in R [75].

4.1. Bodily injury claims

The first example comes from the insurance world. As well-known insurance data are often positive, right-skewed, and leptokurtic \[42\]. Several parametric families of distributions have been considered in the literature to accommodate these peculiarities, including the Pareto, Weibull, log-normal, gamma and tempered positive stable distributions \[7,29,37,50,73\]. However, when insurance data exhibit unusual shapes, such as multiple modes, these distributions may not be a good candidate, as well-argued in Lee and Lin \[45\] and Jeon and Kim \[51\]. In these cases, a more flexible modeling framework, such as a mixture modelling framework, is to be preferred. The flexibility of finite mixtures in accommodating various shapes of insurance data is now widely recognized \[9,17,18,54\]. Among them, mixtures of gamma distributions were successfully considered in Dey \textit{et al}.\[75\].
As we will see in the analysis below, mixtures of rIG distributions, introduced in Section 3.1, represent a valid alternative.

We use insurance data from Rempala and Derrig [76], which are also available in the CAS datasets package [31] for R. The sample represents the bodily injury claims from Massachusetts closed in 2001. We consider the $n = 272$ claims that are coded as ‘other providers’, thus ignoring potentially fraudulent claims; all numbers are in thousand dollars as in the original paper.

The histogram of the data, displayed in Figure 3, shows multimodality and right-skewness. To further explore the characteristics of the empirical PDF, we compute the rIG kernel density estimator introduced in Section 3.2. The smoothing parameter, selected according to the likelihood cross-validation method discussed in Section 3.2.2, is $\hat{\gamma}_{LCV} = 0.431$; the corresponding solid curve is superimposed on the histogram in Figure 3. The nonparametric curve confirms the multimodality suggested by the histogram giving prominence to a clear bimodality.

Motivated by these preliminary findings, we fit mixtures of the following distributions: rIG, unimodal gamma, exponential, Weibull, Burr, and inverse Burr. The models chosen for comparison are widely used in the actuarial literature (see, e.g. [61,68]. The parameters of the competing models are estimated by ML via the EM algorithm, with a number $k$ of mixture components ranging from 1 to 4. We implemented a convenient code to fit mixtures of rIG and unimodal gamma distributions, while we use the `flexmix()` function, of the flexmix package [38], to fit the other models. We select the best value of $k$, as usual in the mixture modelling literature, via the Bayesian information criterion (BIC; [78]). Even though the regularity properties needed for the development of the BIC are not satisfied by mixture models [47,48], it has been used extensively (see, e.g. [24,35]) and performs well in practice. We compute the BIC as

$$BIC = 2l(\hat{\pi}, \hat{\theta}, \hat{\gamma}) - (3k - 1) \ln n,$$

Figure 3. Bodily injury claims. Histogram together with a rIG-kernel density estimator.
where \( l(\hat{\pi}, \hat{\theta}, \hat{\gamma}) \) is the maximized (observed-data) log-likelihood. Note that Bayes factors can be used to compare models that are not nested, and the BIC approximation thereto holds when models are not nested (cf. [74]).

Table 1 shows the obtained BIC values. The BIC suggests \( k = 3 \) components for mixtures of gamma and Weibull distributions, and \( k = 2 \) components for the other mixtures. These results confirm the observation that a single (\( k = 1 \)) parametric model, among those considered, is unable to represent the distribution of the bodily injury claims.

Overall, the best model is the mixture of two rIG distributions. The appropriateness of this model is also corroborated by comparing the empirical and estimated quantiles on the right tail of the distribution; these quantiles are computed numerically over a grid of 51 probabilities equally spaced between 0.965 and 0.999 (see Figure 4). The comparison is focused on the right tail only because it is of particular interest in actuarial applications. To have an idea of this importance, just think that the value at risk (VaR), which is a classical

Table 1. Bodily injury claims. BIC values for the fitted models. Bold numbers refer to the best value of \( k \) for each model.

| Model                      | 1   | 2      | 3      | 4      |
|----------------------------|-----|--------|--------|--------|
| Mixt. of rIG PDFs          | -1169.075 | **-1026.641** | -1031.266 | -1046.069 |
| Mixt. of unimodal gamma PDFs | -1093.122 | -1066.879 | **-1033.998** | -1049.733 |
| Mixt. of exponential PDFs  | -1089.475 | **-1075.402** | -1086.613 | -1097.825 |
| Mixt. of Weibull PDFs      | -1095.071 | -1043.331 | **-1039.256** | -1050.277 |
| Mixt. of Burr PDFs         | -1067.415 | **-1033.547** | -1052.026 | -1072.458 |
| Mixt. of inverse Burr PDFs | -1049.546 | **-1029.501** | -1050.182 | -1071.788 |

Figure 4. Bodily injury claims. Model fitting in terms of VaR on the right tail.
risk measure, summarizes the distribution of possible losses by a quantile on the right tail. The VaR values from the rIG and Weibull mixtures pass through, and seem to be closer to, the empirical ones (cf. Figure 4). On the opposite side, the VaR values from the gamma mixture are systematically greater than the empirical ones, while the VaR values from the inverse Burr mixture are almost always lower. To confirm these considerations, we also compute the sum, over the considered grid, of the absolute differences between each estimated curve and the empirical quantiles. The ranking induced by this discrepancy measure is (from best to worst): rIG mixture (21.128), Weibull mixture (21.169), inverse Burr mixture (44.255), Burr mixture (47.253), exponential mixture (50.006), and unimodal gamma mixture (51.179).

The estimated parameters of the selected rIG mixture are given in Table 2, while its graphical representation is displayed, via a solid line, in Figure 5, with dotted curves showing the component densities multiplied by the corresponding estimated weights \( \hat{\pi}_j, j = 1, 2 \). Group membership of the observations is represented by ticks of different colors (black for group 1 and gray for group 2) on the x-axis. The analogous representation for the other mixture model having unimodal components parameterized with respect to the mode, which is the mixture of \( k = 3 \) unimodal gamma distributions, is displayed in Figure 6. We can note how the additional flat component, with estimated mixing proportion 0.085,

| Component j | \( \hat{\pi}_j \) | \( \hat{\theta}_j \) | \( \hat{\gamma}_j \) |
|-------------|-----------------|-----------------|------------------|
| 1           | 0.507           | 0.175           | 11.901           |
| 2           | 0.493           | 2.527           | 0.262            |

**Table 2.** Bodily injury claims. Estimated parameters for the mixture of two rIG distributions.

![Figure 5. Bodily injury claims. Histogram together with the fitted mixture of \( k = 2 \) rIG densities. Dotted lines show the component densities multiplied by the corresponding weights. Black and gray are used for observations in group 1 and group 2, respectively, as classified by the fitted model.](image_url)
mode approximately equal to 0, and variability parameter 9.648, is attempting to capture the right tail of the distribution of the claims. In these terms, the selected rIG mixture is more parsimonious because it does not require an additional component to reproduce the right tail of the empirical distribution of the claims.

This application emphasizes the importance of the mode-parameterization, which immediately gives an idea of the location, on the $x$-axis, of the losses with the highest probability (see the third column of Table 2). In particular, the first mode suggests that a loss of 175 dollars is the most likely for this dataset. Moreover, the estimated modes can be used to facilitate comparisons across space and time of the two losses more representative of the distribution.

### 4.2. Income of Italian households in 1986

The second example comes from the economic literature and it is related to the estimation of the income distribution. Information from such estimation is used to measure welfare, inequality and poverty, to assess changes in these measures over time, and to compare measures across countries, over time and before and after specific policy changes, designed, for example, to alleviate poverty. Thus, the estimation of the income distribution is of central importance for assessing many aspects of the well being of society (see [81, for further considerations]).

The income distribution has been estimated both parametrically and nonparametrically (see, e.g. [16]). Parametric estimation is convenient because it facilitates subsequent inferences about inequality and poverty measures based on the estimated income distribution.
parameters. A large number of alternative parametric models, including the Weibull and Dagum distributions [32], have been considered in the literature for estimating the income distribution (see [49], for a survey). As well documented in Dagum [22], a convenient parametric model should be: defined on a strictly positive support, unimodal, and positively skewed; moreover, all the parameters of the specified model should have a well-defined economic meaning and, following a principle of parsimony, the model should make use of the smallest possible number of parameters for adequate and meaningful representation. Unfortunately, as emphasized by Van Praag et al. [88], Feser [33] and Cowell and Victoria-Feser [20], real income data are often ‘contaminated’ by outliers (bad incomes) – at one or both ends of the distribution [3,91] – that affect the estimation of the parameters for the chosen model. This in turn will affect the inequality measure computed from the estimated parameters. To remedy this problem, Cowell and Victoria-Feser [21] and Weich et al. [91] trim 1% of the upper and lower tails of the income distribution. However, the approach relies on a subjective choice of the proportion of incomes to be trimmed out and the information contained in these discarded incomes is completely lost. As we will see in the analysis below, the contaminated IG distribution can be a simple remedy to these problems.

We use incomes of Italian households, for 1986, obtained from the Luxembourg Income Study (LIS) database (http://www.lisdatacenter.org/). The data analyzed here are \( n = 6016 \) household incomes, expressed in millions of lire (national currency that was in use at the time of data collection), with corresponding sample weights. The weighted histogram of the data, obtained via the function \texttt{wtd.hist()} of the \texttt{weights} package [65] for R, is displayed in Figure 7. The histogram highlights unimodality and positive skewness and, as quite common in situations like this (see, e.g. [3,21]), outliers appear (see the ticks on the \( x \)-axis) that yield an heavier right tail.

Motivated by these considerations, we fit the rIG and contaminated IG distributions via the MWL approach. To pursue this aim, in both cases, we implemented a convenient R code; while for the rIG distribution we retrace Section 2.2, for the contaminated IG

![Figure 7. Income of Italian households in 1986. Weighted histogram.](image-url)
distribution we follow a direct approach that uses the `optim()` function to (simultaneously) estimate its four parameters. To select the best model, we cannot use classical model selection criteria, such as AIC and BIC, because they are not valid with sample weights. Lumley and Scott [52] recently proposed the dAIC as model selection criterion to be used under complex sampling schemes. According to the notation introduced in Section 2.2, the dAIC can be written as

\[
d\text{AIC} = 2l(\hat{\theta}) - 2\text{tr}(\hat{V}^{-1}\hat{U}),
\]

where \(\hat{\theta}\) is the MVL estimate of the parameter vector \(\theta\) of the considered model,

\[
\hat{V} = -\sum_{i=1}^{n} \frac{w_i}{n} \sum_{j=1}^{n} w_j v(x_i; \hat{\theta})
\]

and

\[
\hat{U} = \sum_{i=1}^{n} \frac{w_i}{n} \sum_{j=1}^{n} w_j \left[ u(x_i; \hat{\theta}) u(x_i; \hat{\theta})' \right].
\]

The dAIC in (26) coincides with the Takeuchi information criterion (TIC; [84]) when there are not sample weights (cf. [52]). With data at hand, the dAIC values are \(-48386.445\) and \(-48383.397\) for the rIG and contaminated IG distributions, respectively; this leads us to prefer the contaminated model.

Figure 8. Income of Italian households in 1986. Weighted histogram together with the fitted contaminated IG density (solid line). Dotted lines show the densities for good and bad incomes multiplied by the corresponding weights. Gray and black are used for good and bad incomes, respectively, as classified by the fitted model.
Figure 9. Income of Italian households in 1986. Estimated posterior probabilities to be good incomes. Gray and black are used for good and bad incomes, respectively, as classified by the fitted model.

The estimated parameters for the contaminated IG distribution are \( \hat{\theta} = 5.389, \hat{\nu} = 6.179, \hat{\alpha} = 0.991, \) and \( \hat{\eta} = 15.726. \) The modal income \( \hat{\theta} \) highlights that the most common income received by the Italian households is about 5.389 millions of lire. The estimated value of \( \alpha \) indicates that about the 0.9% of the incomes can be considered as bad according to the fitted model, with \( \hat{\eta} \) giving the degree of badness (measure of how far the bad incomes are from the bulk of the data). The corresponding estimated curve is represented, via a solid line, in Figure 8, along with the weighted histogram; dotted curves show the densities for good and bad incomes multiplied by the corresponding estimated weights \( \hat{\alpha} \) and \( 1 - \hat{\alpha} \). Maximum a posteriori classification of incomes, as good or bad, is represented by ticks of different colors (gray for good incomes and black for bad incomes) on the x-axis.

To be more precise, according to the decision rule in (17), incomes are classified as good if they lie in the interval delimited by 0.351 and 82.706 millions of lire; refer to formula (19).

Figure 9 reports, for each income \( x_i \), the estimated posterior probability in (16) to be good, \( i = 1, \ldots, n; \) as we can see, the farther the income is from the bulk of the data, as represented by the mode \( \hat{\theta} \), the lower is its probability to be a good income. Such probability is also related to the down-weighting of bad incomes in the estimation of the model parameters, and this is an important aspect for robust estimation (see [70] for a discussion about this topic with reference to the mixture of contaminated normal distributions).

5. Conclusions

In this paper, the problem of fitting the distribution of insurance and economic data was addressed by considering models based on the IG distribution. To define these models, we proposed a mode-based parameterization of the IG distribution, that we call rIG distribution. This distribution was used to define three different models to be applied for
positive (insurance and economic) data: an rIG kernel smoother for nonparametric density estimation (Section 3.2), a contaminated IG distribution for robust density estimation (Section 3.3), and a finite mixture of rIG distributions for clustering and semiparametric density estimation (Section 3.1). The results from the real data analysis discussed in Section 4 offer a clear and coherent framework in the study of the typical insurance and economic data features and confirm that our models give a range of alternatives that are well suited to depict the process that generate the data.

However, the applicability of our parameterization is not restricted to the models discussed herein and the use of the rIG-based models is not limited to insurance and economic data. First of all, the IG distribution parameterized in terms of the mode, instead of the mean, may be handy in specifying a IG-shaped prior in Bayesian data analysis [36]. The rIG could be used as a distribution of the error term in modal linear regression [94]; the modal linear regression models the conditional mode of a response $Y$ given a set of predictors $x$ as a linear function of $x$. Also, in the fashion of Punzo and McNicholas [70,71], Punzo et al. [69], and Mazza and Punzo [60], contaminated IG distributions may be used either as components in the definition of a finite mixture model or as emission distributions for hidden Markov models [53,67]. In reliability theory, the parameterization with respect to the mode may simplify the formulation of the hazard rate related to the IG distribution (cf. [80, Chapter 5.3]).

As an open point for further research, it could be interesting to further generalize our contaminated model with the aim of accommodating ‘groups’ of concentrated outliers. In such a case, contamination in the mode (in addition to the contamination in variability, like we do at the moment) could be considered as well; see, e.g. Aitkin and Wilson [2] and the contaminated (location-shift) normal distribution considered by Verdinelli and Wasserman [90].

**Disclosure statement**

No potential conflict of interest was reported by the author.

**ORCID**

Antonio Punzo http://orcid.org/0000-0001-7742-1821

**References**

[1] A.C. Aitken, A series formula for the roots of algebraic and transcendental equations, Proc. Roy. Soc. Edinb. 45 (1926), pp. 14–22.
[2] M. Aitkin and G.T. Wilson, Mixture models, outliers, and the EM algorithm, Technometrics 22 (1980), pp. 325–331.
[3] E. Álvarez, R.M. García-Fernández, F.J. Blanco-Encomienda, and J.F. Muñoz, The effect of outliers on the economic and social survey on income and living conditions, World Acad. Sci., Eng. Technol., Int. J. Soc., Behav., Educ., Econ., Bus. Ind. Eng. 8 (2014), pp. 3276–3280.
[4] R.K. Amoh, Estimation of parameters in mixtures of inverse Gaussian distributions, Commun. Stat. Theory Methods 13 (1984), pp. 1031–1043.
[5] L. Bagnato and A. Punzo, Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm, Comput. Stat. 28 (2013), pp. 1571–1597.
[6] L. Bagnato, A. Punzo, and M.G. Zoaia, The multivariate leptokurtic-normal distribution and its application in model-based clustering, Can. J. Stat. 45 (2017), pp. 95–119.
L. Barabesi, A. Cerasa, A. Cerioli, and D. Perrotta, "A new family of tempered distributions," Electron. J. Stat. 10 (2016), pp. 3871–3893.

L. Barabesi, A. Cerasa, D. Perrotta, and A. Cerioli, "Modeling international trade data with the Tweedie distribution for anti-fraud and policy support," Eur. J. Oper. Res. 248 (2016), pp. 1031–1043.

M. Bernardi, A. Maruotti, and L. Petrella, "Skew mixture models for loss distributions: A bayesian approach," Insur. Math. Econ. 51 (2012), pp. 617–623.

D. Böhnning, E. Dietz, R. Schaub, P. Schlattmann, and B.G. Lindsay, "The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family," Ann. Inst. Stat. Math. 46 (1994), pp. 373–388.

M.A. Carreira-Perpiñán, "Mode-finding for mixtures of Gaussian distributions," IEEE. Trans. Pattern. Anal. Mach. Intell. 22 (2000), pp. 1318–1323.

M.A. Carreira-Perpiñán, "Reconstruction of sequential data with probabilistic models and continuity constraints," in Advances in Neural Information Processing Systems, S.A. Solla, T.K. Leen, and K.R. Müller, eds., Vol. 12, MIT Press, London, 2000, pp. 414–420.

S.X. Chen, "Probability density function estimation using gamma kernels," Ann. Inst. Stat. Math. 52 (2000), pp. 471–480.

R.C.H. Cheng and N.A.K. Amin, "Maximum likelihood estimation of parameters in the inverse gaussian distribution, with unknown origin," Technometrics 23 (1981), pp. 257–263.

R.S. Chhikara and J.L. Folks, The Inverse Gaussian Distribution: Theory, Methodology, and Applications, Volume 95 of Statistics: A Series of Textbooks and Monographs, Taylor & Francis, New York, 1988.

D. Chotikapanich and W.E. Griffiths, "Estimating income distributions using a mixture of gamma densities," in Modeling Income Distributions and Lorenz Curves, D. Chotikapanich, ed., Economic Studies in Inequality, Social Exclusion and Well-Being, Chapter 16, Springer, New York, 2008, pp. 285–302.

S.B. Choy, J.S. Chan, and U.E. Makov, "Robust bayesian analysis of loss reserving data using scale mixtures distributions," J. Appl. Stat. 43 (2016), pp. 396–411.

S.T.B. Choy and C.M. Chan, "Scale mixtures distributions in insurance applications," ASTIN Bull. 33 (2003), pp. 93–104.

G. Ciuperca, A. Ridolfi, and J. Idier, Scand. J. Stat. 30 (2003), pp. 45–59.

F.A. Cowell and M.-P. Victoria-Feser, "Robustness properties of inequality measures," Econometrica 64 (1996), pp. 77–101.

F.A. Cowell and M.-P. Victoria-Feser, "Welfare rankings in the presence of contaminated data," Econometrica 70 (2002), pp. 1221–1233.

C. Dagum, "A new model of personal income distribution: Specification and estimation," in Modeling Income Distributions and Lorenz Curves, D. Chotikapanich, ed., Volume 5 of Economic Studies in Equality, Social Exclusion and Well-Being, Chapter 1, Springer, New York, 2008, pp. 3–25.

U.J. Dang, A. Punzo, P.D. McNicholas, S. Inggrassia, and R.P. Browne, "Multivariate response and parsimony for Gaussian cluster-weighted models," J. Classif. 34 (2017), pp. 4–34.

A. Dasgupta and A.E. Raftery, "Detecting features in spatial point processes with clutter via model-based clustering," J. Am. Stat. Assoc. 93 (1998), pp. 294–302.

L. Davies and U. Gather, "The identification of multiple outliers," J. Am. Stat. Assoc. 88 (1993), pp. 782–792.

N.E. Day, "Estimating the components of a mixture of normal distributions," Biometrika 56 (1969), pp. 463–474.

J.E.J. Dennis (ed.), Algorithms for Nonlinear Fitting, NATO advanced Research Symposium: Cambridge University Press, Cambridge, 1981.

W.S. DeSarbo and W.L. Cron, "A maximum likelihood methodology for clusterwise linear regression," J. Classif. 5 (1988), pp. 249–282.

L. Devroye and L. James, On simulation and properties of the stable law, Stat. Methods Appt. 23 (2014), pp. 307–343.
[30] D.K. Dey, L. Kuo, and S.K. Sahu, A bayesian predictive approach to determining the number of components in a mixture distribution, Stat. Comput. 5 (1995), pp. 297–305.
[31] C. Dutang and A. Charpentier, CASdatasets: Insurance datasets (Official website). Version 1.0-6 (2016-05-28), 2016.
[32] L. Fattorini and A. Lemmi, The stochastic interpretation of the Dagum personal income distribution: A tale, Statistica 66 (2006), pp. 325–329.
[33] M.P.V. Feser, Robust estimation of personal income distribution models. Research Paper DARP/4, London School of Economics and Political Science, London, 1993.
[34] J.L. Folks and R.S. Chhikara, The inverse Gaussian distribution and its statistical application – A review, J. Royal Stat. Soc. Ser. B (Stat. Methodol.) 40 (1978), pp. 263–289.
[35] C. Fraley and A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (2002), pp. 611–631.
[36] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, Bayesian Data Analysis, 3rd ed., Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, Boca Raton, FL, 2013.
[37] M. Grabchak, Tempered Stable Distributions: Stochastic Models for Multiscale Processes, Springer Briefs in Mathematics. Springer International Publishing, Basel, 2016.
[38] B. Grün and F. Leisch, FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters, J. Stat. Softw. 28 (2008, 10), pp. 1–35.
[39] R.J. Hathaway, A constrained formulation of maximum-likelihood estimation for normal mixture distributions, Ann. Stat. 13 (1985), pp. 795–800.
[40] R.J. Hathaway, A constrained EM algorithm for univariate normal mixtures, J. Stat. Comput. Simul. 23 (1986), pp. 211–230.
[41] C. Hennig, Fixed point clusters for linear regression: computation and comparison, J. Classif. 19 (2002), pp. 249–276.
[42] M. Ibragimov, R. Ibragimov, and J. Walden, Heavy-Tailed Distributions and Robustness in Economics and Finance, Volume 214 of Lecture Notes in Statistics, Springer International Publishing, New York, 2015.
[43] S. Ingrassia, and A. Punzo, Decision boundaries for mixtures of regressions, J. Korean Stat. Soc. 45 (2016), pp. 295–306.
[44] A.J. Izenman, Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, New York, 2008.
[45] Y. Jeon and J.H.T. Kim, A gamma kernel density estimation for insurance loss data, Insur. Math. Econ. 53 (2013), pp. 569–579.
[46] N.L. Johnson and S. Kotz, Continuous Univariate Distributions, Vol. 1, John Wiley & Sons, New York, 1970.
[47] C. Keribin, Estimation consistante de l’ordre de modèles de mélange, Comptes Rendus l’Académie Sci.-Ser. I-Math. 326 (1998), pp. 243–248.
[48] C. Keribin, Consistent estimation of the order of mixture models, Sankhyā: Indian J. Stat. Ser. A 62 (2000), pp. 49–66.
[49] C. Kleiber and S. Kotz, Statistical Size Distributions in Economics and Actuarial Sciences, John Wiley & Sons, New York, 2003. Volume 470 of Wiley Series in Probability and Statistics.
[50] S.A. Klugman, H.H. Panjer, and G.E. Willmot, Loss Models: From Data to Decisions, Wiley, Hoboken, NJ, 2012. Wiley Series in Probability and Statistics.
[51] S.C.K. Lee and X.S. Lin, Modeling and evaluating insurance losses via mixtures of Erlang distributions, N. Am. Actuar. J. 14 (2010), pp. 107–130.
[52] T. Lumley and A. Scott, AIC and BIC for modeling with complex survey data, J. Surv. Stat. Methodol. 3 (2015), pp. 1–18.
[53] A. Maruotti and A. Punzo, Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers, Comput. Stat. Data. Anal. 113 (2017), pp. 475–496.
[54] A. Maruotti, V. Raponi, and F. Lagona, Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure, Biom. J. 58 (2016), pp. 280–302.
[55] A. Mazza and A. Punzo, Discrete beta kernel graduation of age-specific demographic indicators, in New Perspectives in Statistical Modeling and Data Analysis, S. Ingrassia, R. Rocci, and M.
Vichi, eds., Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, 2011, pp. 127–134.

[56] A. Mazza and A. Punzo, *Graduation by adaptive discrete beta kernels*, in *Classification and Data Mining*, A. Giusti, G. Ritter, and M. Vichi, eds., Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, 2013, pp. 243–250.

[57] A. Mazza and A. Punzo, *Using the variation coefficient for adaptive discrete beta kernel graduation*, in *Statistical Models for Data Analysis*, P. Giudici, S. Ingrassia, and M. Vichi, eds., Studies in Classification, Data Analysis and Knowledge Organization, Springer International Publishing, Basel, 2013, pp. 225–232.

[58] A. Mazza and A. Punzo, *DBKGrad: An R package for mortality rates graduation by discrete beta kernel techniques*, J. Stat. Softw. 57 (2014), pp. Code Snippet 1–18.

[59] A. Mazza and A. Punzo, *Bivariate discrete beta kernel graduation of mortality data*, Lifetime Data. Anal. 21 (2015), pp. 419–433.

[60] A. Mazza and A. Punzo, *Mixtures of multivariate contaminated normal regression models*, Stat. Pap. doi:10.1007/s00362-017-0964-y.

[61] T. Miljkovic and B. Grün, *Modeling loss data using mixtures of distributions*, Insur. Math. Econ. 70 (2016), pp. 387–396.

[62] B. Moghaddam and A. Pentland, *Probabilistic visual learning for object representation*, IEEE. Trans. Pattern Anal. Mach. Intell. 19 (1997), pp. 696–710.

[63] J. Nash, *On best practice optimization methods in R*, J. Stat. Softw. 60 (2014), pp. 1–14.

[64] J.P. Nolan, *Parameterizations and modes of stable distributions*, Stat. Probab. Lett. 38 (1998), pp. 187–195.

[65] J. Pasek, *weights: Weighting and Weighted Statistics*. Version 0.85 (2016-02-12), 2016.

[66] A. Punzo, *Discrete beta-type models*, in *Classification as a Tool for Research*, H. Locarek-Junge and C. Weihs, eds., Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, 2010, pp. 253–261.

[67] A. Punzo and A. Maruotti, *Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model*, J. Comput. Graph. Stat. 25 (2016), pp. 1097–1116.

[68] A. Punzo, A. Mazza, and A. Maruotti, *Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions*, J. Appl. Stat. 45 (2018), pp. 2563–2584.

[69] A. Punzo, A. Mazza, and P. McNicholas, *ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions*, J. Stat. Softw. 85 (2018), pp. 1–25.

[70] A. Punzo and P.D. McNicholas, *Parsimonious mixtures of multivariate contaminated normal distributions*, Biom. J. 58 (2016), pp. 1506–1537.

[71] A. Punzo and P.D. McNicholas, *Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model*, J. Classif. 34 (2017), pp. 249–293.

[72] A. Punzo and A. Zini, *Discrete approximations of continuous and mixed measures on a compact interval*, Stat. Pap. 53 (2012), pp. 563–575.

[73] S.T. Rachev, Y.S. Kim, M.L. Bianchi, and F.J. Fabozzi, *Financial Models with Levy Processes and Volatility Clustering*, Wiley, Hoboken, NJ, 2011. Frank J. Fabozzi Series.

[74] A.E. Raftery, *Bayesian model selection in social research*, Sociol. Methodol. 25 (1995), pp. 111–164.

[75] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.

[76] G.A. Rempala and R.A. Derrig, *Modeling hidden exposures in claim severity via the EM algorithm*, N. Am. Actuar. J. 9 (2005), pp. 108–128.

[77] G. Ritter, *Robust Cluster Analysis and Variable Selection*, CRC Press, Boca Raton, 2015. Volume 137 of Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

[78] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978), pp. 461–464.

[79] W. Seidel, K. Mosler, and M. Alker, *A cautionary note on likelihood ratio tests in mixture models*, Ann. Inst. Stat. Math. 52 (2000), pp. 481–487.
Appendices

Appendix 1. Partial derivatives of the log PDF of the rIG distribution

The first-order partial derivatives with respect to $\theta$ and $\gamma$, of the logarithm of the PDF in (4), are

$$\frac{\partial \ln [f(x; \theta, \gamma)]}{\partial \theta} = -\frac{3}{2x} - \frac{\theta}{x\gamma} + \frac{1}{3\gamma + \theta} + \frac{3\gamma}{2(3\gamma + \theta)} + \frac{\sqrt{\theta}}{2\gamma \sqrt{3\gamma + \theta}} + \frac{\sqrt{3\gamma + \theta}}{2\gamma \sqrt{\theta}}$$

and

$$\frac{\partial \ln [f(x; \theta, \gamma)]}{\partial \gamma} = \frac{x}{2\gamma^2} + \frac{\theta^2}{2x\gamma^2} - \frac{\theta}{2\gamma(3\gamma + \theta)} + \frac{3\sqrt{\theta}}{2\gamma \sqrt{3\gamma + \theta}} - \frac{\sqrt{\theta (3\gamma + \theta)}}{\gamma^2}.$$ 

The second-order partial derivatives are

$$\frac{\partial^2 \ln [f(x; \theta, \gamma)]}{\partial \theta^2} = -\frac{1}{4} \left( \frac{4}{x\gamma} + \frac{2}{\theta^2} + \frac{2}{(3\gamma + \theta)^2} + \frac{9\gamma}{\theta^{3/2}(3\gamma + \theta)^{3/2}} \right),$$

$$\frac{\partial^2 \ln [f(x; \theta, \gamma)]}{\partial \theta \partial \gamma} = \frac{\partial^2 \ln [f(x; \theta, \gamma)]}{\partial \gamma \partial \theta}$$

$$= \frac{\theta}{x\gamma^2} + \frac{-27\gamma^3 - 30\gamma \theta^2 - 4\theta^3 - 3\gamma^2 [21\theta + 2\sqrt{\theta (3\gamma + \theta)}]}{4\gamma^2 \sqrt{\theta (3\gamma + \theta)^{5/2}}}.$$
and

\[
\frac{\partial^2 \ln \left[ f(x; \theta, \gamma) \right]}{\partial \gamma^2} = -\frac{x}{\gamma^3} - \frac{\theta^2}{x \gamma^3} + \frac{3 \theta}{2 \gamma (3 \gamma + \theta)^2} - \frac{9 \sqrt{\theta}}{4 \gamma (3 \gamma + \theta)^{3/2}} + \frac{\theta}{2 \gamma^2 (3 \gamma + \theta)} \\
- \frac{3 \sqrt{\theta}}{\gamma^2 \sqrt{3 \gamma + \theta}} + \frac{2 \sqrt{\theta} (3 \gamma + \theta)}{\gamma^3}.
\]

**Appendix 2. First partial derivatives of the log PDF of the contaminated IG distribution**

The first-order partial derivatives with respect to \( \theta, \gamma, \alpha, \) and \( \eta \) of the logarithm of the PDF in (15), are

\[
\frac{\partial \ln \left[ p(x; \theta, \gamma, \alpha, \eta) \right]}{\partial \theta} = \frac{1}{2 \sqrt{2 \pi} \gamma^3 x^4 p(x; \theta, \gamma, \alpha, \eta)} \\
\times \left[ \alpha (3 \gamma + 2 \theta) \left\{ x \left[ \gamma + \sqrt{\theta (3 \gamma + \theta)} \right] - \sqrt{\theta (3 \gamma + \theta)} \right\} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \gamma + \theta)} \right]^2}{2 x \gamma} \right\} \right] \\
+ \left[ (1 - \alpha) (3 \eta \gamma + 2 \theta) \left\{ x \left[ \eta \gamma + \sqrt{\theta (3 \eta \gamma + \theta)} \right] - \sqrt{\theta (3 \eta \gamma + \theta)} \right\} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \eta \gamma + \theta)} \right]^2}{2 x \eta \gamma} \right\} \right].
\]

\[
\frac{\partial \ln \left[ p(x; \theta, \gamma, \alpha, \eta) \right]}{\partial \gamma} = \frac{\theta}{2 \sqrt{2 \pi} \gamma^3 x^4 p(x; \theta, \gamma, \alpha, \eta)} \\
\times \left[ x^2 (3 \gamma + \theta) + \theta^2 (3 \gamma + \theta) - x \left[ \gamma \theta + 3 \gamma \sqrt{\theta (3 \gamma + \theta)} + 2 \theta \sqrt{\theta (3 \gamma + \theta)} \right] \right] \\
\times \left[ \frac{\left[ x - \sqrt{\theta (3 \gamma + \theta)} \right]^2}{2 x \gamma} \right] \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \gamma + \theta)} \right]^2}{2 x \gamma} \right\} \\
+ \left[ (1 - \alpha) \left[ x^2 (3 \eta \gamma + \theta) + \theta^2 (3 \eta \gamma + \theta) - x \left[ 2 \theta \sqrt{\theta (3 \eta \gamma + \theta)} + \eta \gamma \left( \theta + 3 \sqrt{\theta (3 \eta \gamma + \theta)} \right) \right] \right] \right] \\
\times \left[ \frac{\left[ x - \sqrt{\theta (3 \eta \gamma + \theta)} \right]^2}{2 x \eta \gamma} \right] \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \eta \gamma + \theta)} \right]^2}{2 x \eta \gamma} \right\}. 
\]

\[
\frac{\partial \ln \left[ p(x; \theta, \gamma, \alpha, \eta) \right]}{\partial \alpha} = \frac{\sqrt{\theta (3 \gamma + \theta)} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \eta \gamma + \theta)} \right]^2}{2 x \eta \gamma} \right\}}{\alpha \sqrt{x \gamma^3} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \gamma + \theta)} \right]^2}{2 x \gamma} \right\}} \\
- \sqrt{\theta (3 \eta \gamma + \theta)} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \eta \gamma + \theta)} \right]^2}{2 x \gamma} \right\} \exp \left\{ - \frac{\left[ x - \sqrt{\theta (3 \gamma + \theta)} \right]^2}{2 x \gamma} \right\}.
\]
and
\[ \frac{\partial \ln[p(x; \theta, \gamma, \alpha, \eta)]}{\partial \eta} = \exp\left\{ -\frac{[x - \sqrt{\theta (3 \gamma + \theta)}]_+^2}{2xy} \right\} \]
\[ \times \frac{2x^4 \gamma^2 \eta^3 \theta (3 \gamma + \theta)}{x^3 \gamma \eta^3} \]
\[ \times \frac{(1 - \alpha)}{\alpha} \left( (x^2 + \theta^2) \left[ \theta (3 \eta \gamma + \theta) \right]^{3/2} \right) - x \theta^2 \left\{ 9 \gamma^2 \eta^2 + 2 \theta^2 + \eta \gamma \left[ 9 \theta + \sqrt{\theta (3 \eta \gamma + \theta)} \right] \right\} \]
\[ + (1 - \alpha) \sqrt{\theta (3 \eta \gamma + \theta)} \frac{2x \eta \gamma}{x^3 \eta \gamma} \exp\left\{ -\frac{[x - \sqrt{\theta (3 \gamma + \theta)}]_+^2}{2xy} \right\}. \]

**Appendix 3. Contaminated IG distribution: update of \( \alpha \)**

The first-order derivative of \( Q_1(\alpha) \) in (23), with respect to \( \alpha \), is
\[ \frac{dQ_1(\alpha)}{d\alpha} = \frac{n^{(r)}_{\text{good}} - \alpha n}{\alpha(1 - \alpha)}. \]  

Setting (A1) equal to zero, and solving for \( \alpha \), yields the following critical point
\[ \tilde{\alpha} = n^{(r)}_{\text{good}} / n. \]

Now, the derivative in (A1) is formed by three multiplicative terms: \( n^{(r)}_{\text{good}} - \alpha n \), \( \alpha \), and \( (1 - \alpha) \). By simultaneously evaluating the signs of these terms for values of \( \alpha \) lying in the interval \((0, 1)\), we have that (A1) is positive for \( \alpha < \tilde{\alpha} \) (positive slope of \( Q_1 \)) and negative for \( \alpha > \tilde{\alpha} \) (negative slope of \( Q_1 \)). Therefore, \( \tilde{\alpha} \) is a global maximum.

However, we are looking for a global maximum of \( Q_1(\alpha) \), say \( \alpha^{(r+1)} \), in the sub-interval \((0.5, 1)\). In light of the characteristics of \( Q_1(\alpha) \), we can split the search of \( \alpha^{(r+1)} \) as follows: if \( \tilde{\alpha} > 0.5 \), then \( \alpha^{(r+1)} = \tilde{\alpha} \); while if \( \tilde{\alpha} < 0.5 \), then \( \alpha^{(r+1)} = 0.5 \). Combining these results, we obtain the update of \( \alpha \) given in (24).