Insurance: Mathematics and Economics 109 (2023) 94–112

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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

Deep quantile and deep composite triplet regression ☆

Tobias Fissler [a,*], Michael Merz [b], Mario V. Wüthrich [c]

[a] *Institute for Statistics and Mathematics, Department of Finance, Accounting and Statistics, Vienna University of Economics and Business (WU), Welthandelsplatz 1, 1020 Vienna, Austria*
[b] *Faculty of Business Administration, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany*
[c] *RiskLab, Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland*

ARTICLE INFO

Article history:
Received December 2021
Received in revised form October 2022
Accepted 6 January 2023
Available online 18 January 2023

JEL classification:
C14
C31
C450
C510
C520
C580

Keywords:
Elicitability
Consistent loss function
Proper scoring rule
Bregman divergence
Quantile and expected shortfall regression
Conditional tail expectation
Neural network regression
Splicing model

ABSTRACT

A main difficulty in actuarial claim size modeling is that covariates may have different effects on the body of the conditional distribution and on its tail. To cope with this problem, we introduce a deep composite regression model whose splicing point is given in terms of a quantile of the conditional claim size distribution (rather than a constant). This allows us to simultaneously fit different regression models in the two different parts of the conditional distribution function. To facilitate M-estimation for such models, we introduce and characterize the class of strictly consistent scoring functions for the triplet consisting of a quantile, as well as the lower and upper expected shortfall beyond that quantile. In a second step, this elicitability result is applied to fit deep neural network regression models. We demonstrate the applicability of our approach and its superiority over classical approaches on a real data set from accident insurance.

1. Introduction

1.1. Background

In actuarial modeling, we typically use regression models to estimate the expected values of the claim sizes exploiting systematic effects of covariates (also called features or explanatory variables). These estimated expected values are used for insurance pricing, claims forecasting and claims reserving. The most commonly used regression model is the generalized linear model (GLM) with the exponential dispersion family (EDF) as the underlying distributional model. Parameter estimation within this EDF-GLM framework is performed by maximum likelihood estimation (MLE). In this setup, MLE is equivalent to minimizing the corresponding deviance loss function, the latter being a strictly consistent scoring function for the mean. Strictly consistent scoring functions facilitate M-estimation and forecast evaluation; further details are provided below. For actuarial problems, there are the following main issues with this approach:

(1) The EDF only contains light-tailed distribution functions such as the Gaussian, the Poisson, the gamma or the inverse Gaussian distributions. For this reason, the classical EDF-GLM framework is not suitable if we have a mixture of light-tailed and heavy-tailed data.

☆ An earlier version of this paper has been circulated with the title "Deep Quantile and Deep Composite Model Regression".
* Corresponding author.
E-mail addresses: tobias.fissler@wu.ac.at (T. Fissler), michael.merz@uni-hamburg.de (M. Merz), mario.wuethrich@math.ethz.ch (M.V. Wüthrich).

https://doi.org/10.1016/j.insmatheco.2023.01.001

This fact is often disregarded in practice and the EDF-GLM framework is used nevertheless. This may result in over-fitting to large claims and non-robustness in model estimation.

(2) The systematic effects (and the influencing covariates) may be different in different parts of the claim size distribution. E.g., the age of the policyholder variable may be an important variable to explain systematic effects in small claims, but it might be irrelevant to describe such effects in large claims. Such a behavior may implicitly be implied, for instance, by different injury types when going from small to large claims (accident insurance) or by different industry segments when going from small to large sums insured (industry fire insurance); an explicit example is given in Section 5.3 of Fung et al. (2022).

(3) The GLM regression structure may not be suitable or covariate engineering may be too difficult to bring the regression problem into a GLM form. As a result, not all systematic effects such as interactions may be correctly incorporated into the GLM.

(4) The GLM is usually used to estimate expected values. From a risk management perspective, one is also interested in other quantities such as quantiles and expected shortfalls (ES). If the aforementioned distributional model within the EDF is not suitable for the description of the observed data, then a very accurately estimated expected value is not directly helpful, e.g., in estimating quantiles.

Our paper tackles these points; we first discuss the related literature and then we are going to describe our proposal. To deal with the problem of the lack of any simple off-the-shelf distribution that fits the entire range of the data, i.e., the body and tail of the data, one may either uses a composite (splicing) model or a mixture model. These two approaches can be fitted with the expectation-maximization (EM) algorithm. Composite models are studied in Cooray and Ananda (2005), Scollnik (2007), Pigeon and Denuit (2011), Grün and Miljkovic (2019) and Parodi (2020). These papers do not consider a regression structure and covariates. The only actuarial publications that we are aware of considering composite regression models are Gan and Valdez (2018) and Laudagé et al. (2019). Both proposals choose very specific distributional assumptions above and below a given threshold (splicing point) to obtain analytical tractability. In our proposal, we do not fix the threshold itself, but a quantile level instead. This allows us for direct model fitting below and above that quantile, and, moreover, we can have a flexible regression structure for both the body and the tail of the data, i.e., this takes care of items (1) and (2) from above. For mixture models, not considered in this paper, we refer to Fung et al. (2021) and the literature therein. Both approaches, composite and mixture models, typically use the EM algorithm for model fitting. Our approach is based on the (simpler) stochastic gradient descent (SGD) algorithm.

Focusing on point (3) from above, there are several ways in extending (or modifying) a GLM. These include, e.g., generalized additive models (GAMs), regression trees and tree boosting, or feed-forward neural (FN) network regression models; we refer to Hastie et al. (2009). Here, we focus on FN network regression models as a straightforward and powerful extension of GLMs. Moreover, we see that an FN network architecture can easily be designed so that it can simultaneously fulfill different estimation tasks. This is a crucial property that we need in our derivations, as we jointly estimate different regression models for the body and tail of the data.

Concerning point (4) from above, quantile regression has gained quite some popularity in the machine learning community to quantify uncertainty, see Meinshausen (2006) and Takeuchi et al. (2006). Quantile regression has been introduced by Koenker and Bassett (1978), and it is widely used in statistical modeling these days; for a recent monograph we refer to Uribe and Guillén (2019). Our proposal extends the approach of Richman (2022) who combines joint estimation of a quantile and the mean within an FN network architecture, see Listing 8 in Richman (2022).

1.2. Proposed method

We consider two different, though closely related problems in this paper. On the one hand, we extend the work of Richman (2022) by proposing an FN network architecture that allows for consistent multiple quantile regression estimation, respecting monotonicity of the quantiles at different levels. On the other hand, this FN network architecture is at the basis of the extension of the joint quantile and ES regression considered in Guillén et al. (2021). The main building blocks in our estimation procedure are strictly consistent scoring functions for our target functionals. These are functions of the estimated model prediction and the response variable which are minimized in expectation by the correctly specified model. Strictly consistent scoring functions are at the core of consistent M-estimation (where M stands for minimization) in a regression setup, see Dimitriadis et al. (2020). If a target functional admits a strictly consistent scoring function, it is called elicitable. The elicitability of the mean and the quantile thus allow for mean and quantile regression, using the squared loss (or a Bregman divergence) or the pinball loss (or a generalized piecewise linear loss), respectively. In model selection and, more generally, forecast ranking, Gneiting (2011a) and Gneiting and Raftery (2007) advocate for the usage of strictly consistent scoring functions since they honor honest and truthful forecasts.

In a first step, we perform multiple quantile regression using the sum of pinball losses as strictly consistent scoring functions. Since quantiles are monotone in their probability levels, we modify the FN network architecture of Richman (2022) such that we can guarantee this monotonicity.

In a second step, we extend the work of Guillén et al. (2021) in several ways. Besides knowing quantiles, one is also interested in estimating ES, which reflects the average outcome beyond the quantile. Since the ES is not elicitable Gneiting (2011a), stand-alone ES regression is generally not possible. Fissler and Ziegel (2016) showed that the pair of ES together with the quantile at the same probability level is elicitable. This paved the way to joint quantile–ES regression, see Dimitriadis and Bayer (2019) and Guillén et al. (2021) for (G)LM approaches. We first extend the result of Fissler and Ziegel (2016), showing the elicitability of the *composite triplet* consisting of a quantile together with the lower and upper ES at the same probability level, and we characterize the entire class of strictly consistent scoring functions for it. Hence, we are in the position to estimate a full composite regression model in a one-step procedure. This is in contrast to the two-step estimation approach of Barendse (2020), first estimating the quantile and then, in a second step, the ES below and above the quantile. Such a two-step estimation approach becomes problematic in the presence of common parameters in the three components of the regression model. Second, we motivate the particular choice of the strictly consistent scoring function in a data-driven manner, striving for efficient parameter estimation. To this end, we use optimality results known for mean regression which are akin to the optimality results derived in Dimitriadis et al. (2020) for the pair of the quantile and ES. The third novelty we provide is the FN network architecture that respects the necessary monotonicity property of the composite triplet in an estimation context. This architecture can directly be fitted using the stochastic gradient descent (SGD) algorithm. This fitting approach does neither require a two-step fitting approach nor the EM algorithm as it is used in related problems. Moreover, our fitting turns out to be robust, and we do not encounter the stability issues as in the two-step approaches and the EM algorithm.

We emphasize that estimating the conditional composite triplet is not only beneficial in a risk management context. In our application, it is used for directly estimating the conditional mean, since this conditional mean can be written as the weighted sum of the lower and the upper ES. Hence, we may directly construct a regression function for the mean, reflecting the potentially different effects of the covariates on the body and on the tail of the conditional distribution; this relates to item (2) in Section 1.1. In particular, we propose this approach for insurance claims modeling, and for this we typically do not choose an extreme quantile level (as in risk management), but rather a quantile level such that we have sufficiently many observations on both sides of the quantile so that we can establish and estimate a regression model.

Finally, we remark that our estimation approach for the composite triplet is semi-parametric. That means that we do not fully estimate or specify the whole conditional distribution function parametrically. We are simply interested in estimating the splicing point – in form of a quantile – as well as the conditional mean below and above this splicing point. However, having a thoughtful choice of a full distributional model may provide us with a canonical candidate of a strictly consistent scoring function for M-estimation.

Organization of this manuscript. In Section 2 we review the concepts of strictly consistent scoring functions used in estimation and forecast evaluation. We recall known characterization results of strictly consistent scores for the mean and for the quantile. We extend the results of Fissler and Ziegel (2016) introducing the class of strictly consistent scoring functions for the composite triplet of the quantile, lower ES and upper ES at the same probability level. In Section 3 we discuss how these quantities can be estimated within a neural network regression framework, and we also discuss asymptotically efficient choices of scoring functions. In Section 4 we give a real data example which demonstrates the suitability of our proposal. Section 5 concludes and provides an outlook to related open problems.

2. Statistical learning

2.1. A review of strictly consistent scoring functions

We start from the decision-theoretic approach of forecast evaluation developed by Gneiting (2011a) and Gneiting and Raftery (2007), also used for M-estimation in Dimitriadis et al. (2020). This provides us with suitable choices of scoring functions for model fitting, model selection, and forecast evaluation. Denote by $\mathcal{F}$ the class of considered distribution functions $F \in \mathcal{F}$. Let $\mathbb{Y} \subseteq \mathbb{R}$ be an interval, a halfline or the whole real line $\mathbb{R}$, such that the support of each $F \in \mathcal{F}$ is contained in $\mathbb{Y}$. Choose an action space $\mathbb{A} \subseteq \mathbb{R}$ from which we select actions $a \in \mathbb{A}$ to estimate a statistics $A(F) \in \mathbb{A}$ of $F \in \mathcal{F}$. That is, we have a functional

$$A : \mathcal{F} \to \mathbb{A} \qquad F \mapsto A(F), \tag{2.1}$$

that we try to estimate. Commonly used functionals are the mean functional $A(F) = \mathbb{E}_F[Y]$ for $Y \sim F \in \mathcal{F}$ and quantiles. Given a probability level $\tau \in (0, 1)$, the τ-quantile of $F \in \mathcal{F}$ is given by the (left-continuous) generalized inverse

$$A(F) = F^{-1}(\tau) = \inf\{y \in \mathbb{R};\ F(y) \geq \tau\}. \tag{2.2}$$

To receive an intuitive understanding we usually speak about a functional (2.1) that attains a single value $A(F)$ in the action space $\mathbb{A}$. Often this functional is obtained by an optimization (M-estimator), or by finding the roots of a given function (Z-estimator). Since, in general, we do not want to assume uniqueness of such a solution, we should understand this functional as a set-valued map

$$A : \mathcal{F} \to \mathcal{P}(\mathbb{A}) \qquad F \mapsto A(F) \subset \mathbb{A},$$

where $\mathcal{P}(\mathbb{A})$ is the power set of $\mathbb{A}$. E.g., the set-valued τ-quantile of a distribution F is given by

$$q_\tau(F) = \{t \in \mathbb{R} :\ F(t-) \leq \tau \leq F(t)\}, \tag{2.3}$$

where $F(t-) = \lim_{x \uparrow t} F(x)$. This defines a closed interval and its lower endpoint corresponds to the left-continuous generalized inverse $F^{-1}(\tau)$ given in (2.2). To keep notation simple, we identify $\widehat{a}$ and $A(F)$ if $A(F) = \{\widehat{a}\}$ is a singleton.

In order to evaluate the accuracy of actions a for the statistics $A(F)$ (for unknown F) we consider a scoring function (also called loss function)

$$L : \mathbb{Y} \times \mathbb{A} \to \mathbb{R}, \qquad (y; a) \mapsto L(y; a).$$

This describes the loss of an action $a \in \mathbb{A}$ if a realization y of $Y \sim F$ materializes. To incentivize truthful forecasts, Gneiting (2011a) advocates that this scoring function L should be strictly consistent for the functional $F \mapsto A(F)$ of interest.

Definition 2.1 *(Strict consistency).* A scoring function $L : \mathbb{Y} \times \mathbb{A} \to \mathbb{R}$ is $\mathcal{F}$-consistent for a given functional $A : \mathcal{F} \to \mathcal{P}(\mathbb{A})$ if $\mathbb{E}_F[|L(Y; a)|] < \infty$ for all $Y \sim F \in \mathcal{F}$ and for all $a \in \mathbb{A}$ and if

$$\mathbb{E}_F[L(Y; \widehat{a})] \ \leq \ \mathbb{E}_F[L(Y; a)], \tag{2.4}$$

for all $Y \sim F \in \mathcal{F}$, $\widehat{a} \in A(F)$ and $a \in \mathbb{A}$. It is strictly $\mathcal{F}$-consistent if it is $\mathcal{F}$-consistent and equality in (2.4) implies that $a \in A(F)$.

Gneiting (2011a) shows that strictly consistent scoring functions are linked to proper scoring rules, which serve to evaluate probabilistic forecasts (i.e., actions a taking the form of probability distributions) and which are discussed in detail in Gneiting and Raftery (2007). On the estimation side, we also need (strict) consistency of scoring functions to obtain consistency of M-estimators, see Chapter 5 in Van der Vaart (1998) and Dimitriadis et al. (2020). This raises the question of which functionals $A : \mathcal{F} \to \mathbb{A}$ admit strictly consistent scoring functions L, i.e., are elicitable.

Definition 2.2 *(Elicitable).* The functional $A : \mathcal{F} \to \mathcal{P}(\mathbb{A})$ is elicitable on a given class of distribution functions $\mathcal{F}$ if there exists a scoring function L that is strictly $\mathcal{F}$-consistent for A.

The general question of elicitability is studied and discussed in detail in Gneiting (2011a). For instance, the mean functional $A(F) = \mathbb{E}_F[Y]$ and the τ-quantile q_τ defined in (2.3) are elicitable; see Theorems 2.3 and 2.4, below. We start with useful properties of scoring functions L that are going to be used in the next theorems. Assume $\mathbb{Y} = \mathbb{A} \subseteq \mathbb{R}$ is an interval with a non-empty interior. We set:

(S0) $L(y; a) \geq 0$ and equality holds if and only if $y = a$;
(S1) $L(y; a)$ is measurable in y and continuous in a;
(S2) The partial derivative $\partial_a L(y; a)$ exists and is continuous in a whenever $a \neq y$.

These regularity conditions can often be weakened. Especially the positivity in (S0) can be relaxed. However, it is a nice property to have in optimizations as it gives us a natural lower bound to the minimal score that can be achieved.

The following result goes back to Savage (1971).

Theorem 2.3 *(Gneiting (2011a), Theorem 7). Let $\mathcal{F}$ be a class of distribution functions on an interval $\mathbb{Y} \subseteq \mathbb{R}$ with finite first moment.*

- *The mean functional $F \mapsto A(F) = \mathbb{E}_F[Y]$ is elicitable relative to $\mathcal{F}$.*
- *Assume the scoring function $L : \mathbb{Y} \times \mathbb{A} \to \mathbb{R}_+$ satisfies (S0)–(S2) for an interval $\mathbb{Y} = \mathbb{A} \subseteq \mathbb{R}$ and that $\mathcal{F}$ is the class of compactly supported distributions on $\mathbb{Y}$. The scoring function L is $\mathcal{F}$-consistent for the mean functional if and only if L is of the form*

$$L(y; a) = L_\phi(y; a) = \phi(y) - \phi(a) - \phi'(a)\,(y - a)\,, \tag{2.5}$$

  for a convex function ϕ with (sub-)gradient ϕ' on $\mathbb{Y}$.
- *If ϕ is strictly convex on $\mathbb{Y}$, then scoring function (2.5) is strictly consistent for the mean functional on the class $\mathcal{F}$ of distribution functions F on $\mathbb{Y}$ for which both $\mathbb{E}_F[Y]$ and $\mathbb{E}_F[\phi(Y)]$ exist and are finite.*

The maps L_ϕ defined in (2.5) are called *Bregman divergences* and, basically, the previous theorem says that all strictly consistent scoring functions for the mean functional are Bregman divergences. The most prominent Bregman divergence arises by setting $\phi(y) = y^2$, yielding the squared loss $L(y; a) = (y - a)^2$. On the other hand, the variance functional $\mathbb{V}_F(Y)$ is not elicitable (Gneiting (2011a) and Osband (1985)), i.e., there does not exist any strictly consistent scoring function for estimating the variance according to a minimization (2.4).

The following elicitability results for quantiles (2.3) originate from Thomson (1979) and Saerens (2000).

Theorem 2.4 *(Gneiting (2011a), Theorem 9). Let $\mathcal{F}$ be a class of distribution functions on an interval $\mathbb{Y} \subseteq \mathbb{R}$ and choose $\tau \in (0, 1)$.*

- *The τ-quantile (2.3) is elicitable relative to $\mathcal{F}$.*
- *Assume the scoring function $L : \mathbb{Y} \times \mathbb{A} \to \mathbb{R}_+$ satisfies (S0)–(S2) for an interval $\mathbb{Y} = \mathbb{A} \subseteq \mathbb{R}$ and that $\mathcal{F}$ is the class of compactly supported distributions on $\mathbb{Y}$. The scoring function L is $\mathcal{F}$-consistent for the τ-quantile (2.3) if and only if L is of the form*

$$L_\tau(y; a) = (g(y) - g(a))\left(\tau - \mathbb{1}_{\{y \leq a\}}\right), \tag{2.6}$$

  for a non-decreasing function g on $\mathbb{Y}$.
- *If g is strictly increasing on $\mathbb{Y}$ and $\mathbb{E}_F[g(Y)]$ exists and is finite for all $F \in \mathcal{F}$, then L defined by (2.6) is strictly $\mathcal{F}$-consistent for the τ-quantile (2.3).*

Members of the class (2.6) are called *generalized piecewise linear losses*, see Gneiting (2011b). Basically, it is this theorem that allows us to consider quantile regression, introduced by Koenker and Bassett (1978), as it tells us that quantiles can be estimated from strictly consistent scoring functions. This is going to be outlined below. There is still the freedom of the choice of the strictly increasing function g. The most commonly used choice is the identity function $g(y) = y$. This then provides us with the so-called *pinball loss* for given probability level $\tau \in (0, 1)$

$$L_\tau(y; a) = (y - a)\left(\tau - \mathbb{1}_{\{y \leq a\}}\right) \;\geq\; 0. \tag{2.7}$$

In view of Theorem 2.4, the choice of the pinball loss requires that $\mathcal{F}$ contains only distributions with a finite first moment.

We introduce scoring functions closely related to the pinball loss (2.7) where the positivity assumption (S0) has been relaxed:

$$S_\tau^-(y; a) = \left(\mathbb{1}_{\{y \leq a\}} - \tau\right) a - \mathbb{1}_{\{y \leq a\}}\, y,$$
$$S_\tau^+(y; a) = \left(1 - \tau - \mathbb{1}_{\{y > a\}}\right) a + \mathbb{1}_{\{y > a\}}\, y,$$

for $y, a \in \mathbb{R}$ and for $\tau \in (0, 1)$. Note that $S_\tau^+(y; a) = S_\tau^-(y; a) + y$, moreover, it holds for the pinball loss

$$L_\tau(y; a) = S_\tau^-(y; a) + \tau\, y = S_\tau^+(y; a) - (1 - \tau)\, y.$$

An immediate consequence from Theorem 2.4 is the following corollary.

Corollary 2.5. *Let $\mathcal{F}$ contain only distributions with finite first moments. Then S_τ^- and S_τ^+ are strictly $\mathcal{F}$-consistent for the τ-quantile (2.3), that is,*

$$q_\tau(F) = \underset{a \in \mathbb{R}}{\arg\min}\, \mathbb{E}_F\left[S_\tau^-(Y; a)\right] = \underset{a \in \mathbb{R}}{\arg\min}\, \mathbb{E}_F\left[S_\tau^+(Y; a)\right] = \underset{a \in \mathbb{R}}{\arg\min}\, \mathbb{E}_F\left[L_\tau(Y; a)\right].$$

Observe that the three functions S_τ^- S_τ^+ and L_τ only differ in terms that do not depend on a, and therefore the argument of these minimizations is the same. The pinball loss $L_\tau(y;a)$ has the advantage to satisfy condition (S0).

2.2. Expected shortfall and the composite triplet

From the previous section we know that we can estimate τ-quantiles (2.3) by minimizing (2.4) under the strictly consistent pinball loss (2.7). In actuarial science we are often interested in also considering the lower ES and the upper ES

$$\mathrm{ES}_\tau^-(Y) = \frac{1}{\tau}\int_0^\tau F^{-1}(p)\,\mathrm{d}p \qquad \text{and} \qquad \mathrm{ES}_\tau^+(Y) = \frac{1}{1-\tau}\int_\tau^1 F^{-1}(p)\,\mathrm{d}p,$$

of a random variable $Y \sim F$. Since ES_τ^- and ES_τ^+ are law-determined, we can interpret them as functionals (2.1). The monotonicity of the generalized inverse $p \mapsto F^{-1}(p)$ immediately yields

$$\mathrm{ES}_\tau^-(F) \le F^{-1}(\tau) \le \mathrm{ES}_\tau^+(F). \tag{2.8}$$

Remark 2.6. If the distribution function F is continuous, in particular, if $F(F^{-1}(\tau)) = \tau$, the ES is equal to the more familiar conditional tail expectation (CTE), see Lemma 2.16 in McNeil et al. (2015). Namely, we have

$$\mathrm{ES}_\tau^-(Y) = \mathrm{CTE}_\tau^-(Y) = \mathbb{E}_F\left[Y\,\middle|Y \le F^{-1}(\tau)\right],$$

and

$$\mathrm{ES}_\tau^+(Y) = \mathrm{CTE}_\tau^+(Y) = \mathbb{E}_F\left[Y\,\middle|Y > F^{-1}(\tau)\right].$$

If we can estimate these two CTEs, this allows us to fit a composite model to Y, with the quantile $F^{-1}(\tau)$ giving the splicing point where the lower and upper parts are concatenated. Furthermore, this allows us to estimate the mean of Y from

$$\mathbb{E}[Y] = \tau\,\mathrm{ES}_\tau^-(Y) + (1-\tau)\mathrm{ES}_\tau^+(Y).$$

We come back to this, below, in a regression model context (3.19). Typically, we will not choose an extreme probability level τ, but one that allows us to reliably estimate the lower and upper ES, so that we receive an accurate mean estimate.

The ES, considered as functionals, turn out not to be elicitable on families of distributions $\mathcal{F}$ which offer the necessary flexibility in most modeling situations; see Gneiting (2011a) and Weber (2006). Thus, there is no strictly consistent scoring function suitable for M-estimation of ES. Fissler and Ziegel (2016) have proved that $\mathrm{ES}_\tau^-(Y)$ is *jointly* elicitable with the τ-quantile $F^{-1}(\tau)$, the latter also being called Value-at-Risk (VaR) in risk management. We are going to extend this result in Theorem 2.8, establishing the elicitability of the *composite triplet* $(\mathrm{ES}_\tau^-, q_\tau, \mathrm{ES}_\tau^+)$. We first need the following lemma.

Lemma 2.7. *For any distribution F with finite first moment, it holds that*

$$\mathrm{ES}_\tau^-(F) = -\frac{1}{\tau}\min_{v\in\mathbb{R}} \mathbb{E}_F\left[S_\tau^-(Y;v)\right] \qquad \textit{and} \qquad \mathrm{ES}_\tau^+(F) = \frac{1}{1-\tau}\min_{v\in\mathbb{R}} \mathbb{E}_F\left[S_\tau^+(Y;v)\right].$$

Proof. This follows directly along the lines of Lemmas 2.3 and 3.3 in Embrechts and Wang (2015). □

Theorem 2.8. *Choose $\tau \in (0,1)$ and let $\mathcal{F}$ contain only distributions with finite first moments and supported on an interval $\mathbb{Y} \subseteq \mathbb{R}$. The scoring function $L\colon \mathbb{Y} \times \mathbb{Y}^3 \to \mathbb{R}_+$ of the form*

$$\begin{aligned} L(y;e^-,v,e^+) &= (g(y)-g(v))\left(\tau - \mathbb{1}_{\{y\le v\}}\right) \\ &\quad + \left\langle \nabla\Phi(e^-,e^+), \begin{pmatrix} e^- + \frac{1}{\tau}S_\tau^-(y;v) \\ e^+ - \frac{1}{1-\tau}S_\tau^+(y;v)\end{pmatrix}\right\rangle - \Phi(e^-,e^+) + \Phi(y,y), \end{aligned} \tag{2.9}$$

is (strictly) $\mathcal{F}$-consistent for the composite triplet $(\mathrm{ES}_\tau^-, q_\tau, \mathrm{ES}_\tau^+)$ if $\Phi\colon \mathbb{Y}^2 \to \mathbb{R}$ is (strictly) convex with sub-gradient $\nabla\Phi$ such that for $g\colon \mathbb{Y} \to \mathbb{R}$ and for all $(e^-,e^+) \in \mathbb{Y}^2$ the function

$$G_{e^-,e^+}\colon \mathbb{Y} \to \mathbb{R}, \quad v \mapsto g(v) + \tfrac{1}{\tau}\partial_1\Phi(e^-,e^+)v - \tfrac{1}{1-\tau}\partial_2\Phi(e^-,e^+)v \tag{2.10}$$

is (strictly) increasing, and if $\mathbb{E}_F[|g(Y)|] < \infty$, $\mathbb{E}_F[|\Phi(Y,Y)|] < \infty$ for all $Y \sim F \in \mathcal{F}$. Moreover, $L(y;e^-,v,e^+) \ge L(y;y,y,y) = 0$.

The intuitive interpretation is that the score in (2.9) is a combination of a Bregman divergence (second line) and a generalized piecewise linear loss (first line in combination with S_τ^- and S_τ^+). This structure will be exploited in the proof.

Proof of Theorem 2.8. First, fix some $v \in \mathbb{Y}$. The map

$$(y;e^-,e^+) \mapsto L(y;e^-,v,e^+) = \left\langle \nabla\Phi(e^-,e^+), \begin{pmatrix} e^- + \frac{1}{\tau}S_\tau^-(y;v) \\ e^+ - \frac{1}{1-\tau}S_\tau^+(y;v)\end{pmatrix}\right\rangle - \Phi(e^-,e^+) + b_v(y),$$

where the remainder $b_v(y)$ does not depend on (e^-, e^+), is a Bregman divergence. Hence, if Φ is (strictly) convex, it is (strictly) $\mathcal{F}$-consistent for the functional

$$F \mapsto \big(-\tfrac{1}{\tau}\mathbb{E}_F\big[S_\tau^-(Y;v)\big], \tfrac{1}{1-\tau}\mathbb{E}_F\big[S_\tau^+(Y;v)\big]\big).$$

Second, for fixed $(e^-, e^+) \in \mathbb{Y}^2$, the map

$$(y;v) \mapsto L(y;e^-,v,e^+) = (\mathbb{1}_{\{y\le v\}} - \tau)G_{e^-,e^+}(v) - \mathbb{1}_{\{y\le v\}}G_{e^-,e^+}(y) + b_{e^-,e^+}(y),$$

where the remainder $b_{e^-,e^+}(y)$ does not depend on v, is a generalized piecewise linear loss (2.6) not necessarily satisfying the positivity (S0). Hence, if G_{e^-,e^+} is (strictly) increasing, it is (strictly) $\mathcal{F}$-consistent for q_τ.

Combining these two observations yields the (strict) $\mathcal{F}$-consistency of L.

Finally, $L(y;y,y,y) = 0$ can be verified by a direct computation, and the non-negativity of L follows from its consistency and the fact that for a random variable Y which deterministically equals the constant $y \in \mathbb{R}$, it holds that $\mathrm{ES}_\tau^-(Y) = \mathrm{ES}_\tau^+(Y) = y$ and $q_\tau(Y) = \{y\}$. □

Theorem 2.8 extends Theorem 1 in Frongillo and Kash (2021) asserting that an elicitable functional (in our case the τ-quantile) is jointly elicitable with finitely many associated Bayes risks (here corresponding to ES_τ^- and ES_τ^+).

Theorem 2.9, below, shows that the scores of the form (2.9) are basically the only $\mathcal{F}$-consistent scores for $(\mathrm{ES}_\tau^-, q_\tau, \mathrm{ES}_\tau^+)$. The argument – which can be found in its entirety along with the corresponding assumptions in the Supplementary Section A – uses Osband's principle, which originates from Kent Osband's seminal thesis Osband (1985); see Gneiting (2011a) for an intuitive exposition and Fissler and Ziegel (2016) for a precise technical formulation. It exploits first- and second-order conditions stemming from the optimization problem of (strict) consistency (2.4). Hence, we need to impose smoothness conditions on the expected score, which play the role of condition (S2), but are slightly weaker. This is reflected in Assumption A.6, and the fact that we work within a subclass of $\mathcal{F} \subseteq \mathcal{F}_{\text{cont}}^\tau$, the class of distribution functions on $\mathbb{R}$ which are continuously differentiable and whose τ-quantiles are singletons. The second kind of conditions is richness conditions on the underlying class of distributions $\mathcal{F}$. On the one hand, the first- and second-order conditions only yield local assertions about the expected scores. Hence, $\mathcal{F}$ needs to be rich enough such that the functional maps surjectively to the action domain considered. Recalling monotonicity condition (2.8), this means that we can only provide conditions on action spaces contained in $\{(a_1,a_2,a_3) \in \mathbb{R}^3 : a_1 \le a_2 \le a_3\}$. Second, the richness conditions ensure that the functional 'varies sufficiently' such that it can be distinguished from any other functional.[1] This is reflected in Assumptions A.2 and A.3. Finally, since Osband's principle actually characterizes the *gradient* of the *expected* score, one needs to be in the position to integrate this gradient (Assumption A.5) and to approximate the pointwise values of the score with expectations (Assumption A.4).

Theorem 2.9. *Let $\tau \in (0,1)$ and $\mathcal{F} \subseteq \mathcal{F}_{\text{cont}}^\tau$. Let $L\colon \mathbb{R}\times\mathbb{A} \to \mathbb{R}$, $\mathbb{A} \subseteq \{(a_1,a_2,a_3) \in \mathbb{R}^3 : a_1 \le a_2 \le a_3\}$ be an $\mathcal{F}$-consistent scoring function for the composite triplet $(\mathrm{ES}_\tau^-, q_\tau, \mathrm{ES}_\tau^+)$, satisfying Assumptions A.2–A.5 and $L(y;y,y,y) = 0$ for all $y \in \mathbb{R}$ such that $(y,y,y) \in \mathbb{A}$. Then L is necessarily of the form (2.9) almost everywhere where Φ is convex and where for any fixed $(e^-, e^+) \in \mathbb{R}^2$ such that there is a $v \in \mathbb{R}$ with $(e^-, v, e^+) \in \mathbb{A}$ the function G_{e^-,e^+} in (2.10) is increasing.*

Supplementary Section A provides all technical details. In particular, the proof of Theorem 2.9 can be found in Subsection A.3.

2.3. Particular choices of scoring functions for the composite triplet

There remain the choices of the functions g and Φ in (2.9). Especially, the choice of the strictly convex function Φ such that G_{e^-,e^+} in (2.10) is strictly increasing is not obvious. We discuss the following particularly convenient three choices for given $\tau \in (0,1)$

$$\Phi(e^-,e^+) = \phi_-(e^-) + \phi_+(e^+), \tag{2.11}$$

$$\Phi(e^-,e^+) = \phi(\tau e^- + (1-\tau)e^+) + \phi_+(e^+), \tag{2.12}$$

$$\Phi(e^-,e^+) = \phi(\tau e^- + (1-\tau)e^+) + \phi_-(e^-), \tag{2.13}$$

where $\phi, \phi_+, \phi_-\colon \mathbb{Y} \to \mathbb{R}$ are strictly convex and the (sub-)gradients satisfy $\phi'_+ < 0$, $\phi'_- > 0$. Moreover, if we choose $g\colon \mathbb{Y} \to \mathbb{R}$ to be an increasing function (not necessarily strictly increasing), we obtain for G_{e^-,e^+}, defined in (2.10),

$$G_{e^-,e^+}(v) = \begin{cases} g(v) + \frac{1}{\tau}\phi'_-(e^-)v - \frac{1}{1-\tau}\phi'_+(e^+)v & \text{for } \Phi \text{ in (2.11)},\\ g(v) - \frac{1}{1-\tau}\phi'_+(e^+)v & \text{for } \Phi \text{ in (2.12)},\\ g(v) + \frac{1}{\tau}\phi'_-(e^-)v & \text{for } \Phi \text{ in (2.13)}.\end{cases}$$

Since $\phi'_+ < 0$ and $\phi'_- > 0$, G_{e^-,e^+} is strictly increasing, even if g is constant. This yields the following three types of scores: For Φ in (2.11) we get

$$\begin{aligned} L(y;e^-,v,e^+) = {} & (g(y) - g(v))\left(\tau - \mathbb{1}_{\{y\le v\}}\right) \\ & + \phi'_-(e^-)\left(e^- + \tfrac{1}{\tau}S_\tau^-(y;v)\right) - \phi_-(e^-) + \phi_-(y) \\ & + \phi'_+(e^+)\left(e^+ - \tfrac{1}{1-\tau}S_\tau^+(y;v)\right) - \phi_+(e^+) + \phi_+(y). \end{aligned} \tag{2.14}$$

[1] E.g., on the class of symmetric distributions with positive densities, the mean and the median coincide and cannot be distinguished. Such phenomena need to be excluded.

Similarly, for Φ in (2.12) we receive

$$\begin{aligned} L(y;e^-,v,e^+) = {} & (g(y)-g(v))\left(\tau - \mathbb{1}_{\{y\le v\}}\right) \\ & + \phi_+'(e^+)\left(e^+ - \tfrac{1}{1-\tau} S_\tau^+(y;v)\right) - \phi_+(e^+) + \phi_+(y) \\ & + \phi'(\tau e^- + (1-\tau)e^+)\left(\tau e^- + (1-\tau)e^+ - y\right) - \phi(\tau e^- + (1-\tau)e^+) + \phi(y), \end{aligned} \tag{2.15}$$

and finally for (2.13)

$$\begin{aligned} L(y;e^-,v,e^+) = {} & (g(y)-g(v))\left(\tau - \mathbb{1}_{\{y\le v\}}\right) \\ & + \phi_-'(e^-)\left(e^- + \tfrac{1}{\tau} S_\tau^-(y;v)\right) - \phi_-(e^-) + \phi_-(y) \\ & + \phi'(\tau e^- + (1-\tau)e^+)\left(\tau e^- + (1-\tau)e^+ - y\right) - \phi(\tau e^- + (1-\tau)e^+) + \phi(y). \end{aligned} \tag{2.16}$$

Scores of the form (2.14) can directly be interpreted to be the sum of scores for the pair $(q_\tau, \mathrm{ES}_\tau^-)$ as deduced in Fissler and Ziegel (2016) and for the pair $(q_\tau, \mathrm{ES}_\tau^+)$ as derived in Nolde and Ziegel (2017). On the other hand, (2.15) can be deduced from the sum of a scoring function for $(q_\tau, \mathrm{ES}_\tau^+)$ and for the mean, using the so-called *revelation principle*. This principle originates from Osband's thesis Osband (1985) and it has been made rigorous in Gneiting, 2011a, Theorem 4. It asserts that any bijection of an elicitable functional is elicitable and it makes the corresponding strictly consistent scoring functions explicit. In the case of (2.14), this bijection is

$$(\mathrm{ES}_\tau^-, q_\tau, \mathrm{ES}_\tau^+) \mapsto (q_\tau, \mathrm{ES}_\tau^+, \tau\mathrm{ES}_\tau^- + (1-\tau)\mathrm{ES}_\tau^+) = (q_\tau, \mathrm{ES}_\tau^+, \mathbb{E}).$$

For the scores in (2.16), the corresponding bijection reads similar.

The next section shows how to use strictly consistent scoring functions for parameter estimation via M-estimation in a regression context. Here, the (strict) $\mathcal{F}$-consistency of the corresponding score is crucial to obtain the consistency of the M-estimator, i.e., that the M-estimator converges in probability to the true value as the sample size goes to infinity; see Dimitriadis et al. (2020). Knowing that the estimator is consistent, it is of interest to have a fast convergence, i.e., an efficient estimator. Under asymptotic normality of the estimator, a more efficient estimator has a strictly smaller asymptotic covariance matrix.[2] In the absence of any explanatory information (that is, when estimating an intercept-only model), the choice of the strictly consistent scoring function is immaterial since the intercept estimators under different strictly consistent scores will always coincide on finite samples and correspond to the functional of the empirical distribution function. In a more interesting regression scenario, however, i.e., using feature information, the question of efficiency becomes important, this is discussed in our setup in Section 3.4, below.

3. Deep quantile and deep composite model regressions

3.1. Quantile regression

The previous sections have discussed estimation theory of functionals of (unknown) distribution functions F. We now lift this framework to a regression context where random variables Y are supported by covariates (feature information). Assume that the random variable Y is established with feature information $\boldsymbol{X} \in \mathcal{X} \subset \{1\} \times \mathbb{R}^q$, where $\mathcal{X}$ is the feature space of all potential explanatory variables $\boldsymbol{X}$. We assume for a datum $(Y, \boldsymbol{X})$ that the conditional distribution of Y, given $\boldsymbol{X}$, is described by a distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$, $\boldsymbol{x} \in \mathcal{X}$, or in short $F_{Y|\boldsymbol{x}}$, and the conditional τ-quantile of this random variable is given by the left-continuous generalized inverse

$$F_{Y|\boldsymbol{x}}^{-1}(\tau) = \inf\left\{y \in \mathbb{R};\ F_{Y|\boldsymbol{x}}(y) \ge \tau\right\}.$$

Note that we label distributions with subscripts $Y|\boldsymbol{x}$, now, to clearly indicate that we are considering the conditional distribution of Y, given feature information $\boldsymbol{X} = \boldsymbol{x} \in \mathcal{X}$. This gives us the existence of a regression function $Q_\tau : \mathcal{X} \to \mathbb{R}$ such that

$$\boldsymbol{x} \mapsto Q_\tau(\boldsymbol{x}) = F_{Y|\boldsymbol{x}}^{-1}(\tau).$$

Quantile regression tries to determine this regression function Q_τ from a given function class $\mathcal{Q}$ based on i.i.d. observations $(Y_i, \boldsymbol{X}_i)$, $1 \le i \le n$. The classical approach of Koenker and Bassett (1978) makes a GLM assumption by postulating the existence of a strictly monotone and smooth link function $h : \mathbb{R} \to \mathbb{R}$ such that

$$Q_\tau(\boldsymbol{x}) = h^{-1}\langle \boldsymbol{\beta}_0, \boldsymbol{x}\rangle, \tag{3.1}$$

with regression parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{q+1}$ and $\langle\cdot,\cdot\rangle$ denoting the scalar product in the Euclidean space $\mathbb{R}^{q+1}$. In this case the function class $\mathcal{Q}$ is parametrized by a regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$, that we try to optimally determine from i.i.d. data $(Y_i, \boldsymbol{X}_i)$, $1 \le i \le n$. The choice of the link h acts as a hyper-parameter, that is not part of the optimization process.

We can choose a strictly consistent scoring function L_τ (2.6) for the τ-quantile $F_{Y|\boldsymbol{x}}^{-1}(\tau)$ and estimate the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ by

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}\in\mathbb{R}^{q+1}}{\arg\min}\, \mathbb{E}\left[L_\tau\left(Y; h^{-1}\langle\boldsymbol{\beta}, \boldsymbol{X}\rangle\right)\right] = \underset{\boldsymbol{\beta}\in\mathbb{R}^{q+1}}{\arg\min}\, \mathbb{E}\left[\mathbb{E}\left[L_\tau\left(Y; h^{-1}\langle\boldsymbol{\beta}, \boldsymbol{X}\rangle\right)|\boldsymbol{X}\right]\right], \tag{3.2}$$

[2] As usual, we mean this with respect to the Loewner order. That is, a covariance matrix A is strictly smaller than a covariance matrix B of the same dimension if $A \ne B$ and if $B - A$ is positive semi-definite.

subject to existence. Under the assumption of a correctly specified model (3.1), for ν-a.e. $\boldsymbol{x} \in \mathcal{X}$ (if ν denotes the distribution of $\boldsymbol{X}$), the inner expectation $\mathbb{E}\left[L_\tau\left(Y; h^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle\right) | \boldsymbol{X} = \boldsymbol{x}\right]$ is minimized by the true regression parameter $\boldsymbol{\beta}_0$, invoking the strict consistency of L_τ, applied to the conditional distribution $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ in (2.4). Due to the monotonicity of the expectation, also the outer expectation is thus minimized by $\boldsymbol{\beta}_0$. A generalized linear model typically does not correctly represent the true regression function well. However, thanks to the universality theorems for FN networks (see Section 7.2.2 of Wüthrich and Merz (2023)), a sufficiently complex FN network may arbitrarily well approximate the true regression function subject to regularity conditions on the true regression function; see next subsection. Hence, implicitly exploiting the property of self-calibration of consistent scoring functions (Subsection 2.2 in Fissler and Ziegel (2019)) strictly consistent loss functions allow to asymptotically recover the true approximation function.

Typically, we do not know the true distribution function and, therefore, cannot explicitly evaluate the right-hand side of the above optimization. Using an empirical version based on i.i.d. data $(Y_i, \boldsymbol{X}_i)$, $1 \le i \le n$, motivates the M-estimator

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta} \in \mathbb{R}^{q+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L_\tau\left(Y_i; h^{-1}\langle \boldsymbol{\beta}, \boldsymbol{X}_i\rangle\right). \tag{3.3}$$

3.2. Deep quantile regression

In practice, the GLM structure (3.1) is often too restrictive. This emphasizes to use a more flexible function class $\mathcal{Q}$. FN networks provide the building blocks for such a more flexible function class. This motivates *deep quantile regression* which has been introduced to the actuarial literature by Richman (2022). We replace (3.1) by

$$Q_\tau(\boldsymbol{x}) = h^{-1}\langle \boldsymbol{\beta}_0, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle, \tag{3.4}$$

where $\boldsymbol{z}^{(d:1)} : \mathcal{X} \to \{1\} \times \mathbb{R}^{r_d}$ is an FN network of depth $d \in \mathbb{N}$, regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{r_d+1}$ and link h. The FN network $\boldsymbol{z}^{(d:1)}$ is a composition of d FN layers

$$\boldsymbol{x} \in \mathcal{X} \mapsto \boldsymbol{z}^{(d:1)}(\boldsymbol{x}) = \left(\boldsymbol{z}^{(d)} \circ \cdots \circ \boldsymbol{z}^{(1)}\right)(\boldsymbol{x}) \in \{1\} \times \mathbb{R}^{r_d},$$

with FN layers $\boldsymbol{z}^{(m)}$ for $1 \le m \le d$ involving further parameters $\boldsymbol{w}_j^{(m)} \in \mathbb{R}^{r_{m-1}+1}$, and with $r_{m-1}+1$ describing the input dimension to FN layer $\boldsymbol{z}^{(m)}$; for a detailed exhibition of FN networks we refer to Section 7.2 in Wüthrich and Merz (2023). Altogether this FN network approach (3.4) is parametrized by

$$\boldsymbol{\vartheta} = (\boldsymbol{w}_1^{(1)}, \ldots, \boldsymbol{w}_{r_d}^{(d)}, \boldsymbol{\beta}) \in \mathbb{R}^r \qquad \text{of dimension } r = \sum_{m=1}^{d} r_m(r_{m-1}+1) + (r_d+1).$$

Thus, we fix a depth $d \in \mathbb{N}$, FN layer dimensions $r_1, \ldots, r_d \in \mathbb{N}$, the activation functions in the FN layers and link function h (as hyper-parameters), then the function class $\mathcal{Q}$ is parametrized by $\boldsymbol{\vartheta}$, and the "optimal" member for i.i.d. data $(Y_i, \boldsymbol{X}_i)$, $1 \le i \le n$, is found by

$$\widehat{\boldsymbol{\vartheta}}_\tau = \underset{\boldsymbol{\vartheta} \in \mathbb{R}^{r}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L_\tau\left(Y_i; h^{-1}\langle \boldsymbol{\beta}, \boldsymbol{z}^{(d:1)}(\boldsymbol{X}_i)\rangle\right). \tag{3.5}$$

On purpose "optimal" has been written in quotation marks. On a finite sample of size n the solution to (3.5) will likely in-sample overfit to the learning data $\mathcal{L} = (Y_i, \boldsymbol{X}_i)_{1 \le i \le n}$ because already small networks are fairly flexible. Therefore, this model is usually fit with a SGD algorithm that explores an *early stopping rule*, i.e., that selects an estimate $\widehat{\boldsymbol{\vartheta}}_\tau$ that describes the systematic effects in the data $\mathcal{L}$ and not the noisy part. This fitting approach is state-of-the-art, and it is described in detail in Section 7.2.3 of Wüthrich and Merz (2023). Therefore, we will not repeat it here.

3.3. Deep multiple quantile regression

Often we do not want to estimate quantiles for only one probability level, but we would like to study quantiles at different levels. For illustrative purposes we consider two probability levels $0 < \tau_1 < \tau_2 < 1$, and a generalization to more than two probability levels is straightforward. A naive way is to individually estimate the regression functions $\boldsymbol{x} \mapsto Q_{\tau_l}(\boldsymbol{x}) = F^{-1}_{Y|\boldsymbol{x}}(\tau_l)$ for $l = 1, 2$ using (3.4) and (3.5). We call this a naive approach because these individual estimations may violate the monotonicity property of quantiles, i.e., for all $\boldsymbol{x}$ we require

$$Q_{\tau_1}(\boldsymbol{x}) \le Q_{\tau_2}(\boldsymbol{x}), \tag{3.6}$$

which may be violated in the naive approach. To simplify this outline, we assume that the random variable Y is positive, a.s., which implies that the generalized inverse $\tau \mapsto F^{-1}_{Y|\boldsymbol{x}}(\tau) > 0$ has a positive range. This motivates the choice of a link function h with positive support $\mathbb{R}_+$. For enforcing the monotonicity of quantiles for different levels (3.6), we propose to jointly model these quantiles. For our first proposal we choose two link functions h and h_+ being both positively supported. In analogy to (3.4), this motivates *joint deep quantile regression* for probability levels $\tau_1 < \tau_2$

$$\begin{aligned}\boldsymbol{x} \mapsto &\left(Q_{\tau_1}(\boldsymbol{x}),\ Q_{\tau_2}(\boldsymbol{x})\right)^\top \\ &= \left(h^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle,\ h^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle + h_+^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle\right)^\top \in \mathbb{R}_+^2,\end{aligned} \tag{3.7}$$

for a network parameter $\boldsymbol{\vartheta} = (\boldsymbol{w}_1^{(1)}, \dots, \boldsymbol{w}_{r_d}^{(d)}, \boldsymbol{\beta}_{\tau_1}, \boldsymbol{\beta}_{\tau_2})^\top$. Thus, we choose a common deep FN network $\boldsymbol{z}^{(d:1)}$ that is shared by both quantiles, and the bigger quantile $Q_{\tau_2}(\boldsymbol{x})$ is modeled by a positive difference $h_+^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle \geq 0$ to the smaller quantile $Q_{\tau_1}(\boldsymbol{x})$. We call (3.7) an *additive approach* with base level $Q_{\tau_1}(\boldsymbol{x})$, and in network jargon we say that the FN network learns a common representation $\boldsymbol{z}_i = \boldsymbol{z}^{(d:1)}(\boldsymbol{x}_i)$ of features $\boldsymbol{x}_i$, $1 \leq i \leq n$, which is then used in the two GLMs

$$\boldsymbol{z}_i \;\mapsto\; \left(Q_{\tau_1}(\boldsymbol{z}_i),\; Q_{\tau_2}(\boldsymbol{z}_i)\right)^\top = \left(h^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}_i\rangle,\; h^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}_i\rangle + h_+^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}_i\rangle\right)^\top \;\in \mathbb{R}_+^2.$$

Alternatively, for positive random variables Y, a.s., we can choose the upper quantile $Q_{\tau_2}(\boldsymbol{x})$ as base level, and to ensure positivity we can multiplicatively decrease this upper quantile. For this we choose the sigmoid function for $h_\sigma^{-1}(x) = (1+\exp(-x))^{-1} \in (0,1)$ which motivates the *multiplicative approach* for probability levels $\tau_1 < \tau_2$

$$\begin{aligned} \boldsymbol{x} &\mapsto \left(Q_{\tau_1}(\boldsymbol{x}),\; Q_{\tau_2}(\boldsymbol{x})\right)^\top \\ &= \left(h_\sigma^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle\, h^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle,\; h^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle\right)^\top \;\in \mathbb{R}_+^2. \end{aligned} \tag{3.8}$$

Also in this case monotonicity is guaranteed because by assumption $h_\sigma^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle \in (0,1)$.

Since the learned representations $\boldsymbol{z}_i = \boldsymbol{z}^{(d:1)}(\boldsymbol{x}_i)$ need to fit both quantiles simultaneously we need to learn these representations jointly. This motivates the optimization problem under regression assumption (3.7) or (3.8), respectively, and up to over-fitting (see discussion after (3.5))

$$\widehat{\boldsymbol{\vartheta}}_{\tau_1,\tau_2} \;=\; \underset{\boldsymbol{\vartheta}}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n L_{\tau_1}\left(Y_i; Q_{\tau_1}(\boldsymbol{X}_i)\right) + L_{\tau_2}\left(Y_i; Q_{\tau_2}(\boldsymbol{X}_i)\right). \tag{3.9}$$

Remarks 3.1.

- We remark that both the multiplicative and the additive approach described above ensure monotonicity of the quantiles and rule out quantile crossings *by construction* while still keeping the flexibility that the covariates may influence the different quantiles differently. The literature has addressed the issue of quantile crossings in several different ways. In He (1997), a location-scale model $Y = \mu(\boldsymbol{X}) + \sigma(\boldsymbol{X})e$ with an error e being independent of $\boldsymbol{X}$ is fitted. This results in $Q_\tau(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})F_e^{-1}(\tau)$. This approach implies that all quantiles depend on the covariates grossly in the same manner, apart from one multiplicative factor. This contradicts our very motivation that, empirically, covariates often have a profoundly different influence on the body and on the tail of the distribution, resulting in a different influence for different quantiles. Chernozhukov et al. (2010) consider a rearrangement approach, which is applied as an additional post-processing step to the estimated conditional quantiles. Liu and Wu (2011) consider embeddings into reproducing kernel Hilbert spaces with positive kernels resulting in a constraint optimization approach. On the other hand, the penalization method used by Kellner et al. (2022) promotes monotonicity rather than enforcing it by construction. Despite the wealth of suggestions, we are unaware of any other instance of our simple, yet powerful proposal to enforce monotonicity of quantiles while still keeping the flexibility of different influences of covariates on the different quantiles.
- Our proposals (3.7) and (3.8) only use one single deep FN network $\boldsymbol{x} \mapsto \boldsymbol{z} \overset{\text{def.}}{=} \boldsymbol{z}^{(d:1)}(\boldsymbol{x})$, and $\boldsymbol{z}$ has the interpretation of a new representation of the original covariates $\boldsymbol{x}$. Focusing on (3.7), this new representation $\boldsymbol{z}$ is simultaneously used to model the lower quantile $Q_{\tau_1}(\boldsymbol{x})$ by $h^{-1}\langle \boldsymbol{\beta}_{\tau_1}, \boldsymbol{z}\rangle$ and the upper quantile $Q_{\tau_2}(\boldsymbol{x})$ by a positive spread $h_+^{-1}\langle \boldsymbol{\beta}_{\tau_2}, \boldsymbol{z}\rangle$ to the lower quantile. If the chosen deep FN network $\boldsymbol{z}^{(d:1)}$ is sufficiently flexible, it will provide us with a reasonable representation $\boldsymbol{z}$ of $\boldsymbol{x}$ for both of these two regression tasks. Alternatively, we could also use two different deep FN networks that provide two different representations $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ of $\boldsymbol{x}$, and these two (parallel) deep FN networks become dependent through a joint parameter fitting (3.9).
- Regression structures of the form (3.7)-(3.8) are also called *multi-output networks* that are used to simultaneously perform different prediction tasks. An advantage of our multi-output network approach over the (naive) individual estimation of different quantiles is that the multi-output approach only uses one single model (to be stored), whereas the individual estimation approach results in multiple models (to be maintained).

3.4. Deep composite model regression

A deep composite regression model now only requires little changes compared to the deep multiple quantile regression. Again, we assume that $Y > 0$, a.s. We then aim at estimating the composite triplet

$$\mathrm{ES}_\tau^-(Y|\boldsymbol{x}) \;\leq\; F_{Y|\boldsymbol{x}}^{-1}(\tau) \;\leq\; \mathrm{ES}_\tau^+(Y|\boldsymbol{x}), \tag{3.10}$$

where we highlight the conditional structure of Y, given $\boldsymbol{X} = \boldsymbol{x}$. Thus, in addition to the quantile regression of the previous sections, we aim at estimating regression functions $\boldsymbol{x} \mapsto E_\tau^s(\boldsymbol{x}) = \mathrm{ES}_\tau^s(Y|\boldsymbol{x})$ for $s \in \{-,+\}$. Thanks to Theorem 2.8, we know that we can jointly estimate the triplet $(E_\tau^-(\boldsymbol{x}),\; Q_\tau(\boldsymbol{x}),\; E_\tau^+(\boldsymbol{x}))$ using the strictly consistent scoring function (2.9); see the discussion after (3.2).

Since the composite triplet (3.10) also obeys a natural monotonicity relation just like quantiles at different levels (3.6), we choose positively supported link functions h and h_+ in analogy to (3.7). This motivates *deep composite triplet model regression*

$$\begin{aligned} \boldsymbol{x} &\mapsto \left(E_\tau^-(\boldsymbol{x}),\; Q_\tau(\boldsymbol{x}),\; E_\tau^+(\boldsymbol{x})\right)^\top \\ &= \Big(h^{-1}\langle \boldsymbol{\beta}_1, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle,\; h^{-1}\langle \boldsymbol{\beta}_1, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle + h_+^{-1}\langle \boldsymbol{\beta}_2, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle, \end{aligned} \tag{3.11}$$

$$h^{-1}\langle \boldsymbol{\beta}_1, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle + h_+^{-1}\langle \boldsymbol{\beta}_2 \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle + h_+^{-1}\langle \boldsymbol{\beta}_3, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle \Big)^\top \in \mathbb{R}_+^3,$$

for network parameter $\boldsymbol{\vartheta} = (\boldsymbol{w}_1^{(1)}, \dots, \boldsymbol{w}_{r_d}^{(d)}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)^\top$. Again, we choose a common deep FN network $\boldsymbol{z}^{(d:1)}$ that is shared by the τ-quantile and the lower and upper ES. Alternatively to (3.11), we could also use a multiplicative approach similar to (3.8). We obtain the following M-estimator

$$\widehat{\boldsymbol{\vartheta}}_\tau = \underset{\boldsymbol{\vartheta}}{\arg\min} \frac{1}{n}\sum_{i=1}^n L\left(Y_i; E_\tau^-(\boldsymbol{X}_i),\ Q_\tau(\boldsymbol{X}_i),\ E_\tau^+(\boldsymbol{X}_i)\right), \tag{3.12}$$

where L is a strictly consistent score given in (2.9).

Mutatis mutandis, the same comments apply as in Remark 3.1.

As already discussed at the end of Subsection 2.3, in a regression context, the choice of the strictly consistent scoring function influences the (asymptotic) variance of the M-estimator. Therefore, we would like to provide some guidance on how to choose a score of the form (2.9) in a data driven manner. For simplicity, we will focus on scores of the form (2.14) and (2.15), mainly discussing the choice of ϕ, ϕ_- and ϕ_+. Moreover, motivated by modeling claim sizes, we assume that $Y > 0$, a.s., such that $\mathbb{Y} = (0, \infty)$.

To this end, first recall a classical efficiency result in the context of mean regression; see Newey and McFadden (1994). If $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the estimable regression function and $\sigma(\boldsymbol{x})^2 = \mathbb{V}(Y|\boldsymbol{X} = \boldsymbol{x})$ is the conditional variance, the most efficient Bregman score (2.5) should satisfy

$$\phi''\big(\mu(\boldsymbol{x})\big) = \frac{c}{\sigma(\boldsymbol{x})^2}, \tag{3.13}$$

for some $c > 0$ and for all $\boldsymbol{x} \in \mathcal{X}$. Since the third line of (2.15) is basically a Bregman score for the mean, we use condition (3.13) to come up with a choice for ϕ in (2.15). For ϕ_+ in (2.15) and (2.14), we suggest to exploit a similar relation for the *truncated* variance (recalling that ES_τ^+ is also a truncated mean). In particular, Theorem 4.3 in Dimitriadis et al. (2020) suggests

$$\phi_+''\big(E_\tau^+(\boldsymbol{x})\big) = \frac{c_+}{\sigma_\tau^+(\boldsymbol{x})^2},$$

for some $c_+ > 0$ and for all $\boldsymbol{x} \in \mathcal{X}$, where $\sigma_\tau^+(\boldsymbol{x})^2 = \mathbb{V}(Y|Y > F_{Y|\boldsymbol{x}}^{-1}(\tau),\ \boldsymbol{X} = \boldsymbol{x})$. Similarly,

$$\phi_-''\big(E_\tau^-(\boldsymbol{x})\big) = \frac{c_-}{\sigma_\tau^-(\boldsymbol{x})^2},$$

for some $c_- > 0$ and for all $\boldsymbol{x} \in \mathcal{X}$, where $\sigma_\tau^-(\boldsymbol{x})^2 = \mathbb{V}(Y|Y \le F_{Y|\boldsymbol{x}}^{-1}(\tau),\ \boldsymbol{X} = \boldsymbol{x})$.

To render this approach feasible, we suggest the following. Fit a pre-estimate for the mean $\widehat{\mu}(\boldsymbol{x})$, using a strictly consistent score for mean estimation, and an FN network for $\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x})$. This allows us to study the squared residuals, resulting in a non-parametric regression problem, for $1 \le i \le n$,

$$\big(Y_i - \widehat{\mu}(\boldsymbol{X}_i)\big)^2 = \frac{c}{\phi''\big(\widehat{\mu}(\boldsymbol{X}_i)\big)} + u_i, \tag{3.14}$$

where the error terms u_i should be centered $\mathbb{E}[u_i|\boldsymbol{X}_i] = 0$. The goal is to solve (3.14) for $c > 0$ and ϕ''. We suggest to simplify this problem and to turn it into a parametric problem by considering the following one-dimensional parametric family for ϕ:

$$\phi_b(y) = \begin{cases} \frac{2}{b(b-1)} y^b, & \text{for } b \neq 0 \text{ and } b \neq 1, \\ -2\log(y), & \text{for } b = 0, \\ 2y\log(y) - 2y, & \text{for } b = 1, \end{cases} \tag{3.15}$$

where $y > 0$. This provides us with Bregman divergences, see (2.5),

$$L_{\phi_b}(y; a) = \phi_b(y) - \phi_b(a) - \phi_b'(a)(y - a) = \begin{cases} 2\left[\frac{y^b}{b(b-1)} - y\frac{a^{b-1}}{b-1} + \frac{a^b}{b}\right], & \text{for } b \neq 0 \text{ and } b \neq 1, \\ 2\left[\log(a/y) + (y - a)/a\right], & \text{for } b = 0, \\ 2\left[y\log(y/a) + a - y\right], & \text{for } b = 1. \end{cases}$$

For $b \notin (1, 2)$, these are exactly the deviance losses within Tweedie's (1984) family for power variance parameters $p = 2 - b$, see Example 4.11 in Wüthrich and Merz (2023). The case $b = 2$ corresponds to the Gaussian distribution, $b = 1$ to the Poisson distribution, $b = 0$ to the gamma distribution and $b = -1$ to the inverse Gaussian distribution; for $b \in (1, 2)$ there are no Tweedie's distributions, see Theorem 2 in Jørgensen (1987). Thus, analyzing relation (3.14) will motivate the specific choice of b and of ϕ_b, respectively, such that

$$\big(Y_i - \widehat{\mu}(\boldsymbol{X}_i)\big)^2 \approx \frac{c}{\phi_b''\big(\widehat{\mu}(\boldsymbol{X}_i)\big)} = \frac{c}{2}\,\widehat{\mu}(\boldsymbol{X}_i)^{2-b}, \tag{3.16}$$

and it allows us to select $c > 0$.

For ϕ_+ and ϕ_-, we use a similar, though slightly more complicated, approach. First, we come up with a pre-estimate for the conditional quantile function $\widehat{Q}_\tau(\boldsymbol{x})$, along the lines of Subsection 3.2. Using this pre-estimate $\widehat{Q}_\tau(\boldsymbol{x})$, we split our sample into two distinct sets,

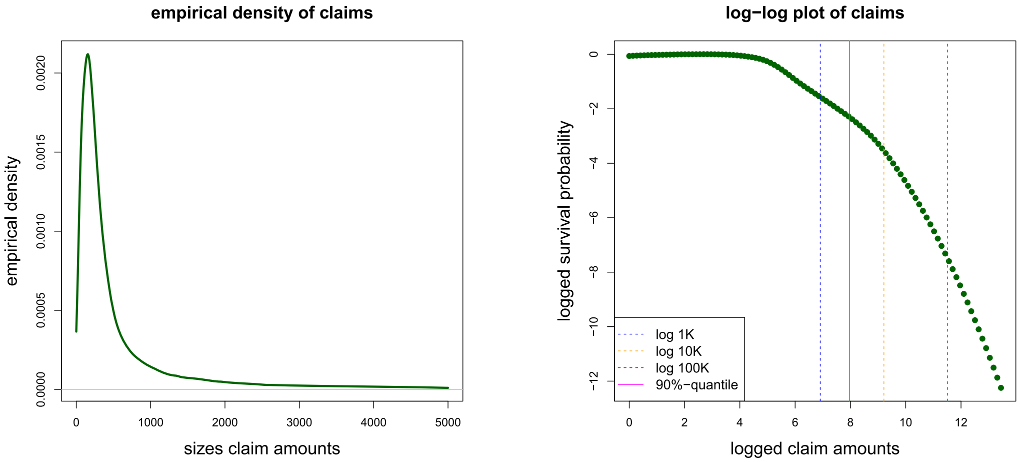

Fig. 1. (lhs) Empirical density (upper-truncated at 5,000), (rhs) log-log plot of the observed compulsory Swiss accident insurance claim amounts. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

introducing the index sets $\mathcal{I}_- = \{i \in \{1,\dots,n\} | Y_i \le \widehat{Q}_\tau(\boldsymbol{X}_i)\}$ and $\mathcal{I}_+ = \{i \in \{1,\dots,n\} | Y_i > \widehat{Q}_\tau(\boldsymbol{X}_i)\}$. We fit conditional mean models $\widehat{E}^-_\tau(\boldsymbol{x})$ and $\widehat{E}^+_\tau(\boldsymbol{x})$ to the two data sets $\mathcal{I}_-$ and $\mathcal{I}_+$. Then, we can analyze

$$\left(Y_i - \widehat{E}^+_\tau(\boldsymbol{X}_i)\right)^2 \approx \frac{c_+}{\phi''_+\left(\widehat{E}^+_\tau(\boldsymbol{X}_i)\right)}, \qquad i \in \mathcal{I}_+, \tag{3.17}$$

$$\left(Y_i - \widehat{E}^-_\tau(\boldsymbol{X}_i)\right)^2 \approx \frac{c_-}{\phi''_-\left(\widehat{E}^-_\tau(\boldsymbol{X}_i)\right)}, \qquad i \in \mathcal{I}_-. \tag{3.18}$$

Again, one needs to estimate the parameters c_- and c_+ as well as the functions ϕ''_- and ϕ''_+, and, for simplicity, we again suggest to use a member of the parametric family (3.15), so that we only need to select b_-, c_-, and b_+, c_+, respectively. Then, one obtains $\phi_- = \phi_{b_-}$ and $\phi_+ = \phi_{b_+}$. However, one should invoke the restrictions that $\phi'_+ < 0$ and $\phi'_- > 0$. Therefore, we have the restriction $b_- > 1$ for ϕ_{b_-} and $b_+ < 1$ for ϕ_{b_+}.

We dispense with a discussion of the optimal choice of g. Equally so, we ignore conditions of the type of equation (4.19) in Dimitriadis et al. (2020), stipulating that for G_{e^-,e^+} in (2.10) it holds that $G'_{e^-,e^+}\left(Q_\tau(\boldsymbol{x})\right) = c_\tau f_{Y|\boldsymbol{x}}\left(Q_\tau(\boldsymbol{x})\right)$ for some $c_\tau > 0$ and for all $\boldsymbol{x} \in \mathcal{X}$. Here, $f_{Y|\boldsymbol{x}}$ is the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$. Unless we use strong parametric assumptions on $f_{Y|\boldsymbol{x}}$ to come up with a reasonable pre-estimate, we would need to resort to kernel density estimation methods here, which amounts to a considerable computational complexity. We think that in the given setting, the precise estimation of the splicing point corresponding to the estimation of the conditional quantile is less important than the precise estimation of the lower and upper conditional ES, yielding a good estimate of the overall expected claim size. Hence, we suggest to either set g to be constant 0 or to use a multiple of the classical pinball loss, arising from $g(y) = c_\tau y$, $c_\tau > 0$.

We conclude that the overall claim size (pure risk premium) can be calculated by

$$\boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{x}] = \tau \mathrm{ES}^-_\tau(Y|\boldsymbol{x}) + (1-\tau)\mathrm{ES}^+_\tau(Y|\boldsymbol{x}). \tag{3.19}$$

The beauty of this approach now is that we can fit different models above and below the splicing point $F^{-1}_{Y|\boldsymbol{x}}(\tau)$, accounting for different properties in the main body and the tail of the data. Thus, we can have different distributions above and below the splicing point, reflected in using different scoring functions above and below $F^{-1}_{Y|\boldsymbol{x}}(\tau)$ (implied by ϕ_- and ϕ_+). This results in different regression functions potentially using covariates $\boldsymbol{x} \in \mathcal{X}$ in different ways, i.e., we can have different regression functions in the tail and the main body of the data.

4. Real claim size example

4.1. Description of data

We present a real data example with claim amounts describing medical expenses in compulsory Swiss accident insurance. In total we have 267,992 claims with positive claim amounts, $Y_i > 0$, and these claim amounts range from 1 to 691,066 CHF. Fig. 1 shows the empirical density and the log-log plot of these claim amounts. The empirical density is unimodal, and from the log-log plot we conclude that the tail is moderately heavy-tailed, but it is not regularly varying (which would correspond to an asymptotic straight line in the log-log plot).

These claim amounts are supported by 7 features. We have 3 categorical features for the 'labor sector' of the injured, the 'injury type' and the 'injured body part', 1 binary feature telling us whether the injury is a 'work or leisure' accident, and 3 continuous features corresponding to the 'age' of the injured, the 'reporting delay' of the claim and the 'accident quarter' (capturing seasonality in claims because leisure activities differ in summer and winter times). A preliminary analysis shows that all these features have predictive power, i.e., they are explaining systematic effects in the claim amounts.

For our analysis, we partition the entire data into a learning data set $\mathcal{L} = (Y_i, \boldsymbol{x}_i)_{1 \le i \le n}$ that is used for model fitting, and a test data set $\mathcal{T} = (Y^\dagger_t, \boldsymbol{x}^\dagger_t)_{1 \le t \le T}$ which we (only) use for an out-of-sample (generalization) analysis. We do this partition stratified w.r.t. the claim

Table 1
Out-of-sample pinball losses L_{τ_j}, $\tau_j \in \{10\%, 50\%, 90\%\}$, on the test data $\mathcal{T}$ of the deep multiple quantile regression using the additive and the multiplicative approaches, and for three individual deep quantile regressions.

	number of parameters	out-of-sample pinball losses		
		$\tau_j = 10\%$	$\tau_j = 50\%$	$\tau_j = 90\%$
additive multi-output approach	834	141.69	622.11	717.33
multiplicative multi-output approach	834	141.60	622.72	716.46
individual deep quantile estimation	2436	141.24	622.71	716.89

amounts and in a ratio of 9 : 1. This results in learning data set $\mathcal{L}$ of size $n = 241{,}193$ and in test data set $\mathcal{T}$ of size $T = 26{,}799$. We hold on to the same partition in all examples studied. For network fitting we further partition the learning sample $\mathcal{L} = (Y_i, \boldsymbol{x}_i)_{1\le i\le n}$ into training data set $\mathcal{U}$ and validation data set $\mathcal{V}$. Thus, $\mathcal{U} \cup \mathcal{V}$ and $\mathcal{T}$ are disjoint (and assumed to be independent) so that we can perform a proper out-of-sample forecast evaluation. For a detailed discussion of such a partition of the data for SGD fitting we refer to Section 7.2.3 and, in particular, to Fig.7.7 in Wüthrich and Merz (2023).

4.2. Example: deep multiple quantile regression

We apply the framework of Section 3.3 to perform a deep multiple quantile regression. We choose 3 probability levels $0 < \tau_1 < \tau_2 < \tau_3 < 1$ that we simultaneously estimate; the specific choices considered are $(\tau_1, \tau_2, \tau_3) = (10\%, 50\%, 90\%)$.

We first discuss pre-processing of the feature components before choosing the deep FN network architecture $\boldsymbol{z}^{(d:1)}$. For the 3 categorical variables we use embedding layers of dimension 2, i.e., they are treated by a contextualized embedding. Embedding layers are explained in Section 7.4.1 of Wüthrich and Merz (2023). We use the R library `keras` Chollet et al. (2017) for our implementation, and these embedding layers are encoded on lines 3-13 of Listing 1 in the supplementary material Section C. The binary variable is encoded by 0-1 and the continuous variables are pre-processed by the MinMaxScaler to ensure that they live on the same scale, see formula (7.30) in Wüthrich and Merz (2023) for the MinMaxScaler. Based on this feature pre-processing we use an FN network of depth $d = 3$ having input dimension $r_0 = 3 \cdot 2 + 1 + 3 = 10$ (for the categorical, binary and continuous feature components). This 10-dimensional variable enters the network in line 15 of Listing 1. For the deep FN network architecture $\boldsymbol{z}^{(d:1)}$ we choose depth $d = 3$ with $(r_1, r_2, r_3) = (20, 15, 10)$ hidden neurons in the hidden layers, and the hyperbolic tangent activation function Ψ. This is encoded on lines 16-18 of Listing 1 and gives us the learned representations $\boldsymbol{z}_i = \boldsymbol{z}^{(d:1)}(\boldsymbol{x}_i) \in \mathbb{R}^{r_d+1}$ of dimension $r_d + 1 = 11$ of the features $\boldsymbol{x}_i$. We remark that for an insurance data of a sample size of roughly 100,000 to 500,000 and with 10 to 20 feature components we have had positive experience by an FN network architecture of this complexity, confirmed by the various examples in Wüthrich and Merz (2023). Therefore, we hold on to this choice.

Next we implement an additive structure (3.7) for the deep multiple quantile regression

$$\boldsymbol{x} \mapsto \left(Q_{\tau_1}(\boldsymbol{x}),\ Q_{\tau_2}(\boldsymbol{x}),\ Q_{\tau_3}(\boldsymbol{x})\right)^\top \in \mathbb{R}^3_+.$$

This requires that we (re-)use the learned representations $\boldsymbol{z}_i = \boldsymbol{z}^{(d:1)}(\boldsymbol{x}_i)$ in the last hidden layer three times and the sequence of quantiles should be monotonically increasing in τ_j. We choose as link functions h and h_+ the log-link which provides us with the exponential function for their inverses. We then model these quantiles recursively to preserve monotonicity. Lines 20-26 of Listing 1 give the corresponding R code, and line 28 outputs these ordered quantiles. The multiplicative approach is rather similar, and the corresponding changes in the R code are shown in Listing 2 in the supplementary material; it uses the log-link for the biggest quantile level, and then we recursively decrease this by the logit-link (inverse of the sigmoid/logistic function).

This network architecture has $r = 834$ network parameters that need to be fitted to the learning data $\mathcal{L}$. We use the pinball loss (2.7) for the probability levels $(\tau_1, \tau_2, \tau_3) = (10\%, 50\%, 90\%)$; Listing 3 in the supplementary material shows the corresponding R code. These pinball losses with the different probability levels enter the compilation of the model on line 8 of Listing 3. Moreover, we use the `nadam` version of SGD which usually has a good fitting performance. We fit the two architectures (additive and multiplicative) to our learning data $\mathcal{L}$, using 80% of the learning data $\mathcal{L}$ as training data $\mathcal{U}$ and 20% as validation data $\mathcal{V}$ to explore the early stopping rule to prevent from over-fitting. To reduce the randomness of SGD fitting we repeat this fitting procedure for 20 different starting points of the SGD algorithm, and we calculate the average predictor over these 20 runs. The results are given in Table 1.

The first two lines of Table 1 give the average out-of-sample pinball losses L_{τ_j} on the test data $\mathcal{T}$ of the two approaches (3.7) and (3.8) for the quantile levels $\tau_j \in \{10\%, 50\%, 90\%\}$. We note that the results of the two approaches are very similar, and we cannot give a clear preference to one of the two. We benchmark these results by fitting individually each quantile, i.e., by fitting three independent networks (with the same network architecture, but a single output) to the three quantiles 10%, 50%, 90%. The results are given on the last line of Table 1. First, because we fit three deep regression models, the number of parameters and the run-time for fitting the models roughly triplicates in this naive approach. In terms of out-of-sample performance, fitting individual deep quantile models does not provide any better out-of-sample results, therefore, we prefer to stay within a single model with multiple outputs for this prediction task, see also the last bullet point of Remark 3.1.

In Section 3.3 we have also argued that the naive approach of estimating the different quantiles in separate quantile regression models may lead to the violation of the monotonicity property (3.6). In the present case, the three quantile levels 10%, 50%, 90% are sufficiently far apart so that a quantile crossing does not occur (out-of-sample). If we choose more similar quantiles, however, quantile crossings have occurred. E.g., the choices 20% and 30% have led to 161 monotonicity violations (out of 26,799 cases) when naively fitting separate models to the two quantiles.

We further illustrate the additive and the multiplicative multi-output results. An important tool in statistical modeling is the identification function that explores whether the models are properly calibrated, see Fissler et al. (2022). In the case of quantiles, this essentially results in checking the empirical out-of-sample coverage ratios (on test data $\mathcal{T}$) of the fitted models. These (empirical) coverage ratios are defined by

Table 2
Out-of-sample empirical coverage ratios $\widehat{\tau}_j$ below the estimated deep quantile estimates $Q_{\tau_j}(\boldsymbol{x}_t^\dagger)$ for $\tau_j \in \{10\%, 50\%, 90\%\}$, see (4.1).

probability levels τ_j	out-of-sample coverage ratios		
	10%	50%	90%
additive approach	10.35%	50.56%	90.22%
multiplicative approach	10.39%	50.74%	90.25%
null model	10.12%	50.00%	90.00%

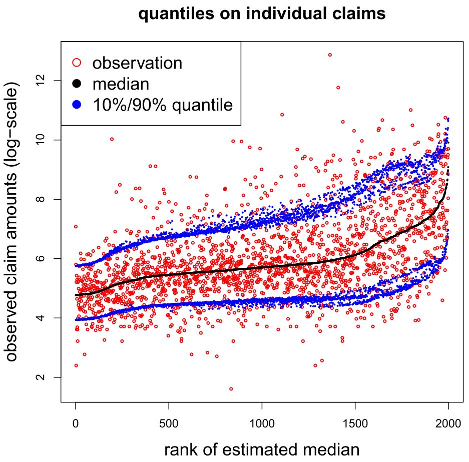

Fig. 2. Estimated quantiles $Q_{\tau_j}(\boldsymbol{x}_t^\dagger)$ of 2,000 randomly selected individual features $\boldsymbol{x}_t^\dagger$ on probability levels $\tau_j \in \{10\%, 50\%, 90\%\}$ (blue, black, blue), and the red dots show the corresponding out-of-sample observations (realizations) $Y_t^\dagger$; the x-axis orders the claims w.r.t. the estimated median $Q_{50\%}(\boldsymbol{x}_t^\dagger)$ (in black); this shows the multiplicative approach.

$$\widehat{\tau}_j = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\left\{Y_t^\dagger \le Q_{\tau_j}(\boldsymbol{x}_t^\dagger)\right\}}, \tag{4.1}$$

where $Q_{\tau_j}(\boldsymbol{x}_t^\dagger)$ are the estimated quantiles for the levels $\tau_j \in \{10\%, 50\%, 90\%\}$ (using either the additive or the multiplicative approach) and evaluated in the features $\boldsymbol{x}_t^\dagger$ of the out-of-sample observations $Y_t^\dagger$, $1 \le t \le T$.

The first two lines of Table 2 verify that the fitted deep multiple quantile regressions find the quantiles (on portfolio level) very well because $\widehat{\tau}_j \approx \tau_j$; we emphasize that these are out-of-sample figures. On the other hand, it seems that there might be a small bias since all coverage ratios have the same sign for $\widehat{\tau}_j - \tau_j > 0$. Asymptotically, this bias should vanish because the pinball loss is a strictly consistent loss function for the quantile, see Theorem 2.4, and if the chosen network architecture is sufficiently flexible to model the true regression function (the latter is related to the universality theorems for networks, see Section 7.2.2 of Wüthrich and Merz (2023)). This strict consistency is an asymptotic statement which may be violated on finite samples. We have tested the same network architecture on a different data set, having a similar complexity. Interestingly, for the quantiles 10% and 90% we have received rather similar coverage ratios as in Table 2, however, the coverage ratio for the median resulted in a value slightly below 50%. Moreover, testing the network architecture on a synthetic data example in Supplement B, we have received fluctuations of the empirical coverage ratios without a consistent sign of the bias; see Table 5 in the supplementary material. This suggests that we do not expect a systematic bias in this estimation approach.

The first two lines of Table 2 give us the coverage ratios on the portfolio level. This does not tell us much about the suitability of the regression function on an individual claim level; the null model (intercept model not using any feature information) achieves a coverage ratio of a similar accuracy, see last line of Table 2. That is, for model selection (after verifying that the coverage ratio (4.1) is sufficiently close to τ_j), we should rely on the out-of-sample pinball losses of Table 1. Individual differences between the claims are illustrated in Fig. 2; we illustrate the multiplicative approach. This figure shows the estimated quantiles $Q_{\tau_j}(\boldsymbol{x}_t^\dagger)$ for individual features $\boldsymbol{x}_t^\dagger$ at the probability levels $\tau_j \in \{10\%, 50\%, 90\%\}$ (blue, black, blue). The individual observations are ordered w.r.t. the estimated median in black. We observe that this ordering does not imply monotonicity for the other quantiles, especially for bigger claim potentials. This indicates heteroskedasticity in our data, and justifies the use of (a more complex) regression model. The red dots show the corresponding (out-of-sample) observed claim amounts $Y_t^\dagger$.

4.3. Preliminary considerations for deep composite regression

To fit a deep composite regression model we need a first preliminary step to explore the relationship (3.16). This preliminary step motivates the choices of ϕ, ϕ_+ and ϕ_- in (2.11)-(2.13); for g we use the identity function $g(y) = y$, giving us the pinball loss. For this preliminary step, we choose exactly the same network architecture (3.4) as for the deep quantile regression, the only change is that we replace the pinball loss by a strictly consistent scoring function for the mean functional, see Theorem 2.4. As Bregman divergence we choose the gamma deviance loss, which corresponds to the choice $b = 0$ in (3.15). We remark that the gamma model is the most popular model for insurance claim size modeling, and it often provides a first reasonable choice for a regression model; this is also the case for this preliminary step.

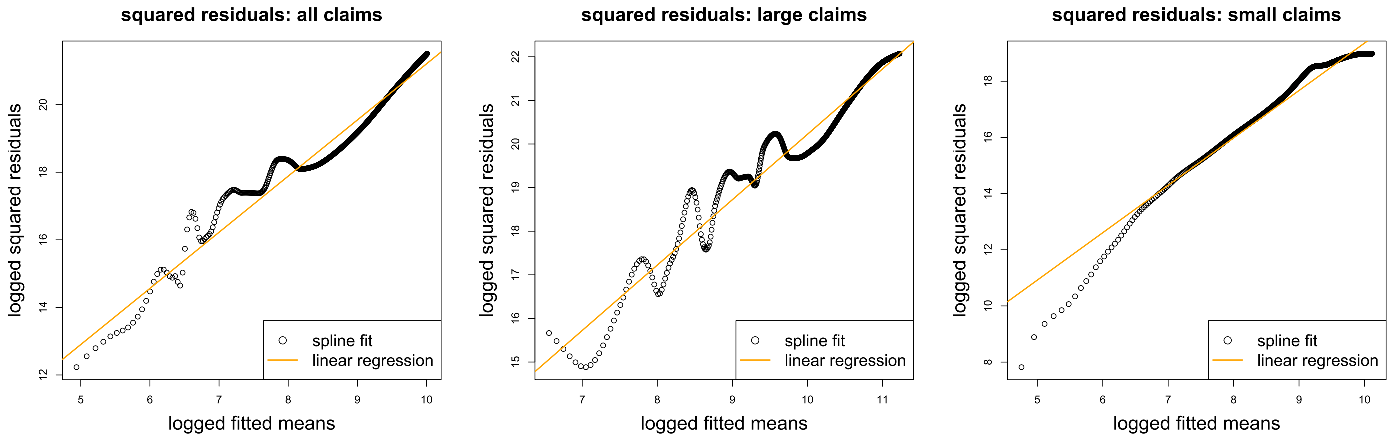

Fig. 3. Spline fit and linear regression to the estimated squared residuals as a function of the estimated means (both axes are on the log-scale): (lhs) all claims, (middle) large claims above the 90%-quantile, and (rhs) small claims below the 90%-quantile.

Table 3
Linear regression parameters for $c\phi_b(\cdot)/2$ from Fig. 3.

	all claims	large claims	small claims
	ϕ	ϕ_+	ϕ_-
intercept $\log(c/2)$	4.592	5.229	2.483
slope $2-b$	1.662	1.499	1.687
parameter b	0.338	0.401	0.313

We fit three networks of depth $d = 3$ with $(r_1, r_2, r_3) = (20, 15, 10)$ neurons and the exponential output activation h^{-1} to (a) all learning data $\mathcal{L}$, (b) the learning data having observations $Y_i > Q_{90\%}(\boldsymbol{x}_i)$, and (c) the learning data having observations $Y_i \le Q_{90\%}(\boldsymbol{x}_i)$, where $Q_{90\%}(\boldsymbol{x}_i)$ is the estimated $\tau = 90\%$ quantile from the previous example of Section 4.2. This allows us to analyze (3.14) providing a choice for ϕ, (3.17) giving a choice for ϕ_+, and (3.18) giving a choice for ϕ_- for a deep composite model regression at the probability level $\tau = 90\%$.

Fig. 3 shows a spline fit and a linear regression to the estimated squared residuals $(Y_i - \widehat{\mu}(\boldsymbol{x}_i))^2$, where $\widehat{\mu}(\cdot)$ is the estimated mean functional and where both axes are on the log-scale. The left-hand side shows all claims, the middle shows the situation where we only fit the network to the claims above the estimated quantile $Q_{90\%}(\boldsymbol{x}_i)$, and the right-hand side only considers the claims below that quantile.

Table 3 gives the linear regression estimates for c and b in (3.16) for the three cases. We observe that in all three cases we receive $b < 1$ which results in derivatives $\phi_b'(y) = y^{b-1}/(b-1) < 0$, $y > 0$. This implies that we can only work under the scoring function (2.15), because ϕ_- requires $b > 1$, e.g., the square loss function with $b = 2$ would work for ϕ_-, but this will not provide optimal convergence rates according to (3.18).

4.4. Deep composite model regression

We fit a deep composite regression model to the claims Y. The first modeling choice is the selection of the probability level τ. This choice is a bit heuristic. On the one hand, $Q_\tau(\boldsymbol{x})$ should separate the tail behavior of the data from the main body of the observations. On the other hand, we do not want to select an extreme quantile level (as in risk management) because we still need sufficient information above the quantile so that we can reliably estimate a regression function for the upper ES. Based on these needs, we start by a quantile level of $\tau = 90\%$. This results in roughly 24,000 observations of the learning data $\mathcal{L}$ being above the τ-quantile and 217,000 observations being below that quantile; this is illustrated by the vertical magenta line in the log-log plot of Fig. 1 (rhs).

We implement the regression function (3.11) and as deep FN network we use the same architecture as in the additive deep multiple quantile regression approach. This regression model is then jointly fitted under the strictly consistent scoring function (2.15) for the τ-quantile and the lower and upper ES (using the parameters of Table 3), we refer to Listing 4 in Supplement C for the encoding of the scoring function.

Fig. 4 shows the fitting results. It gives the out-of-sample estimated lower ES, $E^-_{90\%}(\boldsymbol{x}^\dagger_t)$, and the upper ES, $E^+_{90\%}(\boldsymbol{x}^\dagger_t)$, against the estimated quantiles, $Q_{90\%}(\boldsymbol{x}^\dagger_t)$, of 2,000 randomly selected instances $\boldsymbol{x}^\dagger_t$, and the cyan lines present spline fits to all out-of-sample instances. Basically, the gaps between the cyan lines and the diagonal orange line describe the differences between the ES and the quantile. This gap is roughly of a constant size between the lower ES and the quantile above 7 (on the log-scale) which means that the lower ES is a fixed ratio of the quantile. The structure of the upper ES relative to the quantile is more complicated as the gap is becoming smaller with bigger values for the 90%-quantile.

We analyze the fitted composite triplet regression model in more detail. A simple measure for analyzing the importance of individual feature components is the so-called variable permutation importance (VPI) of Breiman (2001). The VPI is received by randomly permuting one covariate component of $\boldsymbol{x}_i$ (at a time) across the entire portfolio, and then measuring the relative increase in loss received by using the features with the permuted component. Denote the original features by $\boldsymbol{x}_i$, and denote by $\boldsymbol{x}_i^{(j)}$ the features where the j-th component of $\boldsymbol{x}_i$ has been randomly permuted across the entire portfolio $1 \le i \le n$. From this we can calculate the VPI of each feature component j, and separately for lower and upper ES. We define

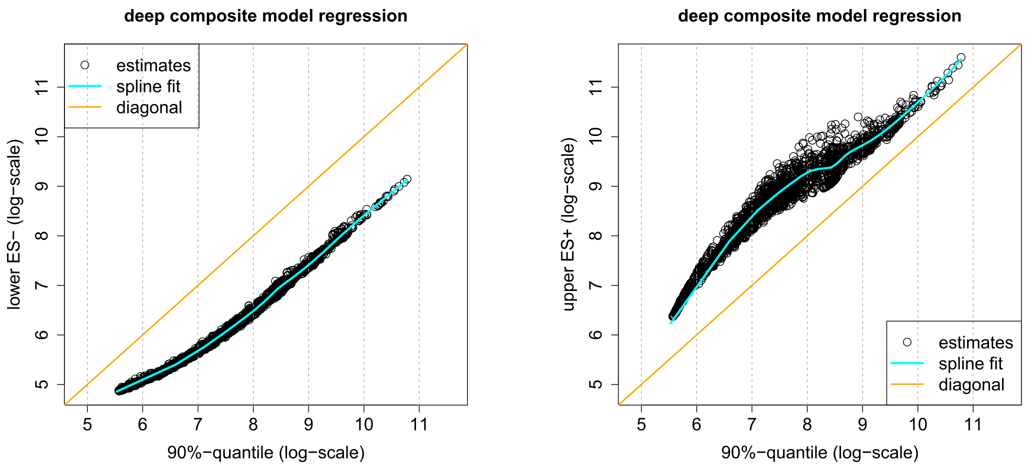

Fig. 4. (lhs) Estimated lower $E^{-}_{90\%}(\boldsymbol{x}^{\dagger}_t)$ vs. estimated quantile $Q_{90\%}(\boldsymbol{x}^{\dagger}_t)$ and (rhs) estimated upper $E^{+}_{90\%}(\boldsymbol{x}^{\dagger}_t)$ vs. estimated quantile $Q_{90\%}(\boldsymbol{x}^{\dagger}_t)$ at probability level $\tau = 90\%$.

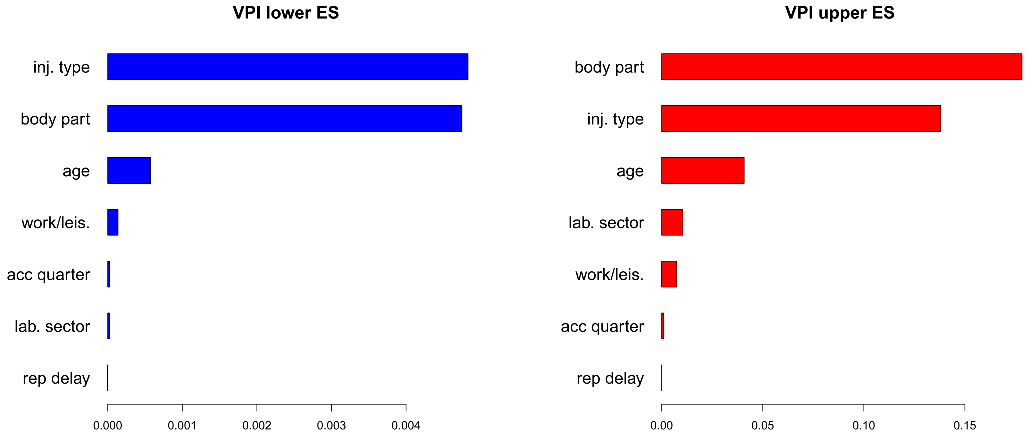

Fig. 5. Variable permutation importance: (lhs) lower VPI^{-}_{j} and (rhs) upper VPI^{+}_{j} by randomly permuting the j-th component of $\boldsymbol{x}_i$ for the lower and upper ES estimation, respectively.

$$\text{VPI}^{-}_{j} = \frac{\sum_{i=1}^{n} L\left(Y_i;\, E^{-}_{90\%}(\boldsymbol{x}^{(j)}_i),\; Q_{90\%}(\boldsymbol{x}_i),\; E^{+}_{90\%}(\boldsymbol{x}_i)\right)}{\sum_{i=1}^{n} L\left(Y_i;\, E^{-}_{90\%}(\boldsymbol{x}_i),\; Q_{90\%}(\boldsymbol{x}_i),\; E^{+}_{90\%}(\boldsymbol{x}_i)\right)} - 1,$$

$$\text{VPI}^{+}_{j} = \frac{\sum_{i=1}^{n} L\left(Y_i;\, E^{-}_{90\%}(\boldsymbol{x}_i),\; Q_{90\%}(\boldsymbol{x}_i),\; E^{+}_{90\%}(\boldsymbol{x}^{(j)}_i)\right)}{\sum_{i=1}^{n} L\left(Y_i;\, E^{-}_{90\%}(\boldsymbol{x}_i),\; Q_{90\%}(\boldsymbol{x}_i),\; E^{+}_{90\%}(\boldsymbol{x}_i)\right)} - 1.$$

Fig. 5 shows these VPIs, on the left-hand side for the lower ES and on the right-hand side for the upper ES. The bars show these relative increases, ordered by their magnitudes. We observe that the ordering is not the same and also their relative magnitudes change between lower and upper ES. E.g., the labor sector seems irrelevant for the lower ES, whereas it has a non-negligible importance for the upper ES. Such a behavior exactly motivates the consideration of models that allow for different regression functions for the body and the tail of the data, and it justifies the use of our deep composite triplet regression model.

Based on the lower and upper ES estimates, we can now determine the expected value of response Y, given feature $\boldsymbol{x}$,

$$\widehat{\mu}(\boldsymbol{x}) = \widehat{\mathbb{E}}[Y|\boldsymbol{x}] = \tau\,\text{ES}^{-}_{\tau}(Y|\boldsymbol{x}) + (1-\tau)\text{ES}^{+}_{\tau}(Y|\boldsymbol{x}).$$

This is the resulting best-estimate insurance price that results from the deep composite triplet regression model. We benchmark these estimated means of the deep composite triplet regression model with the ones obtained from the plain-vanilla deep gamma regression model used in the preparatory Section 4.3; note that the minimization of the gamma deviance loss with $b = 0$ in (3.15) is equivalent to the maximization of the log-likelihood function of the gamma distribution. Comparing the two models is not that straightforward, as we can choose any (strictly consistent) Bregman divergence for assessing the accuracy of the mean functional. Choosing the (standard) rooted mean squared error of prediction (RMSEP) gives a slight preference to the composite triplet model (5,894) over the deep gamma regression model (5,909).

Fig. 6 (lhs) compares the fitted deep gamma model (on the y-axis) against the fitted deep composite triplet regression model (on the x-axis). For small estimated means $\widehat{\mu}(\boldsymbol{x})$ the deep gamma and the deep composite triplet model are rather similar. For large estimated means, however, the deep gamma model provides clearly smaller estimates. The gray dotted lines indicate the maximal estimate $\max_t \widehat{\mu}(\boldsymbol{x}^{\dagger}_t)$ in the deep gamma model. It seems that the deep gamma model under-estimates claims with large expected payments.

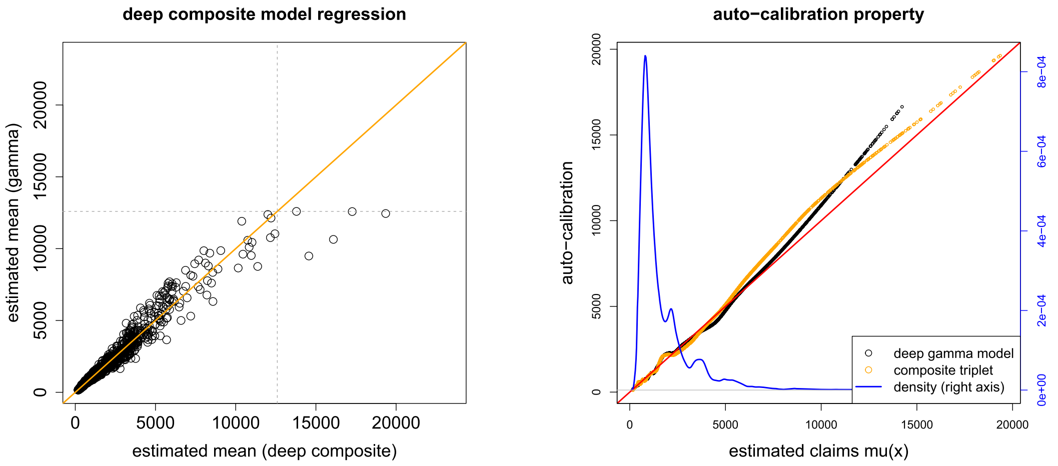

Fig. 6. (lhs) Comparison of the estimated out-of-sample means and (rhs) auto-calibration property of the deep gamma model and the deep composite triplet regression model.

Table 4
Out-of-sample empirical coverage ratios $\widehat{\tau}$ and identification functions $\widehat{v}_-$ and $\widehat{v}_+$ of the three considered models.

	coverage ratio	lower ES identification	upper ES identification
	$\widehat{\tau} = 90\%$	$\widehat{v}_-$	$\widehat{v}_+$
deep composite model	90.13%	1.3	-170.7
deep gamma model	93.63%	5,612.7	-7,962.4

We further investigate the upper tail of the estimated means. A quantity that has attracted attention recently is the so-called auto-calibration property, see Krüger and Ziegel (2021), Denuit et al. (2021) and Section 7.4.2 in Wüthrich and Merz (2023). An auto-calibrated forecast $\widehat{\mu}(\boldsymbol{x})$ for response Y satisfies the property, a.s.,

$$\widehat{\mu}(\boldsymbol{X}) = \mathbb{E}\left[\, Y \,|\, \widehat{\mu}(\boldsymbol{X})\right].$$

This auto-calibration property can be checked empirically by just exploring a spline fit to the observations Y as a function of the estimates $\widehat{\mu}(\boldsymbol{x})$. We use the R package `locfit` Loader et al. (2022) to compute this empirical spline fit, see also Listing 7.8 in Wüthrich and Merz (2023). The results are presented in Fig. 6 (rhs). The black dots show the auto-calibration property of the deep gamma model. A perfectly auto-calibrated model would lie on the diagonal red line. We observe that the auto-calibration property holds in the deep gamma model rather accurately up to an expected claim size of 5,000, and larger expected claim sizes under-estimate the true claim potential because the black dots are above the diagonal red line. The blue graph shows the empirical density of the out-of-sample means $(\widehat{\mu}(\boldsymbol{x}_t^\dagger))_{1\le t\le T}$, and we note that 96% of the claims have an estimated expected mean less than 5,000. The orange dots in Fig. 6 (rhs) show the estimated means of the deep composite triplet regression approach. For small expected claims, the results are very similar to the gamma model. There is also a small under-estimation for expected claims between 6,000 and 15,000, but, in contrast to the deep gamma model, for larger claims the auto-calibration property again holds rather accurately, see upper-right corner of the graph. This better large claims behavior exactly shows the advantage of having the additional modeling flexibility in the upper tail of the data.

Similarly to the deep quantile estimations, we should also check the calibration (identification) of the fitted deep mean estimation. The lower ES and the upper ES are only jointly elicitable with the corresponding τ-quantile, which also caries over to their identifiability; see Supplement A.1 for a brief discussion of identifiability. Hence, we cannot check the calibration of the models $E^-_{90\%}$ and $E^+_{90\%}$ standalone, but we can only evaluate the calibration of these models *jointly* with the corresponding quantile model $Q_{90\%}$. We need to consider the three-dimensional identification function provided in (A.1). For the first and third component, the empirical out-of-sample identification functions are

$$\widehat{v}_- = \frac{1}{T}\sum_{t=1}^{T}\left[E^-_{90\%}(x_t^\dagger) - \frac{Y_t^\dagger}{0.9}\mathbb{1}_{\{Y_t^\dagger \le Q_{90\%}(\boldsymbol{x}_t^\dagger)\}} + \frac{Q_{90\%}(\boldsymbol{x}_t^\dagger)}{0.9}\left(\mathbb{1}_{\{Y_t^\dagger \le Q_{90\%}(\boldsymbol{x}_t^\dagger)\}} - 0.9\right)\right], \tag{4.2}$$

$$\widehat{v}_+ = \frac{1}{T}\sum_{t=1}^{T}\left[E^+_{90\%}(x_t^\dagger) - \frac{Y_t^\dagger}{0.1}\mathbb{1}_{\{Y_t^\dagger > Q_{90\%}(\boldsymbol{x}_t^\dagger)\}} - \frac{Q_{90\%}(\boldsymbol{x}_t^\dagger)}{0.1}\left(0.1 - \mathbb{1}_{\{Y_t^\dagger > Q_{90\%}(\boldsymbol{x}_t^\dagger)\}}\right)\right]. \tag{4.3}$$

For the second component, the empirical version is the empirical coverage defined in (4.1) minus $\tau = 90\%$. If the three-dimensional empirical out-of-sample identification function is close to 0, this indicates a well calibrated model for $(E^-_{90\%}, Q_{90\%}, E^+_{90\%})$.

Table 4 reports the empirical coverage ratios $\widehat{\tau}$, for $\tau = 90\%$, along with the empirical identification functions $\widehat{v}_-$ and $\widehat{v}_+$ for the deep composite triplet regression model and for the deep gamma model. Having a gamma distributional assumption from the deep gamma approach of Section 4.3, we can also calculate the composite triplet under this gamma model assumption. In Section 4.3 we have fitted the means $\widehat{\mu}(\boldsymbol{x}_t^\dagger)$ under the gamma assumption. Moreover, using these means, we can estimate the dispersion parameter. If we use the

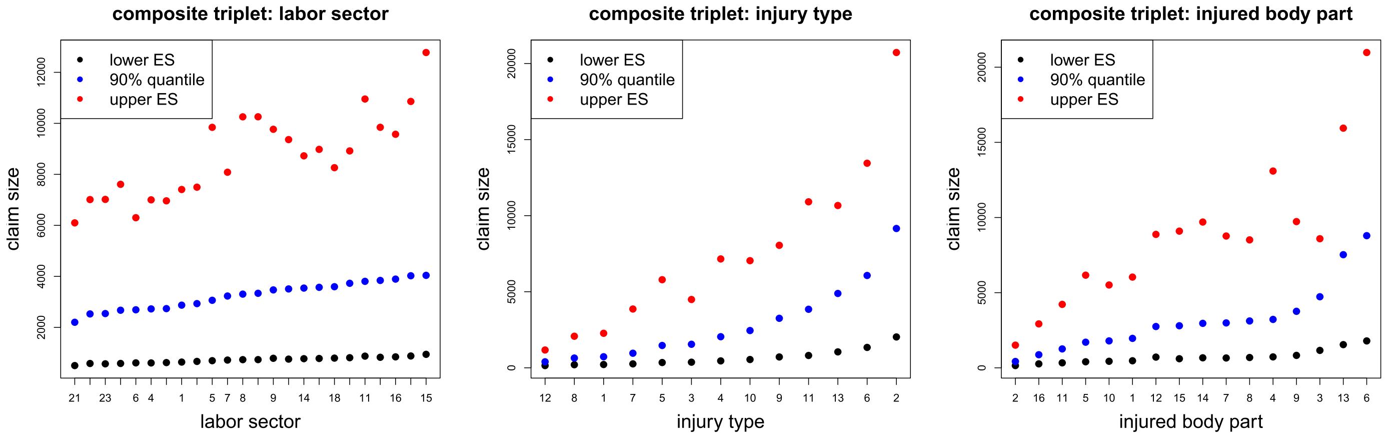

Fig. 7. Average marginal estimations for selected feature levels: (lhs) labor sector, (middle) injury type, (rhs) injured body part; ordered on the x-axis w.r.t. the average quantile estimates $Q_{90\%}$.

deviance dispersion estimate, we receive in this gamma model a dispersion parameter of $1/\gamma = 1.83$. This allows us to calculate the corresponding quantiles and ES in the gamma model. The lower and upper ES in the gamma model are obtained (in closed form) by

$$\mathbb{E}\left[Y \,\middle|\, Y \le \Gamma^{-1}_{\gamma,\mu}(\tau)\right] = \mu\left(\frac{\Gamma_{\gamma+1,\mu}\left(\Gamma^{-1}_{\gamma,\mu}(\tau)\right)}{\tau}\right), \tag{4.4}$$

$$\mathbb{E}\left[Y \,\middle|\, Y > \Gamma^{-1}_{\gamma,\mu}(\tau)\right] = \mu\left(\frac{1-\Gamma_{\gamma+1,\mu}\left(\Gamma^{-1}_{\gamma,\mu}(\tau)\right)}{1-\tau}\right), \tag{4.5}$$

where $Y \sim \Gamma_{\gamma,\mu}$ denotes the gamma distribution with mean $\mu > 0$ and shape parameter $\gamma > 0$. The results are provided in Table 4. We conclude that the deep composite triplet regression model meets the right coverage ratio of 90% very well, whereas the deep gamma model results in a too high coverage ratio which indicates that the gamma distributional assumption does not match the tail of the observed data. This carries over to the identification functions of the lower and upper ES. The deep composite triplet regression model calibrates much better than the deep gamma regression model, which is indicated by values of $\widehat{v}_-$ and $\widehat{v}_+$ being much closer to zero for the composite triplet model.

Fig. 7 gives marginal estimations of the composite triplet of lower ES, the 90% quantile and the upper ES for the features 'labor sector', 'injury type' and 'injured body part'; the x-axis is ordered w.r.t. the average 90% quantiles. We observe monotonicity between the average lower ES and the 90% quantiles (black and blue dots), but this monotonicity gets lost w.r.t. the upper ES (red dots). This indicates that we have different regression functions for the main body and the tail of the data. That is, we have developed a very flexible (deep regression) model where features $\boldsymbol{x} \in \mathcal{X}$ impact predictions differently in the body and the tail of the distributions.

Conclusion of this example.

- We have presented a real data example that is based on a fixed choice of the quantile level $\tau = 90\%$ and a fixed choice of the strictly consistent scoring function L for the composite triplet, see Theorem 2.8. Optimal choices of the scoring function L are motivated by the considerations in Section 4.3, but certainly this needs more exploration because there are (infinitely) many ways of composing such strictly consistent scoring functions, and also an optimal linear transformation may matter, see, e.g., formula (A.2) in the supplementary material.
- Concerning the choice of the quantile level, we have also tried smaller quantiles, and the choices of smaller quantiles (smaller than 90%) have led to more similarity with the deep gamma model because for smaller quantile levels the upper ES considers more and more observations from the body of the data. Certainly this choice also needs more exploration. More mathematically speaking, one could consider test statistics based on the identification functions (4.1), (4.2) and (4.3). It seems feasible (under additional assumptions) to derive their asymptotic behavior as the sample size converges to infinity. This then would allow testing for different quantile levels.
- We could start to benchmark our proposal with further statistical modeling approaches such as mixture density network (MDN), see Delong et al. (2021). We refrain from doing so because most of these more advanced methods have their own caveats. E.g., MDNs are more difficult to fit as they require a combination of gradient descent and the EM algorithm, which often has poor convergence properties and may result in spurious solutions. Moreover, the type and number of mixture components in MDNs is not straightforward, because this choice is not really supported by rigorous statistical tools. Such questions are similar to the choice of the quantile level τ and the loss function L in our problem, however, our proposal has the big advantage that fitting the composite triplet model is straightforward and fast (in one step) using the gradient descent algorithm with a strictly consistent scoring function. Therefore, it is more flexible in exploring different options.
- In Supplement B, we provide a simulation study on synthetic data that verifies the validity of our modeling approach.

5. Summary and outlook

We present a deep quantile and a deep composite triplet regression. For the composite model, we use a conditional quantile as splicing point, not an absolute threshold. This provides flexibility e.g. in the presence of heteroskedasticity. We utilize a multi-output network architecture which respects the natural ordering of quantiles at different probability levels as well as the natural ordering of the composite triplet consisting the lower expected shortfall, the quantile and the upper expected shortfall at the same probability level. Such a multi-output network may promote robustness in estimation. While strictly consistent scoring functions for tuples of quantiles are available, e.g. in the form of sums of pinball losses, and thus M-estimation can be performed for deep quantile models, we first derive and introduce the class of strictly consistent scoring functions for the composite triplet. Addressing the specific choice of the strictly consistent score for the composite triplet, we discuss data-driven choices which potentially increase the efficiency in estimation. The suitability of our methods is illustrated on a real data example with claim amounts describing medical expenses in compulsory Swiss accident insurance.

A relevant extension of our methods is a generalized composite model using two (or even more) splicing points. This allows for modeling the influence of features differently for small claim sizes, large claim sizes and the body of the data. This extension seems particularly beneficial since there is empirical evidence that the (conditional) distribution of claim sizes is different in nature for small claim sizes, the body, and for large claim sizes. Hence, this extension has the potential to increase accuracy of overall claim modeling, which is relevant in insurance pricing. The positive results on the elicitability of the range value at risk (RVaR) – or interquantile expectation – together with the two corresponding quantiles in Fissler and Ziegel (2021) suggest the existence of strictly consistent scoring functions for the "extended composite quintuple" consisting of two quantiles (say, the 10% and 90% quantiles) together with the lower ES, upper ES and the range value at risk. However, the specific data-driven choice of the score motivated by efficiency considerations is far from being clear since the class of scoring functions for RVaR and two quantiles is less flexible than the one involving ES and a quantile. E.g. it is shown in Fissler and Ziegel (2021) that there are no positively homogeneous strictly consistent scores for the former case. Therefore, we defer a detailed study of this extension to future research.

Declaration of competing interest

There is no competing interest.

Data availability

The authors do not have permission to share data.

Acknowledgement

We are grateful to Timo Dimitriadis for valuable discussions and to two anonymous referees for their constructive feedback which helped us greatly to improve the paper. T. Fissler thankfully acknowledges support from the OeNB anniversary fund, project number 17793.

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.insmatheco.2023.01.001.

References

Barendse, S., 2020. Efficiently weighted estimation of tail and interquartile expectations. SSRN Manuscript, 2937665.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Chernozhukov, V., Fernández-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), 1093–1125.
Chollet, F., Allaire, J.J., et al., 2017. R interface to keras. https://github.com/rstudio/keras.
Cooray, K., Ananda, M.M.A., 2005. Modeling actuarial data with composite lognormal-Pareto model. Scandinavian Actuarial Journal 2005 (5), 321–334.
Delong, Ł., Lindholm, M., Wüthrich, M.V., 2021. Gamma mixture density networks and their application to modeling insurance claim amounts. Insurance. Mathematics & Economics 101/B, 240–261.
Denuit, M., Charpentier, A., Trufin, J., 2021. Autocalibration and Tweedie-dominance for insurance pricing in machine learning. Insurance. Mathematics & Economics 101/B, 485–497.
Dimitriadis, T., Bayer, S., 2019. A joint quantile and expected shortfall regression framework. Electronic Journal of Statistics 13 (1), 1823–1871.
Dimitriadis, T., Fissler, T., Ziegel, J.F., 2020. The efficiency gap. arXiv:2010.14146.
Embrechts, P., Wang, R., 2015. Seven proofs for the subadditivity of expected shortfall. Dependence Modeling 3, 126–140.
Fissler, T., Lorentzen, C., Mayer, M., 2022. Model comparison and calibration assessment: user guide for consistent scoring functions in machine learning and actuarial practice. arXiv:2202.12780.
Fissler, T., Ziegel, J.F., 2016. Higher order elicitability and Osband's principle. The Annals of Statistics 44 (4), 1680–1707.
Fissler, T., Ziegel, J.F., 2019. Order-sensitivity and equivariance of scoring functions. Electronic Journal of Statistics 13 (1), 1166–1211.
Fissler, T., Ziegel, J.F., 2021. On the elicitability of range value at risk. Statistics & Risk Modeling 38 (1–2), 25–46.
Frongillo, R., Kash, I., 2021. Elicitation complexity of statistical properties. Biometrika 108 (4), 857–879.
Fung, T.C., Badescu, A.L., Lin, X.S., 2021. A new class of severity regression models with an application to IBNR prediction. North American Actuarial Journal 25 (2), 206–231.
Fung, T.C., Tzougas, G., Wüthrich, M.V., 2022. Mixture composite regression models with multi-type feature selection. North American Actuarial Journal.
Gan, G., Valdez, E.A., 2018. Fat-tailed regression modeling with spliced distributions. North American Actuarial Journal 22 (4), 554–573.
Gneiting, T., 2011a. Making and evaluating point forecasts. Journal of the American Statistical Association 106 (494), 746–762.
Gneiting, T., 2011b. Quantiles as optimal point forecasts. International Journal of Forecasting 27 (2), 197–207.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.
Grün, B., Miljkovic, T., 2019. Extending composite loss models using a general framework of advanced computational tools. Scandinavian Actuarial Journal 2019 (8), 642–660.
Guillén, M., Bermúdez, L., Pitarque, A., 2021. Joint generalized quantile and conditional tail expectation for insurance risk analysis. Insurance. Mathematics & Economics 99, 1–8.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edition. Springer Series in Statistics.

He, X., 1997. Quantile curves without crossing. American Statistician 51 (2), 186–192.
Jørgensen, B., 1987. Exponential dispersion models. Journal of the Royal Statistical Society, Series B 49 (2), 127–145.
Kellner, R., Nagl, M., Rösch, D., 2022. Opening the black box – quantile neural networks for loss given default prediction. Journal of Banking & Finance 134, 1–20.
Koenker, R., Bassett Jr., G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Krüger, F., Ziegel, J.F., 2021. Generic conditions for forecast dominance. Journal of Business & Economic Statistics 39 (4), 972–983.
Laudagé, C., Desmettre, S., Wenzel, J., 2019. Severity modeling of extreme insurance claims for tariffication. Insurance. Mathematics & Economics 88, 77–92.
Liu, Y., Wu, Y., 2011. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. Journal of Nonparametric Statistics 23 (2), 415–437.
Loader, C., Sun, J., Lucent Technologies, Liaw, Liaw, A., 2022. Locfit: local regression, likelihood and density estimation. https://cran.r-project.org/web/packages/locfit/index.html.
McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools, revised edition. Princeton University Press.
Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7, 983–999.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4, Chapter 36. Elsevier, pp. 2111–2245.
Nolde, N., Ziegel, J.F., 2017. Elicitability and backtesting: perspectives for banking regulation. Annals of Applied Statistics 11 (4), 1833–1874.
Osband, K.H., 1985. Providing Incentives for Better Cost Forecasting. PhD thesis. University of California, Berkeley.
Parodi, P., 2020. A generalised property exposure rating framework that incorporates scale-independent losses and maximum possible loss uncertainty. ASTIN Bulletin 50 (2), 513–553.
Pigeon, M., Denuit, M., 2011. Composite lognormal-Pareto model with random threshold. Scandinavian Actuarial Journal 2011 (3), 177–192.
Richman, R., 2022. Mind the gap – safely incorporating deep learning models into the actuarial toolkit. British Actuarial Journal 27, E21.
Saerens, M., 2000. Building cost functions minimizing to some summary statistics. IEEE Transactions on Neural Networks 11, 1263–1271.
Savage, L.J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66 (336), 783–810.
Scollnik, D.P.M., 2007. On composite lognormal-Pareto models. Scandinavian Actuarial Journal 2007 (1), 20–33.
Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.J., 2006. Nonparametric quantile estimation. Journal of Machine Learning Research 7, 1231–1264.
Thomson, W., 1979. Eliciting production possibilities from a well-informed manager. Journal of Economic Theory 20, 360–380.
Tweedie, M.C.K., 1984. An index which distinguishes between some important exponential families. In: Ghosh, J.K., Roy, J. (Eds.), Statistics: Applications and New Directions. Proceeding of the Indian Statistical Golden Jubilee International Conference, Indian Statistical Institute, Calcutta. Indian Statistical Institute, Calcutta, pp. 579–604.
Uribe, J.M., Guillén, M., 2019. Quantile Regression for Cross-Sectional and Time Series Data Applications in Energy Markets Using R. Springer.
Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.
Weber, S., 2006. Distribution-invariant risk measures, information, and dynamic consistency. Mathematical Finance 16, 419–441.
Wüthrich, M.V., Merz, M., 2023. Statistical Foundations of Actuarial Learning and Its Applications. Springer Actuarial.