Astin Bulletin (2024), **54**, pp. 213–238
doi:10.1017/asb.2024.7

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Machine Learning with High-Cardinality Categorical Features in Actuarial Applications

Benjamin Avanzi[1], Greg Taylor[2], Melantha Wang[2] and Bernard Wong[2]

[1]Centre for Actuarial Studies, Department of Economics, University of Melbourne VIC 3010, Australia
[2]School of Risk and Actuarial Studies, UNSW Australia Business School, UNSW Sydney NSW 2052, Australia
Corresponding author: Melantha Wang; Email: wang.melantha@gmail.com

Received: 30 January 2023; **Revised:** 3 February 2024; **Accepted:** 5 February 2024

Keywords: Categorical features; Generalised linear mixed models; Neural networks; Categorical embedding; Random effects; Variational inference; Insurance analytics

JEL codes: C45; C51; C52; C53; G22

MSC Classification: 91G70; 91G60; 62P05

Abstract

High-cardinality categorical features are pervasive in actuarial data (e.g., occupation in commercial property insurance). Standard categorical encoding methods like one-hot encoding are inadequate in these settings.

In this work, we present a novel *Generalised Linear Mixed Model Neural Network* ("GLMMNet") approach to the modelling of high-cardinality categorical features. The GLMMNet integrates a generalised linear mixed model in a deep learning framework, offering the predictive power of neural networks and the transparency of random effects estimates, the latter of which cannot be obtained from the entity embedding models. Further, its flexibility to deal with any distribution in the exponential dispersion (ED) family makes it widely applicable to many actuarial contexts and beyond. In order to facilitate the application of GLMMNet to large datasets, we use variational inference to estimate its parameters—both traditional mean field and versions utilising textual information underlying the high-cardinality categorical features.

We illustrate and compare the GLMMNet against existing approaches in a range of simulation experiments as well as in a real-life insurance case study. A notable feature for both our simulation experiment and the real-life case study is a comparatively low signal-to-noise ratio, which is a feature common in actuarial applications. We find that the GLMMNet often outperforms or at least performs comparably with an entity-embedded neural network in these settings, while providing the additional benefit of transparency, which is particularly valuable in practical applications.

Importantly, while our model was motivated by actuarial applications, it can have wider applicability. The GLMMNet would suit any applications that involve high-cardinality categorical variables and where the response cannot be sufficiently modelled by a Gaussian distribution, especially where the inherent noisiness of the data is relatively high.

1. Introduction

1.1. Background

The advances in machine learning (ML) over the past two decades have transformed many disciplines. Within the actuarial literature, we see an explosion of works that apply and develop ML methods for various actuarial tasks (see Richman, 2021a,b). Meanwhile, there are many distinctive challenges with insurance data that are not addressed by general-purpose ML algorithms; e.g., highly skewed response distributions (Hainaut et al., 2022) and requirements of explainability (Embrechts and Wüthrich, 2022).

One prominent challenge in actuarial ML—which we aim to tackle in this work— is the usual presence of high-cardinality categorical features (i.e., categorical features with many levels or *categories*).

© The Author(s), 2024. Published by Cambridge University Press on behalf of The International Actuarial Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

These features often represent important risk factors in actuarial data. Examples include the occupation in commercial property insurance, or the cause of injury in workers' compensation insurance. Unfortunately, ML algorithms cannot "understand" (process) categorical features on their own.

The classical solution is *one-hot encoding*. It turns a categorical feature of q unique categories into numeric representations by constructing q binary attributes, one for each category; for example, a categorical feature with three unique categories will be represented as [1, 0, 0], [0, 1, 0] and [0, 0, 1].

One-hot encoding works well with a small number of independent categories, which is the case with most applications in the ML literature (e.g., Hastie et al., 2009). However, issues arise as cardinality expands: (i) The orthogonality (i.e., independence) assumption no longer seems appropriate, since the growing number of categories will inevitably start to interact, (ii) the resultant high-dimensional feature matrix entails computational challenges, especially when used with already computationally expensive models such as neural networks; and (iii) the often uneven distribution of data across categories makes it difficult to learn the behaviour of the rare categories.

For a motivating example, consider the workers' compensation scheme of the State of New York (2022). In the four million claims observed from 2000 to 2022, the cause of injury variable has 78 unique categories, with more than 200,000 observations (about 5%) for the most common cause (lifting) and less than 1000 observations (0.025%) for the 10 least common causes (e.g., crash of a rail vehicle, or gunshot). The uneven coverage of claims presents a modelling challenge and has been documented by actuarial practitioners (e.g., Pettifer and Pettifer, 2012). Furthermore, the cause of injury variable alone may not serve well to differentiate claim severities. Injuries from the same cause can result in vastly different claims experiences. It seems natural to consider its interaction with the nature of injury and part of body variables, each with 57 reported categories. Exploring 4446 (78×57) interaction categories is clearly infeasible with one-hot encoding.

The example above highlights that one-hot encoding is an inadequate tool for handling high-cardinality categorical features. A few alternatives exist but also have their own drawbacks, as listed below.

(i) *Manual (or data-guided) regrouping of categories of similar risk behaviours*. The goal here is to reduce the number of categories so that the refined categories can be tackled with the standard one-hot encoding scheme. This is a working approach, but manual regrouping requires significant domain inputs, which are expensive. Furthermore, data-driven methods such as clustering (see e.g., Guiahi, 2017) require that data be first aggregated by categories, and this aggregation gives away potentially useful information available at more granular levels.

(ii) *Entity embeddings from neural networks*. Proposed by Guo and Berkhahn (2016), it seems to be now the most popular approach in the ML-driven stream of actuarial literature (Delong and Kozak, 2021; Shi and Shi, 2021; Kuo and Richman, 2021). Entity embeddings work to extract a low-dimensional numeric representation of each category so that categories closer in distance would observe similar response values. However, as the entity embeddings are trained as an early component of a black-box neural network, they offer little transparency towards their effects on the response.

(iii) *Generalised linear mixed models (GLMM) with the high-cardinality categorical features modelled as random effects*. This is another popular approach among actuaries partly due to its high interpretability (Antonio and Beirlant, 2007; Verbelen, 2019). However, GLMMs as an extension to GLMs also inherit their limitations—the same ones that have motivated actuaries to start turning away from GLMs and exploring ML solutions (see, e.g., Al-Mudafer et al., 2022; Henckaerts et al., 2021).

Some recent work has appeared in the ML literature aiming to extend GLMMs to take advantage of the more capable ML models, by embedding a ML model into the linear predictor of the GLMMs. The most notable examples include GPBoost (Sigrist, 2022) that combines GLMMs with a gradient boosting algorithm and LMMNN (Simchoni and Rosset, 2022) that combines linear mixed models

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

(LMM) with neural networks. Unfortunately, these models present their own limitations in the face of insurance data: GPBoost assumes overly optimistic random effects variance, and LMMNN is limited to a Gaussian-distributed response. While the Gaussian assumption offers computational advantages (in terms of analytical tractability), it is often ill suited to the more general distributions that are required for the modelling of financial and insurance data. Another significant contribution in this lineage, developed by Mandel et al. (2022) in parallel to Simchoni and Rosset (2022), addresses the limitation of LMMNN by proposing a more general framework that allows the exponential family distributions. However, their reliance on Laplace's approximation as an estimation methodology renders their model computationally expensive, which poses challenges for practical applications.

None of the existing techniques appears satisfactory for the modelling of high-cardinality categorical features. This paper seeks an alternative approach that more effectively leverages information in the categories, offers greater transparency than entity embeddings and demonstrates improved predictive performance. In the next section, we introduce our approach.

1.2. Contributions

Our work presents two main contributions.

First, we propose a novel generalised mixed effects neural network called *GLMMNet*. The GLMMNet fuses a deep neural network to the GLMM structure to take advantage of the predictive power of deep learning models and the statistical strength of GLMMs—most notably the ability to produce probabilistic estimates and full transparency on the categorical variables. Compared to the existing methods, GLMMNet offers unique flexibility with the choice of the full range of exponential dispersion (ED) family distributions, including Bernoulli and binomial distributions in the case of classification problems. While close in architecture to the model proposed by Mandel et al. (2022), the GLMMNet formulates the estimation under a Bayesian context and makes use of the highly efficient and scalable variational inference to estimate its parameters—both traditional mean field and versions utilising textual information of the categories.

Second, we provide a systematic empirical comparison of some of the most popular approaches for modelling high-cardinality categorical features. The simulation experiments we present not only show how each method performs but also highlight where certain models excel and where they fall short. Although the difficulty with high-cardinality categorical features has been a long-standing issue in actuarial modelling, to the best of our knowledge, this paper is the first piece of work to extensively compare a range of the existing approaches (including our own GLMMNet). We believe that such a benchmarking study will be valuable, especially to practitioners facing many options available to them.

Importantly, the methodological extensions proposed in this paper, while motivated by actuarial applications, are not limited to this context and can have much wider applicability. The extension of the LMMNN to the GLMMNet would suit any applications where the response cannot be sufficiently modelled by a Gaussian distribution. This unparalleled flexibility with the form of the response distribution, in tandem with the computational efficiency of the algorithm (due to the fast and highly scalable variational inference used to implement GLMMNet), make the GLMMNet a promising tool for many practical ML applications.

1.3. Outline of Paper

In Section 2, we introduce the GLMMNet, which fuses neural networks with GLMMs to enjoy both the predictive power of deep learning models and the statistical strength—specifically, the interpretability and likelihood-based estimation—of GLMMs. In Section 3, we illustrate and compare the proposed GLMMNet against a range of alternative approaches across a spectrum of simulation environments, each to mimic some specific elements of insurance data commonly observed in practice. Section 4 presents a real-life insurance case study, and Section 5 concludes.

The code used in the numerical experiments is available on https://github.com/agi-lab/glmmnet.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

2. GLMMNet: A Generalised Mixed Effects Neural Network

This section describes our GLMMNet in detail. We start by defining some notation in Section 2.1. We then provide an overview of GLMMs in Section 2.2 to set the stage for the introduction of the GLMMNet. Sections 2.3–2.5 showcase the architectural design of the GLMMNet and provide the implementation details.

2.1. Setting and Notation

We first introduce some notation to formalise the problem discussed above.

We write Y to denote the *response variable*, which is typically one of claim frequency, claim severity, or risk premium for most insurance problems. We assume that the distribution of $Y \in \mathcal{Y}$ depends on some *covariates* or *features*. In this work, we draw a distinction between standard features $\mathbf{x} \in \mathcal{X}$ (i.e., numeric features and low-dimensional categorical features) and the high-cardinality categorical features $\mathbf{z} \in \mathcal{Z}$ consisting of K distinct features each with q_i $(i = 1, \cdots, K)$ categories. The latter is the subject of our interest. To clarify, while $\mathbf{z}$ represents the collection of K categorical features, its dimensional space when one-hot encoded is represented as $\mathbb{R}^q$, where $q = \sum_{i=1}^{K} q_i$.

We assume $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$, where $p \in \mathbb{Z}$ and $q \in \mathbb{Z}$ represent the respective dimensions of the vectors. The empirical observations (i.e., data) are denoted $\mathcal{D} = (y_i, \mathbf{x}_i, \mathbf{z}_i)_{i=1}^n$, where n is the number of observations in the dataset. For convenience, let us also write $\mathbf{y} = [y_1, y_2, \cdots, y_n]^\top \in \mathbb{R}^n$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times p}$ and $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^\top \in \mathbb{R}^{n \times q}$. In general, we use bold font to denote vectors and matrices.

The purpose of any predictive model is to learn the conditional distribution $p(y|\mathbf{x}, \mathbf{z})$. It is generally assumed that the dependence of the response Y on the covariates is through a true but unknown function $\mu{:}(\mathcal{X}, \mathcal{Z}) \to \mathcal{Y}$, such that

$$p(y|\mathbf{x}, \mathbf{z}) = p(y|\mu(\mathbf{x}, \mathbf{z}), \boldsymbol{\phi}),$$

where $\boldsymbol{\phi}$ represents a vector of any additional or auxiliary parameters of the likelihood. Most commonly $\mu(\mathbf{x}, \mathbf{z})$ will be the conditional mean $\mathbb{E}(Y|\mathbf{x}, \mathbf{z})$, although it can be interpreted more generally. The predictive model then approximates $\mu(\cdot)$ by some parametric function $\hat{\mu}(\cdot)$. We write $\boldsymbol{\beta}$ to denote the collection of model parameters. Specifically, a linear regression model will assume a linear structure for $\mu(\cdot)$ with $\boldsymbol{\phi} = \sigma_\epsilon^2$ to account for the error variance. Indeed, for most ML models, the function $\mu(\cdot)$ is the sole subject of interest (as opposed to the entire distribution $p(y|\mathbf{x}, \mathbf{z})$). There are many different options for choosing the function $\mu(\cdot)$. We discuss and compare the main ones in this paper; see also Section 3.1.

In what follows, for notational simplicity we will now assume that $\mathbf{z}$ consists of only one (high-cardinality) categorical feature, that is, $K = 1$ and $q = q_1$. The extension to multiple categorical features, however, is very straightforward; see also Section 2.3.2.

2.2. Generalised Linear Mixed Models

Mixed models were originally introduced to model multiple correlated measurements on the same subject (e.g., longitudinal data), but they also offer an elegant solution to the modelling of high-cardinality categorical features (Frees, 2014). We now give a brief review of mixed models in the latter context. For further details, the books by McCulloch and Searle (2001), Gelman and Hill (2007) and Denuit et al. (2019) are all excellent references for this topic.

Let us start with the simplest linear model case. A one-hot encoded linear model assumes

$$y|\mathbf{x}, \mathbf{z} \sim \mathcal{N}(\mu(\mathbf{x}, \mathbf{z}), \sigma_\epsilon^2), \text{ with } \mu(\mathbf{x}, \mathbf{z}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} + \mathbf{z}^\top \boldsymbol{\alpha}, \tag{2.1}$$

where $\mathbf{x}$ are standard features (e.g., sum insured, age), $\mathbf{z}$ is a q-dimensional binary vector representation of a high-cardinality categorical variable (e.g., occupation), $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_p]^\top \in \mathbb{R}^p$, and

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_q]^\top \in \mathbb{R}^q$ with $\sum_{i=1}^q \alpha_i = 0$ are the regression coefficients to be estimated. Note that the additional constraint $\sum_{i=1}^q \alpha_i = 0$ is required to restore the identifiability of parameters.

In high-cardinality settings, the vector $\mathbf{z}$ becomes high-dimensional, and hence so must be $\boldsymbol{\alpha}$. High cardinality therefore introduces a large number of parameters to the estimation problem. Moreover, in most insurance problems, the distribution of categories will be highly imbalanced, meaning that some categories will have much more exposure (i.e., data) than others. The rarely observed categories are likely to produce highly variable parameter estimates. As established in earlier sections, all this makes it extremely difficult to estimate the $\boldsymbol{\alpha}$ accurately (Gelman and Hill, 2007).

Equation (2.1) is also referred to as the *no pooling* model by Antonio and Zhang (2014), as each α_i, $i = 1, 2, \cdots, q$ has to be learned independently. At the other end of the spectrum, there is a *complete pooling* model, which simply ignores the $\mathbf{z}$ and assumes

$$\mu(\mathbf{x}, \mathbf{z}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}. \tag{2.2}$$

In the middle ground between the overly noisy estimates produced by (2.1) and the over-simplistic estimates from (2.2), we can find the (linear) mixed models, which assume

$$\mu(\mathbf{x}, \mathbf{z}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} + \mathbf{z}^\top \mathbf{u}, \tag{2.3}$$

where $\mathbf{u} = [u_1, \cdots, u_q]^\top$ with $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$ are known as the *random effects* characterising the deviation of individual categories from the population mean, in contrast with the *fixed effects* β_0 and $\boldsymbol{\beta}$. Model (2.3) is also known as a *random intercept* model.

The central idea in (2.3) is that instead of assuming some fixed coefficient for each category, we assume that the effects of individual categories come from a distribution, so we only need to estimate (far fewer) parameters that govern the distribution of random effects rather than learning an independent parameter per category. This produces the equivalent effect of partial pooling (Gelman and Hill, 2007): categories with a smaller number of observations will have weaker influences on parameter estimates, and extreme values get regularised towards the collective mean (modelled by the fixed effects, i.e., $\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$).

In the same way that linear mixed models extend linear models, GLMMs extend GLMs by adding random effects capturing between-category variation to complement the fixed effects in the linear predictor.

Suppose that we have q categories in the data, each with n_i ($i = 1, 2, \cdots, q$) observations (so the total number of observations is $n = \sum_{i=1}^q n_i$). A GLMM is then defined by four components:

(1) *Unconditional distribution of random effects*. We assume that the random effects $u_j \overset{iid}{\sim} f(u_j|\boldsymbol{\gamma})$, $j = 1, 2, \cdots, q$ depend on a small set of parameters $\boldsymbol{\gamma}$. It is typically assumed that u_j follows a Gaussian distribution with zero mean and variance σ_u^2, that is, $u_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$.

(2) *Response distribution*. GLMMs assume that the responses Y_i, $i = 1, 2, \cdots, n$, given the random effect $u_{j[i]}$ for the category it belongs to (where $j[i]$ means that the i-th observation belongs to the j-th category, $j = 1, 2, \cdots, q$), are conditionally independent with an ED distribution, that is,

$$p\left(y_i|u_{j[i]}, \theta_i, \phi\right) = \exp\left[\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi} + c\left(y_i, \phi\right)\right], \tag{2.4}$$

where θ_i denotes the location parameter and ϕ the dispersion parameter for the ED density, and $b(\cdot)$ and $c(\cdot)$ are known functions. It follows from the properties of ED family distributions that

$$\mu_i = \mathbb{E}(Y_i|u_{j[i]}) = b'(\theta_i), \tag{2.5}$$

$$\text{Var}\left(Y_i|u_{j[i]}\right) = \phi b''(\theta_i) = \phi V(\mu_i),$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivatives of $b(\cdot)$ and $V(\cdot)$ is commonly referred to as the variance function. Equation (2.5) implies that the conditional distribution

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

of $Y_i|u_{j[i]}$ in (2.4) is completely specified by the conditional mean $\mu_i = \mathbb{E}(Y_i|u_{j[i]})$ and the dispersion parameter ϕ.

(3) *Linear predictor*. The linear predictor includes both fixed and random effects:

$$\eta_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \mathbf{u}, \quad i = 1, \cdots, n, \tag{2.6}$$

where $\mathbf{x}_i$ denotes a vector of fixed effects variables (e.g., sum insured, age) and $\mathbf{z}_i$ a vector of random effects variables (e.g., a binary representation of a high-cardinality feature such as injury code).

(4) *Link function*. Finally, a monotone differentiable function $g(\cdot)$ links the conditional mean μ_i with the linear predictor η_i :

$$g(\mu_i) = g(\mathbb{E}[Y_i|u_{j[i]}]) = \eta_i, \quad i = 1, \cdots, n,$$

completing the model.

As explained earlier, the addition of random effects allows GLMMs to account for correlation within the same category, without overfitting to an individual category as is the case of a one-hot encoded GLM. The shared parameters that govern the distribution of random effects essentially enable information to transfer between categories, which is helpful for learning.

2.3. Model Architecture

In Section 1.1, we briefly reviewed previous attempts at embedding ML into mixed models. In particular, the LMMNN structure proposed by Simchoni and Rosset (2022) and the network proposed by Mandel et al. (2022) are the closest to what we envision for a mixed effects neural network. However, LMMNN suffers from two major shortcomings that restrict its applicability to insurance contexts: first, it assumes a Gaussian response with an identity link function, for which analytical results can be obtained, but which is ill suited to modelling skewed, heavier-tailed distributions that dominate insurance and financial data. Second, LMMNNs model the random effects in a non-linear way (through a subnetwork in the structure). This complicates the interpretation of random effects, which carries practical significance in many applications. In contrast, the mixed neural network proposed by Mandel et al. (2022) uses a computationally less efficient estimation methodology, which limits its application to moderately sized datasets; see Section 2.4.

The GLMMNet aims to address these concerns. The architectural skeleton of the model is depicted in Figure 1. We adopt a very similar structure to that of Simchoni and Rosset (2022), except that we remove the subnetwork that they used to learn a non-linear function of $\mathbf{Z}$. As noted earlier, the main purpose of our modification is to retain the interpretability of random effect predictions from the GLMM's linear predictor. In addition, we also want to avoid an over-complicated structure for the random effects, whose role is to act as a regulariser (Gelman and Hill, 2007).

In mathematical terms, we assume that the conditional $\mathbf{y}|\mathbf{u}$ follows an ED family distribution, with

$$g(\boldsymbol{\mu}) = g(\mathbb{E}[\mathbf{y}|\mathbf{u}]) = f(\mathbf{X}) + \mathbf{Z}\mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $g(\cdot)$ is the link function and $f(\cdot)$ is learned through a neural network.

In the following, we give a detailed description of each component in the architecture plotted above, followed by a discussion on how the collective network can be trained in Section 2.4. We remark that, in this work, we have allowed $f(\cdot)$ to be fully flexible to take full advantage of the predictive power of neural networks. We recognise that in some applications, interpretability of the fixed effects may also be desirable. In such cases, it is possible to replace the fixed effects module (described in Section 2.3.1) by a Combined Actuarial Neural Network (CANN; Wüthrich and Merz 2019) structure.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

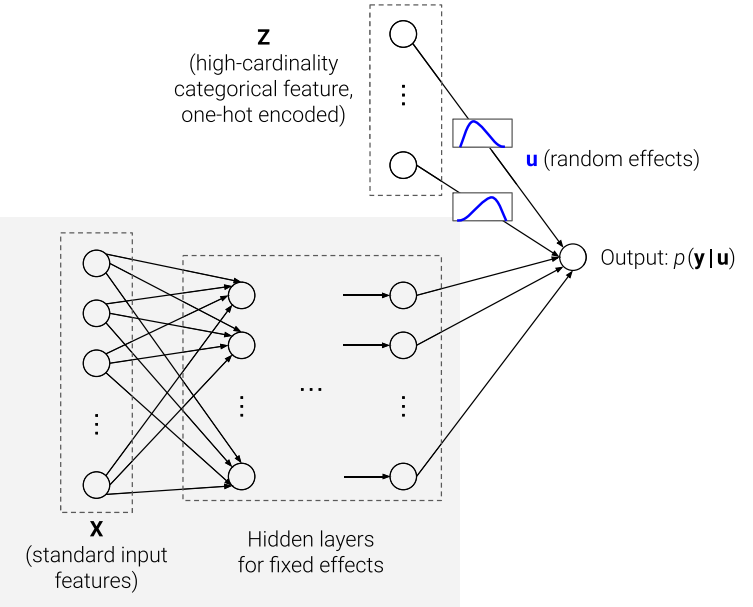


Figure 1. *Architecture of GLMMNet.*

2.3.1. Fixed Effects

The biggest limitation of GLMMs lies in their linear structure in (2.6): similar to GLMs, features need to be hand-crafted to allow for non-linearities and interactions (Richman, 2021a,b). The proposed GLMMNet addresses this issue by utilising a multi-layer network component for the fixed effects, which is represented as the shaded structure in Figure 1. For simplicity, here we consider a fully connected feed-forward neural network (FFNN), although many other network structures, for example, convolutional neural networks (CNNs), can also be easily accommodated.

A FFNN consists of multiple layers of neurons (represented by circles in the diagram) with non-linear activation functions to capture potentially complex relationships between input and output vectors; we refer to Goodfellow et al. (2016) for more details. Formally, for a network with L hidden layers and q_l neurons in the l-th layer for $l = 1, \cdots, L$, the l-th layer is defined in Equation (2.7):

$$\begin{aligned} &\mathbf{a}^{(l)} : \mathbb{R}^{q_{l-1}} \to \mathbb{R}^{q_l}, \\ &\qquad \mathbf{v} \mapsto \mathbf{a}^{(l)}(\mathbf{v}) = \left[a_1^{(l)}(\mathbf{v}), \cdots, a_{q_l}^{(l)}(\mathbf{v})\right]^\top, \end{aligned} \tag{2.7}$$

with

$$a_j^{(l)}(\mathbf{v}) = \varphi^{(l)}\left(b_j^{(l)} + \mathbf{w}_j^{(l)\top}\mathbf{v}\right), \quad j = 1, \cdots, q_l,$$

where $\varphi^{(l)} : \mathbb{R} \to \mathbb{R}$ is called the activation function for the l-th layer, $b_j^{(l)} \in \mathbb{R}$ and $\mathbf{w}_j^{(l)} \in \mathbb{R}^{q_{l-1}}$, respectively represent the bias term (or intercept in statistical terminology) and the network weights (or regression coefficients) for the j-th neuron in the l-th layer, and q_0 is defined as the number of (preprocessed) covariates entering the network.

The network therefore provides a mapping from the (preprocessed) covariates $\mathbf{x}$ to the desired output layer through a composition of hidden layers:

$$\begin{aligned} &f : \mathbb{R}^{q_0} \to \mathbb{R}^{q_{L+1}} \\ &\qquad \mathbf{x} \mapsto f(\mathbf{x}) = \left(\mathbf{a}^{(L+1)} \circ \cdots \circ \mathbf{a}^{(1)}\right)(\mathbf{x}). \end{aligned}$$

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Training a network involves making many specific choices; Appendix B.2 of the supplementary material gives more details.

2.3.2. Random Effects

Below we give a description of the random effects component of GLMMNet, which is the top (unshaded) structure in Figure 1. For illustration, we stay with the case of a single high-cardinality categorical feature with q unique levels. The column of the categorical feature is first one-hot transformed into a binary matrix $\mathbf{Z}$ of size $n \times q$, which forms the input to the random effects component of the GLMMNet.

Here comes the key difference from the fixed effects network: The weights in the random effects layer, denoted by $\mathbf{u} = [u_1, u_2, \cdots, u_q]^\top$, are not fixed values but are assumed to come from a distribution. This way, instead of having to estimate the effects of every individual category—many of which may lack sufficient data to support a reliable estimate, we only need to estimate the far fewer parameters that govern the distribution of $\mathbf{u}$ (in our case, just a single parameter σ_u^2 in (2.8)).

A less visible yet important distinction of our GLMMNet from the LMMNN (Simchoni and Rosset, 2022) is that *our model takes a Bayesian approach to the estimation of random effects (as opposed to an exact likelihood approach in the LMMNN)*. The Bayesian approach helps sidestep some difficult numerical approximations of the (marginal) likelihood function (as detailed in Section 2.4). We should point out that, although the Bayesian approach is popular for estimating GLMMs (e.g. Bürkner, 2017), there are further computational challenges that prevent them from being applied to ML mixed models; we elaborate on this in Section 2.4.

Following the literature (McCulloch and Searle, 2001; Antonio and Beirlant, 2007), we assume that the random effects follow a normal distribution with zero mean and that the random effects are independent across categories:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{I}), \tag{2.8}$$

or equivalently, $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$ for $i = 1, 2, \cdots, q$. This is taken as the prior distribution of the random effects $\mathbf{u}$.

In the context of our research problem, we are interested in the predictions for $\mathbf{u}$, which indicate the deviation of individual categories from the population mean, or in other words, the excess risk they carry. Under the Bayesian framework, *given the posterior distribution* $p(\mathbf{u}|\mathcal{D})$, we can simply take the posterior mode $\hat{\mathbf{u}} = \text{argmax}_{\mathbf{u}}\, p(\mathbf{u}|\mathcal{D})$ as a point estimate, which is also referred to as the maximum a posteriori (MAP) estimate. Alternatively, we can also derive interval predictions from the posterior distribution to determine whether the deviations as captured by $\mathbf{u}$ carry statistical significance.

We should note that the estimation of the posterior distribution, which we treated as given in the above, represents a major challenge in Bayesian inference. We come back to this topic and discuss how we tackle it in Section 2.4.

It should also be pointed out that the extension to multiple categorical features is very straightforward: all we need is to add another random effects layer (and one more variance parameter to estimate for each additional feature).

2.3.3. Response Distribution

As explained earlier, one novel contribution of GLMMNet is its ability to accommodate many non-Gaussian distributions, notably gamma (commonly used for claim severity), (over-dispersed) Poisson (for claim counts), and Bernoulli (for binary classification tasks). Our extension to non-Gaussian distributions is an important prerequisite for insurance applications.

Precisely, we assume that the responses $Y_i, i = 1, 2, \cdots, n$, given the random effect $u_{j[i]}$ for the category it belongs to (where $j[i]$ means that the i-th observation belongs to the j-th category), are conditionally independent with a distribution in the ED family. Equation (2.5) implies that the conditional

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

distribution of $Y_i|u_{j[i]}$ is completely specified by the conditional mean μ_i and the dispersion parameter ϕ, so we can write[1]

$$y_i|u_{j[i]} \sim \text{ED}(\mu_i, \phi).$$

In the GLMMNet, we assume that a smooth and strictly monotone link function $g(\cdot)$ connects the conditional mean μ_i with the non-linear predictor η_i formed by adding together the output from the fixed effects module and the random effects layer:

$$g(\mu_i) = \eta_i = f(\mathbf{x}_i) + \mathbf{z}_i^\top \mathbf{u} = f(\mathbf{x}_i) + u_{j[i]}, \quad i = 1, \cdots, n. \tag{2.9}$$

In implementation, we use the inverse link $g^{-1}(\cdot)$ as the activation function for the final layer, to map the output $[f(\mathbf{x}_i), u_{j[i]}]$ from the fixed effects module and the random effects layer to a prediction for the conditional mean response $\mu_i = g^{-1}(f(\mathbf{x}_i) + u_{j[i]})$.

Finally, as GLMMNet operates under probabilistic assumptions, the dispersion parameter ϕ (independent of covariates in our setting) can be estimated under maximum likelihood as part of the network. This is implemented by adding an input-independent but trainable weight to the final layer of GLMMNet (Chollet et al., 2015), whose value will be updated by (stochastic) gradient descent on the loss function.

2.4. Training of GLMMNet: Variational Inference

In a LMMNN (Simchoni and Rosset, 2022), the network is trained by minimising the negative log-likelihood of $\mathbf{y}$. As the LMMNN assumes a Gaussian density for $\mathbf{y}|\mathbf{u}$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, the marginal likelihood of $\mathbf{y}$ can be derived analytically, that is, $\mathbf{y} \sim \mathcal{N}(f(\mathbf{X}), \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^\top + \sigma_\epsilon^2\mathbf{I})$. The network weights and biases, as well as the variance and covariance parameters σ_ϵ^2 and σ_u^2, are learned by minimising the negative log-likelihood:

$$-\log \mathcal{L} = \frac{1}{2}(\mathbf{y} - f(\mathbf{X}))^\top \mathbf{V}^{-1}(\mathbf{y} - f(\mathbf{X})) + \frac{1}{2}\log\det(\mathbf{V}) + \frac{n}{2}\log(2\pi),$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^\top + \sigma_\epsilon^2\mathbf{I}$ and $\boldsymbol{\Sigma} := \text{Cov}(\mathbf{u}) = \sigma_u^2\mathbf{I}$.

In making the extension to GLMMNet, however, we no longer obtain a closed-form expression for the marginal likelihood:

$$\begin{aligned}\mathcal{L}(\mathbf{y}; \boldsymbol{\beta}, \sigma_u, \phi) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \sigma_u, \phi) = \int \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \mathbf{u}, \phi)\pi(\mathbf{u}|\sigma_u)\mathrm{d}\mathbf{u} \\ &= \int \prod_{i=1}^{n}\prod_{j=1}^{q} p(y_i|\boldsymbol{\beta}, u_j, \phi)\pi(u_j|\sigma_u)\mathrm{d}u_j \end{aligned} \tag{2.10}$$

where $\boldsymbol{\beta}$ denote the weights and biases in the fixed effects component of the GLMMNet (Section 2.3.1), σ_u^2 is the variance parameter of the random effects, and ϕ is the usual dispersion parameter for the ED density for the response. In the work of Mandel et al. (2022), the authors used the Laplace method to approximate the integrand so that the resulting integral can be solved analytically. However, the Laplace method requires calculating and inverting the Hessian matrix, which can be very computationally expensive or even infeasible in high-dimensional settings (Zhang et al., 2019).

We choose to take a Bayesian approach to circumvent the difficult numerical approximations required to compute the integral in Equation (2.10). Under the Bayesian framework, $\pi(u_j) = \pi(u_j|\sigma_u) = \mathcal{N}(0, \sigma_u^2)$, $j = 1, \cdots, q$, is taken as the prior for the random effects. Our goal is to make inference on the posterior of the random effects given by:

[1]The assumption of constant dispersion parameter ϕ is made to comply with the standard assumptions of GLM/GLMMs, but can be relaxed if desired by a slight modification. For the purpose of illustration in this work, we do not pursue this path down further but note that it can be an interesting direction for future research.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

$$p(\mathbf{u}|\mathcal{D}) = \frac{\pi(\mathbf{u})p(\mathcal{D}|\mathbf{u})}{p(\mathcal{D})} = \frac{\pi(\mathbf{u})p(\mathcal{D}|\mathbf{u})}{\int \pi(\mathbf{u})p(\mathcal{D}|\mathbf{u})\mathrm{d}\mathbf{u}}, \tag{2.11}$$

where $\pi(\mathbf{u}) = \prod_{j=1}^{q} \pi(u_j)$ and $\mathcal{D}$ represents the data. Note that we again encounter an intractable integral in (2.11). The traditional solution is to use Markov chain Monte Carlo (MCMC) to sample from the posterior without computing its exact form; however, the computational burden of MCMC techniques restricts its applicability to large and complex models like a deep neural network. This partly explains why Bayesian methods, although frequently used for estimating GLMMs (e.g., Bürkner, 2017), tend not to be as widely adopted among ML mixed models (e.g., Sigrist, 2022; Hajjem et al., 2014).

With our GLMMNet, we apply *variational inference* to solve this problem. Variational inference is a popular approach among ML researchers working with Bayesian neural networks due to its computational efficiency (Blei et al., 2017; Zhang et al., 2019). Variational inference does not try to directly estimate the posterior but proposes a surrogate parametric distribution $q \in \mathcal{Q}$ to approximate it, therefore turning the difficult estimation into a highly efficient optimisation problem. This contrasts with the computational burden of the Laplace approximation used in Mandel et al. (2022): in our experiments, we found that the runtime of the model by Mandel et al. (2022) can easily be on the magnitude of hours for a moderately sized dataset of less than 10,000 rows.

In practice, the choice of the surrogate distribution is often made based on simplicity or computational feasibility. One particularly popular option is to use a mean-field distribution family (Blei et al., 2017), which assumes independence among all latent variables. In this work, we mainly consider a diagonal Gaussian distribution for the surrogate posterior, which is computationally convenient (as we will see below):

$$q_{\boldsymbol{\vartheta}}(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u),$$

where $\boldsymbol{\mu}_u = (\mu_1, \mu_2, \cdots, \mu_q)^\top$ is a q-dimensional vector of the surrogate posterior mean of the q random effects, and $\boldsymbol{\Sigma}_u = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2)$ is a diagonal matrix of the posterior variance. More complex distributions, such as those that incorporate dependencies among latent variables, may provide a more accurate approximation. We explore this topic in Section 4.4.

The vector $\boldsymbol{\mu}_u$ and the diagonal elements of $\boldsymbol{\Sigma}_u$ together form the *variational parameters* of the diagonal Gaussian, denoted by $\boldsymbol{\vartheta}$. The task here is to find the variational parameters $\boldsymbol{\vartheta}^*$ that make the approximating density $q_{\boldsymbol{\vartheta}} \in \mathcal{Q}$ closest to the true posterior, where closeness is defined in the sense of minimising the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). In mathematical terms, this means

$$\begin{aligned}
q_{\boldsymbol{\vartheta}^*}(\mathbf{u}) &= \underset{q_{\boldsymbol{\vartheta}}(\mathbf{u})\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}[q_{\boldsymbol{\vartheta}}(\mathbf{u})\|p(\mathbf{u}|\mathcal{D})] \\
&= \underset{q_{\boldsymbol{\vartheta}}(\mathbf{u})\in\mathcal{Q}}{\operatorname{argmin}}\, \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}[\log q_{\boldsymbol{\vartheta}}(\mathbf{u}) - \log p(\mathbf{u}|\mathcal{D})] \\
&= \underset{q_{\boldsymbol{\vartheta}}(\mathbf{u})\in\mathcal{Q}}{\operatorname{argmin}}\, \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}[\log q_{\boldsymbol{\vartheta}}(\mathbf{u}) - \log \pi(\mathbf{u}) - \log p(\mathcal{D}|\mathbf{u})] \\
&= \underset{q_{\boldsymbol{\vartheta}}(\mathbf{u})\in\mathcal{Q}}{\operatorname{argmin}} \left(\mathrm{KL}\left[q_{\boldsymbol{\vartheta}}(\mathbf{u})\|\pi(\mathbf{u})\right] - \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}[\log p(\mathcal{D}|\mathbf{u})]\right).
\end{aligned}$$

We take the term inside the outermost bracket, which is called the *evidence lower bound* (ELBO) loss or the negative of *variational free energy* in Neal and Hinton (1998), to be the loss function for the GLMMNet:

$$\mathcal{L}_{\mathrm{GLMMNet}} = \mathrm{KL}\left[q_{\boldsymbol{\vartheta}}(\mathbf{u})\|\pi(\mathbf{u})\right] - \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}[\log p(\mathcal{D}|\boldsymbol{\beta}, \mathbf{u}, \phi)] \tag{2.12}$$

$$= \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}\left[\log \frac{q_{\boldsymbol{\vartheta}}(\mathbf{u})}{\pi(\mathbf{u})}\right] - \mathbb{E}_{q_{\boldsymbol{\vartheta}}(\mathbf{u})}[\log p(\mathcal{D}|\boldsymbol{\beta}, \mathbf{u}, \phi)]. \tag{2.13}$$

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

We see that the loss function used to optimise the GLMMNet is a sum of two parts: a first part that captures the divergence between the (surrogate) posterior and the prior, and a second part that captures the likelihood of observing the given data under the posterior. This decomposition echoes the trade-off between the prior beliefs and the data in any Bayesian analysis (Blei et al., 2017). It is also worth pointing out that there is a trade-off between the dispersion parameter ϕ and the random effects variance $\mathbf{\Sigma}_u$ which goes into $\boldsymbol{\vartheta}$. Overly low estimates of ϕ can require compensatory "inflation" in the estimated $\mathbf{\Sigma}_u$ to account for the greater unexplained noise manifesting as increased group heterogeneity. Conversely, overestimated dispersion may prevent the random effects from being accurately estimated.

The loss function in (2.13) can be calculated via Monte Carlo, that is, by generating posterior samples under the current estimates of the variational parameters $\boldsymbol{\vartheta}$, evaluating the respective expressions at each of the sampled values, and taking the average to approximate the expectation. Note that the mean-field Gaussian approximation is mathematically convenient as it allows for the explicit calculation of the first term in Equation (2.12); see Example 11.20 of Wüthrich and Merz (2023, p.535):

$$\mathrm{KL}\left[q_{\boldsymbol{\vartheta}}(\mathbf{u})\|\pi(\mathbf{u})\right] = -\frac{q}{2} + \frac{q}{2}\log\left(\sigma_u^2\right) - \frac{1}{2}\sum_{j=1}^{q}\left(\log\left(\sigma_j^2\right) - \frac{\sigma_j^2 + \mu_j^2}{\sigma_u^2}\right).$$

The derivatives of the loss function —with respect to the network parameters $\boldsymbol{\beta}$, the dispersion parameter ϕ, and the variational parameters $\boldsymbol{\vartheta}$— can be computed via automatic differentiation (Chollet et al., 2015) and used by standard gradient descent algorithms for optimisation (e.g., Kingma and Ba, 2014). Appendix A of the supplementary material discusses the practicalities of implementing and training the GLMMNet.

2.5. Prediction from GLMMNet

Recall that $\mathbf{x}$ denotes the standard input features and $\mathbf{z} \in \mathbb{R}^q$ is a binary representation of the high-cardinality feature. Let $(\mathbf{x}^*, \mathbf{z}^*)$ represent a new data observation for which a prediction should be made. Note that when making predictions, it is possible to encounter a category that was never seen during training, in which case $\mathbf{z}^*$ will be a zero vector (as it does not belong to any of the already-seen categories) and the predictor in Equation (2.9) reduces to $\eta^* = f(\mathbf{x}^*)$.

To make predictions from GLMMNet, we take expectations under the posterior distribution on the random effects (Blundell et al., 2015; Jospin et al., 2022):

$$\begin{aligned} p(y^*|\mathbf{x}^*, \mathbf{z}^*) &= \mathbb{E}_{p(\mathbf{u}|\mathcal{D})}\left[p(y^*|\mathbf{x}^*, \mathbf{z}^*, \mathbf{u})\right] \\ &\approx \mathbb{E}_{q_{\boldsymbol{\vartheta}^*}(\mathbf{u})}\left[p(y^*|\mathbf{x}^*, \mathbf{z}^*, \mathbf{u})\right] \end{aligned} \tag{2.14}$$

where $q_{\boldsymbol{\vartheta}^*}(\cdot)$ denotes the optimised surrogate posterior. As a result of Equation (2.14), any functional of $y^*|\mathbf{x}^*, \mathbf{z}^*$, for example, the expectation, can be computed in an analogous way.

The expectation in (2.14) can be approximated by drawing Monte Carlo samples from the (surrogate) posterior, as outlined in Algorithm 1. When there are multiple high-cardinality features, under the assumption of independence between these features, each will have its own binary vector representation and optimised surrogate posterior, and the expectation in Equation (2.14) will be taken with respect to the joint posterior.

It is also worth pointing out that the model averaging in line 5 has the equivalent effect of training an ensemble of networks, which is usually otherwise computationally impractical (Blundell et al., 2015).

3. Comparison of GLMMNet with Other Leading Models

In this section, we compare the performance of the GLMMNet against the most popular existing alternatives for handling high-cardinality categorical features. We list the candidate models in Section 3.1. The performance metrics we use to evaluate the models are detailed in Appendix B.5 of

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Algorithm 1. Prediction from GLMMNet

Input : A new data observation $(\mathbf{x}^*, \mathbf{z}^*)$; number of Monte Carlo iterations N
Output: Predictive distribution: $p(y^*|\mathbf{x}^*, \mathbf{z}^*)$
for $i = 1$ *to* N **do**
  Simulate $\mathbf{u}_{(i)}$ from the surrogate posterior $q_{\vartheta^*}(\mathbf{u})$;
  Set $p_{(i)} = p(y^*|\mathbf{x}^*, \mathbf{z}^*, \mathbf{u}_{(i)}) = \text{ED}(\mu_{(i)}, \phi)$ where $\mu_{(i)} = g^{-1}\left(f(\mathbf{x}^*) + \mathbf{z}^{*\top}\mathbf{u}_{(i)}\right)$ and ϕ is the dispersion parameter for the ED distribution (learned in the GLMMNet);
end
return $p(y^*|\mathbf{x}^*, \mathbf{z}^*) = \frac{1}{N}\sum_{i=1}^{N} p_{(i)}$;

the supplementary material. Control over the true data generating process via simulation allows us to highlight the respective strengths of the models. We challenge the models under environments of varying levels of complexity, for example, low signal-to-noise ratio, highly imbalanced distribution of categories, and/or skewed response distributions. The simulation environments detailed in Section 3.2 are designed to mimic these typical characteristics of insurance data. Section 3.3 summarises the results. Ensuing insights are used to inform the real data case study in Section 4.

3.1. Models for Comparison

Listed below are the leading models widely used in practice that allow for high-cardinality categorical features. They will be implemented and used for comparison with the GLMMNet in this experiment.

1. GLM with
   - `GLM_ignore_cat`: complete pooling, that is, ignoring the categorical feature altogether. In the notation of Section 2.1, it can be represented as $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(\beta_0 + \mathbf{x}^\top\boldsymbol{\beta})$, where $g(\cdot)$ is the link function, and β_0 and $\boldsymbol{\beta}$ are the regression parameters to be estimated;
   - `GLM_one_hot`: one-hot encoding, which can be represented as $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(\beta_0 + \mathbf{x}^\top\boldsymbol{\beta} + \mathbf{z}^\top\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ are the additional parameters for each category in $\mathbf{z}$;
   - `GLM_GLMM_enc`: GLMM encoding, which can be represented as $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(\beta_0 + \mathbf{x}^\top\boldsymbol{\beta} + z'\alpha)$ where z' represents the encoded value (a scalar) of $\mathbf{z}$ and α the additional parameter associated with it. The scalar encodings are computed as the cross-validated posterior estimates of a random intercept GLMM (more details in Appendix B.1 of the supplementary material);
2. Gradient boosting machine (GBM, with trees as base learners) under the same categorical encoding setup as above:
   - `GBM_ignore_cat`: $\mu(\mathbf{x}, \mathbf{z}) = \mu(\mathbf{x})$ where $\mu(\cdot)$ is a weighted sum of tree learners;
   - `GBM_one_hot`: $\mu(\mathbf{x}, \mathbf{z})$ where $\mu(\cdot, \cdot)$ is a weighted sum of tree learners;
   - `GBM_GLMM_enc`: $\mu(\mathbf{x}, \mathbf{z}) = \mu(\mathbf{x}, z')$ where z' represents the encoded value (a scalar) of $\mathbf{z}$;
3. Neural network with entity embeddings (`NN_ee`). Here, $\mu(\mathbf{x}, \mathbf{z})$ is composed of multiple layers of interconnected neurons, where each neuron receives inputs, performs a weighted sum of these inputs, and applies an activation function to produce its output (see also Section 2.3.1). Entity embeddings add a linear layer between each one-hot encoded input and the first hidden layer and are learned as part of the training process. This is currently the most popular approach for handling categorical inputs in deep learning models and serves as our **target benchmark**;
4. GBM with pre-learned entity embeddings (`GBM_ee`), which is similar to `GBM_GLMM_enc` except with the z' replaced by a vector of entity embeddings learned from `NN_ee` ;

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

5. Mixed models:
   - `GLMM` with $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta} + \mathbf{z}^\top \mathbf{u})$ where $\mathbf{u}$ are the random effects (Section 2.2);
   - `GPBoost` (Sigrist, 2021, 2022) with $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(f(\mathbf{x}) + \mathbf{z}^\top \mathbf{u})$ where $f(\cdot)$ is an ensemble of tree learners;
   - The proposed `GLMMNet`, with $\mu(\mathbf{x}, \mathbf{z}) = g^{-1}(f(\mathbf{x}) + \mathbf{z}^\top \mathbf{u})$ where $f(\cdot)$ is a FFNN (Section 2.3.1).

Each model has its own features and properties, which renders the collection of models suitable for different contexts; Table 1 gives a high-level summary of their different characteristics. In the table, "interpretability" refers to the ease with which a human can understand a model, as well as the ability to derive insights from the model. For instance, GLMs have high interpretability, because GLMs have a small number of parameters and each parameter carries a physical meaning, which contrasts with the case of a standard deep neural network characterised by a large number of parameters that are meaningless on their own. "Operational complexity" refers to the resources and time required to implement, train and fine-tune the model. The last column, "bias-variance focus" gives a taste of the primary focus of each algorithm in regards to the bias variance trad-eoff; note that the descriptions can be subjective and should be interpreted in the context of the entire range of methods presented.

3.2. Simulation Datasets

We generate n observations from a (generalised) non-linear mixed model with q categories ($q \leq n$). We assume that the responses $y_i, i = 1, \cdots, n$, each belonging to a category $j[i] \in \{1, \cdots, q\}$, given the random effects $\mathbf{u} = (u_j)_{j=1}^q$, are conditionally independent with a distribution in the ED family (e.g., gamma), denoted as $y_i | u_{j[i]} \sim \text{ED}(\mu_i, \phi)$, where $\mu_i = \mathbb{E}(y_i | u_{j[i]})$ represents the mean and ϕ represents a dispersion parameter shared across all observations. A link function $g(\cdot)$ connects the conditional mean μ_i with a non-linear predictor in the form of

$$g(\mu_i) = g(\mathbb{E}[y_i | u_{j[i]}]) = f(\mathbf{x}_i) + u_{j[i]}, \quad i = 1, \cdots, n, \tag{3.1}$$

where $\mathbf{x}_i$ is a vector of fixed effects covariates for the i-th observation, $\mathbf{z}_i$ is a vector of random effects variables (in our case, a binary vector representation of a high-cardinality categorical feature), the random effects $u_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$ measure the deviation of individual categories from the population mean, and $f(\cdot)$ is some complex non-linear function of the fixed effects. Equation (3.1) also implies that there is no interaction between the high-cardinality categorical variable and the remaining covariates other than an additive relationship in the non-linear predictor. This can be a strong assumption, and we discuss it in Section 3.3.

For the fixed effects component f, we consider

$$f(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20 (x_3 - 0.5)^2 + 10 x_4 + 5 x_5, \tag{3.2}$$

$$x_i \overset{iid}{\sim} \mathcal{U}(0, 1), \; i = 1, \cdots, 10,$$

which was studied in Friedman (1991) and has become a standard simulation function for regression models (used in, e.g., Kuss et al. 2005; also available in the scikit-learn Python library by Pedregosa et al. 2011). In our application, we use the Friedman function to test the model's ability to filter out irrelevant predictors (note that $x_6, \cdots, x_{10}$ is not used in f) as well as detect interactions and non-linearities of input features, all in the presence of a high-cardinality grouping variable that is subject to random effects.

We set up the desired simulation environments by adjusting the following parameters; refer to Appendix B.4 of the supplementary material for details.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Table 1. *Overview of different characteristics of candidate models*

	Predictive performance			Other considerations		
	Handling non-linearities	*Categorical treatment*	*Response distribution*	*Interpretability*	*Operational complexity*	*Bias-variance focus*
GLM	None (unless performed manually)	Optional categorical encoding methods can be used to preprocess features, for example, one-hot encoding	ED family	High (fully transparent model)	Low	Balanced
GBM	Good	Optional categorical encoding methods can be used to preprocess features, for example, one-hot encoding	Gaussian (alternatives can be accommodated by specifying a different loss function)	Medium (through partial dependence plots and feature importance analysis)	Moderate	Reducing bias
GLMM	None (unless performed manually)	Through random effects, which are an addition to GLMs (see Section 2.2)	ED family	High (fully transparent model)	Low	Balanced
Entity embeddings (neural network)	Excellent	Through entity embeddings	Gaussian (alternatives can be accommodated by specifying a different loss function)	Low	High	Reducing bias
GPBoost	Good	Through random effects (see Section 3.1)	Selected ED family distributions (Bernoulli, Poisson, gamma, and Gaussian)	Medium (transparent random effects)	Moderate	Reducing bias
GLMMNet	Excellent	Through random effects (see Section 2.3)	ED family	Medium (transparent random effects)	High	Reducing bias

Table 2. *Parameters used for the different simulation environments. Bold face indicates changes from the base scenario (i.e., experiment 1).*

Exp ID	Signal-to-noise	Response distribution	Inverse link	Distribution of categories
1 (base)	[4, 1, 1] (high)	Gaussian	Identity	Balanced
2	[4, 1, 1] (high)	**Gamma**	**Exponential**	Balanced
3	[**8, 1, 4**] (low)	Gaussian	Identity	Balanced
4	[**8, 1, 4**] (low)	**Gamma**	**Exponential**	**Skewed**

- *Signal-to-noise ratio*: a three-dimensional vector that captures the relative ratio of signal strength (as measured by μ_f, the mean of $f(\mathbf{x})$), random effects variance (σ_u^2), and variability of the response (σ_ϵ^2; noise, or equivalently the irreducible error, as this component captures the unexplained inherent randomness in the response).
- *Response distribution*: distributional assumption for the response, for example, Gaussian, gamma, or any other member of the ED family.
- *Inverse link*: inverse of the link function, that is, the inverse function of $g(\cdot)$ in Equation (3.1).
- *Distribution of categories*: whether to use balanced or skewed distribution for the allocation of categories. A "balanced" distribution allocates approximately equal number of observations to each category; a "skewed" distribution generates categories from a (scaled) beta distribution.

These simulation parameters allow us to mimic specific characteristics of practical insurance data, such as a low signal-to-noise ratio (the specification of [8,1,4] was selected based on estimated parameters from the real insurance case study in Section 4), an imbalanced distribution across categories, or a highly skewed response distribution for claim severity. Table 2 lists the simulation environments considered in our experiment, along with a brief description below.

- *Experiment 1* simulates the base scenario that adheres to the assumptions of a Gaussian GLMMNet.
- *Experiment 2* simulates a gamma-distributed response, which is often used to model claim severity in lines of business (LoB) such as auto insurance or general liability.
- *Experiment 3* dials up the level of noise in the data and simulate LoBs that are difficult to model by covariates, such as commercial building insurance, catastrophic events, and cyber risks.
- *Experiment 4* represents the most challenging scenario, incorporating the complexities of all the simulation parameters.

For each scenario, we generate 5,000 training observations and 2,500 testing observations, with $q = 100$ categories.

3.3. Results

Figure 2 compares the out-of-sample continuous ranked probability score ("CRPS") of the candidate models (listed in Section 3.1)[2]. The CRPS is a quadratic probabilistic measure of the difference between the forecast predictive distribution and the empirical distribution of the observation. Hence, smaller CRPS values indicate a better fit; we refer to Appendix B.5 of the supplementary material for its mathematical definition. Each simulation experiment is repeated 50 times, with 5,000 training and 2,500

[2]The `GPBoost` results presented below were generated using its R package. We are aware of better results from the Python implementation and using a slightly different hyperparameter tuning approach (although still below that of `NN_ee` and `GLMMNet` in our examples) but have not included these for consistency across the comparisons. We recommend that users consider also `GPBoost` in their analysis, which offers a similar level of interpretability for the random effects as `GLMMNet` and can be particularly competitive on datasets where tree-based approaches tend to outperform neural networks.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

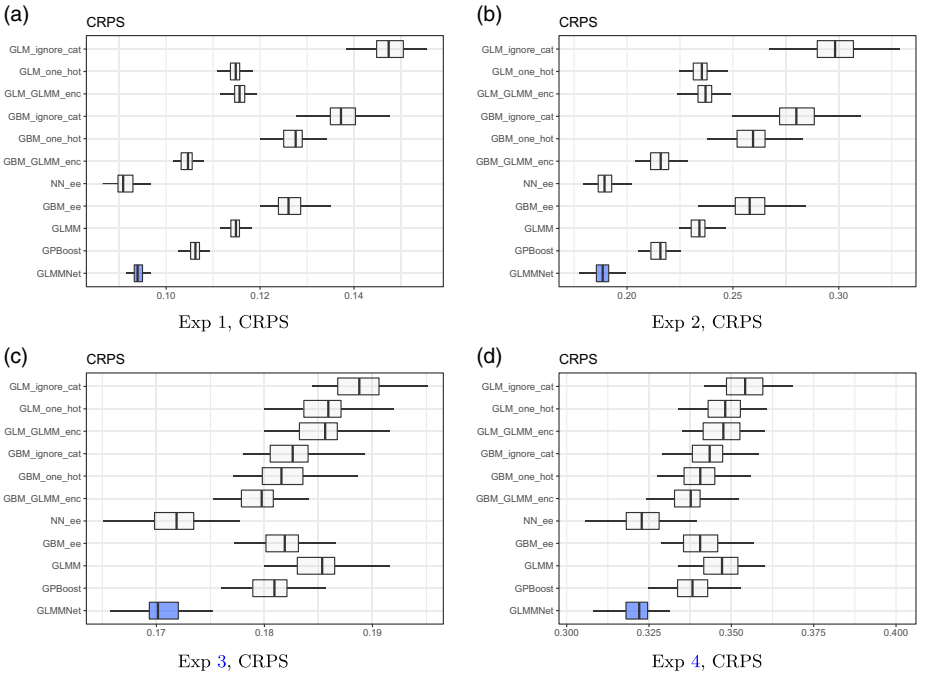


Figure 2. *Boxplots of out-of-sample CRPS of the different models in Experiments 1–4; GLMMNet highlighted in blue.*

testing observations each. The boxplots show the distribution of the out-of-sample metrics from the repeated simulation runs; the proposed GLMMNet is highlighted.

From Figure 2, we see that `GLMMNet` and `NN_ee` (neural network with entity embeddings) are leading in all experiments. In particular, although `GLMMNet` is outperformed by `NN_ee` in the base case experiment 1, it shows an advantage in high-noise environments (experiments 3–4). The random effects part of the GLMMNet acts effectively as a regularising component that helps the network navigate the data better when there is a lot of underlying noise, which is quite typically the scenario in practical actuarial applications.

Besides `GLMMNet` and `NN_ee`, the best-performing models are `GPBoost` and `GBM_GLMM_enc` (a boosting model with GLMM encoding). Both models significantly outperform the linear family of models (GLM or GLMM), showing that a flexible model structure is required to capture the non-linearities in the data. This observation holds regardless of the level of noise in the data, the response distribution, or the distribution of the categorical variable.

More importantly, without a suitable way of accounting for the high-cardinality categorical feature, a flexible model structure alone is not enough to achieve good performance. For example, the struggle of tree-based models in dealing with categorical features has been well documented (Prokhorenkova et al., 2018). In this experiment, with the usual one-hot encoding, the GBM model (`GBM_one_hot`) does not perform much better—indeed, a bit worse in experiments 1 and 2—than its GLM counterpart (`GLM_one_hot`). This confirms the motivation for this research: more flexible ML models, when dealing with financial or insurance data, can often be constrained by their capability to model categorical variables with many levels.

Across all experiments, we find that the `GLMMNet` and `NN_ee` consistently outperform the other models in all scenarios studied. We find the GLMMNet to be an equally competitive model as our target benchmark of an entity-embedded neural network. The GLMMNet is more competitive in higher noise environments or when the response distribution deviates from Gaussian, although it tends to be outperformed by the entity-embedded neural network in lower noise Gaussian environments. We should

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Table 3. *Model comparison: strengths and limitations of the top-performing models*

	Strengths	Limitations
`NN_ee`	• Strong predictive performance, particularly in low-medium noise environments • Flexible structure in which the entity embeddings can interact with continuous features	• Compromised performance in high-noise environments • Limited interpretability
`GBM_GLMM_enc`	• Good predictive performance • Simpler structure and faster to train • Transparency on feature importance	• Compromised performance compared to network models • Limited interpretability of the effects of the categories
`GLMMNet`	• Consistently strong predictive performance, particularly in high -noise environments • Transparency on the category effects (through random effects predictions) • Offers naturally probabilistic estimates	• Compromised performance in low-noise Gaussian environments compared to entity embeddings • Limited interpretability of the fixed effects component • Limited ability to handle interactions between categorical and continuous features

also point out that the simulation experiments considered here do not include any non-trivial interactions between the high-cardinality categorical variable and the remaining features, which pose a specific challenge to the GLMMNet. Unlike the entity embedding model, GLMMNet simplifies this relationship to an additive form in the last network layer in order to preserve the interpretability of the random effects (equivalently, the high-cardinality categorical variable), as we will see in Section 4.3. We refer to Richman and Wüthrich (2023) for an in-depth study on the modelling options available when there exist intricate relationships between categorical and continuous features.

In practice, predictive performance is not the only criterion for model selection. Each model has its own strengths and limitations, and the choice of which model to use will depend on the specific needs of the modelling problem. Table 3 lists some of these considerations. We should also note that `GBM_ee` and `GPBoost` has similar properties to `GBM_GLMM_enc` and is therefore not included in the table for brevity of presentation (except for the fact that `GBM_ee` is more complicated to implement as it requires training an entity-embedded neural network beforehand).

4. Application to Real Insurance Data

In this section, we apply GLMMNet to a real (proprietary) insurance dataset (described in Section 4.1) and discuss its performance relative to the other models we have considered (Section 4.2). We also demonstrate, in Section 4.3, how such an analysis can potentially inform practitioners in the context of technical pricing.

4.1. Description of Data

The data for this analysis were provided by a major Australian insurer. The original data cover 27,351 commercial building and contents reported claims by small- and medium-sized enterprises (SME)

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

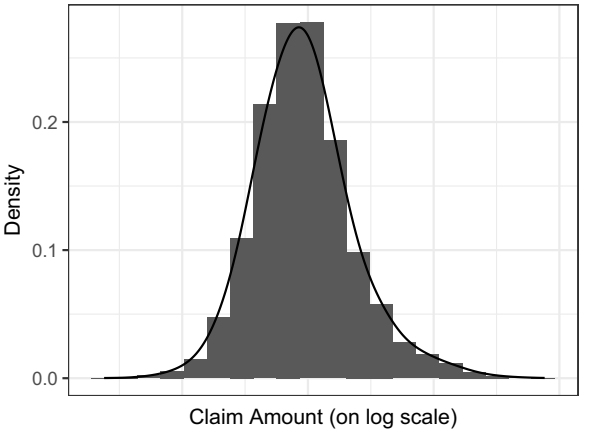


Figure 3. *Histogram and Gaussian kernel density estimate of claim amounts (on log scale). The x-axis numbers have been deliberately removed for confidentiality reasons.*

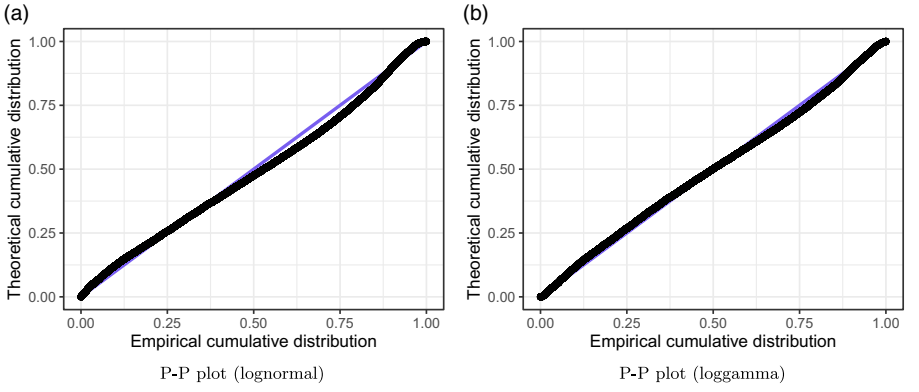


Figure 4. *Probability–probability (P-P) plot of empirical versus fitted lognormal and loggamma (theoretical) distributions. The parametric distributions are fitted to the (unlogged) marginal, that is, without any covariates.*

over the period between 2010 and 2015. The analysis seeks to construct a regression model to predict the ultimate costs of the reported claims based on other individual claim characteristics that can be observed early in the claims process (e.g., policy-level details and claim causes). A description of the (post-engineered) input features is given in Appendix B.6.1 of the supplementary material.

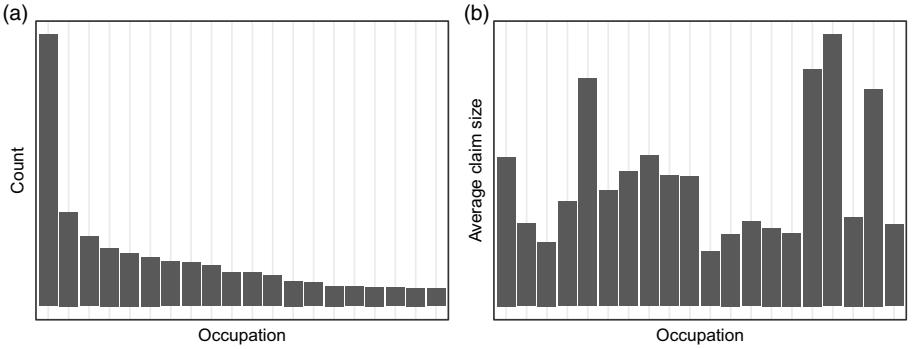
The response variable is the claim amount (claim severity), which has a very skewed distribution, as shown in Figure 3: even on a log scale, we can still observe some degree of positive skewness. In search for a suitable distribution to model the severity, we fit both lognormal and loggamma distributions to the (unlogged) marginal. Figure 4 indicates that both models may be suitable, with loggamma being slightly more advantageous. We will test both lognormal and loggamma distributions in our experiments below (Section 4.2); specifically, we fit Gaussian and gamma models to the log-transformed response.

The high-cardinality categorical variable of interest is the occupation of the business, which is coded through the Australian and New Zealand Standard Industrial Classification (ANZSIC) system. The ANZSIC system hierarchically classifies occupations into Divisions, Subdivisions, Groups, and Classes (from broadest to finest). See Table 4 for an example.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Table 4. *An example of ANZSIC occupation classification*

Division	K	Financial and Insurance Services
Subdivision	63	Insurance and Superannuation Funds
Group	632	Health and General Insurance
Class	6322	General Insurance



Number of claims from the top 20 common occupation classes

Average claim size for the (same) top 20 common occupation classes

Figure 5. *Skewed distribution of occupation class and heterogeneity in experience. The axes labels have been removed to preserve confidentiality.*

We will look at occupations at the Class level.[3] There are over 300 unique levels of such occupation classes in this dataset. Panel (a) of Figure 5 highlights the skewed distribution of occupations: the most common occupation has more than double the number of observations than the second most common, and the number of observations decays rapidly for the rarer classes. Panel (b) shows the heterogeneity in claims experiences between different occupation classes; calculations reveal a coefficient of variation of 160% for the occupation means. One challenge in modelling the occupation variable lies in determining how much confidence we can have in the observed claims experiences; intuitively, we should trust the estimates more when there are more data points and less when there are few data points. As explained in Section 2.2, the mixed effects models are one solution to this problem, which we will explore further in this section.

The dataset under consideration displays many typical characteristics of insurance data that we studied in Section 3 (see Section 3.2 in particular). For example, it is characterised by an extremely skewed response distribution (Figure 3), a low signal-to-noise ratio (there is little information that covariates can add to the marginal fits, as implied by Figure 4) and a highly disproportionate distribution of categories (Figure 5). These features of the dataset best align with those of Experiment 6 in Section 3. If the conclusions generalise, we expect that our GLMMNet will show strong performance.

[3]We chose not to incorporate the hierarchical information. Instead, we directly modelled the lowest level (i.e., Class) as a flat categorical feature. The decision was made in light of (1) the increased complexity of estimation in hierarchical models, and (2) the lack of alignment between the hierarchical system built for general-purpose occupation classification and the factors that differentiate claim severities.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Table 5. *Comparison of lognormal and loggamma model performance (median absolute error, CRPS, negative log-likelihood, RMSE of average prediction per category) on the test (out-of-sample) set. The best values are bolded.*

	Lognormal (out-of-sample)				Loggamma (out-of-sample)			
	MedAE	CRPS	NLL	Category RMSE	MedAE	CRPS	NLL	Category RMSE
GLM_one_hot	4108	0.7931	9.623	0.4569	1946	0.8557	9.751	0.5677
GBM_GLMM_enc	3870	0.7666	9.584	0.4027	**1536**	0.7626	9.578	0.4027
NN_ee	3803	**0.7624**	**9.578**	0.3928	1566	**0.7600**	9.575	0.4007
GBM_ee	3828	0.7665	9.584	0.4144	1549	0.7629	9.579	0.4128
GLMM	3864	0.7666	9.584	0.3978	1570	0.7629	9.577	0.3975
GLMMNet	3736	0.7681	9.587	**0.3892**	1537	0.7706	9.588	0.3997
GLMMNet_l2	**3545**	0.7626	9.579	0.3947	1551	**0.7600**	**9.574**	**0.3961**

4.2. Modelling Results

We consider the same candidate models as before; see Section 3.1. We also consider an ℓ_2-regularised GLMMNet (`GLMMNet_l2`) in the hope that the regularisation would further improve its performance in a high-noise environment.

We split the data into a 90% training set and a 10% test set. Where relevant, the training data are further split into an inner training set and a validation set which is used to select the optimal model hyperparameters. Table 5 presents the out-of-sample metrics for the top-performing lognormal and loggamma models, respectively, compared against a one-hot encoded GLM as a baseline (full results in Appendix C.4 of the supplementary material).[4] The category root mean squared error ("Category RMSE") measures the RMSE of the average prediction for each category (on a log scale). In the present context, the category RMSE can be interpreted as the average accuracy of loss predictions on each sub-portfolio.

The metrics tell consistent stories most of the time. The results for CRPS and NLL are almost perfectly aligned, which is not surprising given that both are measures of how far the observed values depart from the probabilistic predictions. In general, we consider the CRPS and NLL measures to be more reliable estimates of the goodness-of-fit of the models than the MedAE, which only takes into account the point predictions but not the uncertainty associated with them.

The results in Table 5 indicate that the regularised `GLMMNet_l2` and `NN_ee` are equally competitive models and surpass all other models under consideration. The `GLM` family of models performs poorly on this data, much worse than their mixed model counterpart, that is, `GLMM`, confirming that the presence of the high-cardinality categorical variable interferes with their modelling capabilities. We also remark that, while the balance property for GLMs with canonical links (Wüthrich and Buser, 2021) guarantees that the average prediction for any level of the categorical features is unbiased in the training set, this property does not seem to carry over to the testing set according to the category RMSE metrics. At the same time, the outperformance of network-based models also suggests that the dataset contains complex relationships that cannot be captured in a linear structure. One practical consideration to note is that when the environment is highly noisy and the underlying model is uncertain, such as here, the interpretability of the model becomes even more important. In such cases, models that are more interpretable are preferred, which in turn highlights the advantages of `GLMMNet` over `NN_ee`.

An important observation is that adding the ℓ_2 regularisation term clearly helps GLMMNet navigate the data better, which suggests that the original `GLMMNet` overfits to the training data (as can be confirmed

[4] In Table 5 and the relevant tables in Appendix C.4, the median absolute error (MedAE) and NLL are calculated on the original scale of the data (i.e., after the predictions are back-transformed). The CRPS is calculated on the log-transformed scale, because there is no closed-form formula for the loggamma CRPS. As a result of this, the CRPS as presented here will be more tolerant of wrong predictions for extreme observations than the NLL.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

from its deterioration in the test performance). Indeed, the amelioration of the score proves statistically significant ($p < 0.001$) under the Diebold–Mariano test (Gneiting and Katzfuss, 2014). The results hold for both lognormal and loggamma models.

Comparing the lognormal and loggamma models, we find that assuming a loggamma distribution for the response improves the model fit in almost all instances (except the GLM models) and across all three performance metrics. While the CRPS and NLL values are reasonably similar between the two tables, the median absolute error (MedAE) halves when moving from lognormal to loggamma. An intuitive explanation is that the lognormal models struggle more with fitting to the extreme claims and thus tend to over-predict the middle-ranged values in an attempt to boost the predictions. The loggamma models, on the other hand, are more comfortable with the extremes, as those can be captured reasonably well by the long tail. These results confirm the need to consider alternative distributions beyond the Gaussian—one major motivation for our proposed extension of the LMMNN to the GLMMNet. Even within the Gaussian framework, as demonstrated by the lognormal models presented here, the GLMMNet can still be useful.

Overall, as we have seen in the simulation experiments, the differences in model performance are relatively small due to the low signal-to-noise ratio. Only a limited number of risk factors were observed, and they are not necessarily the true drivers of the claims cost. In the experiments above, a simpler structure like GLMM sometimes performs comparably with the more complex and theoretically more capable GLMMNet. That said, the results clearly demonstrate that the GLMMNet is a promising model in terms of predictive performance. We expect to see better results with a greater volume of data, for example, when more information becomes available as a policy progresses and claims develop.

4.3. GLMMNet: Decomposing Random Effects Per Individual Category

One main advantage of mixed effects models in dealing with high-cardinality categorical features is the transparency they offer on the effects of individual categories. This is particularly important as it is often the case that the high-cardinality categorical feature is itself a key risk factor. For example, in this dataset, there is a lot of heterogeneity in claims experience across different occupations, but the lack of data for some makes it hard to determine how much trust can be given to the experience.

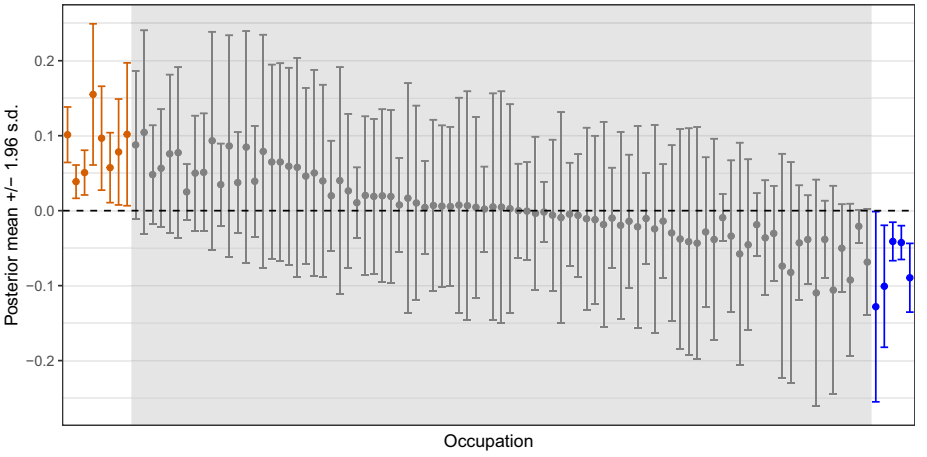


Figure 6. *Posterior predictions of (a randomly selected sample of) the random effects in 95% confidence intervals from the loggamma GLMMNet, ordered by decreasing z-scores. Occupations that do not overlap with the zero line are highlighted in vermillion (if above zero) and blue (if below zero), respectively. Occupations that do overlap with the zero line are in the shaded region. The x-axis labels have been removed to preserve confidentiality.*

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Table 6. *Comparison of loggamma GLMMNet performance on training and test sets.*

	Training			Test (out-of-sample)		
	MedAE	CRPS	NLL	MedAE	CRPS	NLL
GLMMNet_l2	1583	0.7550	9.594	1551	0.7600	9.574
GLMMNet_ANZSIC	1583	0.7584	9.598	1549	0.7612	9.575
GLMMNet_clusters	1596	0.7557	9.595	1581	0.7594	9.574

The GLMMNet can provide useful insights in this regard. Figure 6 shows the posterior predictions for the effects of individual occupations (ordered by decreasing z-scores), plotted with 95% confidence intervals. Unsurprisingly, occupations with a larger number of observations generally have tighter intervals (i.e., lower standard deviations); see Figure C.4 of Appendix C.4 in the supplementary material. This prevents us from overtrusting extreme estimates, which may happen simply due to the small sample size. In the context of technical pricing, such an analysis can be helpful for identifying occupation classes that are statistically more or less risky. This is not possible with the often equally or less high-performing neural networks with entity embeddings. With the latter, we can analyse the embedding vectors to identify occupations that are "similar" to each other in the sense of producing similar output, but we cannot go any further.

4.4. GLMMNet: Beyond Mean-Field Variational Inference

As discussed in Section 2.4, the experiments so far have been restricted to mean-field variational inference; that is, we have always assumed that the posterior estimates of the categories are independent of each other. We relax this assumption here by modelling a non-diagonal covariance matrix in the surrogate posterior. However, rather than estimating a full covariance matrix, which would involve too many parameters, we pre-cluster the categories and only include dependence between categories within the same cluster. The resulting covariance matrix is a block diagonal matrix where each block on the diagonal represents a cluster.

To form the clusters, we experimented with using the original ANZSIC Groups (one level above Class) and using k-means clustering on the textual similarity encodings of the occupations (Cerda et al., 2018). Table 6 reports the performance metrics of the two models under the loggamma assumption. In this example, we found that allowing for posterior dependence did not yield significant improvement in predictive performance. Inspecting the posterior covariance matrix further reveals that there is little correlation between categories in the same cluster. In such cases, it seems sufficient to stick with the simpler-to-implement mean-field Gaussian assumption for the surrogate posterior. However, as is characteristic of any modelling decision, the appropriateness of this choice largely depends on the complexity of the dataset at hand. The inclusion of posterior correlation can be particularly valuable when there is a strong suspicion of high dependence among the categories.

5. Conclusions

High-cardinality categorical features are often encountered in real-world insurance applications. The high dimensionality creates a significant challenge for modelling. As discussed in Section 1.1, the existing techniques prove inadequate in addressing this issue. Recent advancements in ML modelling hold promise for the development of more effective approaches.

In this paper, we took inspiration from the latest developments in the ML literature and proposed a novel model architecture called GLMMNet that targets high-cardinality categorical features in insurance data. The proposed GLMMNet fuses neural networks with GLMMs to enjoy both the predictive power of deep learning models and the transparency and statistical strength of GLMMs.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

GLMMNet is an extension to the LMMNN proposed by Simchoni and Rosset (2022). The LMMNN is limited to the case of a Gaussian response with an identity link function. GLMMNet, on the other hand, allows for the entire class of ED family distributions, which are better suited to the modelling of skewed distributions that we see in insurance and financial data.

Importantly, in making this extension, we have made several modifications to the original LMMNN architecture, including:

- *The removal of a non-linear transformation on the high-cardinality categorical features Z*. This is to ensure the interpretability of random effect predictions, which carries practical significance (as discussed in Section 4.3).
- *The shift from optimising the exact likelihood to a variational inference approach*. As discussed in Section 2.4, the integral in the likelihood function does not generally have an analytical solution (with only a few exceptions, and the Gaussian example is one of those). The shift to a variational inference approach circumvents the complicated numerical approximations to the integral. This opens the door to a flexible range of alternative distributions for the response, including Bernoulli, Poisson, and Gamma—any member of the ED family. This flexibility makes the GLMMNet widely applicable to many insurance contexts (and beyond).

In our experiments, we found that the GLMMNet often performed comparably with and sometimes better than the benchmark model of an entity-embedded neural network, which is the most popular approach among actuarial researchers working with neural networks. Furthermore, the GLMMNet comes with the additional benefits of transparency on the random effects as well as the ability to produce probabilistic predictions (e.g., uncertainty estimates).

To sum up, the proposed GLMMNet has at least the following advantages over the existing approaches:

- Improved predictive performance, especially in high-noise environments ;
- Interpretability with the random effects;
- Flexibility with the form of the response distribution, including the ability to handle categorical target variables in classification problems ;
- Ability to handle large datasets (due to the computationally efficient implementation through variational inference).

However, like any approach, the GLMMNet is not without its limitations. First, we noted from our numerical experiments that the model is not the most competitive in low-noise Gaussian environments. The separate random effects component of the model makes it less flexible compared to an entity-embedded neural network, which allows the categorical feature to interact with all other (continuous) features in the model (Richman and Wüthrich, 2023). Second, while the model offers interpretable random effects, the lack of transparency on the fixed effects component remains, which could pose challenges in contexts requiring a clear understanding of all model components.

While the GLMMNet does not claim to be a perfect solution for every application, it represents a valuable addition to the actuaries' modelling toolbox for handling high-cardinality categorical features.

Acknowledgements. This work was presented at the Australasian Actuarial Education and Research Symposium (AAERS) in November 2022 (Canberra, Australia), the International Congress of Actuaries in May 2023 (Sydney, Australia), and the Insurance Data Science Conference in June 2023 (London, UK). The authors are grateful for the constructive comments received from colleagues present at the event.

The authors are also grateful to Jovana Kolar, for her assistance in coding and running the experiments, and to Mario Wüthrich and three anonymous reviewers, for their insightful comments that led to significant improvements of the paper.

This research was supported under Australian Research Council's Discovery Project DP200101859 funding scheme. Melantha Wang acknowledges financial support from UNSW Australia Business School. The views expressed herein are those of the authors and are not necessarily those of the supporting organisations.

The authors declare that they have no conflicts of interest.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Supplementary material. For supplementary material accompanying this paper visit https://doi.org/10.1017/asb.2024.7

Data and Code. The code used in the numerical experiments, as well as the simulation datasets, is available on https://github.com/agi-lab/glmmnet.

References

Al-Mudafer, M.T., Avanzi, B., Taylor, G., Wong, B., 2022. Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics* **105**, 144–174. https://www.sciencedirect.com/science/article/pii/S0167668722000373, DOI: 10.1016/j.insmatheco.2022.03.010.

Antonio, K., Beirlant, J., 2007. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics* **40**, 58–76. https://www.sciencedirect.com/science/article/pii/S0167668706000552, DOI: 10.1016/j.insmatheco.2006.02.013.

Antonio, K., Zhang, Y., 2014. Linear mixed models, in: Frees, E.W., Meyers, G., Derrig, R.A. (Eds.), Predictive Modeling Applications in Actuarial Science, Volume I: Predictive Modeling Techniques. Cambridge University Press, Cambridge. volume 1 of *International Series on Actuarial Science*, pp. 182–216. https://www.cambridge.org/core/books/predictive-modeling-applications-in-actuarial-science/linear-mixed-models/91FF971A2C418510F2DB8AED3368FF5B, DOI: 10.1017/CBO9781139342674.008.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877. DOI: 10.1080/01621459.2017.1285773. publisher: Taylor & Francis.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural networks, in: *Proceedings of the 32nd International Conference on Machine Learning*, PMLR. pp. 1613–1622. https://proceedings.mlr.press/v37/blundell15.pdf.

Bürkner, P.C., 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80, 1–28. DOI: 10.18637/jss.v080.i01.

Casella, G., 1985. An introduction to empirical bayes data analysis. *The American Statistician* **39**, 83–87.

Cerda, P., Varoquaux, G., Kégl, B., 2018. Similarity encoding for learning with dirty categorical variables. *Machine Learning* **107**, 1477–1494. http://link.springer.com/10.1007/s10994-018-5724-2, DOI: 10.1007/s10994-018-5724-2.

Chollet, F., et al., 2015. Keras. https://keras.io.

Delong, Ł., Kozak, A., 2021. The use of autoencoders for training neural networks with mixed categorical and numerical features. SSRN. https://papers.ssrn.com/abstract=3952470, DOI: 10.2139/ssrn.3952470.

Delong, Ł., Lindholm, M., Wüthrich, M.V., 2021. Gamma mixture density networks and their application to modelling insurance claim amounts. Insurance: Mathematics and Economics, S0167668721001232 https://linkinghub.elsevier.com/retrieve/pii/S0167668721001232, DOI: 10.1016/j.insmatheco.2021.08.003.

Denuit, M., Hainaut, D., Trufin, J., 2019. Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. Springer Actuarial, Springer International Publishing, Cham. http://link.springer.com/10.1007/978-3-030-25820-7, DOI: 10.1007/978-3-030-25820-7.

Embrechts, P., Wüthrich, M., 2022. Recent challenges in actuarial science. Annual Review of Statistics and Its Application 9. https://www.annualreviews.org/doi/10.1146/annurev-statistics-040120-030244, DOI: 10.1146/annurev-statistics-040120-030244.

Ferrario, A., Noll, A., Wüthrich, M., 2020. Insights from inside neural networks. SSRN. https://papers.ssrn.com/abstract=3226852, DOI: 10.2139/ssrn.3226852.

Frees, E.W., 2014. Longitudinal and panel data models, in: Frees, E.W., Meyers, G., Derrig, R.A. (Eds.), Predictive Modeling Applications in Actuarial Science, Volume I: Predictive Modeling Techniques. Cambridge University Press, Cambridge. volume 1 of *International Series on Actuarial Science*, pp. 167–181. https://www.cambridge.org/core/books/predictive-modeling-applications-in-actuarial-science/longitudinal-and-panel-data-models/FA9525C9E531966C9DD65A79C06B7888, DOI: 10.1017/CBO9781139342674.007.

Friedman, J.H., 1991. Multivariate adaptive regression splines. *The Annals of Statistics* **19**, 1–67.

Gelman, A., 2020. Prior choice recommendations. https://github.com/stan-dev/stan.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research, Cambridge University Press, Cambridge ; New York. OCLC: ocm67375137.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*. pp. 249–256. https://proceedings.mlr.press/v9/glorot10a.html. iSSN: 1938-7228.

Gneiting, T., Katzfuss, M., 2014. Probabilistic Forecasting. *Annual Review of Statistics and Its Application* **1**, 125–151. DOI: 10.1146/annurev-statistics-062713-085831.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378. DOI: 10.1198/016214506000001437. publisher: Taylor & Francis.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA. Google-Books-ID: omivDQAAQBAJ.

Guiahi, F., 2017. Applying graphical models to automobile insurance data. *Variance* **11**, 23–44.

Guo, C., Berkhahn, F., 2016. Entity embeddings of categorical variables. arXiv:1604.06737 [cs] http://arxiv.org/abs/1604.06737. arXiv: 1604.06737.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Hainaut, D., Trufin, J., Denuit, M., 2022. Response versus gradient boosting trees, glms and neural networks under tweedie loss and log-link. Scandinavian Actuarial Journal, 1–26. DOI: 10.1080/03461238.2022.2037016. publisher: Taylor & Francis.

Hajjem, A., Bellavance, F., Larocque, D., 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* **84**, 1313–1328. DOI: 10.1080/00949655.2012.741599. publisher: Taylor & Francis.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics. 2nd ed ed., Springer, New York, NY.

Henckaerts, R., Côté, M., Antonio, K., Verbelen, R., 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal* **25**, 255–285. DOI: 10.1080/10920277.2020.1745656. publisher: Routledge.

Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringrules. Journal of Statistical Software 90. http://www.jstatsoft.org/v90/i12/, DOI: 10.18637/jss.v090.i12.

Jospin, L.V., Buntine, W., Boussaid, F., Laga, H., Bennamoun, M., 2022. Hands-on bayesian neural networks – a tutorial for deep learning users. *IEEE Computational Intelligence Magazine* **17**, 29–48. http://arxiv.org/abs/2007.06823, DOI: 10.1109/MCI.2022.3155327. arXiv:2007.06823 [cs, stat].

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980 [Cs] http://arxiv.org/abs/1412.6980.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22, 79–86. https://www.jstor.org/stable/2236703. publisher: Institute of Mathematical Statistics.

Kuo, K., Richman, R., 2021. Embeddings and attention in predictive modeling. arXiv:2104.03545 [q-fin, stat] http://arxiv.org/abs/2104.03545. arXiv: 2104.03545.

Kuss, M., Pfingsten, T., Csato, L., Rasmussen, C., 2005. Approximate Inference for Robust Gaussian Process Regression. Technical Report 136. Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Lakshmanan, V., Robinson, S., Munn, M., 2020. Machine learning design patterns: solutions to common challenges in data preparation, model building, and MLOps. First edition ed., O'Reilly Media, Sebastopol, CA. OCLC: on1178649818.

Mandel, F., Ghosh, R.P., Barnett, I., 2022. Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics*. DOI: 10.1111/biom.13615.

McCulloch, C.E., Searle, S.R., 2001. Generalized, linear, and mixed models. Wiley series in probability and statistics. Applied probability and statistics section, John Wiley & Sons, New York.

Neal, R.M., Hinton, G.E., 1998. A view of the em algorithm that justifies incremental, sparse, and other variants, in: Jordan, M.I. (Ed.), Learning in Graphical Models. Springer Netherlands, Dordrecht, pp. 355–368. http://link.springer.com/10.1007/978-94-011-5014-9_12, DOI: 10.1007/978-94-011-5014-9_12.

Pargent, F., Pfisterer, F., Thomas, J., Bischl, B., 2022. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*. DOI: 10.1007/s00180-022-01207-6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Pettifer, A., Pettifer, J., 2012. A practical guide to commercial insurance pricing, in: Australian Actuaries Institute General Insurance Seminar, Sydney. pp. 1–40. https://www.actuaries.asn.au/Library/Events/GIS/2012/GIS2012PaperAlinaPettifer.pdf.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems, Curran Associates, Inc. pp. 6639–6649. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Richman, R., 2021a. AI in actuarial science – a review of recent advances – part 1. *Annals of Actuarial Science* **15**, 207–229. https://www.cambridge.org/core/product/identifier/S1748499520000238/type/journal_article, DOI: 10.1017/S1748499520000238.

Richman, R., 2021b. AI in actuarial science – a review of recent advances – part 2. *Annals of Actuarial Science* **15**, 230–258. https://www.cambridge.org/core/product/identifier/S174849952000024X/type/journal_article, DOI: 10.1017/S174849952000024X.

Richman, R., Wüthrich, M., 2023. High-cardinality categorical covariates in network regressions. SSRN. https://ssrn.com/abstract=4549049, DOI: 10.2139/ssrn.4549049.

Richman, R., Wüthrich, M.V., 2021. A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science* **15**, 346–366. https://www.cambridge.org/core/product/identifier/S1748499519000071/type/journal_article, DOI: 10.1017/S1748499519000071.

Shi, P., Shi, K., 2021. Nonlife insurance risk classification using categorical embedding. SSRN. https://papers.ssrn.com/abstract=3777526, DOI: 10.2139/ssrn.3777526.

Sigrist, F., 2021. Gaussian process boosting. arXiv:2004.02653 [cs.LG] http://arxiv.org/abs/2004.02653. arXiv:2004.02653.

Sigrist, F., 2022. Latent gaussian model boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence In press, 1–12. DOI: 10.1109/TPAMI.2022.3168152. arXiv:2105.08966.

Simchoni, G., Rosset, S., 2022. Integrating random effects in deep neural networks. arXiv:2206.03314 [cs, stat] http://arxiv.org/abs/2206.03314. arXiv:2206.03314.

State of New York, 2022. Assembled workers' compensation claims: Beginning 2000. https://data.ny.gov/Government-Finance/Assembled-Workers-Compensation-Claims-Beginning-20/jshw-gkgu.

Verbelen, R., 2019. There is (not) enough data! https://www.finity.com.au/publication/commercial-lines-seminar-2019-there-is-not-enough-data.

Wüthrich, M., Buser, C., 2021. Data Analytics for Non-Life Insurance Pricing. Technical Report. Rochester, NY. https://papers.ssrn.com/abstract=2870308, DOI: 10.2139/ssrn.2870308.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press

Wüthrich, M., Merz, M., 2019. Editorial: Yes, we CANN! *ASTIN Bulletin* **49**, 1–3. DOI: 10.1017/asb.2018.42.
Wüthrich, M.V., Merz, M., 2023. Statistical Foundations of Actuarial Learning and Its Applications. Springer Actuarial, Springer International Publishing, Cham. DOI: 10.1007/978-3-031-12409-9.
Zhang, C., Bütepage, J., Kjellströööm, H., Mandt, S., 2019. Advances in variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2008–2026. DOI: 10.1109/TPAMI.2018.2889774. conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

https://doi.org/10.1017/asb.2024.7 Published online by Cambridge University Press