European Actuarial Journal (2021) 11:185–226
https://doi.org/10.1007/s13385-021-00264-3

ORIGINAL RESEARCH PAPER

Making Tweedie's compound Poisson model more accessible

Łukasz Delong[1] · **Mathias Lindholm**[2] · **Mario V. Wüthrich**[3]

Received: 9 June 2020 / Revised: 23 November 2020 / Accepted: 21 January 2021 /
Published online: 13 February 2021

Abstract
The most commonly used regression model in general insurance pricing is the compound Poisson model with gamma claim sizes. There are two different parametrizations for this model: the Poisson-gamma parametrization and Tweedie's compound Poisson parametrization. Insurance industry typically prefers the Poisson-gamma parametrization. We review both parametrizations, provide new results that help to lower computational costs for Tweedie's compound Poisson parameter estimation within generalized linear models, and we provide evidence supporting the industry preference for the Poisson-gamma parametrization.

Keywords Compound Poisson model · Gamma claim sizes · Tweedie's distribution · Exponential dispersion family · Regression model · Generalized linear model · Neural network

1 Introduction

The most commonly used regression model in general insurance pricing is the compound Poisson model with gamma claim sizes. State-of-the-art industry practice fits two separate generalized linear models (GLMs) to the two parts of this model, namely, a Poisson GLM to claim counts and a gamma GLM to claim amounts. Both the Poisson and the gamma distributions belong to the exponential dispersion family

✉ Mario V. Wüthrich
mario.wuethrich@math.ethz.ch

Łukasz Delong
lukasz.delong@sgh.waw.pl

Mathias Lindholm
lindholm@math.su.se

[1] SGH Warsaw School of Economics, Institute of Econometrics, Warsaw, Poland

[2] Department of Mathematics, Stockholm University, Stockholm, Sweden

[3] RiskLab, Department of Mathematics, ETH Zurich, Zurich, Switzerland

(EDF). It has been noted by Tweedie [19] that the compound Poisson model with i.i.d. gamma claim sizes itself belongs to the EDF and, in fact, it closes the interval of power variance functions between the Poisson model and the gamma model, see Section 3 in Jørgensen [5]. As a result of Tweedie's and Jørgensen's findings we obtain two different parametrizations of the compound Poisson model with i.i.d. gamma claim sizes. Selection between these two different parametrizations has been explored in the work of Jørgensen–de Souza [6] in the context of GLM insurance pricing. Interestingly, to predict total claim amounts we need to fit two GLMs in the compound Poisson-gamma parametrization, whereas one GLM is sufficient to get the corresponding predictions within Tweedie's EDF parametrization. This indicates that in GLM applications these two parametrizations are not fully consistent. This point has been raised by Smyth–Jørgensen [17] who propose to use a double generalized linear model (DGLM) in Tweedie's parametrization to simultaneously model mean and dispersion parameters within the EDF.

The main purpose of this article is to revisit the work of Smyth–Jørgensen [17], and to give properties under which the two GLMs for claim counts and claim sizes and the DGLM with Tweedie's EDF parametrization lead to the same predictive model; this involves a discussion about choices of covariate spaces and GLM link functions. Based on this, our first main contribution provides a new result for Tweedie's DGLM that substantially reduces computational costs in calibrations of power variance parameters.

The second point that we explore is whether the insurance industry's preference of using the Poisson-gamma parametrization can be justified. A priori it is not clear whether either of the two ways lead to better predictive models. This part of our work is based on GLMs and on their neural network extensions. We receive evidence that supports the industry preference, in particular, under the choice of neural network regression models the Poisson-gamma parametrization is simpler in calibration and leads to more robust results.

We close this introduction with a number of remarks. First, we mention the recent survey paper of Quijano Xacur–Garrido [12], which has similar goals to the present paper. This survey only considers the single GLM case of Tweedie's parametrization, similar to Jørgensen–de Souza [6]. We emphasize that the full picture can only be obtained by comparing the Poisson-gamma parametrization to the DGLM case introduced in Smyth–Jørgensen [17]. Therefore, we revisit and extend this latter reference to receive a comprehensive comparison. Our view is supported by examples. These examples provide a proof of concept for situations with claims that are not too heavy tailed. However, these examples also highlight the weaknesses of this model on real insurance data, which often exhibits heavier tails than what is suitable under a gamma assumption. We remark that in our discussion we use the terminology of general insurance pricing, however, as commonly the case in general insurance, all our findings can be translated one-to-one to claims reserving problems.

Organization of the paper In Sect. 2 we introduce the compound Poisson model with i.i.d. gamma claim sizes and we derive its corresponding Tweedie parametrization. In Sect. 3 we embed both approaches into a GLM framework. We present the two GLMs needed for the Poisson-gamma parametrization, and we discuss a single GLM and a DGLM parametrization for Tweedie's approach. Our main results,

Theorems 3.6 and 3.8, give conditions under which the different GLM parametrizations lead to identical predictive models. These theorems provide a remarkable property that allows us to lower calibration costs in Tweedie's DGLMs. In Sect. 4 we give insights and intuition based on numerical examples both under GLMs and neural network regression models. In Sect. 5 we conclude, and the "Appendix" gives a short summary of GLMs and describes the data used.

2 Tweedie's compound Poisson model

In Sect. 2.1 we introduce the compound Poisson model with i.i.d. gamma claim sizes, and in Sect. 2.2 we revisit its Tweedie counterpart. For simplicity, in these two sections, we think of using these models for modeling one single insurance policy only. In Sect. 3, below, we consider multiple insurance policies also allowing for heterogeneity between policies.

2.1 Compound Poisson model with i.i.d. gamma claim sizes

Let N be the number of claims and let $(Z_j)_{j\geq 1}$ be the corresponding claim sizes. We assume that the number of claims, N, is Poisson distributed with mean λw, where $\lambda > 0$ is the expected claim frequency relative to a given exposure $w > 0$; we write $N \sim \text{Poi}(\lambda w)$. We assume that the claim sizes Z_j, $j \geq 1$, are i.i.d. and independent of N having a gamma distribution with shape parameter $\gamma > 0$ and scale parameter $c > 0$; we write $Z_1 \sim \mathcal{G}(\gamma, c)$ for this gamma distribution. The moment generating function of the gamma claim sizes is given by, see Section 3.2.1 in [20],

$$\mathbb{E}\big[\exp\{rZ_1\}\big] = \left(\frac{c}{c-r}\right)^{\gamma}, \qquad \text{for r<c.}$$

The compound Poisson model with i.i.d. gamma claim sizes (CPG) is then defined by $S = \sum_{j=1}^{N} Z_j$; we use notation $S \sim \text{CPG}(\lambda w, \gamma, c)$. The moment generating function of S is given by

$$\mathbb{E}\big[\exp\{rS\}\big] = \exp\left\{\lambda w\left(\left(\frac{c}{c-r}\right)^{\gamma} - 1\right)\right\}, \qquad \text{for r<c,} \tag{2.1}$$

we refer to Proposition 2.11 in [20].

2.2 Tweedie's compound Poisson model

Following [5, 6, 17, 19] we select a particular model within the EDF. A random variable Y belongs to the EDF if its density has the following form (w.r.t. a σ-finite measure on $\mathbb{R}$)

$$Y \sim f(y;\theta, w/\phi) = \exp\left\{\frac{y\theta - \kappa(\theta)}{\phi/w} + a(y;w/\phi)\right\}, \tag{2.2}$$

with $w > 0$ is a given exposure (weight, volume), $\phi > 0$ is the dispersion parameter, $\theta \in \mathbf{\Theta}$ is the canonical parameter in the effective domain $\mathbf{\Theta}$, $\kappa : \mathbf{\Theta} \to \mathbb{R}$ is the cumulant function, $a(\cdot;\cdot)$ is the normalization, *not* depending on the canonical parameter θ.

For properties of the EDF we refer to "Appendix A", below. Tweedie's compound Poisson (CP) model is obtained by choosing for $p \in (1,2)$ the cumulant function

$$\kappa(\theta) = \kappa_p(\theta) = \frac{1}{2-p}((1-p)\theta)^{\frac{2-p}{1-p}}, \quad \text{on effective domain } \theta \in \mathbf{\Theta} = \mathbb{R}_- = (-\infty, 0). \tag{2.3}$$

We use notation $Y \sim \text{Tweedie}(\theta, w, \phi, p)$. The first two derivatives of the cumulant function provide the first two moments of Y, see also (A.1) in the "Appendix",

$$\mu = \mathbb{E}[Y] = \kappa_p'(\theta) = ((1-p)\theta)^{\frac{1}{1-p}}, \tag{2.4}$$

$$\text{Var}(Y) = \frac{\phi}{w}\kappa_p''(\theta) = \frac{\phi}{w}((1-p)\theta)^{\frac{p}{1-p}} = \frac{\phi}{w}\mu^p. \tag{2.5}$$

Hyper-parameter $p \in (1,2)$ allows us to model the power variance functions $V(\mu) = \mu^p$ between the Poisson boundary case $p = 1$ and the gamma boundary case $p = 2$, we refer to Sect. 3.1, below, for the boundary cases. $\mu \mapsto \theta = (\kappa_p')^{-1}(\mu)$ gives the canonical link of Tweedie's CP model.

We calculate the moment generating function of the exposure scaled Tweedie's CP random variable wY, see also Corollary 7.21 in [20],

$$\begin{aligned}\mathbb{E}\big[\exp\{rwY\}\big] &= \exp\left\{\frac{w}{\phi}\big(\kappa_p(\theta + r\phi) - \kappa_p(\theta)\big)\right\} \\ &= \exp\left\{\frac{w}{\phi}\,\kappa_p(\theta)\left(\left(\frac{-\theta/\phi}{-\theta/\phi - r}\right)^{\frac{2-p}{p-1}} - 1\right)\right\}, \quad \text{for } r < -\theta/\phi.\end{aligned}$$

Note that this is a CPG model in a different parametrization; we call the model under this EDF parametrization Tweedie's CP model. The following proposition follows by comparing the corresponding moment generating functions.

Proposition 2.1 *Choose* $S \underset{(\mathrm{d})}{\sim} \text{CPG}(\lambda w, \gamma, c)$ *and* $Y \sim \text{Tweedie}(\theta, w, \phi, p)$. *We have identity in distribution* $S/w \overset{(\mathrm{d})}{=} Y$ *under parameter identification*

$$\gamma = \frac{2-p}{p-1} \Leftrightarrow p = \frac{\gamma+2}{\gamma+1} \in (1,2), \tag{2.6}$$

$$c = -\theta/\phi, \tag{2.7}$$

$$\lambda = \frac{1}{\phi}\,\kappa_p(\theta). \tag{2.8}$$

Formula (2.8) can be rewritten in different ways. We have, using the canonical link of Tweedie's CP model, $\theta = (\kappa_p')^{-1}(\mu) = \mu^{1-p}/(1-p)$ and $\kappa_p(\theta) = \kappa_p((\kappa_p')^{-1}(\mu)) = \mu^{2-p}/(2-p)$. This implies, using (2.7) in the second step and (2.6) in the last step,

$$\lambda = \frac{1}{\phi}\,\kappa_p(\theta) \;=\; \frac{c}{-\theta}\,\kappa_p(\theta) \;=\; c\,\frac{p-1}{\mu^{1-p}}\,\frac{\mu^{2-p}}{2-p} \;=\; \frac{c}{\gamma}\,\mu. \tag{2.9}$$

The latter says that, of course, the expected claim frequency λ is obtained by dividing the expected total claim amount $\mathbb{E}[Y] = \mu$ by the average claim size $\mathbb{E}[Z_1] = \gamma/c$.

Thus, under parameter identification scheme (2.6)–(2.8) the two models are identical:

$$\text{Tweedie}(\theta, w, \phi, p) \overset{(d)}{=} \text{CPG}\left(\frac{w}{\phi}\kappa_p(\theta), \frac{2-p}{p-1}, \frac{-\theta w}{\phi}\right), \qquad \text{or}$$

$$\text{CPG}(\lambda w, \gamma, c) \overset{(d)}{=} \text{Tweedie}\left((\kappa_p')^{-1}\left(\lambda w\frac{\gamma}{c}\right), w, \frac{-w}{c}(\kappa_p')^{-1}\left(\lambda w\frac{\gamma}{c}\right), \frac{\gamma+2}{\gamma+1}\right).$$

This illustrates that there is a one-to-one correspondence between the CPG parametrization and Tweedie's CP parametrization, i.e. the two models are identical and only differ in interpretation of parameters. The next section will demonstrate that these subtle differences can be crucial for GLM regression modeling, and resulting models can be rather different as functions of explanatory covariates, see Sect. 3.3 below.

3 Generalized linear models and parameter estimation

In this section we study multiple insurance policies $i = 1, \ldots, n$ having claim distributions $\text{CPG}(\lambda_i w_i, \gamma, c_i)$ and $\text{Tweedie}(\theta_i, w_i, \phi_i, p)$, respectively. We allow for heterogeneity between the policies in all parameters that have a lower index i. We describe modeling and parameter estimation within GLMs: we consider two GLMs to model λ_i (Poisson) and $-c_i/\gamma$ (gamma) in the former case, and we consider a DGLM to model θ_i and ϕ_i in the latter case. There is a slight difference between "two GLMs" and "double GLM", the former considers two independent GLMs, the latter does a simultaneous consideration of two GLMs. The volumes w_i are assumed to be known and do not need any modeling. The shape parameter $\gamma > 0$ and the power variance parameter $p = (\gamma+2)/(\gamma+1)$, see (2.6), are assumed to be the same for all policies i, this is a standard assumption in state-of-the-art use of these GLMs. An overview of GLMs and their parameter estimation within the EDF is given in "Appendix A".

3.1 Compound Poisson model with i.i.d. gamma claim sizes

We begin with the CPG model. Since the log-likelihood function of the CPG model decouples into two separate parts for claim counts and claim sizes, maximum likelihood estimation (MLE) of claim counts and claim size models can be

done independently from each other. We start from n independent random variables $S_i \sim \text{CPG}(\lambda_i w_i, \gamma, c_i)$ with

$$S_i = \sum_{j=1}^{N_i} Z_{i,j}, \qquad \text{for insurance policies } i = 1, \dots, n.$$

The joint log-likelihood function of this model, given observations $(N_i)_i$ and $(Z_{i,j})_{i,j}$ and weights $(w_i)_i$, is given by

$$\begin{aligned} &\ell((\lambda_i)_{i=1,\dots,n}, \gamma, (c_i)_{i=1,\dots,n}) \\ &\quad = \sum_{i=1}^{n} \Bigg(-\lambda_i w_i + N_i \log(\lambda_i w_i) - \log(N_i!) \\ &\quad + \sum_{j=1}^{N_i} \gamma \log(c_i) - \log\Gamma(\gamma) + (\gamma - 1)\log(Z_{i,j}) - c_i Z_{i,j} \Bigg), \end{aligned} \tag{3.1}$$

where the term on the second line is zero for $N_i = 0$. Remark that in this log-likelihood function (for parameter estimation) we treat $(N_i)_i$ and $(Z_{i,j})_{i,j}$ as known observations; for notational convenience we do not use small fonts for observations. From (3.1) we now see that we can estimate the Poisson parameters λ_i and the gamma parameters γ and c_i independently from each other; the former uses observations (N_i) and the latter observations $(N_i)_i$ and $(Z_{i,j})_{i,j}$.

Furthermore, we assume that each insurance policy $i = 1, \dots, n$ is established with covariate information $\boldsymbol{x}_i = (x_{i,0}, \dots, x_{i,d})' \in \mathcal{X} \subset \{1\} \times \mathbb{R}^d$, having initial component $x_{i,0} = 1$ for modeling the intercept component.

GLM for claim counts: Assume that the expected frequencies $\lambda_i = \lambda(\boldsymbol{x}_i)$ of policies $i = 1, \dots, n$ can be modeled by a log-linear regression function

$$\lambda : \mathcal{X} \to \mathbb{R}_+, \qquad \boldsymbol{x} \mapsto \lambda(\boldsymbol{x}) = \exp\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \exp\left\{ \beta_0 + \sum_{k=1}^{d} \beta_k x_k \right\}, \tag{3.2}$$

with regression parameter $\boldsymbol{\beta} = (\beta_0, \dots, \beta_d)' \in \mathbb{R}^{d+1}$. Assuming that the design matrix $\boldsymbol{\mathfrak{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times (d+1)}$ has full rank $d + 1$ we find the unique MLE $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ by the (unique) solution of

$$\boldsymbol{\mathfrak{X}}' \operatorname{diag}(w_1, \dots, w_n) \left(\left(\frac{N_1}{w_1}, \dots, \frac{N_n}{w_n} \right)' - \exp\{\boldsymbol{\mathfrak{X}} \boldsymbol{\beta}\} \right) = \mathbf{0}. \tag{3.3}$$

Remark that the Poisson distribution has an EDF representation with cumulant function $\kappa(\cdot) = \kappa_1(\cdot) = \exp\{\cdot\}$. The lower index $p = 1$ in the cumulant function $\kappa_1(\cdot)$ indicates that we have variance function $V(\lambda) = \lambda$ in the Poisson case, see also (2.5). The choice (3.2) corresponds to the canonical link $(\kappa_1')^{-1}(\cdot) = \log(\cdot)$ in the Poisson GLM. The choice of the canonical link implies that we receive an unbiased portfolio estimate, see [21]. The score Eq. (3.3) is solved numerically, for details see (A.3) in "Appendix A".

GLM for gamma claim sizes: Consider only insurance policies i which have claims, i.e. with $N_i > 0$. All subsequent considerations in this paragraph are conditional on N_i. The average claim amount on policy i has a conditional gamma distribution

$$\bar{Z}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{i,j} \Big|_{\{N_i\}} \sim \mathcal{G}(\gamma N_i, c_i N_i), \tag{3.4}$$

with shape parameter γN_i and scale parameter $c_i N_i$ (note that γ is not policy i dependent). This gamma distributed random variable has conditional mean and variance given by

$$\zeta_i = \mathbb{E}[\bar{Z}_i|N_i] = \frac{\gamma}{c_i} \quad \text{and} \quad \text{Var}(\bar{Z}_i|N_i) = \frac{\gamma}{c_i^2 N_i} = \frac{1}{\gamma N_i}\left(\frac{\gamma}{c_i}\right)^2 = \frac{1}{\gamma N_i}\zeta_i^2.$$

This model belongs to the EDF (2.2) with cumulant function $\kappa_2(\theta) = -\log(-\theta)$ for $\theta \in \boldsymbol{\Theta} = \mathbb{R}_-$, dispersion parameter $\phi = 1/\gamma$ and exposure $w_i = N_i$. The conditional mean and variance are

$$\zeta_i = \mathbb{E}[\bar{Z}_i|N_i] = \kappa_2'(\theta_i) = -\frac{1}{\theta_i} \quad \text{and} \quad \text{Var}(\bar{Z}_i|N_i) = \frac{1}{\gamma N_i}\kappa_2''(\theta_i) = \frac{1}{\gamma N_i}\left(-\frac{1}{\theta_i}\right)^2.$$

This is the boundary case $p = 2$ in Tweedie's CP model with power variance function $V(\zeta) = \zeta^2$, see (2.5).

We set up a second GLM for gamma claim size modeling. This second GLM does not necessarily need to rely on the same covariate space $\mathcal{X}$ as the Poisson GLM (3.2) for claim counts modeling. To emphasize this point, we introduce a new covariate space containing covariate information $\boldsymbol{z}_i = (z_{i,0}, \ldots, z_{i,q})' \in \mathcal{Z} \subset \{1\} \times \mathbb{R}^q$ having initial component $z_{i,0} = 1$ modeling the intercept. We interpret the choices $\mathcal{X}$ and $\mathcal{Z}$ as follows: both covariates $\boldsymbol{x}_i \in \mathcal{X}$ and $\boldsymbol{z}_i \in \mathcal{Z}$ should belong to the same insurance policy i, however, inclusion of individual covariate components and pre-processing of these components may differ in the two different regression models. This is a result of aiming at optimizing the predictive performance of both regression models.

We make the following regression assumption: choose a suitable link function $g_2(\cdot)$ to receive the linear predictor, see also "Appendix A",

$$g_2(\zeta_i) = g_2(\mathbb{E}[\bar{Z}_i|N_i]) = g_2(\kappa_2'(\theta_i)) = g_2(-1/\theta_i) = \eta_i = \langle \boldsymbol{\alpha}, \boldsymbol{z}_i \rangle, \tag{3.5}$$

for regression parameter $\boldsymbol{\alpha} \in \mathbb{R}^{q+1}$. Formula (3.5) explains the relationship between mean $\zeta = \mathbb{E}[\bar{Z}|N] = \kappa_2'(\theta)$, canonical parameter θ and linear predictor $\eta = \eta(\boldsymbol{z})$. Usually, one does not select the canonical link in the gamma GLM because the negativity constraint on the canonical parameter $\theta \in \boldsymbol{\Theta} = \mathbb{R}_-$ may be too restrictive in choosing a linear functional regression form; this is in contrast to the Poisson GLM (3.2). Therefore, the choice of the link function $g_2(\cdot)$ has to be done carefully, because we require $1/\theta_i = -g_2^{-1}(\eta_i) = -g_2^{-1}\langle \boldsymbol{\alpha}, \boldsymbol{z}_i \rangle < 0$ for all policies $i = 1, \ldots, n$, otherwise the canonical parameter θ_i is not in the effective domain $\boldsymbol{\Theta}$. Below, we will choose the log-link for g_2, which is a common choice for gamma GLMs.

These choices imply for the log-likelihood function, only considering policies $i = 1, \ldots, m$ with $N_i > 0$,

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \gamma N_i \big(\bar{Z}_i \theta_i - \kappa_2\big(\theta_i\big)\big) + a(\bar{Z}_i; \gamma N_i). \tag{3.6}$$

The MLE $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is found by solving the score equation, see "Appendix A",

$$\nabla_{\boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}) = \mathbf{0} \;\Leftrightarrow\; \mathfrak{Z}' W_2 \boldsymbol{R} = \mathbf{0}, \tag{3.7}$$

with design matrix $\mathfrak{Z} = (z_1, \ldots, z_m)' \in \mathbb{R}^{m \times (q+1)}$, diagonal working weight matrix (using $V(\zeta_i) = \zeta_i^{-2}$)

$$W_2 = \gamma \operatorname{diag}\left(\left(\frac{\partial g_2(\zeta_i)}{\partial \zeta_i} \right)^{-2} N_i \zeta_i^{-2} \right)_{i=1,\ldots,m}, \tag{3.8}$$

and with working residual vector $\boldsymbol{R} = \left(\frac{\partial g_2(\zeta_i)}{\partial \zeta_i}(\bar{Z}_i - \zeta_i)\right)_{i=1,\ldots,m}$.

Remarks 3.1

- Shape parameter γ may be treated as a hyper-parameter, and the explicit choice of γ does *not* influence parameter estimation because it cancels in the score Eq. (3.7).
- MLE (3.6)–(3.7) is expressed in sufficient statistics $\bar{Z}_i$, and we receive the same regression parameter estimate $\widehat{\boldsymbol{\alpha}}$ if we perform MLE directly on the individual claim sizes $Z_{i,j}$. This is an important property, namely, the gamma GLM can be fit solely on the number of claims N_i and the total claim amount $\bar{Z}_i$ on each policy i. Moreover, this estimated model still allows us to simulate individual claim sizes $Z_{i,j}$. Thus, GLM regression parameter estimation does not differ whether we consider total claim amounts $\bar{Z}_i$ or individual claim sizes $Z_{i,j}$. On the other hand, the process of model and variable selection might give different results in the two estimation cases ($\bar{Z}_i$ vs. $Z_{i,j}$) because the log-likelihood functions and the estimates for γ differ, this is, e.g., important for model selection using likelihood ratio tests or Akaike's information criterion, see Remarks 3.10, below.
- If we model claim counts and claim sizes separately, we use maximal available information N_i and $Z_{i,j}$. Moreover, we can design covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ in an optimal way, and independently from each other.
- If (3.5) is not based on the canonical link of the gamma model, the balance property will not be fulfilled, see [22]. This should be corrected by shifting the intercept parameter β_0 correspondingly. Often one chooses the log-link for $g_2(\cdot)$, under the log-link choice we can also reformulate the regression problem by replacing the average claim amount response (3.4) by the (conditional) total claim amount $S_i|_{\{N_i\}}$ and treating $\log(N_i)$ as a known offset in the linear predictor.
- Shape parameter $\gamma < 1$ leads to an over-dispersed model with strictly decreasing density, and for $\gamma > 1$ the density is uni-modal. Above γ is treated as a hyper-parameter, and below we discuss MLE of γ.

- If shape parameter γ_i needs explicit modeling as a function of i, then (3.7)–(3.8) will no longer have such a simple structure, and MLE of $\boldsymbol{\alpha}$ will depend on the explicit choices of γ_i. In this case, one can either use a gamma DGLM or one can rely on the 2-dimensional exponential family. The latter model is less tractable numerically. It considers cumulant function $\kappa(\theta_1,\theta_2)=\log\Gamma(\theta_2)-\theta_2\log(-\theta_1)$ for scale parameter $c=-\theta_1>0$ and shape parameter $\gamma=\theta_2>0$. This gives inverse link function, see [21],

$$\nabla_{(\theta_1,\theta_2)}\kappa(\theta_1,\theta_2)=\left(\frac{\theta_2}{-\theta_1},\frac{\Gamma'(\theta_2)}{\Gamma(\theta_2)}-\log(-\theta_1)\right)',$$

  the first component being the mean of the gamma distributed random variable Z, and the second component being the mean of $\log(Z)$. We do not further follow up this approach because we would lose the connection to Tweedie's CP approach with a policy independent power variance parameter, see next section.

There remains estimation of shape parameter γ for given MLE $\widehat{\boldsymbol{\alpha}}$. One could either use Pearson's dispersion estimate for $1/\gamma$ or directly calculate the MLE of γ. In view of (3.6), the MLE is obtained from score equation $\frac{\partial}{\partial\gamma}\ell(\widehat{\boldsymbol{\alpha}},\gamma)=0$, which yields

$$\sum_{i=1}^m N_i\Big(\bar{Z}_i\widehat{\theta}_i-\kappa_2(\widehat{\theta}_i)\Big)+N_i\log(\gamma N_i)+N_i+N_i\log(\bar{Z}_i)-N_i\frac{\Gamma'(\gamma N_i)}{\Gamma(\gamma N_i)}=0, \quad (3.9)$$

where we set $\widehat{\theta}_i=-\exp\{-\langle\widehat{\boldsymbol{\alpha}},\boldsymbol{z}_i\rangle\}$. Either we solve this score equation numerically using the Newton-Raphson algorithm, or we plot the one-dimensional log-likelihood function $\gamma\mapsto\ell(\widehat{\boldsymbol{\alpha}},\gamma)$ and determine the MLE $\widehat{\gamma}$ from this plot, see Fig. 1, below, for an example.

We conclude by calculating Fisher's information matrix for $(\boldsymbol{\alpha},\gamma)$ in our gamma GLM. We have, see "Appendix A",

$$-\mathbb{E}\left[\nabla^2_{\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha},\gamma)\Big|N_1,\ldots,N_m\right]=\mathfrak{Z}'W_2\mathfrak{Z}.$$

For the second derivative of the γ term we have

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\gamma^2}\ell(\boldsymbol{\alpha},\gamma)\Big|N_1,\ldots,N_m\right]=-\sum_{i=1}^m\frac{N_i}{\gamma}-N_i^2\psi'(x)\Big|_{x=\gamma N_i},$$

where the second order derivative $\psi'(x)=\frac{d^2}{dx^2}\log\Gamma(x)$ of the log-gamma function is known as the trigamma function, see [10, Sec. 5.15]. The trigamma function is directly available in the statistical software R [13]. For the off-diagonal terms we have

$$-\mathbb{E}\left[\nabla_{\boldsymbol{\alpha}}\frac{\partial}{\partial\gamma}\ell(\boldsymbol{\alpha},\gamma)\Big|N_1,\ldots,N_m\right]=-\sum_{i=1}^m N_i\mathbb{E}\big[\bar{Z}_i-\kappa_2'(\theta_i)\big|N_1,\ldots,N_m\big]\nabla_{\boldsymbol{\alpha}}\theta_i=\mathbf{0}.$$

This gives us the following Fisher's information matrix for the gamma claim size modeling

$$\mathcal{I}(\boldsymbol{\alpha},\gamma)=\begin{pmatrix} \mathfrak{Z}'W_2\mathfrak{Z} & \mathbf{0} \\ \mathbf{0}' & -\sum_{i=1}^m N_i/\gamma - N_i^2\psi'(x)|_{x=\gamma N_i} \end{pmatrix}.$$

3.2 Tweedie's compound Poisson generalized linear model

3.2.1 Homogeneous dispersion case

From Sect. 2.2 we know that Tweedie's CP model belongs to the EDF, thus, GLM is straightforward. In this subsection we start with the homogeneous dispersion parameter $\phi > 0$ case; this case will not be supported in Remarks 3.2, below. We assume having n independent random variables $Y_i \sim \text{Tweedie}(\theta_i, w_i, \phi, p)$, and we choose hyper-parameter $p = (\gamma+2)/(\gamma+1) \in (1,2)$ to make Tweedie's CP model consistent with the CPG case, see Proposition 2.1. Choosing a suitable link function $g_p(\cdot)$ we make the following regression assumption for the linear predictor

$$g_p(\mu_i) = g_p(\mathbb{E}[Y_i]) = g_p(\kappa_p'(\theta_i)) = \eta_i = \left\langle \boldsymbol{\beta}^*, \boldsymbol{x}_i^* \right\rangle, \tag{3.10}$$

where $\boldsymbol{x}_i^* \in \mathcal{X}^* \subset \{1\} \times \mathbb{R}^{d^*}$ are the covariates of policy i and $\boldsymbol{\beta}^*$ is the regression parameter. We change the covariate notation compared to Sect. 3.1 because covariate pre-processing might be done differently for Tweedie's CP model compared to the CPG case (because we consider different responses). In complete analogy with the above, MLE requires solving the score equations

$$\nabla_{\boldsymbol{\beta}^*}\ell(\boldsymbol{\beta}^*) = \mathbf{0} \;\Leftrightarrow\; \mathfrak{X}'W_p\boldsymbol{R} = \mathbf{0}, \tag{3.11}$$

with design matrix $\mathfrak{X} = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)'$, diagonal working weight matrix (using $V(\mu_i) = \mu_i^p$)

$$W_p = \frac{1}{\phi}\,\text{diag}\left(\left(\frac{\partial g_p(\mu_i)}{\partial \mu_i}\right)^{-2} w_i \mu_i^{-p}\right)_{i=1,\ldots,n}, \tag{3.12}$$

and working residual vector $\boldsymbol{R} = \left(\frac{\partial g_p(\mu_i)}{\partial \mu_i}(Y_i - \mu_i)\right)_{i=1,\ldots,n}$.

Remarks 3.2 There are a couple of crucial differences between Tweedie's CP approach with homogeneous dispersion ϕ and the CPG approach of the previous section:

1. The CPG approach of the previous section uses all available information of claim counts N_i and claim sizes $\bar{Z}_i$, whereas Tweedie's CP approach with homogeneous dispersion parameter only uses total claim cost information Y_i.
2. The former approach allows us to consider different covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ for claim counts and claim size modeling, whereas the latter approach only relies on one version of the covariate space $\mathcal{X}^*$.
3. The mean estimates $\widehat{\lambda}_i\widehat{\zeta}_i$ in the CPG case do *not* rely on the particular choice of the shape parameter γ, whereas in the homogeneous dispersion Tweedie's CP

approach the mean estimates $\widehat{\mu}_i$ rely on the specific choice of power variance parameter $p = (\gamma + 2)/(\gamma + 1)$ through the working weight matrix W_p, see (3.12).

4. In general, the dispersions resulting from CPG$(\lambda_i w_i, \gamma, c_i)$ are *not* constant:

$$\begin{aligned}\mathrm{Var}(S_i/w_i) =& w_i^{-2}\mathbb{E}[N_i]\mathbb{E}[Z_{i,1}^2] \;=\; w_i^{-1}\lambda_i\left(\frac{\gamma}{c_i^2}+\frac{\gamma^2}{c_i^2}\right) \;=\; w_i^{-1}\lambda_i\zeta_i\,\frac{1+\gamma}{c_i}\\ =& w_i^{-1}\mu_i^p\,\frac{\mu_i^{1-p}}{c_i(p-1)} = w_i^{-1}\left(\frac{-\theta_i}{c_i}\right)\mu_i^p = \frac{\phi_i}{w_i}\,\mu_i^p.\end{aligned}$$

The dispersion can only be constant if $\phi_i = -\theta_i/c_i$ does not depend on i. Typically, this is not the case, see also Conclusions and Remarks 3.9, below. Therefore, we need to extend the homogeneous dispersion case of Tweedie's CP model to a DGLM Tweedie's CP model, otherwise it cannot be compared to the CPG case, which is more flexible in dispersion modeling. For more analysis of the homogeneous dispersion case see [12].

3.2.2 Heterogeneous dispersion case

As stated in Remarks 3.2, the homogeneous dispersion Tweedie's CP approach does not use full information of claim counts and claim costs and it does not allow for flexible dispersion modeling ϕ_i. In Section 2 of [17], the authors raise the point that in applications of Tweedie's CP model to insurance claim data it is important to use full information so that also the dispersion parameter ϕ_i is modeled flexibly. As a consequence, the dispersion parameter cannot be factored out as in (3.12), and it does not cancel in optimization (3.11). Therefore, [17] propose to use the framework of DGLMs which was introduced and developed by [8, 15, 18]. DGLMs allow for simultaneous modeling of both mean and dispersion parameters by using a second GLM for the dispersion parameter ϕ_i. The two GLMs are jointly calibrated using claim count *and* claim cost information. The joint density of a single case (N, Y) has been derived in formula (11) of [6]:

$$(N, Y) \;\sim\; f(n, y;\theta, w/\phi) = \exp\left\{\frac{y\theta - \kappa_p(\theta)}{\phi/w} + a(n, y;w/\phi)\right\}, \tag{3.13}$$

with $p = (\gamma + 2)/(\gamma + 1)$, $\kappa_p(\cdot)$ given in (2.3), and

$$\exp\{a(n, y;w/\phi)\} = \left(\frac{(w/\phi)^{\gamma+1}y^\gamma}{(p-1)^\gamma(2-p)}\right)^n \frac{1}{n!\Gamma(n\gamma)y}.$$

If we re-parametrize this joint distribution using mean parameter $\mu = \kappa_p'(\theta) = ((1-p)\theta)^{1/(1-p)}$ for total claim costs we arrive at the log-likelihood function

$$\ell(\mu,\phi)=\begin{cases}\frac{w}{\phi}\left(Y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right)+N\log\left(\frac{(w/\phi)^{\gamma+1}Y^{\gamma}}{(p-1)^{\gamma}(2-p)}\right)-\log\left(N!\Gamma(N\gamma)Y\right) & \text{for } N>0,\\ -\frac{w}{\phi}\frac{\mu^{2-p}}{2-p} & \text{for } N=0.\end{cases}$$

In complete analogy with the above we determine the score equations w.r.t. μ and ϕ

$$\frac{\partial}{\partial\mu}\ell(\mu,\phi)=0\Leftrightarrow\frac{w}{\phi}\frac{1}{V(\mu)}(Y-\mu)=0, \tag{3.14}$$

$$\frac{\partial}{\partial\phi}\ell(\mu,\phi)=0\Leftrightarrow-\frac{w}{\phi^2}\left(Y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right)-\frac{1}{\phi}\frac{N}{p-1}=0, \tag{3.15}$$

with variance function $V(\mu)=\mu^p$.

Proposition 3.3 *Fisher's information contribution in the heterogeneous dispersion Tweedie's CP model w.r.t.* (μ,ϕ) *is given by*

$$\mathcal{I}(\mu,\phi)=-\mathbb{E}\left[\nabla^2_{(\mu,\phi)}\ell(\mu,\phi)\right]=\begin{pmatrix}\frac{w}{\phi}\frac{1}{V(\mu)} & 0\\ 0 & \frac{w\mu^{2-p}}{(p-1)(2-p)}\frac{1}{\phi^3}\end{pmatrix}. \tag{3.16}$$

Moreover, we have

$$\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial p}\ell(\mu,\phi)\right]=0.$$

Remark that in the above proposition we talk about Fisher's information *contribution* because the statement considers only one single random variable (Y, N). This is in contrast to (3.27) where we calculate Fisher's information *matrix* over the entire portfolio.

Joint MLE of μ and ϕ requires solving score Eqs. (3.14)-(3.15). This can be done by any suitable root search or gradient descent algorithm. In [17], this root search problem is approached using a slightly different representation, namely, by introducing a dispersion response variable D. This allows for a reformulation of the model in a DGLM form. We revisit [17] after proving Proposition 3.3.

Proof of Proposition 3.3 We start by calculating the means of the terms of the score in (3.15). We have

$$\mathbb{E}\left[Y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right]=\frac{\mu^{2-p}}{1-p}-\frac{\mu^{2-p}}{2-p}=\frac{1}{1-p}\frac{\mu^{2-p}}{2-p}=\frac{1}{1-p}\kappa_p(\theta),$$

and for the second term we receive

$$\mathbb{E}\left[\frac{1}{\phi}\frac{N}{p-1}\right] = \frac{1}{\phi^2}\frac{\phi\lambda w}{p-1} = \frac{1}{\phi^2}\frac{-\theta}{c}\frac{\lambda w}{p-1} = \frac{1}{\phi^2}\frac{\mu^{1-p}}{2-p}\frac{\gamma}{c}\frac{\lambda w}{p-1}$$
$$= \frac{w}{\phi^2}\frac{1}{p-1}\frac{\mu^{2-p}}{2-p} = -\frac{w}{\phi^2}\frac{1}{1-p}\kappa_p(\theta).$$

From these two formulas it follows that, indeed, the score in (3.15) is a residual with mean zero. The cross-covariance terms are easily obtained by noting that also the score in (3.14) is a zero mean residual. This implies

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial\phi}\ell(\mu,\phi)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial p}\ell(\mu,\phi)\right] = 0. \tag{3.17}$$

There remain the diagonal terms. For the first one we have, using integration by parts,

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ell(\mu,\phi)\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\ell(\mu,\phi)\right)^2\right] = \frac{w^2}{\phi^2}\frac{1}{V(\mu)^2}\mathrm{Var}(Y) = \frac{w}{\phi}\frac{1}{V(\mu)}.$$

For the second diagonal term we have, this provides the variance of the zero mean score in (3.14),

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\phi^2}\ell(\mu,\phi)\right] = -\mathbb{E}\left[2\frac{w}{\phi^3}\left(Y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right) + \frac{N(\gamma+1)}{\phi^2}\right] = \frac{w}{\phi^3}\frac{\mu^{2-p}}{(p-1)(2-p)}.$$

This finishes the proof of Proposition 3.3. □

Thus, for MLE of μ and ϕ we need to consider the scores in (3.14)–(3.15), the latter one defining (unscaled) residuals w.r.t. the dispersion given by

$$\mathcal{E}_d = \frac{\partial}{\partial\phi}\ell(\mu,\phi) = \frac{1}{\phi^2}\left[-w\left(Y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right) - \phi\frac{N}{p-1}\right].$$

As mentioned above, solving score Eqs. (3.14)–(3.15) produce the MLEs for μ and ϕ; basically, this finishes the MLE problem. In the remainder of this section, following [17], we rewrite this MLE problem. This different representation introduces a new (dispersion) response variable D, such that the root search problem can directly be related to Fisher's scoring method in a DGLM form. Choose square variance function $V_d(\phi) = \phi^2$ and dispersion-prior weights

$$v = \frac{2w}{\phi}\frac{\mu^{2-p}}{(p-1)(2-p)} > 0. \tag{3.18}$$

This allows us to define so-called dispersion responses

$$D = \frac{2}{v}\left(-w\left(Y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right) - \phi\frac{N}{p-1}\right) + \phi = \frac{2}{v}V_d(\phi)\mathcal{E}_d + \phi, \tag{3.19}$$

having $\mathbb{E}[D] = \phi$, $\text{Var}(D) = \frac{2}{v} V_d(\phi)$ and scores w.r.t. ϕ

$$\frac{\partial}{\partial \phi} \ell(\mu, \phi) = \frac{v}{2} \frac{1}{V_d(\phi)} (D - \phi). \tag{3.20}$$

Fisher's information contribution (3.16) then reads as

$$\mathcal{I}(\mu, \phi) = -\mathbb{E}\left[\nabla^2_{(\mu,\phi)} \ell(\mu, \phi)\right] = \begin{pmatrix} \frac{w}{\phi} \frac{1}{V(\mu)} & 0 \\ 0 & \frac{v}{2} \frac{1}{V_d(\phi)} \end{pmatrix}.$$

As emphasized by [16], orthogonality of μ and (ϕ, p), see (3.17), typically leads to fast convergence in estimation algorithms.

Remarks 3.4

- We start from the joint distribution of (N, Y), given in (3.13), for estimating (μ, ϕ). This estimation problem is modified by considering a new response vector (Y, D), instead. The new dispersion response D, defined in (3.19), is not gamma distributed, but in view of score (3.20) we bring it into a gamma EDF structure with weight $v > 0$, dispersion parameter 2 and square variance function $V_d(\phi) = \phi^2$, see also (2.5). In [17] it is mentioned that these definitions of v and D are somewhat artificial, but they bring this estimation problem into a DGLM form; note that this requires to include one dispersion term ϕ into the weight v and the response D, this means that we have an approximate score equation equivalence with a gamma MLE problem. In view of Proposition 3.3, we could also define dispersion response D differently by choosing an inverse Gaussian power variance function, i.e. $V_d(\phi) = \phi^3$, and defining the dispersion-prior weight correspondingly. This provides the same numerical solution for MLE, using an approximate score equation equivalence with an inverse Gaussian MLE problem. However, in this latter version the weights do not provide the right scaling for a distribution within the EDF.
- Alternatively, we could try to estimate dispersion ϕ using Tweedie's deviance residuals

$$\mathcal{E} = \text{sgn}(Y - \mu) \sqrt{2w \left(Y \frac{Y^{1-p} - \mu^{1-p}}{1-p} - \frac{Y^{2-p} - \mu^{2-p}}{2-p} \right)}.$$

  Following [17], the squared residuals $\mathcal{E}^2$ are approximately $\phi \chi^2_1$ distributed for ϕ sufficiently small, thus, they can be approximated by a gamma distribution with mean ϕ and variance $2\phi^2$. Section 3.1 of [17] discusses this estimation approach. We do not further follow these lines because this approach does not use any claim count information and, therefore, does not benefit from full information (N, Y) as the CPG case.
- There is a third alternative of including a dispersion estimation, and this third one is the one implemented in the R package `dglm`. This requires that the dispersion parameter is made policy dependent and then a DGLM is explored on

$(Y, \mathcal{E})$ by alternating the corresponding score updates. Also this approach does not benefit from full information (N, Y) (in contrast to the CPG model), and it is therefore not further explored in this manuscript.

3.2.3 Double generalized linear model in the heterogeneous Tweedie case

We use the heterogeneous dispersion Tweedie's CP approach and bring it into a DGLM form as described in the previous section. Choosing a suitable link function $g_p(\cdot)$ we make the following regression assumption for the linear predictor of the mean

$$g_p(\mu_i) = g_p(\mathbb{E}[Y_i]) = g_p(\kappa_p'(\theta_i)) = \eta_i = \left\langle \boldsymbol{\beta}^*, \boldsymbol{x}_i^* \right\rangle, \tag{3.21}$$

upper indices * distinguishing the parametrization in Tweedie's CP GLM case from the individual models in Sect. 3.1. For the modeling of the dispersion parameter we choose a second link function $g_d(\cdot)$ such that we have the linear predictor

$$g_d(\phi_i) = \left\langle \boldsymbol{\alpha}^*, z_i^* \right\rangle, \tag{3.22}$$

where the covariates $z_i^* \in \mathcal{Z}^* \subset \{1\} \times \mathbb{R}^{q^*}$ are potentially differently pre-processed than the ones $\boldsymbol{x}_i^* \in \mathcal{X}^* \subset \{1\} \times \mathbb{R}^{d^*}$, but still belong to the same policy i. MLE of $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)$ requires solving the score equations, see (3.14) and (3.20),

$$\nabla_{\beta^*} \ell(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \mathbf{0} \Leftrightarrow \sum_{i=1}^n \frac{w_i}{\phi_i} \frac{1}{V(\mu_i)} \big(Y_i - \mu_i\big) \nabla_{\beta^*} \mu_i = \mathfrak{X}' W_p \boldsymbol{R} = \mathbf{0}, \tag{3.23}$$

$$\nabla_{\alpha^*} \ell(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \mathbf{0} \Leftrightarrow \sum_{i=1}^n \frac{v_i}{2} \frac{1}{V_d(\phi_i)} \big(D_i - \phi_i\big) \nabla_{\alpha^*} \phi_i = \mathfrak{Z}' W_d \boldsymbol{R}_d = \mathbf{0}, \tag{3.24}$$

with design matrices $\mathfrak{X} = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)'$ and $\mathfrak{Z} = (z_1^*, \ldots, z_n^*)'$, working weight matrices

$$W_p = \text{diag}\left(\left(\frac{\partial g_p(\mu_i)}{\partial \mu_i} \right)^{-2} \frac{w_i}{\phi_i} \frac{1}{V(\mu_i)} \right)_{i=1,\ldots,n},$$

$$W_d = \text{diag}\left(\left(\frac{\partial g_d(\phi_i)}{\partial \phi_i} \right)^{-2} \frac{v_i}{2} \frac{1}{V_d(\phi_i)} \right)_{i=1,\ldots,n},$$

and working residual vectors $\boldsymbol{R} = (\frac{\partial g_p(\mu_i)}{\partial \mu_i}(Y_i - \mu_i))_{i=1,\ldots,n}$ and $\boldsymbol{R}_d = (\frac{\partial g_d(\phi_i)}{\partial \phi_i}(D_i - \phi_i))_{i=1,\ldots,n}$. For the definition of the dispersion-prior weights $v_i = v_i(\phi_i)$ and the dispersion responses D_i we refer to (3.18)–(3.19). Using Fisher's scoring method for estimating $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$, see "Appendix A", we explore the scoring updates

$$\boldsymbol{\beta}_t^* \mapsto \boldsymbol{\beta}_{t+1}^* = \left(\mathfrak{X}' W_p \mathfrak{X}\right)^{-1} \mathfrak{X}' W_p \left(\boldsymbol{R} + g_p(\boldsymbol{\mu})\right), \tag{3.25}$$

$$\boldsymbol{\alpha}_t^* \mapsto \boldsymbol{\alpha}_{t+1}^* = \left(\mathfrak{Z}' W_d \mathfrak{Z}\right)^{-1} \mathfrak{Z}' W_d \left(\boldsymbol{R}_d + g_d(\boldsymbol{\phi})\right), \tag{3.26}$$

where all terms on the right-hand side are evaluated at algorithmic time t, that is, $W_p = W_p(\boldsymbol{\beta}_t^*, \boldsymbol{\alpha}_t^*)$, $W_d = W_d(\boldsymbol{\alpha}_t^*)$, $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\beta}_t^*)$, $\boldsymbol{R}_d = \boldsymbol{R}_d(\boldsymbol{\beta}_t^*, \boldsymbol{\alpha}_t^*)$, $g_p(\boldsymbol{\mu}) = g_p(\boldsymbol{\mu}(\boldsymbol{\alpha}_t^*))$ and $g_d(\boldsymbol{\phi}) = g_d(\boldsymbol{\phi}(\boldsymbol{\alpha}_t^*))$. This also indicates how the two sets of parameters interact. Since parameters $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ are orthogonal, alternating the updates leads to fast convergence. Standard errors are obtained from the inverse of Fisher's information matrix

$$\mathcal{I}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \begin{pmatrix} \mathfrak{X}' W_p \mathfrak{X} & 0 \\ 0 & \mathfrak{Z}' W_d \mathfrak{Z} \end{pmatrix}. \tag{3.27}$$

There remains estimation of p. This is usually done by considering the profile log-likelihood for p, given optimal estimates of $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)$, that is, we study $p \mapsto \ell(\widehat{\boldsymbol{\beta}}^*(p), \widehat{\boldsymbol{\alpha}}^*(p), p)$ where, in general, the MLEs $\widehat{\boldsymbol{\beta}}^*(p)$ and $\widehat{\boldsymbol{\alpha}}^*(p)$ depend on the explicit choice of the power variance parameter p; for an example of a profile log-likelihood we refer to Fig. 1, below.

Remarks 3.5

- We emphasize that covariates may be chosen and pre-processed differently in the CPG and in Tweedie's CP models; this is indicated by choosing different notation for the covariate spaces $(\mathcal{X}, \mathcal{Z})$ and $(\mathcal{X}^*, \mathcal{Z}^*)$, respectively. Different pre-processing of covariates might be necessary because we aim at optimally modeling different responses in the two models. This optimal modeling also includes good choices of link functions which may even imply that a CPG GLM does not lead to a Tweedie CP DGLM counterpart (or vice versa) because the linear predictor structure does not necessarily carry through general choices of link functions. In Sect. 3.3 we fully rely on log-links which allow for a one-to-one identification scheme between the different GLM frameworks.
- The calculation of the terms of Fisher's information matrix involving p are a bit cumbersome, for this reason we do not give them explicitly.
- As usual in MLE, typically, the dispersion parameters ϕ_i will be under-estimated because MLE is not unbiased for variance parameter estimation, we refer to [17], Sects. 3.2 and 4.3. Using both total claim costs Y and claim counts N, the bias is often small, see [17].

We close this subsection by considering the special case of log-links for g_p and g_d. This special choice provides working weight matrices W_p and W_d

$$W_p = \operatorname{diag}\left(\frac{w_i}{\phi_i}\mu_i^{2-p}\right)_{i=1,\dots,n} = (p-1)(2-p)\operatorname{diag}\left(\frac{v_i}{2}\right)_{i=1,\dots,n} = (p-1)(2-p)W_d,$$

and working residual vectors $\boldsymbol{R} = ((Y_i/\mu_i - 1))_{i=1,\ldots,n}$ and $\boldsymbol{R}_d = ((D_i/\phi_i - 1))_{i=1,\ldots,n}$. This provides us with score equations

$$\nabla_{\boldsymbol{\beta}^*} \ell(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \mathbf{0} \Leftrightarrow \mathfrak{X}' W_p \boldsymbol{R} = \mathbf{0}, \tag{3.28}$$

$$\nabla_{\boldsymbol{\alpha}^*} \ell(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \mathbf{0} \Leftrightarrow \mathfrak{Z}' W_p \boldsymbol{R}_d = \mathbf{0}, \tag{3.29}$$

thus, in both cases we can use the same working weight matrix W_p.

Theorem 3.6 *Assume Tweedie's CP DGLM holds with covariate spaces $\mathcal{X}^* = \mathcal{Z}^*$ and covariate choices $\boldsymbol{x}_i^* = \boldsymbol{z}_i^*$ for all insurance policies $i = 1, \ldots, n$. Moreover, assume that for both GLMs we choose log-links for g_p and g_d. The MLE $\widehat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$ does not depend on the explicit choice of the power variance parameter $p \in (1, 2)$, and also the corresponding mean estimates $\widehat{\mu}_i = \exp\langle \widehat{\boldsymbol{\beta}}^*, \boldsymbol{x}_i^* \rangle$ are p-independent. Assume that $\widehat{\mu}_i$ and $\widehat{\phi}_i(p)$ solve the score Eqs. (3.28)–(3.29) for power variance parameter $p \in (1, 2)$. The dispersion parameter estimates scale as a function of power variance parameters $q \in (1, 2)$ as*

$$\widehat{\phi}_i(q) \;=\; \frac{2-p}{2-q}\, \widehat{\phi}_i(p)\, \widehat{\mu}_i^{p-q} \qquad \text{for all insurance policies } i.$$

Remarks 3.7

- Theorem 3.6 is a very useful and strong result. In general, we have to run Fisher's scoring method for every power variance parameter $p \in (1, 2)$ to find optimal MLEs $\widehat{\boldsymbol{\beta}}^*(p)$ and $\widehat{\boldsymbol{\alpha}}^*(p)$. In a second step, the optimal power variance parameter is found by considering the profile log-likelihood in p. Under the assumptions of Theorem 3.6 we only need to run Fisher's scoring method once to receive MLEs $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\alpha}}^*(p)$ for a fixed power variance parameter p. All dispersion estimates for different power variance parameters are then directly obtained from Theorem 3.6, and mean parameter estimates do not vary in p. That is, we can directly maximize function $q \mapsto \ell(\widehat{\mu}_i, \widehat{\phi}_i(q), q)$ where the dispersion $\widehat{\phi}_i(q)$ scales in q according to Theorem 3.6.
- Theorem 3.6 also highlights that the heterogeneous dispersion case is fundamentally different from the homogeneous one. The mean estimates in the homogeneous case depend on the choice of the power variance parameter p through the working weight matrix W_p in (3.12). In contrast to the heterogeneous dispersion case, a constant dispersion parameter does not leave any room to balance different p's through portfolio varying dispersions. On the other hand, under the assumptions of Theorem 3.6, the mean estimates are not p sensitive, which is equivalent to the CPG case.

Proof of Theorem 3.6 The score equations for $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ are under log-link choices provide, see (3.14)–(3.15),

$$\nabla_{\boldsymbol{\beta}^*}\ell(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)=\sum_{i=1}^n \frac{w_i}{\phi_i}\mu_i^{1-p}\big(Y_i-\mu_i\big)\boldsymbol{x}_i^* \;=\; \mathbf{0},$$

$$\nabla_{\boldsymbol{\alpha}^*}\ell(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)=-\sum_{i=1}^n \frac{w_i}{\phi_i}\left(Y_i\frac{\mu_i^{1-p}}{1-p}-\frac{\mu_i^{2-p}}{2-p}-\frac{\phi_i}{1-p}\frac{N_i}{w_i}\right)\boldsymbol{x}_i^* \;=\; \mathbf{0}.$$

Assume that $\mu_i=\mu_i(p)$ and $\phi_i=\phi_i(p)$ solve the above score equations for given power variance parameter $p\in(1,2)$. Next, we choose power variance parameter $q\neq p$, and define $\widetilde{\phi}_i=k\phi_i\mu_i^{p-q}$ for some $k>0$. We plug μ_i and $\widetilde{\phi}_i$ into the first score equation for power variance parameter q

$$\sum_{i=1}^n \frac{w_i}{\widetilde{\phi}_i}\mu_i^{1-q}\big(Y_i-\mu_i\big)\boldsymbol{x}_i^*=\frac{1}{k}\sum_{i=1}^n \frac{w_i}{\phi_i}\mu_i^{1-p}\big(Y_i-\mu_i\big)\boldsymbol{x}_i^*=\mathbf{0},$$

thus, the pairs $(\mu_i,\widetilde{\phi}_i)$ fulfill the first score equation. We now need to massage these pairs through the second score equation for power variance parameter q

$$\begin{aligned}
&-\sum_{i=1}^n \frac{w_i}{\widetilde{\phi}_i}\left(Y_i\frac{\mu_i^{1-q}}{1-q}-\frac{\mu_i^{2-q}}{2-q}-\frac{\widetilde{\phi}_i}{1-q}\frac{N_i}{w_i}\right)\boldsymbol{x}_i^*\\
&\quad=\frac{-1}{k}\sum_{i=1}^n \frac{w_i}{\phi_i}\left(Y_i\frac{\mu_i^{1-p}}{1-q}-\frac{\mu_i^{2-p}}{2-q}-\frac{k\phi_i}{1-q}\frac{N_i}{w_i}\right)\boldsymbol{x}_i^*\\
&\quad=\frac{p-1}{1-q}\sum_{i=1}^n \frac{w_i}{\phi_i}\left(Y_i\frac{\mu_i^{1-p}}{1-p}\frac{1}{k}-\frac{\mu_i^{2-p}}{2-p}\frac{2-p}{2-q}\frac{1-q}{k(1-p)}-\frac{\phi_i}{1-p}\frac{N_i}{w_i}\right)\boldsymbol{x}_i^*.
\end{aligned}$$

Next we apply that the pairs (μ_i,ϕ_i) solve the score equations for p. This provides for the score function of $\boldsymbol{\alpha}^*$

$$\begin{aligned}
&-\sum_{i=1}^n \frac{w_i}{\widetilde{\phi}_i}\left(Y_i\frac{\mu_i^{1-q}}{1-q}-\frac{\mu_i^{2-q}}{2-q}-\frac{\widetilde{\phi}_i}{1-q}\frac{N_i}{w_i}\right)\boldsymbol{x}_i^*\\
&\quad=\frac{p-1}{1-q}\sum_{i=1}^n \frac{w_i}{\phi_i}\left(Y_i\frac{\mu_i^{1-p}}{1-p}\left(\frac{1}{k}-1\right)-\frac{\mu_i^{2-p}}{2-p}\left(\frac{2-p}{2-q}\frac{1-q}{k(1-p)}-1\right)\right)\boldsymbol{x}_i^*.\\
&\quad=\frac{1-k}{k}\frac{1}{q-1}\sum_{i=1}^n \frac{w_i}{\phi_i}\mu_i^{1-p}\left(Y_i-\mu_i\frac{(2-p)(1-q)-k(1-p)(2-q)}{(2-q)(2-p)(1-k)}\right)\boldsymbol{x}_i^*.
\end{aligned}\tag{3.30}$$

Now we still have one parameter $k>0$ that we can choose. We require

$$\frac{(2-p)(1-q)-k(1-p)(2-q)}{(2-q)(2-p)(1-k)}=1 \quad\Leftrightarrow\quad k=\frac{2-p}{2-q}.$$

This choice implies that (3.30) is equal to zero which follows from the fact that the pairs (μ_i,ϕ_i) solve the score equations for $\boldsymbol{\beta}^*=\boldsymbol{\beta}^*(p)$. This finishes the proof. In Remarks 3.10 we give a shorter proof. □

3.3 Relation between the two GLM approaches

We compare the CPG model to its counterpart being parametrized through Tweedie's CP model. To start off, recall formulas (2.6)–(2.9). The first formula gives relationship $p = (\gamma + 2)/(\gamma + 1) \in (1, 2)$. Since these two parameters are not modeled insurance policy dependent, we directly identify them. We start with the gamma claim size GLM of Sect. 3.1 using identification (2.7). The means are given by, see (3.5),

$$g_2^{-1}\langle \boldsymbol{\alpha}, z\rangle \;=\; \zeta \;=\; \frac{\gamma}{c} \overset{(2.7)}{=} \frac{\gamma\phi}{-\theta} \;=\; \frac{(2-p)\phi}{\mu^{1-p}} \;=\; (2-p)\frac{g_d^{-1}\langle \boldsymbol{\alpha}^*, z^*\rangle}{\left(g_p^{-1}\langle \boldsymbol{\beta}^*, \boldsymbol{x}^*\rangle\right)^{1-p}}, \tag{3.31}$$

where we have used canonical link $\theta = (\kappa_p')^{-1}(\mu) = -\mu^{1-p}/(p-1)$. From identification (2.9) we have

$$\exp\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle \;=\; \lambda \overset{(2.9)}{=} \frac{1}{\phi}\kappa_p(\theta) \;=\; (2-p)^{-1}\frac{\left(g_p^{-1}\langle \boldsymbol{\beta}^*, \boldsymbol{x}^*\rangle\right)^{2-p}}{g_d^{-1}\langle \boldsymbol{\alpha}^*, z^*\rangle}. \tag{3.32}$$

From identities (3.31)–(3.32) we conclude that for general link functions it is nontrivial to derive one parametrization from the other, i.e. this requires quite some feature engineering to bring the models in line (if possible at all). If we choose log-links for g_2, g_p and g_d (these are not the canonical links in all three cases but they are convenient because they preserve the right sign convention on the canonical scale) we can directly compare the linear predictors

$$\begin{aligned}\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle &= -\log(2-p) - \langle \boldsymbol{\alpha}^*, z^*\rangle + (2-p)\langle \boldsymbol{\beta}^*, \boldsymbol{x}^*\rangle,\\ \langle \boldsymbol{\alpha}, z\rangle &= \log(2-p) + \langle \boldsymbol{\alpha}^*, z^*\rangle - (1-p)\langle \boldsymbol{\beta}^*, \boldsymbol{x}^*\rangle.\end{aligned}$$

Formulating this differently gives us the following theorem.

Theorem 3.8 *Assume all link functions in* (3.2), (3.5), (3.21) *and* (3.22) *are chosen to be the log-links. The CPG GLM having constant shape parameter $\gamma > 0$ and Tweedie's CP DGLM with variance parameter $p = (\gamma + 2)/(\gamma + 1) \in (1, 2)$ can be identified by (i.e. the resulting two models are equal under) the following equations for the linear predictors*

$$\begin{aligned}\langle \boldsymbol{\beta}^*, \boldsymbol{x}^*\rangle &= \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle + \langle \boldsymbol{\alpha}, z\rangle,\\ \langle \boldsymbol{\alpha}^*, z^*\rangle &= -\log(2-p) - (p-1)\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle + (2-p)\langle \boldsymbol{\alpha}, z\rangle.\end{aligned}$$

Conclusions and Remarks 3.9

- If we have found a good parametrization for the Poisson claim counts GLM and the gamma claim size GLM involving covariates $\boldsymbol{x} \in \mathcal{X}$ and $z \in \mathcal{Z}$, then Tweedie's CP model should include all components present in $\boldsymbol{x} \cup z$, and $\boldsymbol{x}^*$ and z^* should only differ if some components of $\boldsymbol{x} \cup z$ cancel out by a particular choice

of regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The same holds true if we exchange the roles of the two models.

- From the second identity of Theorem 3.8 we see that dispersion ϕ_i is constant over all policies i if and only if

$$(p-1)\left\langle \boldsymbol{\beta}^{(-0)}, \boldsymbol{x}_i^{(-0)} \right\rangle = (2-p)\left\langle \boldsymbol{\alpha}^{(-0)}, \boldsymbol{z}_i^{(-0)} \right\rangle \qquad \text{for all i,} \tag{3.33}$$

  the upper indices $^{(-0)}$ indicate that we exclude the intercept components $x_{i,0} = z_{i,0} = 1$ in these scalar products. Identity (3.33) gives the condition under which the assumptions of Sect. 3.2.1 are justified. However, in many practical insurance pricing examples, we find that the covariate space $\mathcal{Z}$ for claim sizes is strictly smaller than $\mathcal{X}$ used for claim counts modeling because certain factors only influence claim frequencies but are not significant for claim severities. In addition, often there are covariates that have opposite signs for claim counts and claim sizes. In all these cases (3.33) is not satisfied, and working under a constant dispersion assumption cannot be justified.
- We believe that covariate pre-processing is more easily done within the CPG model. The reason being, as stated above, that claim counts and claim sizes often behave differently w.r.t. covariate information. Covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ allow us to explore such differences individually. In Tweedie's CP model everything is merged together which makes it more difficult to choose good covariates and to separate the different systematic effects.
- Tweedie's CP model calibrated with MLE will typically differ from the corresponding CPG model if we follow Theorem 3.8. The CPG model involves $|\mathcal{X}| + |\mathcal{Z}| = d + q + 2$ parameters. This typically results in a Tweedie CP model with $|\mathcal{X}^*| + |\mathcal{Z}^*| = 2|\mathcal{X}^*|$ parameters, which is bigger than $d + q + 2$ if $\mathcal{X} \neq \mathcal{Z}$. Thus, in Tweedie's CP model there are more parameters to be estimated if we follow the above guidance.

We close this section by giving the log-likelihoods of Tweedie's CP DGLM and of the CPG GLM under log-link choices. The one of Tweedie's CP DGLM is given by

$$\begin{aligned}\ell_{\mathrm{Tw}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = \sum_{i=1}^{n} w_i \left(Y_i \frac{e^{(1-p)\langle \boldsymbol{\beta}^*, \boldsymbol{x}_i^* \rangle - \langle \boldsymbol{\alpha}^*, \boldsymbol{z}_i^* \rangle}}{1-p} - \frac{e^{(2-p)\langle \boldsymbol{\beta}^*, \boldsymbol{x}_i^* \rangle - \langle \boldsymbol{\alpha}^*, \boldsymbol{z}_i^* \rangle}}{2-p} \right) - \frac{N_i}{p-1} \left\langle \boldsymbol{\alpha}^*, \boldsymbol{z}_i^* \right\rangle \\ + N_i \log \left(\frac{w_i^{\gamma+1} Y_i^{\gamma}}{(p-1)^{\gamma}(2-p)} \right) - \log \left(N_i! \Gamma(N_i \gamma) Y_i \right).\end{aligned} \tag{3.34}$$

To make the log-likelihood of the CPG GLM directly comparable to (3.34), we make a change of variables $(N_i, \bar{Z}_i) \mapsto (N_i, Y_i)$ by setting $Y_i = N_i \bar{Z}_i / w_i$. This gives us log-likelihood

$$\ell_{\text{CPG}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n} w_i \left(Y_i \frac{e^{\log(2-p)-\langle \boldsymbol{\alpha}, z_i\rangle}}{1-p} - \frac{e^{\log(2-p)+\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle}}{2-p} \right) - \frac{N_i}{p-1}(2-p)\langle \boldsymbol{\alpha}, z_i\rangle + \frac{N_i}{p-1}(p-1)\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle + N_i \log\left(w_i^{\gamma+1}\gamma^{\gamma} Y_i^{\gamma}\right) - \log\left(N_i!\Gamma(\gamma N_i) Y_i\right). \tag{3.35}$$

Assuming covariate relationship $\boldsymbol{x}_i^* = z_i^*$ we can re-parametrize the first log-likelihood (3.34) by setting $\boldsymbol{\beta}^+ = (2-p)\boldsymbol{\beta}^* - \boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^+ = -(1-p)\boldsymbol{\beta}^* + \boldsymbol{\alpha}^*$, this gives us (we drop irrelevant terms)

$$\ell_{\text{Tw}}(\boldsymbol{\beta}^+,\boldsymbol{\alpha}^+) \propto \sum_{i=1}^{n} w_i \left(Y_i \frac{e^{-\langle \boldsymbol{\alpha}^+, \boldsymbol{x}_i^*\rangle}}{1-p} - \frac{e^{\langle \boldsymbol{\beta}^+, \boldsymbol{x}_i^*\rangle}}{2-p} \right) - \frac{N_i}{p-1}\left\langle \left[(2-p)\boldsymbol{\alpha}^+ - (p-1)\boldsymbol{\beta}^+\right], \boldsymbol{x}_i^*\right\rangle.$$

This proves under $\boldsymbol{x}_i^* = z_i^* = \boldsymbol{x}_i \cup z_i$ that the CPG model is nested in Tweedie's CP model and we have for given $p = (\gamma+2)/(\gamma+1)$

$$\max_{(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)} \ell_{\text{Tw}}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*) \;\geq\; \max_{(\boldsymbol{\beta},\boldsymbol{\alpha})} \ell_{\text{CPG}}(\boldsymbol{\beta},\boldsymbol{\alpha}), \tag{3.36}$$

this explicitly uses that we have the same data representation $(N_i, Y_i)_i$ in both log-likelihoods.

Remarks 3.10

- Under the assumptions of Theorem 3.8 and additionally assuming that $\boldsymbol{x}_i = z_i = \boldsymbol{x}_i^* = z_i^*$, we receive an identity in (3.36). Since the mean estimates in the CPG case do not depend on the particular choice of the shape parameter γ, the same must hold true for Tweedie's CP DGLM model under identical covariates $\boldsymbol{x}_i = z_i = \boldsymbol{x}_i^* = z_i^*$. Using Proposition 2.1 we then receive the dispersion scaling of Theorem 3.6, thus, this gives us a second shorter proof for Theorem 3.6.
- If $\boldsymbol{x}_i^* = z_i^* = \boldsymbol{x}_i \cup z_i$ and $\boldsymbol{x}_i \neq z_i$, the CPG model is strictly nested in Tweedie's CP model and, in general, we do not get an identity in (3.36). In that case, Theorem 3.8 reflects an ideal world because noise in the data prevents MLE estimated parameters (estimated separately in both models) from strictly satisfying the identities in Theorem 3.8.
- To perform model selection in the general case we can use Akaike's information criterion (AIC) [1]. This corrects both sides of (3.36) by the number of regression parameters involved, thus, with AIC the model with the smaller value should be preferred from either

$$\begin{aligned} &-2\max_{(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)} \ell_{\text{Tw}}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*) + 2(d^*+q^*+2) \quad \text{or} \\ &-2\max_{(\boldsymbol{\beta},\boldsymbol{\alpha})} \ell_{\text{CPG}}(\boldsymbol{\beta},\boldsymbol{\alpha}) + 2(d+q+2). \end{aligned} \tag{3.37}$$

AIC applies because in both models we use the same data representation $(N_i, Y_i)_i$ and both models are evaluated in the MLEs for the corresponding parameters. We emphasize that for estimating the CPG GLM we use in (3.36) sufficient

statistics $Y_i = N_i \bar{Z}_i / w_i$. If, instead, we use the individual claim sizes $Z_{i,j}$ to estimate the CPG GLM, AIC does not apply because the log-likelihoods to be compared use the available data in different ways.

4 Numerical examples

We study two numerical examples to benchmark the two modeling approaches of Theorem 3.8. First, we design a synthetic data example that fully meets the assumptions of Theorem 3.8. Thus, there is no model uncertainty involved in this first (synthetic) example about underlying distributions, covariate spaces and link functions, and we can fully focus on estimating parameters with MLE in the CPG GLM and in Tweedie's CP DGLM. These results are then compared to neural network regression approaches on the same synthetic data. In contrast to GLMs, neural networks explore optimal covariate selection themselves. This is done in Sect. 4.2.2. Our second example in Sect. 4.3 is a real data example. This additionally raises the issue of model uncertainty because the real data has not been generated by a CPG model. Both examples are based on the motorcycle insurance data `swmotorcycle` used in [11], this data is available through the R package `CASdatasets` [3], see Listing 1 for an excerpt of the data. For the synthetic data we sample a portfolio of covariates from the original data, and then generate claims with a CPG GLM designed according to the assumptions of Theorem 3.8. For the real data example we fully rely on the `swmotorcycle` data and we use the corresponding claim observations.

4.1 Description of motorcyle data

We briefly describe the data, for more information we refer to "Appendix B", below. The data comprises comprehensive insurance for motorcycles which covers loss or damage of motorcycles other than collision, for instance, caused by theft, fire or vandalism. The data is aggregated on insurance policy level for years 1994–1998. The data is shown in Listing 1. We have applied some pre-processing, e.g., we have dropped all policies that have an exposure equal to zero.

Listing 1: Swedish motorcycle data `swmotorcycle` of [11] from the R package `CASdatasets` [3].

```
1  'data.frame':   62036 obs. of  9 variables:
2   $ Age       : num  36 52 25 50 45 24 52 47 30 32 ...
3   $ Gender    : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 1 1 1 ...
4   $ Zone      : Factor w/ 7 levels "Zone 1","Zone 2",..: 4 4 4 3 3 1 4 4 4 4 ...
5   $ McClass   : int  1 4 7 4 6 3 4 4 3 3 ...
6   $ McAge     : num  12 19 9 14 11 2 16 17 16 16 ...
7   $ Bonus     : int  6 4 3 1 7 6 2 7 1 4 ...
8   $ Exposure  : num  0.0274 0.4986 0.3863 1.9507 1.5014 ...
9   $ ClaimNb   : int  0 0 0 0 0 0 0 0 0 0 ...
10  $ ClaimCosts: int  0 0 0 0 0 0 0 0 0 0 ...
```

We briefly describe the variables, the following enumeration refers to lines 2–10 of Listing 1:

2. `Age`: age of motorcycle owner in $\{18, \dots, 70\}$ years (we cap at 70 because of scarcity above);
3. `Gender`: gender of motorcycle owner either being `Female` or `Male`;
4. `Zone`: seven geographical Swedish zones being (1) central parts of Sweden's three largest cities, (2) suburbs and middle-sized towns, (3) lesser towns except those in zones (5)–(7), (4) small towns and countryside except those in zones (5)–(7), (5) Northern towns, (6) Northern countryside, and (7) Gotland (Sweden's largest island);
5. `McClass`: seven ordered motorcycle classes received from the so-called EV ratio defined as (Engine power in kW $\times$ 100)/(Vehicle weight in kg + 75 kg);
6. `McAge`: age of motorcycle in $\{0, \dots, 30\}$ years (we cap at 30 because of scarcity beyond);
7. `Bonus`: ordered bonus-malus class from 1 to 7, entry level is 1;
8. `Exposure`: total exposure in yearly units in interval [0.0274, 31.3397], the shortest entry referring to 1 day and the longest one to more than 31 years;[1]
9. `ClaimNb`: number of claims N on the policy;
10. `ClaimCosts`: total claim costs $S = \sum_{j=1}^{N} Z_j$ on the policy (thus, we do not have information about individual claims Z_j but only about sufficient statistics $\bar{Z}$ on each policy).

The data is illustrated in "Appendix B".

4.2 Synthetic data example

This section is based on synthetic (simulated) data from a CPG GLM.

4.2.1 A generalized linear model approach

We start by describing the simulation of the synthetic data. We randomly choose $n = 250'000$ insurance policies from `dat=swmotorcycle` using the R code:

```
portfolio <- dat[sample(x = c(1 : nrow(dat)), size = 250,000, replace = TRUE), ]
```

Based on this `portfolio` we generate claims (N, Y) using two GLMs that fulfill the CPG assumptions of Theorem 3.8, the modeling details are specified in columns 1–3 of Table 1. We especially emphasize that the covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ differ for claim counts and claim sizes.

CPG GLM We estimate the Poisson claim counts GLM and the gamma claim amounts GLM separately, according to Sect. 3.1 and under log-link choices. The results are presented in column 'estimated CPG' of Table 1, the brackets provide one estimated standard deviation received from the inverse of Fisher's information matrix. Note that we can estimate all regression parameters β_k and α_k *without*

[1] For a rigorous pricing exercise one should truncate longer exposures, say, to one accounting year, otherwise one implicitly considers a survival bias on policies with longer exposures, supposed that people give up motorcycling more likely after a claim.

Table 1 Synthetic CPG GLM example: the first 3 columns show the chosen (true) model; column 'estimated CPG' shows the resulting MLEs (with estimated std.dev. brackets)

Variable	Parameter	True Value	Estimated Param. CPG	Standard deviation
Intercept	β_0	13.80	12.99	(1.46)
Age	β_1	−0.180	−0.188	(0.010)
Age^2	β_2	1.70×10^3	1.74×10^3	(0.12×10^3)
Gender	β_3	0.30	0.36	(0.07)
Zone2	β_4	−0.60	−0.58	(0.05)
Zone3	β_5	−1.10	−1.07	(0.06)
Zone4	β_6	−1.50	−1.46	(0.05)
Zone5	β_7	−1.60	−1.39	(0.10)
McClass	β_8	−14.50	−13.42	(1.72)
McClass^2	β_9	2.30	2.16	(0.27)
log(McClass)	β_{10}	12.60	11.52	(1.58)
McClass^3	β_{11}	−0.140	−0.134	(0.017)
McAge	β_{12}	−0.140	−0.147	(0.008)
McAge^2	β_{13}	2.60×10^3	2.86×10^3	(0.36×10^3)
Intercept	α_0	8.650	8.650	(0.143)
Age	α_1	0.110	0.113	(0.008)
Age^2	α_2	-1.40×10^3	-1.44×10^3	(0.10×10^3)
McClass	α_3	8.0×10^2	7.3×10^2	(1.0×10^2)
McAge^2	α_4	-2.80×10^2	-2.90×10^2	(0.13×10^2)
McAge^3	α_5	1.80×10^3	1.91×10^3	(0.12×10^3)
McAge^4	α_6	-3.0×10^5	-3.3×10^5	(0.5×10^5)
Shape param.	γ	1.50	1.56	(0.04)

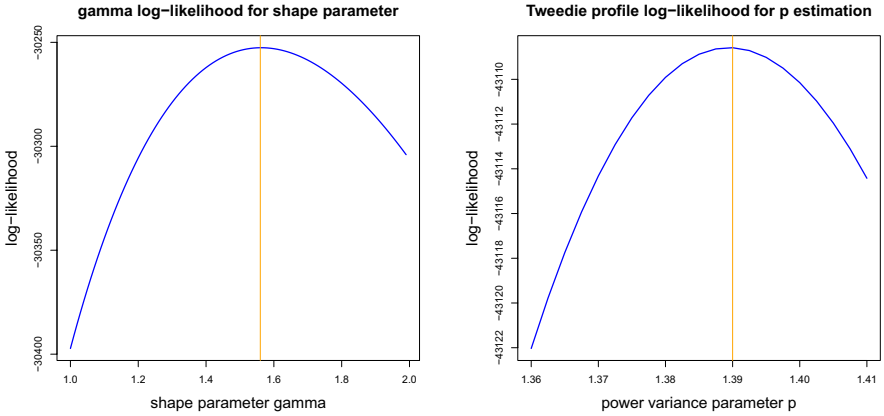

Fig. 1 (lhs) Log-likelihood $\gamma \mapsto \ell(\widehat{\boldsymbol{\alpha}}, \gamma)$ of the gamma GLM to estimate shape parameter γ for given $\widehat{\boldsymbol{\alpha}}$; (rhs) Tweedie profile log-likelihood $p \mapsto \ell(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\alpha}}^*(p), p)$ to estimate p

specifying shape parameter $\gamma > 0$ explicitly. Most estimated parameters are within one standard deviation of the true parameter values. The true parameters have been chosen such that they resemble the true data `swmotorcycle`. The true data has an observed claim frequency of only 1.05%, see "Appendix B". In the present example, claims are scarce too, and the gamma claim size GLM has been estimated on (only) 2'795 claims. The parameter estimates are remarkably accurate (we do not have model uncertainty here, only parameter estimation uncertainty). We conclude that this model can be calibrated well using the separate approach for claim counts and claim amounts.

Figure 1 (lhs) considers the log-likelihood function $\gamma \mapsto \ell(\widehat{\boldsymbol{\alpha}}, \gamma)$ of the gamma GLM to estimate shape parameter γ, we also refer to score Eq. (3.9). From this we find MLE $\widehat{\gamma} = 1.56$, and the inverse of Fisher's information matrix provides an estimated standard deviation of 0.04 for this estimate. Thus, the estimated shape parameter is slightly too high, though still within two standard deviations of the true value of $\gamma = 1.5$. We again highlight that this estimate is based on only 2'795 claims. Moreover, we remark that $\widehat{\gamma}$ has been used in the standard deviation estimates of Table 1, see (3.8).

Tweedie's DGLM Next we turn our attention to Tweedie's CP case. The true values $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and γ as well as their MLE counterparts $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\gamma}$ from the CPG model are transformed with Theorem 3.8 to receive the same model in Tweedie's CP parametrization, this is illustrated in the first four columns of Table 2. In a first calibration step for Tweedie's CP model, we choose $p = 1.39$ which is the optimal power variance parameter estimate of the CPG model, see last line in column 4 of Table 2. We then calibrate Tweedie's CP DGLM model for this power variance parameter p with Fisher's scoring method (3.25)–(3.26); as starting values for the algorithm we use the estimates from the CPG model (in italic in Table 2). Fisher's scoring method converges in 7 iterations with these initial values. Due to (3.36) we receive a model that has a bigger log-likelihood than its CPG counterpart (we include all constants in this consideration so that the log-likelihoods are directly comparable).

In the next step, we optimize over the power variance parameter p. Therefore, we use Theorem 3.6, which says that the mean estimates $\widehat{\mu}_i$ do not depend on p, and which provides the p-scaling for dispersion parameter MLEs $\widehat{\phi}_i(p)$. This allows us to directly plot the profile log-likelihood $p \mapsto \ell(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\alpha}}^*(p), p)$ as a function of $p \in (1.36, 1.41)$, see Fig. 1 (rhs). From this figure, we find maximizing value $\widehat{p} = 1.39$, which is close to the true value of $p = 1.4$. The second last column in Table 2 shows the resulting MLEs $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\alpha}}^*(\widehat{p})$ of the optimal Tweedie's CP model. A first observation is that the parameter estimates from Tweedie's CP model are not as close to the true values as the MLEs from the CPG model. However, model selection should not be based on this observation: note that the (true) CPG model has 22 parameters and Tweedie's CP model has 33 parameters, therefore, we expect some differences in model calibration.

We summarize the two estimated models in Table 3. On row (a) we compare the log-likelihoods $\ell_{\text{CPG}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{p})$ and $\ell_{\text{Tw}}(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\alpha}}^*, \widehat{p})$ of the estimated models CPG and Tweedie's CP, see also (3.36), to the one of the true model $\ell(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, p)$: we observe that both models slightly overfit to the data, with Tweedie's CP model having a

Table 2 Synthetic example: the first 3 columns show the chosen (true) model; column 'estimated CPG' (in italic) shows estimated parameters from the CPG model; column 'estimated Tweedie's CP' shows the MLEs from Tweedie's CP model (with estimated std.dev. in brackets)

Variable	Parameter	True Value	Estimated Param. CPG	Estimated Param. Tweedie's CP	Standard Deviation
Intercept	β_0^*	22.45	21.64	19.50	(1.87)
Age	β_1^*	-7.0×10^2	-7.5×10^2	-7.5×10^2	(1.3×10^2)
Age^2	β_2^*	3.0×10^4	3.1×10^4	3.0×10^4	(1.6×10^4)
Gender	β_3^*	0.30	0.36	0.40	(0.09)
Zone2	β_4^*	− 0.60	− 0.58	− 0.52	(0.07)
Zone3	β_5^*	− 1.10	− 1.07	− 0.99	(0.08)
Zone4	β_6^*	− 1.50	− 1.46	− 1.42	(0.07)
Zone5	β_7^*	− 1.60	− 1.39	− 1.36	(0.12)
McClass	β_8^*	− 14.42	− 13.35	− 10.79	(2.20)
McClass^2	β_9^*	2.30	2.16	1.72	(0.35)
log(McClass)	β_{10}^*	12.60	11.52	9.39	(2.03)
McClass^3	β_{11}^*	− 0.140	− 0.134	− 0.103	(0.0221)
McAge	β_{12}^*	− 0.140	− 0.147	− 0.196	(0.047)
McAge^2	β_{13}^*	-2.54×10^2	-2.62×10^2	-1.98×10^2	(0.77×10^2)
McAge^3	β_{14}^*	1.80×10^3	1.91×10^3	1.63×10^3	(0.49×10^3)
McAge^4	β_{15}^*	-3.0×10^5	-3.3×10^5	-2.9×10^5	(0.8×10^5)
intercept	α_0^*	0.1808	0.6909	− 0.5702	(0.9114)
Age	α_1^*	0.1380	0.1420	0.1429	(0.0061)
Age^2	α_2^*	-1.5×10^3	-1.6×10^3	-1.6×10^3	(0.1×10^3)
Gender	α_3^*	− 0.120	− 0.140	− 0.115	(0.043)
Zone2	α_4^*	0.240	0.228	0.269	(0.034)
Zone3	α_5^*	0.440	0.417	0.462	(0.037)
Zone4	α_6^*	0.60	0.57	0.59	(0.03)
Zone5	α_7^*	0.640	0.543	0.558	(0.060)
McClass	α_8^*	5.848	5.288	6.714	(1.074)
McClass^2	α_9^*	− 0.920	− 0.845	− 1.094	(0.169)
log(McClass)	α_{10}^*	− 5.040	− 4.500	− 5.654	(0.990)
McClass^3	α_{11}^*	5.6×10^2	5.2×10^2	6.9×10^2	(1.1×10^2)
McAge	α_{12}^*	5.6×10^2	5.7×10^2	5.2×10^2	(2.3×10^2)
McAge^2	α_{13}^*	-1.78×10^2	-1.88×10^2	-1.74×10^2	(0.38×10^2)
McAge^3	α_{14}^*	1.1×10^3	1.2×10^3	1.0×10^3	(0.2×10^3)
McAge^4	α_{15}^*	-2×10^5	-2×10^5	-2×10^5	(0.4×10^5)
variance param.	p	1.40	1.39	1.39	

slightly larger overfit [this is consistent with (3.36)]. Therefore, we penalize in AIC the log-likelihoods of the models by the number of parameters involved, see (3.37). The AIC values are given on row (b) of Table 3, and we give preference to the CPG calibration. Performing a likelihood-ratio test having the CPG model as null

Table 3 Synthetic example: summary statistics of fitted CPG and Tweedie's CP GLMs

	True Parameters	Estimated CPG	Estimated Tweedie CP
(a) log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, p)$	-43'125	-43'115	-43'109
(b) Akaike information criterion AIC		86'273.56	86'283.27
(c) Rooted mean square error (RMSE)		58.30	80.49
(d) Average of means $w_i\mu_i = w_i\lambda_i\zeta_i$	340	346	347
(e) Std. dev. in means $w_i\mu_i = w_i\lambda_i\zeta_i$	835	853	857
(f) Average of dispersions ϕ_i	4530	4724	4746
(g) Std. dev. in dispersions ϕ_i	1847	1898	1924

hypothesis model nested in Tweedie's CP model, gives a p-value of 34%, thus, we do not reject the null hypothesis on a 5% significance level. This gives support that we should go for the smaller CPG model in this example. Row (c) of Table 3 gives the rooted mean square error (RMSE) between the true model means $w_i\mu_i$ and their estimated counterparts $w_i\widehat{\mu}_i = w_i\widehat{\lambda}_i\widehat{\zeta}_i$; rows (d)–(g) show average means and dispersions as well as the corresponding standard deviations. We observe that these figures match the true values quite well. Recall that these figures are based on one simulation from the true model for each insurance policy, thus, they involve simulation error (but they do not involve model error because we only assume parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and p as unknown in this example). Moreover, we remark that the dispersion is not under-estimated, here, we also refer to the last bullet point of Remark 3.5.

Finally, in Fig. 2 we plot the predicted means $\widehat{\mu}_i$ against the true values μ_i. The left-hand side compares the two estimated models against the true model, and the right-hand side compares the two estimated models against each other. From these plots we conclude that both models are very accurate, the CPG estimated one (orange) being slightly closer to the true model than its Tweedie's CP counterpart (green). Summarizing: This synthetic example gives evidence supporting industry practice on focusing on the CPG model. Specifying covariate spaces is easier in the CPG case because systematic effects of claim counts and claim amounts are clearly separated, and in our example accuracy is slightly higher because Tweedie's CP seems to slightly overfit in our example.

4.2.2 A neural network regression approach

Next we explore neural network regression models on the same synthetic data. Neural networks have the capability of representation learning which means that they can perform covariate engineering themselves, we refer to Sections 4 and 5 of [21]. Therefore, covariates can be provided in their raw form to neural networks. The neural networks then, at the same time, pre-process these covariates and predict the response variables. Starting from a GLM, the required changes to achieve this representation learning are comparably small. We illustrate this in the present

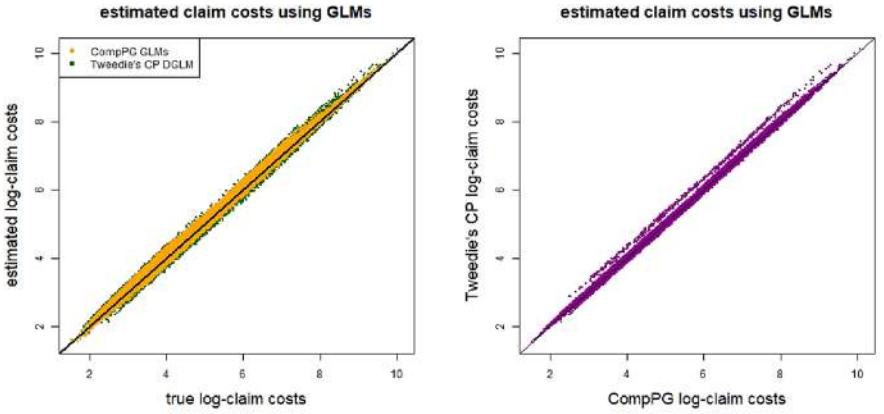

Fig. 2 (lhs) Comparison of estimated means $\widehat{\mu}_i$ versus true means μ_i: CPG GLM (orange) and Tweedie CP DGLM (green), (rhs) CPG GLM means versus Tweedie CP DGLM means over all $i = 1, \ldots, n$ policies

section. Alternatively, one may also be interested in using generalized additive models (GAMs). GAMs are more flexible in modeling different functional forms in the components of the covariates compared to GLMs, however, they do not automatically allow for flexible interaction modeling between covariate components. For this reason, we favor neural networks over GAMs.

We first define the (raw) covariate space $\mathcal{X}^\dagger$ which is going to be used throughout this section:

$$\boldsymbol{x}^\dagger = (\texttt{Age}, \texttt{Gender}, \texttt{Zone}, \texttt{McClass}, \texttt{McAge}) \in \mathcal{X}^\dagger \subset \mathbb{R}^8, \tag{4.1}$$

where we use dummy coding for the categorical variable $\texttt{Zone} \in \{0,1\}^4$. In contrast to Table 1, we do not specify the continuous variables in all its functional forms, but we let the neural network find these functional forms. A neural network is a function

$$\psi : \mathcal{X}^\dagger \to \mathbb{R}^d, \quad \boldsymbol{x}^\dagger \mapsto \boldsymbol{x} = \psi(\boldsymbol{x}^\dagger), \tag{4.2}$$

that consists of a composition of a fixed number of hidden network layers, each of them having a certain number of hidden neurons. For an explicit mathematical definition we refer to Section 3.1 in [21]. $\boldsymbol{x}^\dagger$ has the interpretation of being the raw covariate, and $\boldsymbol{x} = \psi(\boldsymbol{x}^\dagger) \in \mathbb{R}^d$ can be interpreted as the (network) pre-processed covariate. These pre-processed covariates $\psi(\boldsymbol{x}^\dagger)$ are then used in a classical GLM, e.g., for claim counts we may set for the log-link choice, see (3.2),

$$\lambda : \mathcal{X}^\dagger \to \mathbb{R}_+, \qquad \boldsymbol{x}^\dagger \mapsto \lambda(\boldsymbol{x}^\dagger) = \exp\left\langle \boldsymbol{\beta}, \psi(\boldsymbol{x}^\dagger) \right\rangle = \exp\left\{ \beta_0 + \sum_{k=1}^{d} \beta_k \psi_k(\boldsymbol{x}^\dagger) \right\}, \tag{4.3}$$

note that we use a slight abuse of notation here because strictly speaking $\psi(\boldsymbol{x}^\dagger)$ does not include an intercept term for β_0, so this always needs to be added. Neural

network regression function (4.3) involves regression parameters $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ as well as network weights $\vartheta \in \mathbb{R}^r$ which parametrize network function $\psi = \psi_\vartheta$. The dimension r of ϑ depends on the complexity of the chosen network ψ. Network fitting now trains at the same time network parameter ϑ for an optimal covariate pre-processing as well as GLM parameter $\boldsymbol{\beta}$ for response prediction. State-of-the-art fitting uses variants of the gradient descent algorithm, and a good performance depends on the complexity of ψ, we just mention the universal approximation property of appropriately designed neural networks. For more information, we refer to the relevant literature, in particular, to [21]. Based on this reference we explore (4.3) and its counterparts for claim counts and Tweedie's CP model. In all three prediction problems we use the identical covariate space $\mathcal{X}^\dagger$, and only network function ψ will differ in the weights ϑ to bring covariates into the appropriate form for the corresponding prediction task.

Poisson claim counts We start by modeling claim counts using neural network approach (4.3). We use the R library `keras` to implement this, and we use exactly the same architecture as in Listing 4 of [14], the only thing that changes is the dimension of $\mathcal{X}^\dagger$ from 40 on line 1 of Listing 4 in [14] to 8 in the present example, see (4.1). This results in $r = 655$ and $d + 1 = 11$ parameters. We fit these parameters in the usual way by considering 80% of the data for training and 20% of the data for out-of-sample validation to track overfitting in the gradient descent algorithm (we run 100 epochs on batch size 5000). We then choose the parameter that has the best out-of-sample performance on the validation data. To this network solution we apply the bias regularization step of Listing 5 in [21] to make the model unbiased.

On rows (a1)–(a2) of Table 4 we present the results for the claim counts neural network model. We provide the Poisson deviance losses of the true model λ_i (which is known here because we simulate from this model), the intercept model that does not use covariate information (i.e. is only based on intercept parameter β_0), the claim counts GLM (upper part of Table 1) and its neural network counterpart. We observe that both regression models slightly overfit to the data $8.4366 \cdot 10^{-2}$ and $8.4393 \cdot 10^{-2}$, respectively, compared to the true model loss of $8.4431 \cdot 10^{-2}$.

On row (a2) we provide the RMSE between the true model means λ_i and the estimated ones $\widehat{\lambda}_i$. We note that the Poisson GLM has a smaller RMSE than the neural network Poisson regression model. This is not surprising because the Poisson GLM uses the right functional form (no model uncertainty) and only estimates regression parameter $\boldsymbol{\beta}$ whereas the neural network regression model also determines this functional form for the raw covariates $\boldsymbol{x}^\dagger$. In Fig. 3 (lhs) we compare the resulting estimated frequencies to the true ones on all individual insurance policies $i = 1, \ldots, n$. From this plot we conclude that both models do a fairly good job because the dots lie more or less on the diagonal (which reflects the perfect model).

Gamma claim sizes Next we consider a neural network approach for the gamma claim sizes. This essentially means that we replace linear predictor (3.5) by the following neural network predictor (under a log-link choice for g_2)

Table 4 Comparison of deviance losses and RMSEs of the true (synthetic) model, the intercept model not using covariate information, the GLM approaches and the neural network approaches

	True model	Intercept Model	GLM model	Neural network Model
(a1) Poisson deviances for N_i (in 10^{-2})	8.4431	9.8052	8.4366	8.4393
(a2) RMSE between λ_i and $\widehat{\lambda}_i$		2.06%	0.18%	0.55%
(b1) Gamma deviances for $\bar{Z}_i$	0.7058	1.1442	0.7052	0.7015
(b2) RMSE between ζ_i and $\widehat{\zeta}_i$		27'210	693	4'460

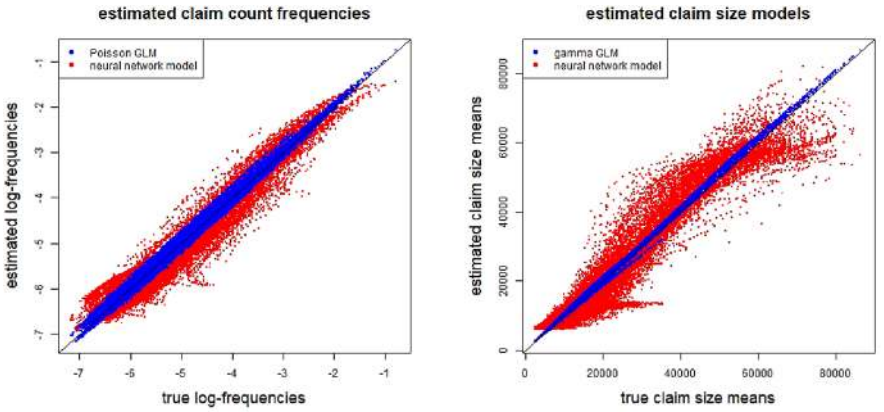

Fig. 3 (lhs) Comparison of estimated models versus true model: (lhs) Poisson claim counts models for N_i; (rhs) gamma claim size models for $\bar{Z}_i$

$$\zeta : \mathcal{X}^\dagger \to \mathbb{R}_+, \quad \boldsymbol{x}^\dagger \mapsto \zeta(\boldsymbol{x}^\dagger) = \exp\left\langle \boldsymbol{\alpha}, \psi(\boldsymbol{x}^\dagger) \right\rangle = \exp\left\{ \alpha_0 + \sum_{k=1}^{d} \alpha_k \psi_k(\boldsymbol{x}^\dagger) \right\}, \tag{4.4}$$

where ψ is a neural network function (4.2) that may have the same structure as the one used for the Poisson regression model (4.3), but typically differs in network weights ϑ. For simplicity, we use exactly the same neural network architecture as in the Poisson case, only the exposure offset is dropped and the Poisson deviance loss function is changed to the gamma deviance loss function (including weights), in line with the distributional assumptions made.

The results are presented on rows (b1)–(b2) of Table 4 and Fig. 3 (rhs) (we run 1000 epochs on batch size 5000 and we callback the model with the smallest validation loss). Again we receive reasonably good results from the network approach, i.e., covariate engineering on $\mathcal{X}^\dagger$ is done quite well by the network, we emphasize that these results are based on only 2'795 claims. But we also see from Fig. 3 (rhs) that individual predictions spread more around the diagonal than in the gamma GLM case (where we assume perfect knowledge about the functional form of the

Table 5 Synthetic example: summary statistics of the fitted CPG and Tweedie's CP neural network models

	True Parameters	Estimated CPG	Estimated Tweedie CP
(a) Network log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, p)$	-43'125	-43'110	-43'079
(b) Power variance parameter p	1.400	1.389	1.390
(c) Rooted mean square error (RMSE)		190.33	225.50
(d) Average of means $w_i\mu_i = w_i\lambda_i\zeta_i$	340	335	357
(e) Std. dev. in means $w_i\mu_i = w_i\lambda_i\zeta_i$	835	812	872
(f) Average of dispersions ϕ_i	4530	4595	5264
(g) Std. dev. in dispersions ϕ_i	1847	1812	1950

regression function). Better accuracy can only be achieved by having more claim observations.

Next, we estimate the shape parameter γ. This is done analogously to the gamma GLM case by plotting the corresponding log-likelihood $\ell(\gamma)$ as a function of γ. This gives estimate $\widehat{\gamma} = 1.57$, which is slightly too large but still reasonable compared to the true value of $\gamma = 1.5$. A too high shape parameter implies a too low dispersion, which is a sign of over-fitting to the observations.

We conclude with the summary statistics for the neural network approaches in Table 5 column 'estimated CPG', which look fairly similar to the GLM ones in Table 3. We obtain a larger RMSE, which is not surprising because we have more model uncertainty due to missing covariate knowledge, this is also obvious from Fig. 3.

Tweedie's compound Poisson neural network approach First, we remark that, in general, there is no simple comparison between a CPG and a Tweedie CP neural network approach similar to (3.34)–(3.35). The relation (3.34)–(3.35) is strongly based on the fact that we can directly compare linear predictors under suitable choices of covariate spaces. Since the networks given (4.2) transform covariates in a non-trivial way under non-linear activation functions, there is no hope to get an easy comparison between the models unless the network architectures are chosen in a very specific way, i.e. artificial way, so to say. Therefore, we do not aim to nest the CPG neural network into Tweedie's CP neural network model, but we directly focus on modeling the latter. This essentially implies that we have to replace linear predictors (3.21)–(3.22) by the following two-dimensional neural network predictors (under log-link choices for g_p and g_d)

$$(\mu, \phi) : \mathcal{X}^\dagger \to \mathbb{R}_+^2, \quad \boldsymbol{x}^\dagger \mapsto (\mu, \phi)(\boldsymbol{x}^\dagger) = \left(\exp\left\langle \boldsymbol{\beta}^*, \psi(\boldsymbol{x}^\dagger)\right\rangle, \exp\left\langle \boldsymbol{\alpha}^*, \psi(\boldsymbol{x}^\dagger)\right\rangle\right),$$

where ψ is a neural network function (4.2). The first component of $(\mu, \phi)(\boldsymbol{x}^\dagger) \in \mathbb{R}_+^2$ predicts the total claim costs Y and the second component estimates the dispersion parameter ϕ. We use one network ψ to simultaneously perform this prediction task for mean and dispersion parameter. We implement this in the R library `keras` and we use the same architecture as in Listing 4 of [14], but we need to change the

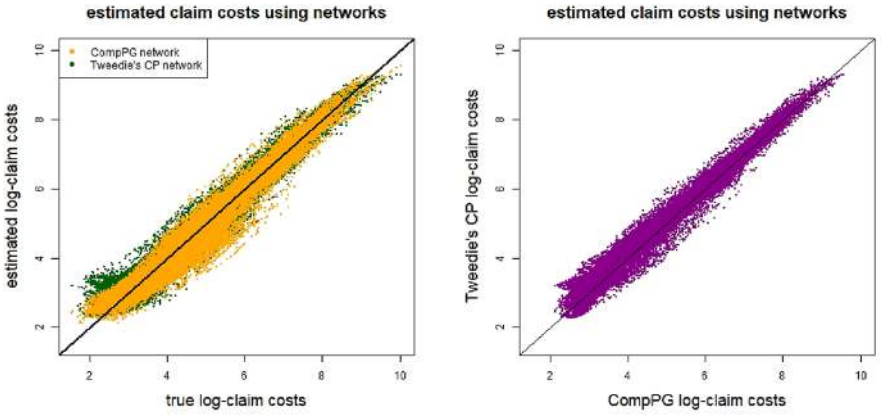

Fig. 4 (lhs) comparison of estimated means $\widehat{\mu}_i$ versus true means μ_i: CPG neural network (orange) and Tweedie CP neural network (green), (rhs) CPG network means versus Tweedie CP neural network means over all $i = 1, \ldots, n$ policies

input dimension to 8 and the output dimension to 2. The exposures w_i are treated as weights as follows

$$\ell(\mu, \phi) \propto \sum_{i=1}^{n} w_i \left[\frac{1}{\phi_i} \left(Y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) - \frac{N_i/w_i}{p-1} \log \phi_i \right].$$

This requires a custom made loss function in `keras` for parameter estimation, the details are provided in Listing 2 in the "Appendix". We fit this model with the gradient descent algorithm exactly using the same methodology as outlined above (callback of the lowest validation loss model after 100 epochs on batch sizes 5000).

In order to come up with the optimal neural network model we need to fit neural networks for multiple power variance parameters p, because there is no result similar to Theorem 3.6 that allows for a shortcut. Of course, this disadvantages Tweedie's CP neural network model from a computational point of view. We come up with an optimal power variance parameter estimate of $\widehat{p} = 1.390$, which yields then the results in the last column of Table 5. From the figures on rows (c)–(g) we conclude that Tweedie's CP approach is not fully competitive with the CPG fitting. These differences are also illustrated in Fig. 4 with the CPG approach being slightly closer to the true model means. Nevertheless, all these estimates look very reasonable and the estimated neural network seems to capture the crucial features of the true model.

Conclusions from our synthetic data example Our findings support industry practice of focusing on the CPG parametrization. Our estimated models based on this parametrization are closer to the true model than the ones obtained from Tweedie's CP parametrization. If we work under GLM assumptions we need to pre-process covariates which is easier in the CPG parametrization because systematic effects of claim counts and claim amounts can be separated. If we work under neural network regression models, model calibration is not efficient under Tweedie's CP parametrization because we need to run gradient descent algorithms on multiple power

Table 6 Comparison of gamma deviance losses on real data: intercept model and gamma neural network regression model

	Intercept Model	Neural network Model
Gamma deviances losses for $\bar{Z}_i$	2.0854	1.5863

variance parameters p to find the optimal model. Moreover, in our example, the CPG case leads to more accurate predictive models.

4.3 Real data example: an outlook

In view of the previous example everything seems to be fairly clear. However, our synthetic data is based on the very strong property of having gamma claim sizes with constant shape parameter γ over the whole insurance portfolio. This assumption may be critical in real insurance applications. We briefly analyze it in terms of our real data example given in "Appendix B", and we give an outlook in case this assumption is not fulfilled. We keep this section very short, and we mainly view it as a motivation to conduct future research in this direction.

There are two possibilities in which the constant shape parameter assumption may fail, either the claim sizes are gamma distributed, but the shape parameter γ_i is also insurance policy i dependent, or the gamma distribution is inappropriate due to that the claim sizes exhibit too heavy tails. We explore this on the real data example provided in "Appendix B". For this it suffices to focus on the gamma claim size model, i.e. we do not study claim counts in this real data example. Moreover, to minimize covariate pre-processing we explore a gamma neural network regression model on these claim sizes, the chosen model architecture is identical to the one used in (4.4), in particular, it does covariate engineering itself.

Table 6 shows the (in-sample) gamma deviance losses of the intercept model and the neural network regression model. Obviously, the neural network approach has a better performance (note that the network model has been received by a proper training-validation analysis as described above). Using the resulting mean estimates $\widehat{\zeta}_i$ we can estimate the (constant) shape parameter γ. This is illustrated in Fig. 5 (lhs): we estimate $\widehat{\gamma} = 0.75$. Thus, we receive a shape parameter smaller than 1, which provides over-dispersion $1/\widehat{\gamma} = 1.33 > 1$, i.e., the estimated gamma densities are strictly decreasing. This fact requires further examination because there might be two situations: either the true shape parameter is smaller than 1 (and everything is fine), or the claim sizes are more heavy tailed than a gamma distribution allows. This is typically compensated by over-dispersion in the estimated model. We analyze this warning signal on our real data.

Figure 5 (rhs) gives the Tukey–Anscombe plot of the gamma deviance residuals against the fitted means. This plot supports the model choice because we cannot see any particular structure in the figure, it also supports the constant shape parameter assumption on γ. Figure 6 gives the QQ-plot and it compares the observed claims

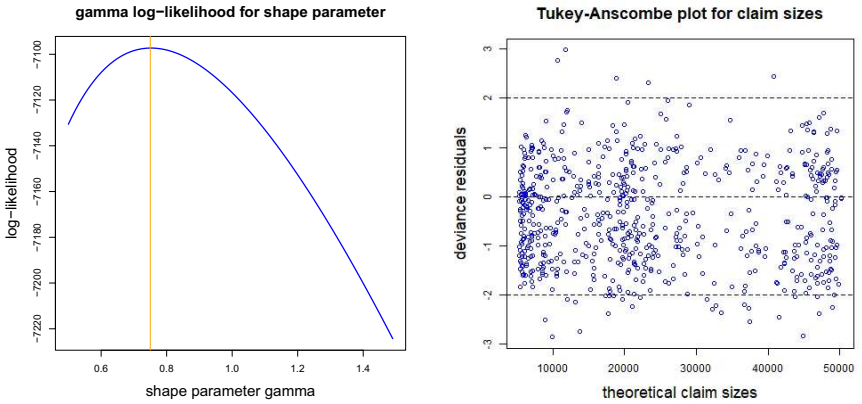

Fig. 5 (lhs) Log-likelihood $\gamma \mapsto \ell(\gamma)$ of the gamma neural network to estimate shape parameter γ on the real data, (rhs) Tukey–Anscombe plot giving gamma deviance residuals against fitted means

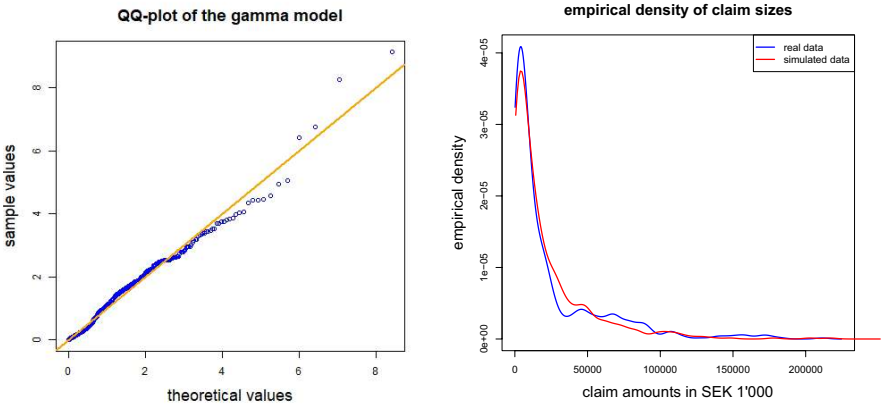

Fig. 6 (rhs) QQ-plot of the estimated gamma model for claim sizes, (rhs) density of real data compared to one simulation from the estimated model

against one simulation from the fitted model. Also these two plots look quite reasonable, one may only question the upper tail of the QQ-plot.

Conclusions The short analysis on the real data has shown that for the motorcycle claims data the gamma claim size model is fairly reasonable, thus, supporting the CPG model. On different data, one may relax the constant shape parameter assumption on γ. This may result in a DGLM for gamma claim sizes (which is known in industry) and a Poisson GLM for claim counts. Again this model can easily be fitted in the Poisson-gamma parametrization, however, this approach does not have a Tweedie's CP counterpart relying on a fixed parameter p, giving more support to the industry preference of choosing the Poisson-gamma parametrization.

5 Conclusion

We have revisited the compound Poisson model with i.i.d. gamma claim sizes. This model allows for two different parametrizations, namely, the Poisson-gamma parametrization and Tweedie's compound Poisson parametrization. We have provided results for GLMs illustrating when the two parametrizations are identical, and we have provided a theorem that allows for efficient fitting of power variance parameters in Tweedie's parametrization (under log-link choices for the GLMs).

In the applied section, we have analyzed why the insurance industry gives preference to the Poisson-gamma parametrization. Based on examples, we find that, indeed, this parametrization is easier to fit, and results turn out to be more accurate in our examples. In particular, under neural network regression models we give a clear preference to the Poisson-gamma parametrization because Tweedie's version does not possess an easy and efficient way in estimating the power variance parameter. That is, the Tweedie version is computationally clearly lacking behind the Poisson-gamma case.

For our real data example it turns out that the gamma claim size model with constant shape parameter is quite reasonable. However, in many other applications this is not the case. Therefore, insurance industry explores double GLMs for a flexible modeling of shape parameters of claim sizes; on the other hand, a case-dependent p modeling in Tweedie's compound Poisson parametrization is not (easily) feasible. For modeling more heavy tailed claim sizes, mixture models are a promising proposal.

Appendix

A Generalized linear models

GLMs have been introduced in [9], and they have been studied in the monograph [7]. GLMs are based on the EDF (2.2). The EDF has been studied extensively in [2, 4, 5], and its properties have been revisited in [21]. The original introduction of EDF distributions (2.2) is constructive from which it follows that the effective domain $\boldsymbol{\Theta}$ is a convex set and that the cumulant function κ is a smooth and convex function on the interior of the effective domain $\mathring{\boldsymbol{\Theta}}$. Moreover, we get the following moments for Y having EDF distribution (2.2)

$$\begin{aligned} \mu =& \mathbb{E}[Y] = \kappa'(\theta), \quad \operatorname{Var}(Y) = \frac{\phi}{w}\kappa''(\theta) \quad \text{and} \\ \mathbb{E}[\exp\{rY\}] =& \exp\left\{\frac{w}{\phi}(\kappa(\theta + r\phi/w) - \kappa(\theta))\right\}, \end{aligned} \tag{A.1}$$

for $|r|$ sufficiently small such that $\theta + r\phi/w \in \mathring{\boldsymbol{\Theta}}$ for $\theta \in \mathring{\boldsymbol{\Theta}}$. Convexity of κ implies existence of the canonical link providing canonical parameter and variance function, respectively,

$$\theta = (\kappa')^{-1}(\mu) \quad \text{and} \quad V(\mu) = \big(\kappa'' \circ (\kappa')^{-1}\big)(\mu).$$

GLMs are based on a linear predictor η for modeling the mean parameter $\mu = \mathbb{E}[Y]$. Assume we have $(d+1)$-dimensional covariates $\boldsymbol{x} \in \mathcal{X} = \{1\} \times \mathbb{R}^d$. The linear predictor $\eta = \eta(\boldsymbol{x})$ is received by choosing a suitable link function $g(\cdot)$ such that the following relationship holds

$$g(\mu) = \eta = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle,$$

for a given regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$. We need to ensure to have a well-defined GLM by

$$\theta = (\kappa')^{-1}(\mu) = \big((\kappa')^{-1} \circ g^{-1}\big)(\eta) = \big((\kappa')^{-1} \circ g^{-1}\big)\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \ \in \ \mathring{\boldsymbol{\Theta}}. \tag{A.2}$$

This might be a challenge for (one-sided) bounded effective domains $\boldsymbol{\Theta}$ and may require a careful choice of the link function $g(\cdot)$.

Assume we have n independent pairs of random variable and covariates $(Y_i, \boldsymbol{x}_i)$ following an EDF distribution (2.2) with the same cumulant function κ; we choose the same link function $g(\cdot)$ to receive linear predictors $\eta_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$. The log-likelihood function of this model is

$$\boldsymbol{\beta} \ \mapsto \ \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i}{\phi_i} \big(Y_i \theta_i - \kappa\big(\theta_i\big)\big) + a(Y_i; w_i/\phi_i),$$

with canonical parameter $\theta_i = (\kappa')^{-1}(\mu_i) = \big((\kappa')^{-1} \circ g^{-1}\big)(\eta_i)$. The score w.r.t. $\boldsymbol{\beta}$ is obtained by the gradient

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i}{\phi_i} \, \frac{\partial \big(Y_i \theta_i - \kappa\big(\theta_i\big)\big)}{\partial \theta_i} \, \frac{\partial \theta_i}{\partial \mu_i} \, \frac{\partial \mu_i}{\partial \eta_i} \, \nabla_{\boldsymbol{\beta}} \eta_i \\ &= \sum_{i=1}^{n} \frac{w_i}{\phi_i} \big(Y_i - \mu_i \big) \frac{1}{V(\mu_i)} \, \frac{\partial \mu_i}{\partial \eta_i} \, \boldsymbol{x}_i. \end{aligned}$$

We define the diagonal working weight matrix W and working residual vector $\boldsymbol{R}$ by

$$\begin{aligned} W &= \operatorname{diag}\left(\left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-2} \frac{w_i}{\phi_i} \, \frac{1}{V(\mu_i)} \right)_{i=1,\ldots,n} \in \mathbb{R}^{n \times n}, \\ \boldsymbol{R} &= \left(\frac{\partial g(\mu_i)}{\partial \mu_i} (Y_i - \mu_i) \right)_{i=1,\ldots,n} \in \mathbb{R}^n. \end{aligned}$$

This allows us to write the score equation for finding the MLE of regression parameter $\boldsymbol{\beta}$ by

$$\nabla_\beta \ell(\boldsymbol{\beta}) = \mathbf{0} \;\Leftrightarrow\; \mathfrak{X}' W \boldsymbol{R} = \mathbf{0}, \tag{A.3}$$

with design matrix $\mathfrak{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n\times(d+1)}$. MLE system (A.3) is solved either using Fisher's scoring method or the iteratively re-weighted least squares (IRLS) algorithm, see [7, 9]. For Fisher's scoring method we explore the scoring updates

$$\boldsymbol{\beta}_t \;\mapsto\; \boldsymbol{\beta}_{t+1} = \left(\mathfrak{X}' W \mathfrak{X}\right)^{-1} \mathfrak{X}' W (\boldsymbol{R} + g(\boldsymbol{\mu})), \tag{A.4}$$

where all terms on the right-hand side are evaluated for algorithmic time t. It has been pointed out by an anonymous referee that the R command `glm()` does not directly calculate the inverse of the matrix $\mathfrak{X}' W \mathfrak{X}$ in (A.4), but, instead, solves a linear system for $\boldsymbol{\beta}_{t+1}$. The motivation for this approach is that in high-dimensional covariate spaces or in the situation of multiple categorical variables with many labels (implemented by dummy coding), the matrix $\mathfrak{X}' W \mathfrak{X}$ may be close to singular and, henceforth, inversion of this matrix may lead to unstable results.

Standard errors are obtained from the inverse of Fisher's information matrix

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}\left[\nabla_\beta \ell(\boldsymbol{\beta}) \left(\nabla_\beta \ell(\boldsymbol{\beta})\right)'\right] = -\mathbb{E}\left[\nabla_\beta^2 \ell(\boldsymbol{\beta})\right] = \mathfrak{X}' W \mathfrak{X} \;\in\; \mathbb{R}^{(d+1)\times(d+1)},$$

where ∇_β^2 denotes the Hessian w.r.t. $\boldsymbol{\beta}$. The IRLS algorithm replaces the inverse Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})^{-1} = (\mathfrak{X}' W \mathfrak{X})^{-1}$ in the scoring updates by the inverse of the observed information matrix

$$\mathcal{J}(\boldsymbol{\beta})^{-1} = -\left(\nabla_\beta^2 \ell(\boldsymbol{\beta})\right)^{-1}.$$

B Motorcycle data example

We start with a descriptive and exploratory analysis of the Swedish motorcycle data of Listing 1. We have $n = 62'036$ insurance policies with positive exposures $w_i > 0$. The empirical claim frequency is $\bar{\lambda} = \sum_{i=1}^n N_i / \sum_{i=1}^n w_i = 1.05\%$, and the average claim size is $\bar{\xi} = \sum_{i=1}^n \sum_{j=1}^{N_i} Z_{i,j} / \sum_{i=1}^n N_i = 24'641$ Swedish crowns SEK.

Figure 7 shows a boxplot over all exposures w_i and the claim counts N_i on all insurance policies. We note that insurance claims are rare events for this product, because the claim frequency is only $\bar{\lambda} = 1.05\%$.

Figures 8 and 9 give the marginal total exposures (split by gender), the marginal claim frequencies and the marginal average claim amounts for the covariate components `Age`, `Zone`, `McClass`, `McAge` and `Bonus`. The first observation is that we have a very imbalanced portfolio between genders, only 11% of the total exposure is coming from females. The empirical claim frequency of females is 0.86% and the one of males is 1.08%. We note that the female claim frequency comes from (only) 61 claims (based on an exposure of female of 7'094 accounting years, versus 57'679 for male). Therefore, it is difficult to analyze females separately, and all marginal claim frequencies and claim sizes in Figs. 8 and 9 (middle and rhs) are analyzed

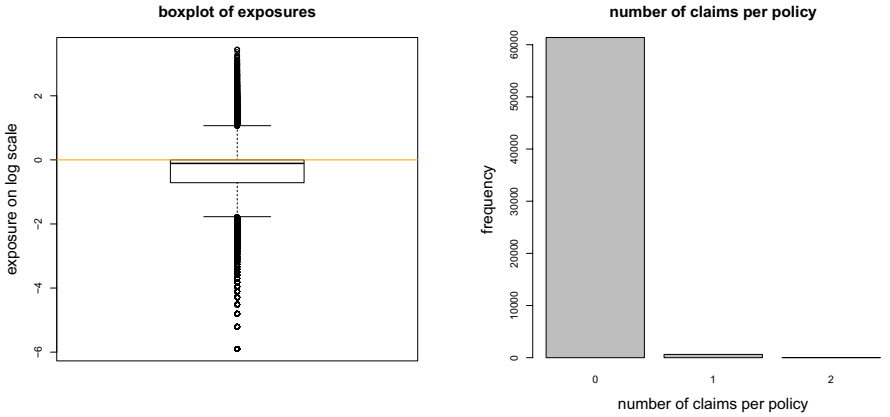

Fig. 7 (lhs) Boxplot of exposures w_i on the log scale (the orange line corresponds to 1 accounting year), (rhs) histogram of the number of observed claims N_i per policy.

jointly for both genders. Average claim sizes are 18'237 SEK and 25'270 SEK for female and male, respectively.

The empirical marginal frequencies in Figs. 8 and 9 (middle) are complemented with confidence bounds of two standard deviations (blue dotted lines) and the empirical overall frequency $\bar{\lambda} = 1.05\%$ (orange color). From the plots we conclude that we should keep the explanatory variables `Age`, `Zone`, `McClass` and `McAge`, but the variable `Bonus` does not seem to have any predictive power. At the first sight, this seems surprising because the bonus-malus level encodes the past claims history. The reason that the bonus-malus level is not needed for our claims is that we consider comprehensive insurance for motorcycles covering loss or damage of motorcycles other than collision (for instance, caused by theft, fire or vandalism), and the bonus-malus level encodes collision claims. The situation for average claim amounts is a bit more difficult to understand, but we make a similar conclusion, namely, that we can drop the covariate `Bonus`. Moreover, we merge Zones 5–7 because of small exposures and similar behavior.

Figure 10 shows the correlations between the covariates: (lhs) correlations between continuous covariates, (plots rhs), dependence between continuous covariates and the categorical `Zone` covariate. We have some dependence, for instance, in `Zone 1` (three largest Swedish cities) motorcycles are more light (`McClass`) and less old. Older people drive less heavy motorcycles that are more old, and older motorcycles are less heavy.

Figure 11 gives the empirical density, empirical distribution and log-log plot of average claim amounts $\bar{Z}_i$. From the log-log plot we conclude that the average claim amounts are not heavy tailed, which does not reject the use of gamma claim size distributions at that stage.

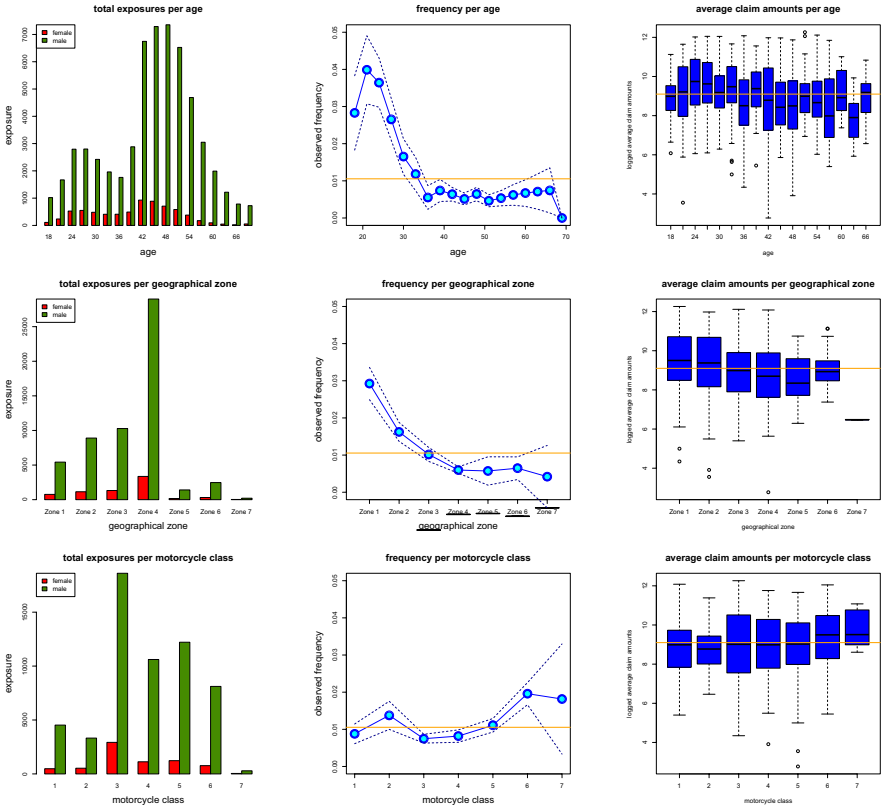

Fig. 8 (top, middle and bottom rows) `Age`, `Zone`, `McClass`: (lhs) histogram of exposures (split by gender), (middle) observed claim frequency, (rhs) boxplot of observed average claim amounts $\bar{Z}_i$ of policies with $N_i > 0$ (on log-scale)

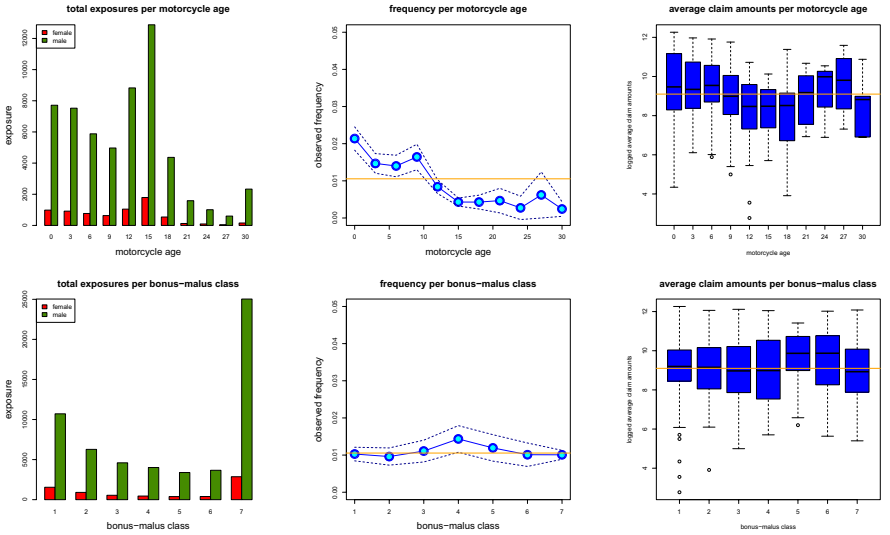

Fig. 9 (top and bottom rows) `McAge`, `Bonus`: (lhs) histogram of exposures (split by gender), (middle) observed claim frequency, (rhs) boxplot of observed average claim amounts $\bar{Z}_i$ of policies with $N_i > 0$ (on log-scale)

	Age	McClass	McAge
Age		-11%	7%
McClass	-10%		-19%
McAge	6%	-12%	

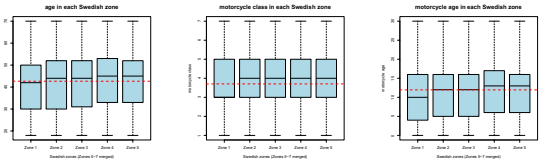

Fig. 10 (lhs) correlations: top-right shows Pearson's correlation; bottom-left shows Spearman's rho; (rhs) boxplots of `Age`, `McClass`, `McAge` versus `Zone`

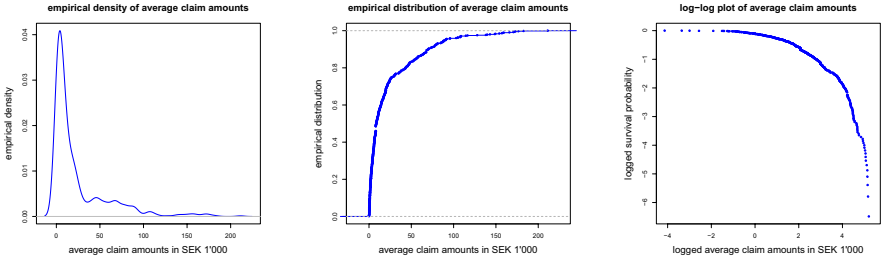

Fig. 11 (lhs) Empirical density (middle) empirical distribution and (rhs) log-log plot of average claim amounts $\bar{Z}_i$ of policies with $N_i > 0$

C R code

Listing 2: R code for Tweedie's CP neural network model.

```
library(keras)
#
network.Tweedie <- function(seed){
    set.seed(seed)
    use_session_with_seed(seed)
    design  <- layer_input(shape=c(8), dtype='float32', name='design')
    #
    output = design  %>%
        layer_dense(units=20, activation='tanh', name='hidden1') %>%
        layer_dense(units=15, activation='tanh', name='hidden2') %>%
        layer_dense(units=10, activation='tanh', name='hidden3') %>%
        layer_dense(units=2,  activation='exponential', name='output')
    #
    model <- keras_model(inputs=list(design), outputs=c(output))
    model
    }
#
p <- 1.4
Tweedie_loss <- function(y_true, y_pred)
      - k_mean( y_true[,3]*((y_true[,1]*y_pred[,1]^(1-p)/(1-p)-
                y_pred[,1]^(2-p)/(2-p))/y_pred[,2]-y_true[,2]*log(y_pred[,2])/(p-1)) )}
#
model <- network.Tweedie(seed=200)
model %>% compile(loss = Tweedie_loss, optimizer = 'nadam')
#
XX <- as.matrix(dat[,c("Age","Gender","Zone2","Zone3","Zone4","Zone5","McClass","McAge")])
YY <- as.matrix(cbind(dat$ClaimCosts/dat$Exposure,dat$ClaimNb/dat$Exposure,dat$Exposure))
#
fit  <- model %>% fit(list(XX), YY, validation_split=0.2, batch_size=5000, epochs=200)
dat$predict <- model %>% predict(list(XX))
```

Funding Open Access funding provided by ETH Zurich.

References

1. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
2. Barndorff-Nielsen O (2014) Information and exponential families. In: Statistical theory. John Wiley & Sons, Chichester, UK
3. Dutang C, Charpentier A (2019) CASdatasets R package vignette. Reference manual, November 13, 2019. Version 1.0-10
4. Jørgensen B (1986) Some properties of exponential dispersion models. Scand J Stat 13(3):187–197
5. Jørgensen B (1987) Exponential dispersion models. J R Stat Soc Ser B (Methodol) 49/2:127–145

6. Jørgensen B, de Souza MCP (1994) Fitting Tweedie's compound Poisson model to insurance claims data. Scand Actuar J 1994(1):69–93
7. McCullagh P, Nelder JA (1983) Generalized linear models. Chapman & Hall, London
8. Nelder JA, Pregibon D (1987) An extended quasi-likelihood function. Biometrik 74:221–231
9. Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A (Gen) 135/3:370–384
10. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/ Release 1.0.28 of 2020-09-15. In: Olver FWJ, Olde Daalhuis AB, Lozier DW, Schneider BI, Boisvert RF, Clark CW, Miller BR, Saunders BV, Cohl HS, McClain MA (eds)
11. Ohlsson E, Johansson B (2010) Non-life insurance pricing with generalized linear models. Springer, Berlin
12. Quijano Xacur OA, Garrido J (2015) Generalised linear models for aggregate claims: to Tweedie or not? Eur Actuar J 5(1):181–202
13. R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
14. Schelldorfer J, Wüthrich MV (2019) Nesting classical actuarial models into neural networks. SSRN Manuscript ID 3320525. Version of January 22, 2019
15. Smyth GK (1989) Generalized linear models with varying dispersion. J R Stat Soc Ser B (Methodol) 51:47–60
16. Smyth GK (1996) Partitioned algorithms for maximum likelihood and other nonlinear estimation. Stat Comput 6:201–216
17. Smyth GK, Jørgensen B (2002) Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modeling. ASTIN Bull 32(1):143–157
18. Smyth GK, Verbyla AP (1999) Adjusted likelihood methods for modelling dispersion in generalized linear models. Environments 10:696–709
19. Tweedie MCK (1984) An index which distinguishes between some important exponential families. In: Ghosh JK, Roy J (eds) Statistics: applications and new directions. Proceeding of the Indian statistical golden jubilee international conference. Indian Statistical Institute, Calcutta, pp 579–604
20. Wüthrich MV (2013) Non-life insurance: mathematics & statistics. SSRN Manuscript ID 2319328. Version of January 7, 2020
21. Wüthrich MV (2019) From generalized linear models to neural networks, and back. SSRN Manuscript ID 3491790. Version of April 3, 2020
22. Wüthrich MV (2020) Bias regularization in neural network models for general insurance pricing. Eur Actuar J 10(1):179–202