risks

MDPI

Article

On Two Mixture-Based Clustering Approaches Used in Modeling an Insurance Portfolio

Tatjana Miljkovic [1,*] and and Daniel Fernández [2,3]

[1] Department of Statistics, Miami University, Oxford, OH 45056, USA
[2] Research and Development Unit, Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu, CIBERSAM, Sant Boi de Llobregat, Barcelona 08830, Spain; df.martinez@pssjd.org or dfdez23@outlook.com
[3] School of Mathematics and Statistics, Victoria University of Wellington, Wellington 6140, New Zealand
* Correspondence: miljkot@miamioh.edu; Tel.: +513-529-3299

Received: 7 March 2018; Accepted: 14 May 2018; Published: 17 May 2018

Abstract: We review two complementary mixture-based clustering approaches for modeling unobserved heterogeneity in an insurance portfolio: the generalized linear mixed cluster-weighted model (CWM) and mixture-based clustering for an ordered stereotype model (OSM). The latter is for modeling of ordinal variables, and the former is for modeling losses as a function of mixed-type of covariates. The article extends the idea of mixture modeling to a multivariate classification for the purpose of testing unobserved heterogeneity in an insurance portfolio. The application of both methods is illustrated on a well-known French automobile portfolio, in which the model fitting is performed using the expectation-maximization (EM) algorithm. Our findings show that these mixture-based clustering methods can be used to further test unobserved heterogeneity in an insurance portfolio and as such may be considered in insurance pricing, underwriting, and risk management.

Keywords: generalized linear model; cluster-weighted model; ordered stereotype model; ordinal data

JEL Classification: C02; C40; C60

1. Introduction

The multivariate classification ratemaking has made significant advances in the past decade with the introduction of different types of statistical methods for pricing individual risks. These new techniques bring important benefits to insurers such as more equatable pricing of the individual risks, better competitive advantage, and protection against adverse selection and support informed decisions about the type of risk that the insurer is willing to write. Werner and Modlin (2016) introduced many of these techniques with a primary focus on generalized linear models (GLMs), which are used in pricing and reserving.

According to the Actuarial Review (Baribeau 2016) "the predictive modeling is advancing far beyond its general linear model (GLM)-based roots due to the explosion of new data sources, technological innovation and advanced analytics." Recent literature on extensions of GLMs with mixture modeling has been a part of this innovation process, and this is proposed by several researchers. A finite mixture of Poisson regression models with an application to insurance ratemaking was studied by Bermúdez (2012) for count data. The authors recognized that unobserved heterogeneity in this type of data requires more structure in the modeling techniques and that the current model fitting can be improved by using finite mixture of bivariate Poisson regressions. The unobserved heterogeneity in the zero-inflated type of data is attributed to the fact that variance is often larger than the mean. In 2015, another Poisson mixture model for count data was considered by Brown and Buckley (2015)

for modeling heterogeneity in a Group Life insurance portfolio. The authors showed violation of heterogeneity across groups and suggested that putting similar groups together is necessary for further analysis. Later, a non-parametric Bayesian approach was considered in modeling this mixture distribution by incorporating a Dirichlet process prior and using reversible-jump Markov chain Monte Carlo (Green 1995) to estimate the number of components. Garrido et al. (2016) and Shi et al. (2015) consider relaxing the GLM assumption of independence between the number and the size of claims. The multi-modality of univariate insurance losses was recently modeled using mixtures by Miljkovic and Grün (2016), Verbelen et al. (2014), Lee and Lin (2010), and Klugman and Rioux (2006). The findings of these studies indicate a need to explore the modeling of unobserved heterogeneity in a regression setting where the amount of claims is linked to several other covariates and the goal is to find a finite number of sub-populations of policyholders in an insurance portfolio.

In predictive modeling with a GLM setting, unobserved heterogeneity may occur when important covariates have been omitted and their influence is not accounted for in the analysis (Grün and Leisch 2008). The unobserved heterogeneity may not be fully captured when a single component GLM is used to model the data set. While these techniques have been explored in other fields, they have not been fully adopted in the actuarial field. Actuaries may consider relaxing the assumption that the regression coefficients and dispersion parameters are the same for all observations. In this case, a goal is to find groups of policy holders with similar regression coefficients.

Multiple techniques have been developed which deal with the grouping of heterogeneous data such as hierarchical clustering (Johnson 1967; Kaufman and Rousseeuw 1990), association analysis (Manly 2005), and *k*-means clustering algorithm (Jobson 1992; Lewis et al. 2003; McCune and Grace 2002). There are a number of clustering methods based on mathematical techniques such as association indices (Chen et al. (2011); Wu et al. (2008)), distance metrics (Everitt et al. (2001)), and matrix decomposition (Manly (2005); Quinn and Keough (2002); Wu et al. (2007)). However, these algorithms do not have a likelihood-based formulation and therefore do not provide a reliable method of model selection or assessment. A particularly powerful likelihood-based approach to one-dimensional clustering based on finite mixtures, with the variables in the columns being utilized to group the objects in the rows, is provided by McLachlan and Basford (1988), McLachlan and Peel (2004), Everitt et al. (2001), Böhning et al. (2007), Wu et al. (2008), and Melnykov and Maitra (2010).

The objective of this article it to review two recently proposed mixture-based clustering approaches for modeling unobserved heterogeneity in an insurance portfolio, which, to the best of our knowledge, have not yet been exploited by the practitioner in the actuarial science area. Moreover, we focus on modeling severity of losses as a function of several covariates arising from different sub-populations. These sub-populations are grouped into clusters based on a similar historical experience or well-defined similarity rules and the results of each group can be considered in underwriting and ratemaking. We particularly consider two different data modeling frameworks: mixture modeling of ordinal variables (for classified data based on the level of risk) and the mixture of GLMs for a mixed type of covariates (for individual data). As an illustration, we show how these two mixture models can be used to model an insurance portfolio and have complementary properties. Similarities and differences of both approaches are discussed in the context of the data structure available for the selected insurance portfolio.

In an insurance context, ordinal variables are often defined based on a risk classification with intrinsic order. For example, driver age can be treated as an ordinal outcome with the youngest drivers being associated with the highest risk propensity. Similarly, losses can be treated on an ordinal scale based on the intensity of claims. There are a variety of approaches to the modeling of ordinal data that properly respect the ordinal nature of the data. Liu and Agresti (2005) and Agresti (2010) described various proportional odds version models using adjacent–categories logits, cumulative logits McCullagh (1980), and continuation–ratio logits McCullagh and Nelder (1989). Our article focuses on the ordered stereotype model (OSM) introduced by Anderson (1984), which is more flexible than the

most common models as a result of adding additional score parameters associated with the distance among ordinal categories.

The generalized linear mixed cluster-weighted model (known as CWM) was recently proposed by Ingrassia et al. (2015) as a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of mixed-type covariates with the assumption that continuous covariates come from Gaussian distribution. The CWM method does not consider ordinal data; thus, we are also interested in the mixture-based clustering for OSM built to handle ordinal data previously proposed by Fernández et al. (2016). The ordinal data in insurance setting can be considered if the policyholders have been previously classified based on the level of risk (e.g., 1 is the lowest level of risk and 5 is the highest). This classification is often obtained for underwriting and risk management purposes. Both methods assume that an inherent clustering effect is present in the data; therefore, each sub-population of the variable of interest, such as claims, can be modeled separately. One of the main advantages of these approaches over other non model-based clustering techniques is their likelihood-based foundation because maximum likelihood theory provides estimators and model selection. Another advantage of these two methods is that they complement each other, i.e., they allow for flexibility in terms of data collection and categorization. If the collected data is all organized based on the ordinal levels corresponding to several risk classifications, then a mixture-based clustering method for ordered stereotype model would be a suitable tool to further test the heterogeneity in the data. For insurance data sets that are currently analyzed using existing GLMs, the CWM approach would allow for detecting an unobserved heterogeneity in the data by testing if more than a single component GLM fits the data better. If the CWM model provides a better fit to the data, then this model should replace a single component GLM, where applicable.

The remainder of this paper is organized as follows. Section 2 introduces the formulation and model estimation for both mixture-based methods as well as the model selection criteria. Section 3 is devoted to the presentation of the data set used in this study and the results obtained using these two approaches. Conclusions are summarized in Section 4. The appendix provides additional results with respect to the analysis conducted as part of Section 3.

2. Methodology

2.1. Mixture-Based Clustering for the Ordered Stereotype Model

Fernández et al. (2016) proposed an extension of the model-based approach proposed in Pledger and Arnold (2014) for ordinal responses. This approach considered finite mixture models to define a clustering structure and used the OSM introduced by Anderson (1984) with the aim of formulating the ordinal procedure. The stereotype model has the advantage that it requires a smaller number parameters to be estimated than the more commonly used baseline-category logit model or the multinomial logistic regression model (Agresti 2010, Section 4.3.1). Moreover, this model estimates from the data possibly unequal spacing among the levels of the ordinal responses in the form of a set of ordered score parameters. These scores are directly interpretable as measures of similarity between neighboring levels of the variable. This ease of interpretation is an advantage over other ordinal-based models such as the proportional odds model and the continuation-ratio model.

We illustrate here the model formulation for the row clustering version. The analysis for the column clustering version is basically the same, but exchanging the parameters related to rows by their equivalent column parameters. For the row clustering version, the probability that the ordinal response response $\{y_{ij}\}$ ($i = 1, \ldots, n$ and $j = 1, \ldots, m$) takes the category k ($k = 1, \ldots, q$) is represented by the following log odds:

$$\log\left(\frac{P\left[y_{ij} = k \mid i \in g\right]}{P\left[y_{ij} = 1 \mid i \in g\right]}\right) = \mu_k + \phi_k(\alpha_g + \beta_j), \quad k = 2, \ldots, q, \quad g = 1, \ldots, G, \quad j = 1, \ldots, m \tag{1}$$

where the inclusion of the monotone increasing constraint $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$ ensures that the variable response $\boldsymbol{Y} = (Y_1, \ldots, Y_n) = \{y_{ij}\}$ is ordinal (see Anderson (1984)). For simplicity, we assume that the ordinal responses all have the same number of categories q so that $y_{ij} \in \{1, \ldots, q\}$, and they correspond to the policy holders groups that are already classified based on some underwriting criteria. The parameters $\{\mu_2, \ldots, \mu_q\}$ are the *cut points*, and $\{\phi_2, \ldots, \phi_q\}$ are the parameters which can be interpreted as the "scores" for the categories of the response variable Y_{ij}. $G \leq n$ is the number of row groups, and $i \in g$ means row i is classified in the row cluster g. The set of parameters $\{\beta_1, \ldots, \beta_m\}$ quantify the main effects of the m. We restrict $\mu_1 = \phi_1 = 0$, $\phi_q = 1$, $\sum_{g=1}^{G} \alpha_g = \sum_{j=1}^{m} \beta_j = 0$ to ensure identifiability. It is important to note that the actual membership of the rows among the G row-clusters is unknown; therefore, it is considered missing information. Further, we define $\{\tau_1, \ldots, \tau_G\}$ as the (unknown) proportions of rows in each row group, with $\sum_{g=1}^{G} \tau_g = 1$. We can view $\{\tau_g\}$ as the *a priori* row membership probabilities.

The probability of the data response y_{ij} being equal to the category k conditional on a given clustering is

$$\begin{aligned} &\theta_{gjk}(\Omega) = \Pr\left[y_{ij} = k \mid i \in r\right] = \frac{\exp(\mu_k + \phi_k(\alpha_g + \beta_j))}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell(\alpha_g + \beta_j))}, \\ &k = 1, \ldots, q, \quad g = 1, \ldots, G, \quad j = 1, \ldots, m \end{aligned} \tag{2}$$

where Ω is the parameter vector $\{\{\mu_k\}, \{\phi_k\}\{\alpha_g\}, \{\beta_j\}\}$.

Likelihood functions: The (incomplete) likelihood of the data is

$$L(\Omega, \{\tau_g\} \mid \{y_{ij}\}) = \prod_{i=1}^{n} \left[\sum_{g=1}^{G} \tau_g \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{gjk}\right)^{I(y_{ij}=k)} \right]$$

where θ_{gjk} is the probability of the data response defined in Equation (2).

We define the unknown row group memberships through the following indicator latent variables,

$$Z_{ig} = I(i \in g) = \begin{cases} 1 & \text{if } i \in g \\ 0 & \text{if } i \notin g \end{cases} \qquad i = 1, \ldots, n, \quad g = 1, \ldots, G$$

where $i \in r$ indicates that row i is in row group g. It follows that $\sum_{g=1}^{G} Z_{ig} = 1$ $(i = 1, \ldots, n)$, and, since their *a priori* row membership probabilities are $\{\tau_g\}$,

$$(Z_{i1}, \ldots, Z_{ig}) \sim \text{Multinomial}(1; \tau_1, \ldots, \tau_G), \qquad i = 1, \ldots, n.$$

Consequently, the complete data log-likelihood of this model using the known data $\{y_{ij}\}$ and the unknown data $\{z_{ig}\}$ is as follows:

$$l_c(\Omega, \{\tau_g\} \mid \{y_{ij}\}, \{z_{ig}\}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(\tau_g) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{g=1}^{G} z_{ig} I(y_{ij} = k) \log\left(\theta_{gjk}\right).$$

Parameter estimation: The parameter estimation for a fixed number of components G is performed using the maximum likelihood estimation approach fulfilled by means of the expectation-maximization (EM) algorithm proposed by Dempster et al. (1977) and used in most finite mixture problems discussed by McLachlan and Peel (2004).

The EM algorithm consists of two steps: expectation (E-step) and maximization (M-step). As part of the E-step, a conditional expectation of the complete data log-likelihood function is obtained given the observed data and current parameter estimates. In the finite mixture model, the latent data corresponds to the component identifiers. As part of the E-step, the expectation taken

with respect to the conditional posterior distribution of the latent data, given the observed data and the current parameter estimates, is referred to as the posterior probability that response y_{ij} comes from the gth mixture component, computed at each iteration of the EM algorithm. The remaining part of the M-step requires finding component-specific parameter estimates Ω by solving numerically the maximum likelihood estimation problem for each of the different component distributions.

The E-step and M-step alternate until the relative increase in the log-likelihood function is no bigger than a small pre-specified tolerance value, when the convergence of the EM algorithm is achieved. In order to find an optimal number of components, maximum likelihood estimation is obtained for each number of groups G, and the model is selected based on a chosen model selection criterion.

In this model, the EM algorithm performs a fuzzy assignment of rows to clusters based on the posterior probabilities. The EM algorithm is initialized with an estimate $\{\widehat{\Omega}^{(0)}, \{\widehat{\tau}_g^{(0)}\}\}$ of the parameters and proceeds by alternation of the E-step and M-step to estimate the missing data $\{\widehat{Z}_{ig}\}$ and to update the parameter estimates. In this section, we develop the E-step and M-step for row clustering. This development follows closely Fernández et al. (2016) (Section 3).

E-Step: In the tth iteration of the EM algorithm, the E-Step evaluates the expected values $\widehat{Z}_{ig}$ of the unknown classifications Z_{ig} conditional on the data $\{y_{ij}\}$ and the previous estimates of the parameters $\{\widehat{\Omega}^{(t-1)}, \{\widehat{\tau}_g^{(t-1)}\}\}$. The conditional expectation of the complete data log-likelihood at iteration t is given by

$$\begin{aligned} Q(\Omega, \{\tau_g\} \mid \widehat{\Omega}^{(t-1)}, \{\widehat{\tau}_g^{(t-1)}\}) &= E_{\{Z_{ig}\}|\{y_{ij}\},\Omega^{(t-1)}} \left[\ell_c(\Omega, \{\tau_g\} \mid \{y_{ij}\}, \{Z_{ig}\})\right] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \log(\widehat{\tau}_g^{(t-1)}) E\left[z_{ig} \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{g=1}^{G} I(y_{ij}=k) \log\left(\theta_{gjk}^{(t-1)}\right) E\left[z_{ig} \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right]. \end{aligned} \tag{3}$$

The random variable Z_{ig} is Bernoulli distributed, so that $E\left[Z_{ig} \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right] = \Pr\left[Z_{ig}=1 \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right]$, and, applying Bayes' rule, we obtain

$$\begin{aligned} \widehat{Z}_{ig}^{(t)} = \Pr\left[Z_{ig}=1 \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right] &= \frac{\Pr\left(\{y_{ij}\}, \widehat{\Omega}^{(t-1)} \mid z_{ig}=1\right)\Pr\left(z_{ig}=1\right)}{\sum_{\ell=1}^{G}\Pr\left(\{y_{ij}\}, \widehat{\Omega}^{(t-1)} \mid z_{i\ell}=1\right)\Pr\left(z_{i\ell}=1\right)} \\ &= \frac{\widehat{\tau}_g^{(t-1)}\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\widehat{\theta}_{gjk}^{(t-1)}\right)^{I(y_{ij}=k)}}{\sum_{\ell=1}^{G}\left\{\widehat{\tau}_\ell^{(t-1)}\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\widehat{\theta}_{\ell jk}^{(t-1)}\right)^{I(y_{ij}=k)}\right\}}. \end{aligned}$$

Finally, we complete the E-step by substituting the previous expression in the complete data log-likelihood at iteration t expressed in Equation (3),

$$\begin{aligned} \widehat{Q}(\Omega, \{\tau_g\} \mid \widehat{\Omega}^{(t-1)}, \{\widehat{\tau}_g^{(t-1)}\}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{Z}_{ig}^{(t)} \log(\widehat{\tau}_g^{(t-1)}) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{g=1}^{G} \widehat{Z}_{ig}^{(t)} I(y_{ij}=k) \log\left(\widehat{\theta}_{gjk}^{(t-1)}\right). \end{aligned} \tag{4}$$

M-step: The M-step of the EM algorithm is the global maximization of the log-likelihood (4) obtained in the E-step, now conditional on the complete data $\{\{y_{ij}\}, \{\widehat{Z}_{ig}\}\}$. For the case of finite mixture

models, the updated estimations of the term containing the row-cluster proportions $\{\tau_1, \ldots \tau_G\}$ and the one containing the rest of the parameters Ω are computed independently. Thus, the M-step has two separate parts.

The maximum-likelihood estimator for the parameter τ_g where the data Z_{ig} are unobserved is

$$\widehat{\tau}_g^{(t)} = \frac{1}{n}\sum_{i=1}^{n} E\left[Z_{ig} \mid \{y_{ij}\}, \widehat{\Omega}^{(t-1)}\right] = \frac{1}{n}\sum_{i=1}^{n} \widehat{Z}_{ig}^{(t)}, \qquad g = 1, \ldots, G.$$

To estimate the remaining parameters Ω, we must numerically maximize the conditional expectation of the complete data log-likelihood in Equation (3). In the case of row clustering,

$$\widehat{\Omega}^{(t)} = \underset{\Omega}{\operatorname{argmax}} \left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{g=1}^{G} \widehat{Z}_{ig} I(y_{ij} = k) \log\left(\theta_{gjk}(\Omega)\right)\right]$$

where the maximization is conditional on the parameter constraints following Equation (1).

2.2. The General Linear Cluster-Weighted Model

In this section, we summarize the CWM model proposed by Ingrassia et al. (2015) starting with the relevant background. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ is a vector of independent random variables with the density function of a distribution from the exponential family given by

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

where θ_i is a natural parameter and ϕ is a scale parameter. Consider that $\boldsymbol{Y}$ is related to a set of variables $\boldsymbol{x} = (x_1, \ldots, x_p)$ through the following linear relationship

$$\eta = g[E(\boldsymbol{y}|\boldsymbol{x})] = \boldsymbol{X}\boldsymbol{\beta} \tag{5}$$

where $\boldsymbol{\beta}$ is $p \times 1$ vector of parameters and $\boldsymbol{X}$ is an $N \times p$ design matrix, η is the linear predictor and $g[E(\boldsymbol{y}|\boldsymbol{x})]$ is a link function, and it is considered as a simple mathematical function in the original formulation of GLMs by Nelder and Wedderburn (1972). In this model, $\boldsymbol{Y}$ is referred to as a response variable and $\boldsymbol{x}$ is a vector of continuous and discrete type explanatory variables (e.g., covariates).

Let $(\boldsymbol{X}', \boldsymbol{Y})'$ be a vector defined on some space $\Im$ with values in $\mathcal{X} \times \mathcal{Y}$. Further, assume that there exist G partitions of $\Im$, defined as $\Im_1, \ldots, \Im_G$. Gershenfeld (1997) has introduced CWM based on Gaussian mixtures with the joint distribution, $f(\boldsymbol{x}, y)$ of $(\boldsymbol{X}', \boldsymbol{Y})'$, expressed as follows

$$f(\boldsymbol{x}, y) = \sum_{g=1}^{G} \tau_g f(y|\boldsymbol{x}, \Im_g) f(\boldsymbol{x}, \Im_g) \tag{6}$$

where $f(y|\boldsymbol{x}, \Im_g)$ and $f(\boldsymbol{x}, \Im_g)$ are conditional and marginal distributions of $(\boldsymbol{X}', \boldsymbol{Y})'$ and τ_g represents the weight of the g-th component s.t. $\sum_{g=1}^{G} \tau_g = 1$, $\tau_g > 0$.

Ingrassia et al. (2015) introduced a broader family of CWMs that allows for the component conditional distributions to belong to the exponential family and for the mixed ype of covariates. The joint distribution of a random vector $(\boldsymbol{X}', \boldsymbol{Y})'$ is obtained by splitting the covariates into continuous and discrete as $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{W}')'$ under the assumption of independence. In this setting, the model in Equation (6) can be expressed as

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{g=1}^{G} \tau_g f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g) f(\boldsymbol{x}, \boldsymbol{\theta}_g) = \sum_{g=1}^{G} \tau_g f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g) f(\boldsymbol{v}, \boldsymbol{\theta}_g') f(\boldsymbol{w}, \boldsymbol{\theta}_g'') \tag{7}$$

where $f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g)$ is a conditional density of $y|\boldsymbol{x}, \Im_g$ with parameter ϑ_g, $f(\boldsymbol{v}, \boldsymbol{\theta}_g^{'})$ is the marginal distribution of $\boldsymbol{v}$ with parameter $\theta_g^{'}$, $f(\boldsymbol{w}, \boldsymbol{\theta}_g^{''})$ is the marginal distribution of $\boldsymbol{w}$ with parameter $\boldsymbol{\theta}_g^{''}$, and $\boldsymbol{v}$ and $\boldsymbol{w}$ are the vectors of continuous and discrete covariates respectively. Further, all model parameters are defined as $\boldsymbol{\Phi} = (\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\vartheta})$. The conditional distribution $f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g)$ is assumed to belong to the exponential family.

Modeling for $f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g)$ and $f(\boldsymbol{x}, \boldsymbol{\theta}_g)$: The CWM model is based on the assumption that $f(y|\boldsymbol{x}, \boldsymbol{\vartheta}_g)$ belongs to the exponential family of distributions that are strictly related to GLMs. The link function in Equation (5) relates the expected value $g(\mu_g) = \beta_{0g} + \beta_{1g}x_1, \ldots, +\beta_{pg}x_p$. We are interested in estimation of the vector $\boldsymbol{\beta}_g$, so the distribution of $y|\boldsymbol{x}, \Im_g$ is denoted by $f(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \lambda_g)$, where λ_g denotes an additional parameter associated with a two-parameter exponential family.

The marginal distribution $f(\boldsymbol{x}, \theta_g)$ has the following components: $f(\boldsymbol{v}, \theta_g^{'})$ and $f(\boldsymbol{w}, \theta_g^{''})$. The first component is modeled as p-variate Gaussian density with mean $\boldsymbol{\mu_g}$ and covariance matrix $\boldsymbol{\Sigma}_g$ as $\phi(\boldsymbol{v}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma}_g)$.

The marginal density $f(\boldsymbol{w}, \theta_g^{''})$ assumes that each finite discrete covariate W is represented as a vector $\boldsymbol{w}^r = (w^{r1}, \ldots, \boldsymbol{w}^{rc_r})^{'}$, where $w^{rs} = 1$ is w_r, which has the value s, s.t. $s \in \{1, \ldots, c_r\}$, and $w^{rs} = 0$ otherwise.

$$f(\boldsymbol{w}, \boldsymbol{\gamma_g}) = \prod_{r=1}^{q}\prod_{s=1}^{c_r}(\gamma_{grs})^{w^{rs}}, g = 1, \ldots, G \tag{8}$$

where $\boldsymbol{\gamma_g} = (\gamma_{g1}^{'}, \ldots, \gamma_{gq}^{'})^{'}$, $\gamma_{gr} = (\gamma_{gr1}^{'}, \ldots, \gamma_{grc_q}^{'})^{'}$, $\gamma_{grs} > 0$, and $\sum_{s=1}^{c_r}\gamma_{grs}$, $r = 1, \ldots, q$. The density $f(\boldsymbol{w}, \boldsymbol{\gamma_g})$ represents the product of q conditionally independent multinomial distributions with parameters $\boldsymbol{\gamma_{gr}}$, $r = 1, \ldots, q$. Considering these assumptions, the model in Equation (7) can be stated as

$$f(\boldsymbol{x}, y; \Phi) = \sum_{g=1}^{G}\tau_g f(y|\boldsymbol{x}; \boldsymbol{\beta}_g, \lambda_g)\phi(\boldsymbol{v}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)f(\boldsymbol{w}, \gamma_g). \tag{9}$$

If the CWM models allow for the conditional distribution $f(y|.)$ to be Binomial or Poisson, then they are referred to as the Binomial CWM or the Poisson CWM, respectively. The CWM are also built to handle Gaussian, log-normal, and gamma distributions of $f(y|.)$. In the next subsection, we will discuss the parameter estimation of the model in Equation (9).

Parameter Estimation: The EM algorithm discussed in the previous section is used to estimate parameters of this model. Let $(\boldsymbol{x}_1^{'}, y_1)^{'}, \ldots, (\boldsymbol{x}_n^{'}, y_n)^{'}$ be a sample of n independent pairs observations drawn from the model in Equation (9). For this sample, the complete data likelihood function, $L(\boldsymbol{\Phi})$, is given by

$$Ł_c(\boldsymbol{\Phi}) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\tau_g f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, \lambda_g)\phi(\boldsymbol{v}_i, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)f(\boldsymbol{w}_i, \boldsymbol{\gamma}_g)\right]^{z_{ig}} \tag{10}$$

where z_{ig} is the latent indicator variable with $z_{ig} = 1$, indicating that observation $(\boldsymbol{x_i}, y_i)$ originated from the j-th mixture component, and $z_{ig} = 0$ otherwise.

By taking the logarithm of Equation (10), the complete data log-likelihood function, $\ell_c(\boldsymbol{\Phi})$, is expressed as

$$\ell_c(\boldsymbol{\Phi}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log(\tau_g) + \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, \lambda_g) + \log\phi(\boldsymbol{v}_i, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \log f(\boldsymbol{w}_i, \boldsymbol{\gamma}_g)\right]. \tag{11}$$

It follows that, at the t-th iteration, the conditional expectation of Equation (11) on the observed data and the estimates from the $(t-1)$-th iteration results in

$$Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{(t-1)}) = \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(t-1)}\left[\log(\tau_g) + \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,\lambda_g) + \log\phi(\boldsymbol{v}_i,\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) + \log f(\boldsymbol{w}_i,\gamma_g)\right].$$

The idea behind the EM algorithm is to generate a sequence of the estimates from the maximum likelihood estimation starting from an initial solution $\hat{\boldsymbol{\Phi}}^{(1)}$ and iterating it with the following steps until convergence:

E-step: The posterior probability that $(\boldsymbol{x_i}', y_i)'$ comes from the g-th mixture component is calculated at the t-th iteration of the EM algorithm as

$$\tau_{ig}{}^{(t)} = E[z_{ig}|(\boldsymbol{x_i}', y_i)', \boldsymbol{\Phi}^{(t)}] = \frac{\tau_g{}^{(t)} f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g^{(t)},\lambda_g{}^{(t)})\phi(\boldsymbol{v}_i,\boldsymbol{\mu}_g^{(t)},\boldsymbol{\Sigma}_g^{(t)})f(\boldsymbol{w}_i,\gamma_g{}^{(t)})}{\sum_{g'=1}^{G} f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_{g'}^{(t)},\lambda_{g'}^{(t)})\phi(\boldsymbol{v}_i,\boldsymbol{\mu}_{g'}^{(t)},\boldsymbol{\Sigma}_{g'}^{(t)})f(\boldsymbol{w}_i,\gamma_{g'}^{(t)})\tau_{g'}^{(t)}}. \quad (12)$$

M-step: The Q-function is maximized with respect to $\boldsymbol{\Phi}$, which is done separately for each term on the right hand side in Equation (9). As a result, the parameter estimates $\hat{\tau}_g$, $\hat{\boldsymbol{\mu}}_g$, $\hat{\boldsymbol{\Sigma}}_g$, and $\hat{\gamma}_g$, are obtained on the $(t+1)$-th iteration of the EM algorithm:

$$\begin{aligned}
\hat{\tau}_g^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\tau_{ig}^{(t)}\\
\hat{\boldsymbol{\mu}}_g^{(t+1)} &= \frac{1}{\sum_{i=1}^{n}\tau_{ig}^{(t)}}\sum_{i=1}^{n}\tau_{ig}^{(t)}\boldsymbol{v}_i\\
\hat{\boldsymbol{\Sigma}}_g^{(t+1)} &= \frac{1}{\sum_{i=1}^{n}\tau_{ig}^{(t)}}\sum_{i=1}^{n}\tau_{ig}^{(t)}(\boldsymbol{v}_i-\hat{\boldsymbol{\mu}}_g^{(t+1)})(\boldsymbol{v}_i-\hat{\boldsymbol{\mu}}_g^{(t+1)})'\\
\hat{\gamma}_{gr}^{(t+1)} &= \frac{\sum_{i=1}^{n}\tau_{ig}^{(t)}v_i^{rs}}{\sum_{i=1}^{n}\tau_{ig}^{(t)}},
\end{aligned}$$

while the estimates of $\boldsymbol{\beta}$ are computed by maximizing each of the G terms

$$\sum_{i=1}^{n}\tau_{ig}^{(t)}\log f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,\lambda_g). \quad (13)$$

Maximization of Equation (13) is performed by numerical optimization in the **R** language (R Core Team 2016) in a similar framework as the mixture of generalized linear models are implemented. For additional details about this implementation, the reader is referred to Wedel and De Sabro (1995) and Wedel (2002).

2.3. Model Selection Criterion

In mixture-based clustering, the model selection is often made based on the Akaike information criterion (Akaike 1974) and the Bayesian information criterion (Schwarz 1978) using the following formulas

$$\begin{aligned}
AIC &= 2\ell + 2k\\
BIC &= 2\ell + k\ln(n)
\end{aligned}$$

where l represents the value of the log-likelihood function, k represents the number of estimated parameters in the model, and n is the sample size. Calculation of AIC and BIC is completed for each selected number of mixture components, G. The best model is then selected based on the lowest value of AIC and BIC with the corresponding value of G. We compute both AIC and BIC for the models

discussed in this section. However, we make the final decision based on BIC only since the previous literature suggested that BIC should be preferred over AIC in mixture modeling (for discussion, refer to Fraley and Raftery 2002).

3. Application

3.1. Data

The French motor claims data set, on 413,169 motor third-party liability policies, was accessed from the CASdatasets ((Dutang and Charpentier 2016); Miljkovic (2017)) in R, for the purpose of our analysis. Both claim number and corresponding losses are available from the same portfolio of policyholders. Charpentier (2014) used the same data set to illustrate the modeling of frequency and severity of claims using various single component GLMs. In order to illustrate how OSM and CWM models are applied, we focus our analysis on Region 24 (R24), engine power **f**, and car brand category "Renault, Nissan, or Citroen" with the sample size of 1269 claims. The R24 has about 39% of the total French policies written, with engine power type **f** and "Renault, Nissan, or Citroen" being the most popular cars.

The variables of interest for our analysis are loss amount, driver age, car age, density, and exposure. When the CWM method is employed, loss amount, density, and exposure are treated as numerical continuous variables, while driver age and car age are modeled as categorical variables. The driver age and car age, coded as 99 (unknown), are excluded from the analysis. Table 1 provides the summary of all variables used in both models.

We categorized the numerical variables into ordinal variables with the aim of applying the OSM model over the same data set. The ordinal variables correspond to level of risk, where 1 is the lowest level of risk and 5 is the highest. We assume that this classification has already been determined by the underwriting practices and our goal is to test if there is a need for further classification due to unobserved heterogeneity in the data. The OSM method is sufficiently flexible that it can be adopted to the different variables and a different number of ordinal levels, depending on the data.

Table 1. Summary of the variables used in the cluster-weighted model (CWM) and the ordered stereotype model (OSM).

CWM	
Variable Name	**Description with Categorical Levels in Parenthesis**
Driver Age	<23 (1), [23, 27) (2), [27, 43) (3), [43, 75) (4), and [75+ (5)
Car Age	<1 (1), [1, 5) (2), [5, 10) (3), [10, 15) (4), and 15+ (5)
Density	continuous
Exposure	continuous
Losses	continuous
OSM	
Variable Name	**Description with Ordinal Levels in Parenthesis**
Driver Age	<23 (5), [23, 27) (4), [27, 43) (3), [43, 75) (2), and [75+ (1)
Car Age	<1 (1), [1, 5) (2), [5, 10) (3), [10, 15) (4), and 15+ (5)
Exposure	<0.25 (1), [0.25, 0.50) (2), [0.50, 0.75) (3), [0.75, 1.00) (4), and >1.00+(5)
Density	<40 (1), [40, 200) (2), [200, 500) (3), [500, 4500) (4), and 4500+ (5)
Losses	<1000 (1), [1000, 2000) (2), [2000, 50,000) (3), [50,000, 100,000) (4), and 100,000+ (5)

In the following subsections, we present the analysis and the results based on the two modeling approaches presented in this paper.

3.2. OSM Results

An array of row (claim) clustering models (1) with different numbers of clusters G = 1:5 were fitted, and the information criteria measures AIC and BIC were computed. The results are summarized in Table 2.

Table 2. Model selection. Ordinal variables

G	Loglik	AIC	BIC
1	−12,155	24,453	24,599
2	−12,081	24,188	24,276
3	**−11,777**	**23,584**	**23,685**
4	−12,773	25,580	25,695
5	−12,851	25,641	25769

Both the AIC and the BIC indicate that the best model is the OSM version, including row (claim) clustering with $G = 3$ clusters with AIC = 23,584 and BIC = 23,685. Each row is allocated to the group to which the claim belongs with the highest posterior probability, where the mixing probabilities are $\tau = (0.60, 0.29, 0.11)$. Figure 1 displays the resultant $G = 3$ clustering structures and their profiles. The scatter plot (left) displays the average fitted scores over the 5 variables, using a weighted average which accounts for the fitted spacings. Appendix B explains the details of the calculation of these average scores.

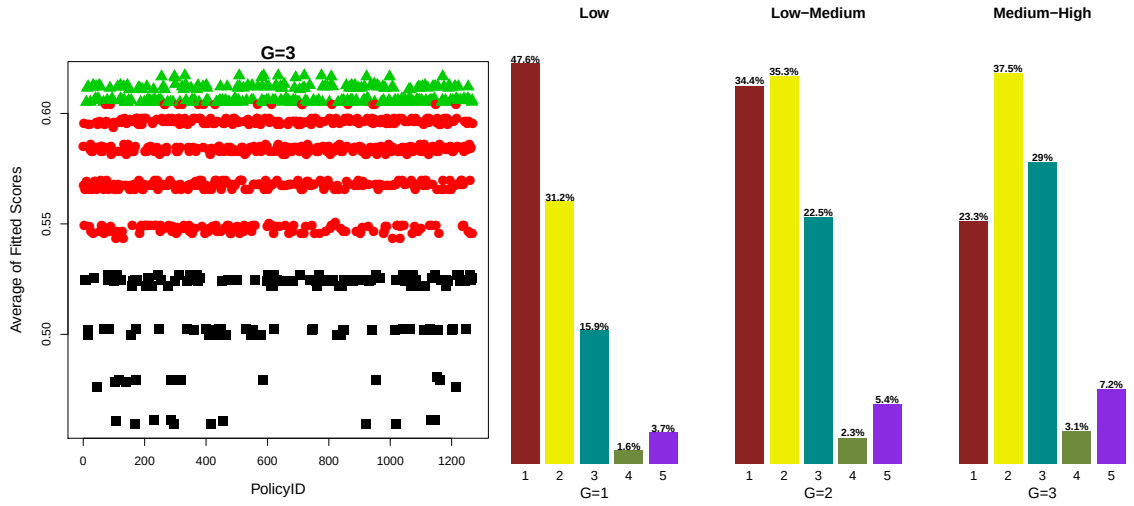

Figure 1. Scatter plot depicting the clustering composition for $R = 3$ (**left**) claim clusters. Different color and shape symbols represent the clusters: Cluster 1 (square), Cluster 2 (circle), and Cluster 3 (triangle). The bar plot (**right**) displays the profile of the claims in each cluster. The percentage represents the probability θ_{gjk} in each category (Equation (2)).

Different color and shape points and color bars represent the resultant $G = 3$ claim clustering settings. Three groups are clearly distinguished in the scatter plot. This graph illustrates the conclusion from AIC/BIC that, among the row clustering models, the model with three claim groups is the best for our data. The bar plot (right) displays the profiles of the different clusters according to the estimated probability of the data response y_{ij} being equal to the category k in Equation (2). We might conclude that the claims classified in the first group correspond to those with the lowest risk regarding the five categories, the ones in the second group have a more low-to-moderate risk, and the claims in the third group are those with more risk. An attractive feature of the stereotype regression model is that it allows one to test whether two adjacent categories are distinguishable. Since each ordinal

response category k ($k = 1, \ldots, 5$) is associated with a score parameter ϕ_k, the spacing between adjacent ϕ_k values shows us how similar or different the categories are in terms of the effect of claims (see Agresti (2010) (Section 4.3.5) and Fernández et al. (2016) (Section 1.2.2)). Table A1 in Appendix A shows the parameter estimates and their uncertainty for the model with $G = 3$ clusters. For this data set, the fitted score parameters were $\phi_1 = 1$, $\widehat{\phi}_2 = 3.636$, $\widehat{\phi}_3 = 4.855$, $\widehat{\phi}_4 = 4.990$, and $\phi_5 = 5$. Therefore, the distance between ordinal categories "1" and "2" (2.636) is greater than that between other adjacent categories. However, there is almost no spacing between categories "4" and "5". This implies that the relative frequencies in these two categories are independent of the clustering structure. Therefore, retaining the distinction between "4" and "5" is not informative about the clustering structure. In that case, the model still holds with the same scores if the ordinal scale is collapsed by combining those two adjacent categories into one single response category. This spacing setting is illustrated in the spaced mosaic plot (Fernández et al. 2014) in Figure 2.

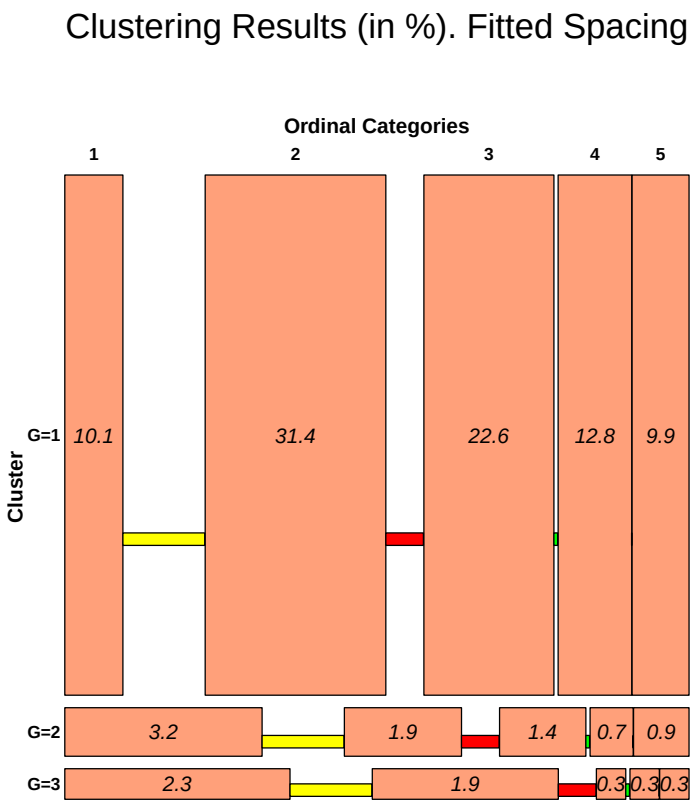

Figure 2. Spaced mosaic plot for the row clustering model $G = 3$. The height of each block is proportional to the number of claims in each claim cluster; the width is proportional to the numbers of each ordinal response within each cluster. The area represents the frequency of each combination, also shown numerically in each block. The relative spacing between ordinal categories (e.g., 2.636 between 0 and 1, shown by the yellow, red, and green bars) has been determined by the data.

Common location measures in continuous response variables such as the mean or the median are not directly applicable in ordinal responses. However, the estimates of the scores parameters $\{\widehat{\phi}_k\}$, $k = 1, \ldots, q$ provide a continuous scale in $[0, 1]$. Re-scaling $\{\widehat{\phi}_k\}$ conveniently as $v_1 = 1$, $v_q = q$, and $v_k = 1 + (q - 1) \times \hat{\phi}_k$ gives us a continuous scale in the range $[1, q]$, which we can use as a new numerical value in our data set, replacing the original ordinal response values $\{y_{ij}\}$ with the new adjusted spacing $\{v_k\}$. For instance, if initially $y_{ij} = k$, the replacement would be $y^{\star}_{ij} = v_k$. We can then calculate average over this new data $\{y^{\star}_{ij}\}$ to determine the different profiles of each cluster from our original heterogeneous data. Table 3 shows the mean in the data set.

Table 3. Mean of the Ordinal Variables by Cluster ($G = 3$).

G	Loss	Driver Age	Exposure	Car Age	Density
1	3.22	4.63	4.52	4.57	3.57
2	1.95	4.81	3.96	4.38	1.88
3	1.78	4.90	3.20	3.23	1.94

We can see that the cluster $G = 1$ is composed by the claims with the largest losses, which corresponds to the youngest drivers, oldest cars, and largest density. The clusters $G = 2$ and $G = 3$ have very similar values apart from the car age, which makes them different.

3.3. CWM Results

Tools for generalized linear CWM are available in the statistical R package **flexCWM**, developed by Mazza et al. (2017). The log-normal CWM was fitted to the following covariates: driver age, car age, density, and exposure. The model selection procedure based on the AIC and the BIC found three mixture components with their corresponding mixing probabilities as follows: 0.52, 0.43, and 0.05. Table 4 shows the summary results for log-likelihood, AIC, and BIC. The CWM function selects the best model based on the minimum value of BIC. In our analysis, the best model is detected when $G = 3$ and these results are shown in bold in Table 4. The number of selected components is consistent with the OSM approach.

Table 4. Model selection. CWM.

G	Loglik	AIC	BIC
1	−12,495	25,025	25,112
2	−11,956	23,229	23,394
3	**−11,064**	**22,222**	**22,464**
4	−10,801	22,200	22,519

This analysis is done for Region 24, engine power **f**, and car brand category "Renault, Nissan, or Citroen" with the sample size of 1269 claims. Within these settings, we can observe that the effect of clustering is present in the portfolio and three subpopulation of drivers are found to have similar characteristics. As a result 3-component GLM seems to provide better fit than a single component GLM based on BIC criteria.

The left display of Figure 3 shows the distribution of losses by group and driver age while the right display shows the distribution of losses by group and car age. It is apparent that the highest losses are generated by the youngest drivers (age group 18–22), classified in Group 1. Additionally, the second highest losses are reported by drivers aged 43–74 in Group 1. Old cars, 10–15 years old, are associated with the largest losses in Group 1. A combination of young drivers (less than 23 years old) and old cars (10–15 years) are indicators for the largest losses in this portfolio. A classification of drivers in R24 can be done based on a level of risk to which they are exposed.

Table A2 in Appendix A shows the results for the estimated coefficients in each component of the CWM model. Driver age and car age are coded as factors with Level 1 being omitted. Thus, the coefficients shown in the output for driver's age by category are estimated relative to the youngest drivers (less than 23 years old). We can observe that the significance of the coefficients in each cluster vary greatly across these clusters. The significance is coded as ≈ 0 (***), 0.001 (**), 0.01 (*), and 0.05 (.), corresponding to the p-value of the specific coefficient. Coefficients in Cluster 3 are all significant, while most of the coefficients associated with driver age and car age are not significant in the first two clusters. Coefficients associated with driver age in Cluster 3 are all positive and increase with age showing that losses increase with age. The 75+ age group generate on average more losses compared

to the <23 age group. In Clusters 1 & 2, we observe the negative coefficients for drivers age, indicating a reverse trend in losses for these two groups. Here, we can see that, in these groups, younger drivers generate the highest losses, and, as age increases, losses tend to decrease on average, holding other variables constant. Similarly, the coefficients for car's age by category are estimated relative to the newest cars (less than 1 year old). These coefficients are all significant in Cluster 3 and their values are positive. We observe that older cars generate more losses on average, compared to newer cars in Cluster 3. In Cluster 1, the car age coefficients are mostly negative while in Cluster 2 they are positive. The sign and magnitude of the coefficients change depending on the relative center of each cluster. Density variable is highly significant in all three clusters. Exposure variable is significant in Cluster 1 only. Here, the finite mixture modeling allows us to model a natural representation of heterogeneity in three latent groups of policyholders in this insurance portfolio. Based on the BIC, the model fit is improved by using three component GLM rather than one component. The interpretation of the coefficients can be done for each cluster separately to account for the "observed" heterogeneity within each subpopulation of the corresponding cluster.

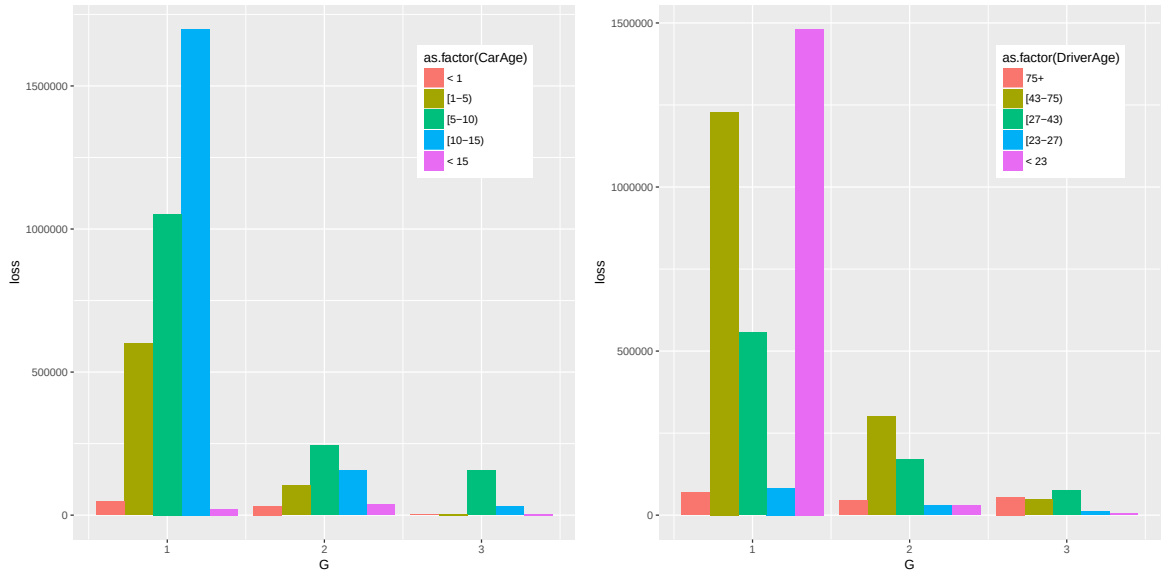

Figure 3. The bar plot (**left**) displays the profile of the losses in each cluster G = 1:3, by driver age. The bar plot (**right**) displays the profile of the losses in each cluster G = 1:3, by car age.

4. Conclusions

This article reviews two recently developed mixture-based clustering approaches applied to an insurance portfolio in order to find the optimal number of clusters and test the assumption of heterogeneity. While these approaches have been used in different fields, to the best of our knowledge, they have so far not been applied in the actuarial field for testing the unobserved heterogeneity in the insurance portfolio. These two methods allow for flexibility in terms of data collection and categorization and allow for further assessment of the underwriting risk. If all the collected data are organized based on the ordinal levels corresponding to several risk classifications, then a mixture-based clustering method for OSM would be a suitable tool to further test heterogeneity in the data. For the insurance data sets that are currently analyzed using existing GLMs, the CWM approach would allow for finding multiple GLM components when modeling frequency and severity of claims. The following family of distributions are supported by CWM with their link functions: binomial, poisson, t-distribution, log-normal, gamma, and inverse gaussian. If the multi-component CWM model provides a better fit to the data then a single component GLM, we recommend that this clustering effect is considered in practice. The OSM

and CWM approaches can be used by practitioners and academics in the actuarial area. We have shown that they have complementary properties.

In a mixture-based clustering method for OSM, when the data is organized based on ordinal levels, the effect of a few extreme losses needs to be tempered down. Allocating these extreme losses in an ordinal, "high-risk" category makes them less influential in the model fitting, giving broad categories which enable us to detect major overall patterns. This enables the inclusion of all of the different levels of dispersion across ordinal categories. Additionally, assigning scores to ordinal categories provides an easy way of showing the descriptive statistics. If practitioners have knowledge about the score for each of the ordered categories, assigning scores might be the best way to analyze the data, because ordinary linear models can be applied. However, if practitioners do not have any predetermined idea about the spacing between adjacent categories, the use of an ordinal stereotype model is convenient, as the data dictates non-equally spaced scores among ordinal outcomes. The estimation of the spacing among ordinal responses is an improvement over other ordinal data models, such as the proportional odds model and continuation-ratio model, although more research in performance comparison with other equivalent methods is needed.

For illustration of these methods, we analyzed a small data set of French motor insurance claims in Region 24 using the methods discussed in this paper. We found three clusters in the data using both the CWM and OSM approaches and therefore show the presence of unobserved heterogeneity. Due to the nature of an insurance portfolio, unobserved heterogeneity is most likely to exist in almost all insurance portfolios. Thus, current modeling techniques should be extended to account for mixture-based clustering that will allow practitioners to detect additional sub-populations of the policy holders based on the same set of characteristics within an existing portfolio, and make the appropriate pricing and risk management decisions for each group. These two methods can also assist practitioners with the underwriting selection process as they have good complementary properties.

Finally, the future research may focus on comparing clustering structures resulting from different methodologies. Over the same data set, one can apply three measures in common that are used to compare clusterings: the Adjusted Rand Index (ARI), the Variation of Information (VI), and the Normalized Information Distance (NID) proposed by Hubert and Arabie (1985), Meila (2005), and Kraskov et al. (2005), respectively.

Appendix A. Model Fitting

Table A1 shows the parameter estimates and their uncertainty for the model with $G = 3$ clusters as a result of fitting the OSM (1) for the French motor claims by policy data set.

Table A1. Results of fitting the OSM (1) for the French motor claims by policy data set.

Coefficient	Estimation	S.E.	95% C.I.
$\widehat{\mu}_2$	0.551	0.148	(0.261, 0.841)
$\widehat{\mu}_3$	−0.219	0.171	(−0.554, 0.116)
$\widehat{\mu}_4$	2.533	0.224	(2.094, 2.972)
$\widehat{\mu}_5$	−1.702	0.160	(−2.016, −1.388)
$\widehat{\alpha}_1$	1.096	0.210	(0.684, 1.508)
$\widehat{\alpha}_2$	0.044	0.125	(−0.201, 0.289)
$\widehat{\beta}_1$	−2.188	0.143	(−2.468, −1.908)
$\widehat{\beta}_2$	−2.631	0.199	(−3.021, −2.241)
$\widehat{\beta}_3$	−0.002	0.190	(−0.374, 0.370)
$\widehat{\beta}_4$	1.673	0.172	(1.336, 2.010)
$\widehat{\phi}_2$	3.636	0.209	(3.226, 4.046)
$\widehat{\phi}_3$	4.855	0.193	(4.477, 5.233)
$\widehat{\phi}_4$	4.990	0.154	(4.688, 5.292)

Table A2. Results for the CWM model. The significance of the p-values are shown with the corresponding level of significance as defined ≈ 0 (***), 0.001 (**), 0.01 (*), and 0.05 (.) for each estimated coefficient.

Cluster 1				
Coefficient	**Estimation**	**S.E.**	***p*-Value**	
Intercept	7.3496	0.2765	$<2.2\times10^{-16}$	***
DriverAge2	−0.3275	0.2137	0.1634	
DriverAge3	−0.2088	0.1963	0.2877	
DriverAge4	−0.0451	0.1925	0.8146	
DriverAge5	0.4828	0.2812	0.0863	.
CarAge2	−0.1575	0.2119	0.4574	
CarAge3	0.0086	0.2001	0.9653	
CarAge4	−0.1845	0.2088	0.3770	
CarAge5	−0.4929	0.2939	0.0938	.
Density	0.0004	0.0001	5.008×10^{-05}	***
Exposure	−0.8287	0.1401	4.332×10^{-09}	***
Cluster 2				
Coefficient	**Estimation**	**S.E.**	***p*-Value**	
Intercept	7.0694	0.0088	$<2.2\times10^{-16}$	***
DriverAge2	−0.0244	0.0084	0.0381	**
DriverAge3	−0.0157	0.0066	0.0177	*
DriverAge4	−0.0095	0.0074	0.1412	
DriverAge5	0.0008	0.0078	0.9186	
CarAge2	0.0051	0.7986	0.4246	
CarAge3	0.0118	2.0162	0.0439	*
CarAge4	0.0101	0.0060	0.0970	.
CarAge5	0.0113	0.0077	0.1440	
Density	3.2818×10^{-06}	3.0435×10^{-06}	0.2811	
Exposure	−0.0051	0.0047	0.2815	
Cluster 3				
Coefficient	**Estimation**	**S.E.**	***p*-Value**	
Intercept	3.3979	0.2561	$<2.2\times10^{-16}$	***
DriverAge2	1.2945	0.1588	8.84×10^{-16}	***
DriverAge3	1.2333	0.1382	$<2.2\times10^{-16}$	***
DriverAge4	1.1096	0.1295	$<2.2\times10^{-16}$	***
DriverAge5	2.6965	0.2028	$<2.2\times10^{-16}$	***
CarAge2	0.6748	0.2105	0.0013	**
CarAge3	1.9939	0.18853	$<2.2\times10^{-16}$	***
CarAge4	1.8501	0.19410	$<2.2\times10^{-16}$	***
CarAge5	2.7567	0.26130	$<2.2\times10^{-16}$	***
Density	1.7878×10^{-04}	4.3673×10^{-05}	4.520×10^{-05}	***
Exposure	6.2711×10^{-02}	0.5266	0.5985	

Appendix B. Average Scores for Scatter Plots

The average score (along the x-axis) shown in Figure 1 is calculated in the following way. First, we compute the fitted response probabilities with the estimated parameters over the G row clusters and the q response categories,

$$P[y_{ij} = k \mid i \in r] = \frac{\exp(\widehat{\mu}_k + \widehat{\phi}_k(\widehat{\alpha}_g + \widehat{\beta}_j))}{\sum_{\ell=1}^{q} \exp(\widehat{\mu}_\ell + \widehat{\phi}_\ell(\widehat{\alpha}_g + \widehat{\beta}_j))},$$
$$i = 1, \ldots, n, \qquad j = 1, \ldots, m, \qquad k = 1, \ldots, q, \qquad g = 1, \ldots, G.$$

From the previous probabilities, we can compute the weighted average over the q categories for each row cluster

$$\overline{\phi}_{rj} = \sum_{k=1}^{q} \widehat{\phi}_k P[y_{ij} = k \mid i \in r],$$
$$i = 1, \ldots, n, \qquad j = 1, \ldots, m, \qquad g = 1, \ldots, G.$$

From here, we can calculate the mean response level of claim i to variable j, conditional on its (fuzzy) allocation to the row clusters:

$$\overline{\phi}_{(ij)} = \sum_{g=1}^{G} \widehat{z}_{ig} \overline{\phi}_{gj}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m. \tag{A1}$$

This is a numerical measure of the typical response to variable j for claims of row cluster g, appropriately adjusting for the uneven spacing of the levels of the ordinal response. Finally, we determine the mean of the previous weighted averages over the m columns in order to obtain the average fitted scores of claim i across all of the variables

$$\overline{\phi}_{(i.)} = \frac{1}{m} \sum_{j=1}^{m} \overline{\phi}_{(ij)}, \qquad i = 1, \ldots, n.$$

Author Contributions: Both authors equality contributed in developing this manuscript.

Acknowledgments: This research has been partially supported by the Marsden grant number E2987-3648 (Royal Society of New Zealand). We thank four anonymous reviewers whose comments and suggestions helped improve the clarify and of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

Agresti, Alan. 2010. *Analysis of Ordinal Categorical Data*, 2nd ed. Wiley Series in Probability and Statistics. New York: Wiley.

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–23. [CrossRef]

Anderson, John A. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society Series B (Methodological)* 46: 1–30.

Baribeau, Annmarie Geddes 2016. Predictive modeling: The quest for data gold. *Actuarial Review*. Available online: https://www.casact.org/pubs/New-AR/AR_Nov-Dec_2016.pdf (accessed on 7 March 2018).

Bermúdez, Lluís, and Dimitris Karlis. 2012. A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis* 56: 3988–99.

Böhning, Dankmar, Wilfried Seidel, Marco Alfò, Bernard Garel, Valentin Patilea, and Günther Walther. 2007. Advances in mixture models. *Computational Statistics & Data Analysis* 51: 5205–10.

Brown, Garfield O., and Winston S. Buckley. 2015. Experience rating with poisson mixtures. *Annals of Actuarial Science* 9: 304–21. [CrossRef]

Charpentier, Arthur. 2014. *Computational Actuarial Science with R*. Boca Raton: CRC Press.

Chen, Lien-Chin, Philip S. Yu, and Vincent S. Tseng. 2011. A weighted fuzzy-based biclustering method for gene expression data. *International Journal of Data Mining and Bioinformatics* 5: 89–109. [CrossRef] [PubMed]

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B* 39: 1–38.

Dutang, Christophe, and Arthur Charpentier. 2016. *CASdatasets*. R package version 1.0-6. Available online: http://cas.uqam.ca/ (acessed on: 6 March 2018)

Everitt, Brian S., Morven Leese, and Sabine Landau. 2001. *Cluster Analysis*, 4th ed. London: Hodder Arnold Publication.

Fernández, Daniel, Richard Arnold, and Shirley Pledger. 2016. Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis* 93: 46–75.

Fernández, Daniel, Shirley Pledger, and Richard Arnold. 2014. *Introducing Spaced Mosaic Plots*. Research Report Series 14-3. Wellington: School of Mathematics, Statistics and Operations Research, VUW. ISSN: 1174-2011.

Fraley, Chris, and Adrian E. Raftery. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97: 611–31. [CrossRef]

Garrido, José, Christian Genest, and Juliana Schulz. 2016. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics* 70: 205–15. [CrossRef]

Gershenfeld, Neil. 1997. Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences* 808: 18–24. [CrossRef]

Green, Peter J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–32. [CrossRef]

Grün, Bettina, and Friedrich Leisch. 2008. *Finite Mixtures of Generalized Linear Regression Models*. Berlin: Springer.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2: 193–218. [CrossRef]

Ingrassia, Salvatore, Antonio Punzo, Giorgio Vittadini, and Simona C. Minotti. 2015. The Generalized Linear Mixed Cluster-Weighted Model. *Journal of Classification* 32: 85–113. [CrossRef]

Jobson, John D. 1992. *Applied Multivariate Data Analysis: Categorical and Multivariate Methods*. Springer Texts in Statistics. Berlin: Springer.

Johnson, Stephen C. 1967. Hierarchical clustering schemes. *Psychometrika* 2: 241–54. [CrossRef]

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data an Introduction to Cluster Analysis*. New York: Wiley.

Klugman, Stuart, and Jacques Rioux. 2006. Toward a unified approach to fitting loss models. *North American Actuarial Journal* 10: 63–83. [CrossRef]

Kraskov, Alexander, Harald Stögbauer,Ralph Gregor. Andrzejak, and Peter Grassberger. 2005. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)* 70: 278–84. [CrossRef]

Lee, Simon C. K., and X. Sheldon Lin. 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal* 14: 107–30. [CrossRef]

Lewis, S. J. G., Thomas Foltynie, Andrew D. Blackwell, Trevor W. Robbins, Adrian M. Owen, and Roger A. Barker. 2003. Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery and Psychiatry* 76: 343–48. [CrossRef] [PubMed]

Liu, Ivy, and Alan Agresti. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 14: 1–73. [CrossRef]

Manly, Bryan F.J. 2005. *Multivariate Statistical Methods: A Primer*. Boca Raton: Chapman & Hall/CRC Press.

Mazza, Angelo, Antonio Punzo, and Salvatore Ingrassia. 2017. *flexCWM*. R package version 1.7. Available online: https://cran.r-project.org/web/packages/flexCWM/index.html (acessed on 6 March 2018)

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society* 42: 109–42.

McCullagh, Peter, and John A Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

McCune, Bruce, James B. Grace. 2002. *Analysis of Ecological Communities*. Gleneden Beach: MjM Software Design, vol. 28.

McLachlan, Geoffrey, and David Peel. 2004. *Finite Mixture Models*. Hobuken: John Wiley & Sons.

McLachlan, Geoffrey J., and Kaye E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering*. Statistics, Textbooks and Monographs. New York: M. Dekker.

Meila, Marina. 2005. Comparing clusterings: An axiomatic view. Paper presented at the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, August 7–11, pp. 577–84.

Melnykov, Volodymyr, and Ranjan Maitra. 2010. Finite mixture models and model-based clustering. *Statistics Surveys* 4: 80–116. [CrossRef]

Miljkovic, Tatjana, and Bettina Grün. 2016. Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics* 70: 387–96. [CrossRef]

Miljkovic, T. 2017. Computational Actuarial Science With R. *Journal of Risk and Insurance* 84: 267.

Nelder, John Ashworth, and Robert W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A (General)* 135: 370–84. [CrossRef]

Pledger, Shirley, and Richard Arnold. 2014. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics and Data Analysis* 71: 241–61. [CrossRef]

Quinn, Gerry P., and Michael J. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Team, R. Core. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6: 461–64. [CrossRef]

Shi, Peng, Xiaoping Feng, and Anastasia Ivantsova. 2015. Dependent frequency severity modeling of insurance claims. *Insurance: Mathematics and Economics* 64: 417–28. [CrossRef]

Verbelen, Roel, Lan Gong, Katrien Antonio, Andrei Badescu, and Sheldon Lin. 2014. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulltin* 45: 729–58. [CrossRef]

Wedel, Michel. 2002. Concominat variables in finite mixture modeling. *Statistica Neerlandica* 56: 362–75. [CrossRef]

Wedel, Michel, and Wayne S. DeSarbo. 1995. A mixture likelihood approach for generalized linear models. *Journal of Classification* 12: 21–55. [CrossRef]

Werner, Geoff, and Claudine Modlin. 2016. *Basic Ratemaking*. Arlington: Casualty Actuarial Society.

Wu, Han-Ming, ShengLi Tzeng, and Chun-houh Chen. 2007. Matrix visualization. In *Handbook of Data Visualization*. Berlin: Springer, pp. 681–708.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, and et al. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14: 1–37. [CrossRef]