Credibility and Penalized Regression

Mattia Casotto,[a] **Marco Banterle,**[a] **Guillaume Beraud-Sudreau**[a]

[a] *Akur8, France*
E-mail: mattia.casotto@akur8.com, marco.banterle@akur8.com, guillaume.beraud@akur8.com

Abstract: In recent years a number of extensions to Generalized Linear Models (GLMs) have been developed to address some limitations, such as their inability to incorporate Credibility-like assumptions. Among these adaptations, Penalized regression techniques, which blend GLMs with Credibility, are widely adopted in the Machine Learning community but are not very popular within the actuarial world. While Credibility methods and GLMs are part of the standard actuarial toolkit of predictive modeling, the actuarial literature describing how Penalized regression blends Credibility with GLMs is not equally developed. The aim of this whitepaper is to provide practitioners with key concepts and intuitions that demonstrate how Penalized regression blends GLM with Credibility-like assumptions. By walking through a simple example, we will explore how Penalized regression (and Lasso in particular) can be interpreted from the perspective of both Credibility and GLM frameworks. The whitepaper objective is to familiarize practitioners with Penalized regression as an extension of established actuarial techniques, instead of considering it one among several new modeling techniques from the Machine Learning and Data Science literature.

Contents

1 Introduction

Predictive modeling has a large number of operational applications within the insurance world. Actuaries have access to a wide toolkit of modeling techniques that best address the various use cases arising in insurance applications. Among those modeling techniques, the majority of predictive models used in retail insurance rating plans are Generalized Linear Models (GLMs). GLMs gained popularity in insurance applications as they introduced significant improvements over univariate-based rating plans. Their main benefit compared to univariate-based modeling is the ability to consider the correlations and dependencies between the selected input variables.

The change from univariate modeling to the multivariate GLM approach comes with one obvious shortcoming - there is no statistically straightforward, consistent way of incorporating actuarial credibility into a GLM. [Klinker, 2010]

In the recent *Whitepaper of Regulatory Review of Predictive Models* by the NAIC [NAIC, 2020], such shortcomings are described as

> *"GLMs effectively assume that the underlying datasets are 100% credible, no matter their size. If some segments have little data, the resulting uncertainty would not be reflected in the GLM parameter estimates themselves. (Although it might be reflected in the standard errors, confidence intervals, etc.)"*
>
> Whitepaper of Regulatory Review of Predictive Models [NAIC, 2020]

This means that, even if the GLM warns the user of the instability of the parameter estimate by providing a big standard error to it, it does not inherently adjust the coefficient to take into account this large volatility, leaving it up to the practitioner to perform ad-hoc adjustments to consider the lack of exposure in a specific segment.

Still, these "volatility adjustments" are available in a univariate scenario, where estimations of risk have to be provided for each **level** (or **factor**) of a class (or any discrete variable). Credibility methods allow the observed risk averages by class to blend with external information called "complement of credibility." Such complement of credibility is assumed to be a more stable yet less precise estimate of the effect. For example, the complement of credibility could be the grand average of risk across the classes, or the current estimates used in a model built on other data. The way observations and complement of credibility are blended together is through a weighted average. The weight Z depends on, among other things, the number of observations per class. Adaptively weighting the observed data by class has been proven to lead to more stable and predictive models. At first glance, it might seem that this "credibility-weighting" technique can be applied only in the domain of univariate modeling. This whitepaper will investigate the theoretical framework underlying the credibility approach, and demonstrate how it can easily be generalized to multivariate modeling. In particular, it will reveal how Penalized regression is exactly equivalent to "credibility-weighting" of GLMs, and how this approach creates more sound and stable estimates of risk.

Section 2 reviews basic concepts of a GLM and provides intuition insight into why GLMs offer 100% credibility to the data by introducing a simple actuarial use case predicting workers' compensation losses.

Section 3 provides an overview of Credibility and Penalized regression, together with how the resulting estimates compare to GLMs.

Section explores how Penalized regression blends Credibility and GLM by describing the connections of the different methodologies through Bayesian statistics. Such connections will be proved both from a theoretical and applied standpoint.

2 Full credibility of GLMs

To define the notations and concepts used in the whitepaper, it is necessary to introduce the basic definitions and concepts of a GLM. For a proper comprehensive introduction to GLMs, the reader can refer to [Goldburd et al., 2016]

Essentially, a GLM consists of three elements [NAIC, 2020]:

1. A target variable, Y, which is a random variable that is independent and is assumed to follow a probability distribution from the exponential family, defined by a selected variance function and dispersion parameter.
2. A linear predictor, $\eta = X\beta$ where X is the design matrix and β is the vectors of coefficients.
3. A monotonic link function g, such that $\mathbb{E}(Y) = \mu = g^{-1}(\eta)$.

These elements have established connections to common insurance concepts.

Y represents the risk. Depending on the nature of the risk, different statistical assumptions are made. For example, accident frequency may be modeled via the Poisson probability distribution. In the same way, accident severity may be modeled via a Gamma distribution.

The matrix X represents the information about the user (or policy or company) and any covariate relevant for predicting the considered risk. Typically, each row of X represents a unit of risk, specified by the modeler (for instance a contract, or a year of observation of the risk). The columns of the matrix provide a numerical representation of the available information on the risk (covariates).

Depending on the nature of the information, the columns of the matrix X will have different representations. In the case of numerical information, at its simplest the column will describe the value itself, and a simple linear dependency will be assumed. If a more adaptive parametric representation is preferred (for example, a polynomial), the columns of X may be extended by transformations of the original values to allow for more complex shapes which can still be modeled in a linear fashion.

In the case of categorical variables, such as classes, the information is usually represented via binary encoding. Each column of the X matrix corresponds to each specific level of the class, with a value of 1 if the row belongs to such level, 0 otherwise.

The value of the coefficient, β, defines how the covariates are linearly combined together to estimate the risk.

Finally, μ represents the actual predictions of risk. The link g represents how the linear combination of the features $X\beta$ are transformed: $\mu = g^{-1}(\eta)$ Standard choices of the link function are logarithm, which gives a multiplicative model, identity or logit. The specific choice of the link is often related to the choice of the probability distribution chosen at modeling.

The process of building (or fitting) a GLM requires the specification of y (and its statistical assumptions) together with the covariates X. The output is a set of coefficients β, which maximizes the likelihood of the data under the statistical hypothesis (the likelihood of the observations being the probability of observing the value y of the target, given X and β). The way that the coefficients are computed is at the core of why *"GLMs effectively assume that the underlying datasets are 100% credible, no matter their size"*.

To provide intuitions of this matter, it is best to define a simplified actuarial use case that will be used repeatedly throughout the whitepaper.

The use case consists of building estimates of loss costs for companies in a workers' compensation insurance. The modeler has access to a dataset of historical data, where each row represents the total loss observation for a specific company in a fixed 1-year period. On top of that, the class code information is available for each company. There are several class codes and for most of them the number of observations is limited. One choice for the practitioner is to ignore the class code information, estimating a constant value for all companies - the grand average - regardless of their properties. This approach allows for a robust but not precise estimate and can be improved. For this reason, the modeler decides to focus the attention on the deviations of the losses from the grand average, using methodologies from the standard actuarial toolkit.

GLMs can of course be applied to the above use case and we will use this example to introduce the following notation that will be used throughout the whitepaper.

The database contains n observations. The information on the companies is encoded in the matrix X, which provides the binary representation of the information of the p class codes in the database. The coordinate x_{ij} of the matrix X will be 1 if company i belongs to class j, 0 otherwise. X_i will denote the row vector of matrix X of size p. In general, index i will be used to represent a line in the matrix/database, and i takes values from 1 to n. In the same way, index j will be used to represent columns of the matrix, and j takes values from 1 to p.

Y represents the vector of the observed losses, meaning that Y_i will represent the observed loss for company i. The grand average is given by $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. y represents the differences of the observed losses from the grand average $\bar{Y}$, that is $y_i = Y - \bar{Y}$ (so y is centered on zero).

n_j denotes the number of observations belonging to class j. The set J represents the set of rows i belonging to class j, that is, i such that $x_{ij} = 1$. This implies that the cardinality of J is n_j . The constant $\bar{y}_j$ represents the observed average loss deviation by class code j, that is $\frac{1}{n_j}\sum_{i\in J} y_i$.

To complete the specification of a Generalized Linear Model the distributional assumptions, the link and the target have to be defined. For the sake of simplicity we will assume a normal distribution of the losses deviations, with constant variance σ^2 and an identity

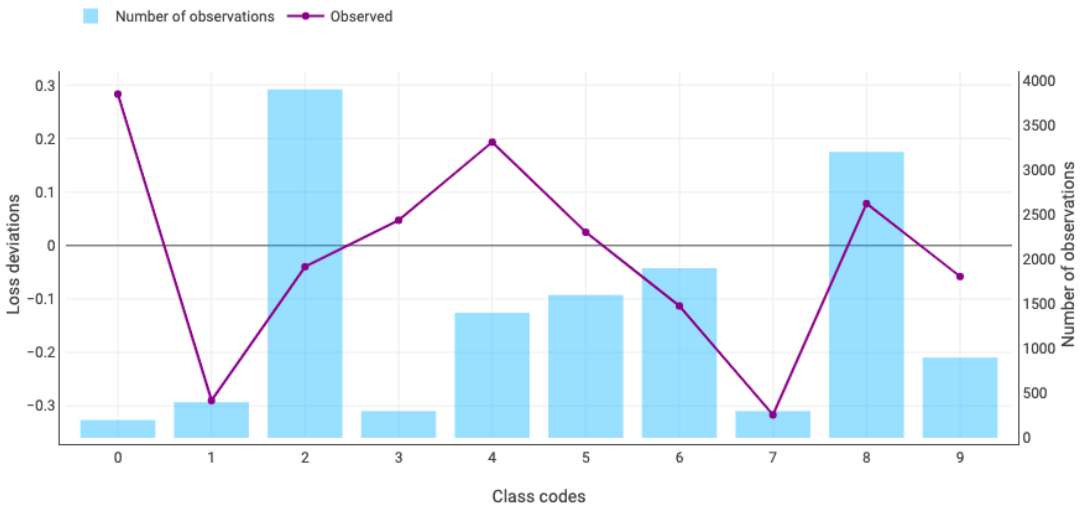

Figure 1: The purple line represents the behavior of the observed average loss deviation $\bar{y}_j$ by class code j on synthetic data. The blue bars represent the total number of observations n_j by class code j.

link. Under these assumptions, the natural target variable is the vector y representing the differences of the observed losses from the grand average [1] .

The GLM coefficients β are estimated by maximizing the log-likelihood, or, equivalently, by minimizing the *negative* log-likelihood (NLL)

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\text{argmax}}\text{Log-Likelihood}(y, X, \beta) \\ &= \underset{\beta}{\text{argmin}}\text{NLL}(y, X, \beta)\end{aligned} \tag{1}$$

Which in the case of a Gaussian distribution with an identity link becomes

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i\beta)^2 \tag{2}$$

Since σ is constant, it doesn't impact the optimization problem, and we can ignore it. We can proceed to compute the gradient only of the summation, by differentiating with respect to each coefficient. This solution can then be found by setting the gradient to zero. The gradient can be simplified

$$\frac{\partial}{\partial \beta_j} [\sum_{i=1}^{n} (y_i - X_i\beta)^2] = \sum_{i=1}^{n} (X_i\beta - y_i)x_{ij} \tag{3}$$

[1] We stress that modeling on the centered vector of differences y instead of the observed losses Y is equivalent to using the grand average $\bar{Y}$ as an **offset** of a Gaussian GLM with identity link, and to use Y as the target. Viewing the modeling from an offset perspective is much more generic and allows us to use different assumptions than a Gaussian GLM with identity link, but it would make the notation much heavier.

by application of the chain rule of the derivative,

$$\sum_{i=1}^{n}(X_i\beta - y_i)x_{ij} = \sum_{i\in J}^{n}(X_i\beta - y_i) \tag{4}$$

the summand is not null only for $i \in J$ since $x_{ij} = 0$ otherwise.

$$\sum_{i\in J}^{n}(X_i\beta - y_i) = n_j\beta_j - \sum_{i\in J} y_i = n_j\beta_j - n_j\bar{y}_j \tag{5}$$

The first addend $n_j\beta_j$ appears since, for all $i \in J$, $X_i\beta = \beta_j$; the $n_j\bar{y}_j$ term is by the definition of the average $\bar{y}_j = \sum_{i\in J} y_i/n_j$.

Setting the gradient to zero, implies that

$$0 = n_j\beta_j - n_j\bar{y}_j \quad \leftrightarrow \quad \beta_j = \bar{y}_j \tag{6}$$

This proves that the coefficients β_j maximizing the log-likelihood are exactly matching the average $\bar{y}_j$ of each class.

Figure 2: The green compares the GLM estimates $\hat{\beta}_j$ with the data as represented in Figure 1. The observed (purple line) coincides with the GLM estimates (green line).

The proof above highlights two important features of the GLMs.

1. When modeling a single class in a binary encoding, a univariate GLM will output the average of the observations for such a class. This is true as well for all distributions when the canonical link is used, as it will be further discussed in Section 4.
2. The condition 4 applies as well on multivariate settings. Whenever a discrete variable is added in a binary format, the coefficients will be estimated so that the average of the predictions $X_i\beta$ on every level j will match the average of the observations, **regardless of the underlying exposure.** This equality is a key property of the GLMs, and will be further discussed in Section 4.

In this sense, **any fitting procedure** that computes estimates by maximizing the likelihood of the data (or minimizing the deviance) **alone**, *effectively* ***assumes*** *that the underlying datasets are 100% credible, no matter their size.*

Applying a simple GLM on the workers' compensation use case provides the empirical averages of the different classes as estimates. These estimates unfortunately are volatile for all class codes whose number of observations is limited. This behavior will be highlighted by GLM statistics such as the p-values or confidence intervals, which will be abnormally high. In this sense, the GLM provides only warnings to the user to exercise considerable care in interpreting that coefficient but doesn't otherwise adjust the estimated coefficient to take into account the low volume of data [Klinker, 2010].

The next section will introduce the Credibility and Penalized regression frameworks, and apply them to the same workers' compensation use case to provide an intuitive comparison between them and with the simple GLM approach.

3 Credibility and Penalized regression

Credibility techniques allow to blend the "goodness of fit" of the model with considerations on the underlying amount of data used to compute the estimates. We will show that integrating credibility in estimates does in fact "shrink" the GLM estimates. Penalized regression, introduced in this section, shrinks the coefficients as well. By computing the Penalized regression estimates on the workers' compensation use case, we will lay down the fundamentals to prove the equivalence of Credibility and Penalized regression in the third section.

3.1 Credibility

> *"Credibility, simply put, is the weighting together of different estimates to come up with a combined estimate."*
>
> Foundations of Casualty Actuarial Science [Casualty Actuarial Society, 2001]

In the context of ratemaking, Credibility provides a framework to combine an estimate based on observed experience (observed losses, frequencies, or loss ratios), subject to volatility, with a more stable yet less individualized estimate (the "complement of credibility"). This combination aims to improve on both estimates to create better predictions of future values.

The estimates are blended together via the credibility factor, called Z, a factor between 0 and 1 which will give more or less weight to the observed experienced or the complement of credibility:

$$\text{Estimate} = Z * \text{Observed Experience} + (1 - Z)\ \text{Complement of credibility} \tag{7}$$

There are two main types of Credibility found in the literature: *Classical* and *Buhlmann*. Even if they differ in terms of underlying hypothesis and formulation of the factor Z (see Table 1), they share the same basic credibility property: the credibility factor increases with

the number of observations n (*i.e.* the exposure). In this sense, unlike simple GLMs, the Credibility framework allows to incorporate the information of the number of observations directly into the estimates.

The Appendix B provides a summary of the underlying hypothesis of Classical and Buhlmann credibility and their comparison.

Classical Credibility	**Buhlmann Credibility**
$Z_j = \min(\sqrt{\frac{n_j}{N_{\text{full}}}}, 1)$	$Z_j = \frac{n_j}{n_j+k}$
Additional parameters	Additional parameters
$N_{\text{full}} = N_{\text{full}}(k,p)$ number observation to reach full credibility	K = ratio of $\tau^2_{PV}/\sigma^2_{HM}$, where
P probability that the observation are within estimated risk	τ^2_{PV} = Expected Process variance - Within class variance
K tolerance to error, as % of risk	σ^2_{HM} = Variance of Hypothetical Means - Between class variance

Table 1: Comparison of credibility factor and additional parameters for Buhlmann and Classical Credibility

The credibility formula can be easily applied to the workers' compensation use case: to compute the estimates, the complement of credibility and the credibility factor Z have to be computed.

In this case, the natural complement of credibility is the grand average of the losses $\bar{Y}$ and the observed experience would be the average of loss by class codes. Then, for each of the class codes j a credibility factor Z_j must be computed as a function of the number of observations n_j and other parameters. Finally, the Credibility estimate $\hat{y}_j$ for the class code j will become

$$\begin{aligned} \hat{y}_j &= Z_j(\bar{Y} + \bar{y}_j) + (1 - Z_j)(\bar{Y}) \\ &= \bar{Y} + Z_j\ \bar{y}_j \end{aligned} \tag{8}$$

To draw a parallel with the GLM example above, the same credibility approach can be applied as well to model directly the differences between the class codes losses and the grand average, as the complement of credibility for the loss deviations is zero:

$$\hat{\beta}_{\text{Credibility},j} = Z_j\ \bar{y}_j = Z_j\ \beta_{\text{GLM},j} \tag{9}$$

The credibility estimates on the workers' compensation use case $\hat{\beta}$ differ from the GLM estimates $\bar{y}_j$ by a multiplicative factor Z_j between 0 and 1. This means that the GLM estimates are shrunk towards the complement of credibility (in this case, zero) by a constant that depends, among others, to the number of observations of the class. This makes intuitive sense: the more evidence is collected, the more the estimates can deviate from the initial

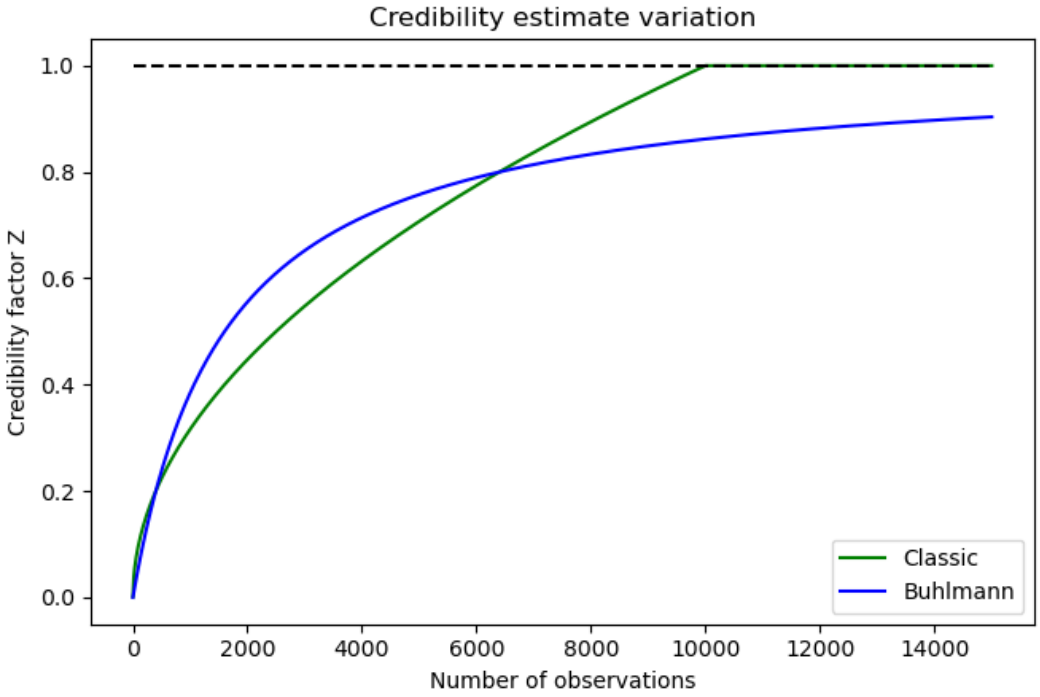

Figure 3: Evolution of the credibility factor Z_j for a given class code j as a function of the number of observations n_j. The Z for the Classic Credibility is computed as $Z_j = \min(\sqrt{\frac{n_j}{N}}, 1)$, with $N = 15000$. The respective formula for Buhlmann credibility is $Z_j = \frac{n_j}{n_j+k}$, $k = 1600$.

assumption which is the complement of credibility. In this specific case, the more exposure is available, the more the coefficients can deviate from the grand average.

Figure 4 provides a visual depiction of an application of Credibility vs the GLM estimates.

The example above illustrates empirically the application of credibility adjustments to a GLM.

First, the complement of credibility (here, the grand average) is computed. Then, the GLM estimates are computed using the complement of credibility as an offset. The coefficients β will represent the adjustment between the complement of credibility and the data. Finally, for each variable of interest, the credibility factors Z_j will be computed and then multiplied by the coefficients β, effectively shrinking the predictions toward the complement of credibility, by a constant dependent on the exposure.

The procedure above is correct as long as the model contains a single variable; the more we add variables in the model, the more it loses efficiency, since correlations across variables play a more important effect.

Figure 4: Illustration of "shrinkage" after application of credibility weighting. The blue line represents the Credibility estimate under the Buhlmann model, where $Z_j = \frac{n_j}{n_j+k}$, $k = 225$. The yellow line represents the Credibility estimates under the classical assumption $Z_j = \min(\sqrt{\frac{n_j}{N}}, 1)$ $N = 1500$. The other quantities are as defined in Figure 2. For class codes with enough observations (e.g. $j = 2, 8$) the classical Credibility estimate are equal to the GLM estimates.

There are of course other, less simplistic ways, to adjust GLM estimates in order to blend credibility assumptions, but they all share the drawbacks highlighted in [Klinker, 2010]

> *Some actuaries have been known to apply an ad hoc credibility adjustment to coefficients output by a GLM. In some cases this even produces results similar to those arrived at by more statistically rigorous methods. If so, then what is so wrong with the ad hoc credibility adjustment of GLM output? [. . .] This gets back to the old issue that a sequence of steps, each optimal individually, may not be optimal in the aggregate*

In order to obtain a statistically rigorous multivariate modeling technique that is able to blend credibility, there are at least three necessary properties to be satisfied:

1. The estimation of the parameters shall not rely on maximizing the log-likelihood (or variance) alone: any technique with this property will inevitably assign 100% credibility to the data

2. When a complement of credibility is used as an offset, the estimates will be shrunk compared to GLMs and the amount of shrinkage will depend on the number of observations

3. To consider correlations, the "credibility-weighting" of the coefficients must be a part of the fitting procedure, not a post-processing step on top of a GLM

Penalized regression satisfies these three necessary properties. To understand why (and how), the next paragraph will introduce the main concept of this modeling approach and

illustrate it on the same workers' compensation use case as above.

3.2 Penalized regression

The intuition behind Penalized regression is simple: maximizing the likelihood (or minimizing the deviance) alone will always give 100% credibility to the data, making the model prone to overfitting; by adding a "penalization" term to the problem it is possible to **jointly** optimize the tradeoff between pure goodness of fit (as GLM do) and a desirable structure of the coefficient induced by the penalty. This allows to "sacrifice" some deviance in the training data, in favor of a model whose parameters have better desirable properties, such as shrinkage when the credibility of a segment is limited.

The general formula of a Penalized regression is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \ \mathrm{NLL}(y, X, \beta) + \lambda \ \mathrm{Penalty}(\beta) \tag{10}$$

where $\lambda \geq 0$ is a positive number. The λ parameter plays the role of a knob which gives more or less importance to the goodness of fit of the model in the train database or to the a priori structure of the coefficients.

Penalized regression is, at the core, very similar to GLMs in their mathematical specification [Goldburd et al., 2016]. The same assumptions can be made on the relationship between the coefficients and the response via a linear predictor and the link function, and the same distributions can be assumed as generating the target. The only difference from a GLM is the definition of the penalty function and the penalization parameter λ. When λ is null, Penalized regression is exactly a GLM.

In Penalized regressions, the choice of the penalty function is generic, but there are two main established types of penalizations. The **Ridge** penalty is given by sum of squares (or l2 squared norm) of β, that is $\sum_{j=1}^{p} \beta_j^2$.

The **Lasso** penalty is given by the sum of the absolute values (or l1 norm) of β, that is $\sum_{j=1}^{p} |\beta_j|$. [2]

As depicted in figure 5, the farther from zero the coefficient beta, the higher the value of the penalty. When solving a Penalized regression problem, the penalty term will counterbalance the effect of perfectly maximizing the likelihood thus estimating coefficients which are shrinked compared to GLM estimates. The amount of shrinkage will depend on the magnitude of the parameter λ. We thus see that Penalized regression (with Lasso or Ridge penalty) satisfies the necessary conditions outlined above for a modeling technique that can blend Credibility assumptions.

Choosing between a Ridge or a Lasso penalty will provide coefficients which have different properties, as it will be shown by applying those two penalties in the workers' compensation use case.

[2] The genericity of the Penalized regression framework allows to blend the Ridge and the Lasso in the Elastic Net. Elastic Net in actuarial application can be often referred to as GLMnet, which is the name of the standard R package used for fitting Elastic Nets [Hastie and Qian, 2014]

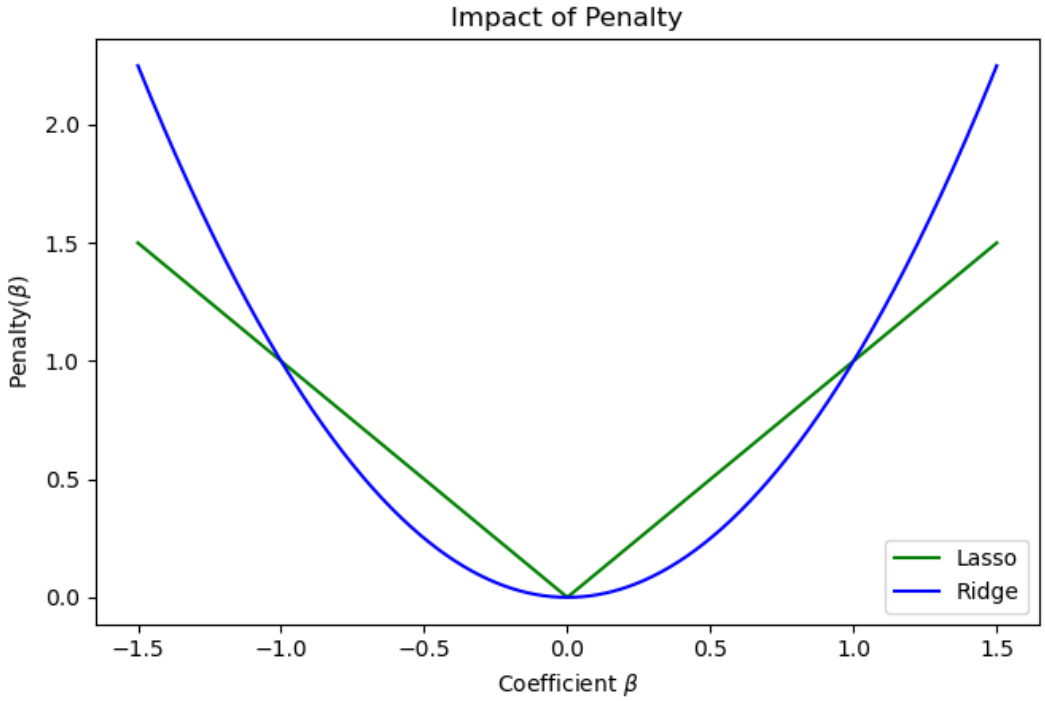

Figure 5: Visualization of function $\beta \to \beta^2$ (1-dimensional Ridge penalty) and $\beta \to |\beta|$ (1-dimensional Lasso penalty).

3.2.1 Ridge

The Ridge regression is a Penalized regression having the Ridge penalty. The estimation for Ridge problem is given by [3]

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \text{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{11}$$

The formulation of the Ridge penalty can be traced to [Hoerl and Kennard, 1970], and its success is determined by its ability to handle correlated variables. Differently than GLMs, where the presence of highly correlated variables determines numerical instabilities of the optimization routines (aliasing), the penalty term provides protection against the coefficients "blowing up" as they might in a GLM [Goldburd et al., 2016].

By applying to the workers' compensation use case we can see how the Ridge compares

[3] Different formulations of the Penalized regression can be found in the literature. For example, often the negative log-likelihood (NLL) is normalized by the number of observations n. In all cases, the formulas are equivalent after a reparametrization of the parameter λ (for example $\lambda \to \frac{\lambda}{n}$).

with the GLM solution. To compute the Ridge estimate, we need to compute the solution $\hat{\beta}$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y - X\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{12}$$

This solution can be found by means of computing the gradient and setting it to zero, similarly to the GLM. The differential of the optimization formula above (denoted f below) is

$$\begin{aligned} \frac{\partial f}{\partial \beta_j}[\ldots] &= \sum_{i=1}^{n} (X_i\beta - y_i)x_{ij} + 2\lambda\ \beta_j \\ &= \sum_{i \in J}^{n} (X_i\beta - y_i) + 2\lambda\ \beta_j \\ &= n_j\beta_j - \sum_{i \in J} y_i + 2\lambda\ \beta_j \\ &= n_j\beta_j - n_j\bar{y}_j + 2\lambda\ \beta_j \end{aligned} \tag{13}$$

The solution is given by the vector which sets the gradient to zero, that is

$$\hat{\beta}_j = \frac{n_j}{n_j + 2\lambda} \bar{y}_j \tag{14}$$

The addition of the penalty term effectively "shrinks" the observed estimates $\bar{y}_j$ by a number which is dependent on the number of observations of the segment in question.

3.2.2 Lasso

The Lasso regression is a Penalized regression having the Lasso penalty $\sum_{j=1}^{p} |\beta_j|$. The estimation for Lasso problem is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \mathrm{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{15}$$

The Lasso model was introduced by the seminal paper [Tibshirani, 1996]. The Lasso achieves *sparsity*, that is, the ability to set exactly to zero those coefficients which are non-significant, as part of the fitting procedure. This means that the Lasso is able to automate variable selection, as well as to build models where the number of features p outgrows the number of observations n.

The sparsity property of the Lasso is at the root of its wide success in various applications. It is therefore even more important to understand **how** the Lasso achieve sparsity and naturally estimates coefficients β with some null coordinates, $\beta_j = 0$. The Appendix A explains with simple arguments why introducing the absolute value function $|\beta|$, which is non-differentiable, leads to sparsity and variable selection.

On the workers' compensation use case, the associated Lasso regression becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y - X\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{16}$$

We prove in the Appendix A that the solution $\hat{\beta}$ for the workers' compensation use case has a similar structure to the simplified example illustrated above

$$\hat{\beta}_j = \begin{cases} \bar{y}_j - \frac{\lambda}{n_j} & \text{if } \bar{y}_j > \frac{\lambda}{n_j} \\ \bar{y}_j + \frac{\lambda}{n_j} & \text{if } \bar{y}_j < -\frac{\lambda}{n_j} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

This piecewise function is called "soft-thresholding" in the literature, and is represented in Figure 6. In particular, whenever the quantity $n_j \bar{y}_j = \sum_{i \in J} y_i < \lambda$, the coefficient for the class j will be set equal to zero, hence assigning no credibility to the data if the quantity of signal is not relevant enough.

Comparison Figure 6 displays how the different estimates β_j differ in the workers' compensation use case for fixed λ, n_j

Compared to the GLM estimates ($\hat{\beta}_j = \bar{y}_j$), the Lasso reduces the coefficient by a factor λ/n_j accordingly to the sign of $\bar{y}_j$. If the value of $\bar{y}_j$ is lower than such a threshold, then such value is set to zero. Compared to the Ridge estimates ($\hat{\beta}_j = \frac{n_j}{n_j+2\lambda}\bar{y}_j$), the Lasso shrinks less high values of the observed $\bar{y}_j$, (since Lasso penalty grows linearly with the coefficients, and Ridge quadratically), but does the opposite for small values of the observed (setting them to exactly zero).

4 Bayesian interpretation of Credibility

The previous section showed that applying Credibility to a GLM and estimating risks through a Penalized regression lead to similar results, averaging the pure GLM estimates with the grand average of the data (the complement of credibility). In this section, we will present a Bayesian interpretation of both the Credibility and Penalized regression approaches, effectively describing why they can be set to be equivalent.

We will provide examples that show how Penalized regression incorporates Credibility in the fitting procedure.

4.1 Penalized regression is a Credibility Framework

> *"The Bayesian and classical (frequentist, NdR) versions have a lot in common, but they have a philosophical difference in that in classical statistics parameters are constants, but for Bayesians they have distributions."*
>
> Gary Venter, Bayesian Regularization for Class Rates [Venter, 2018]

Before any GLM modeling, two assumptions need to be made. First, on the data generating distribution, a probabilistic description of the observed response must be defined. Then a link function must be chosen to describe the relationship between the linear predictor (made up by parameters and covariates) and the target.

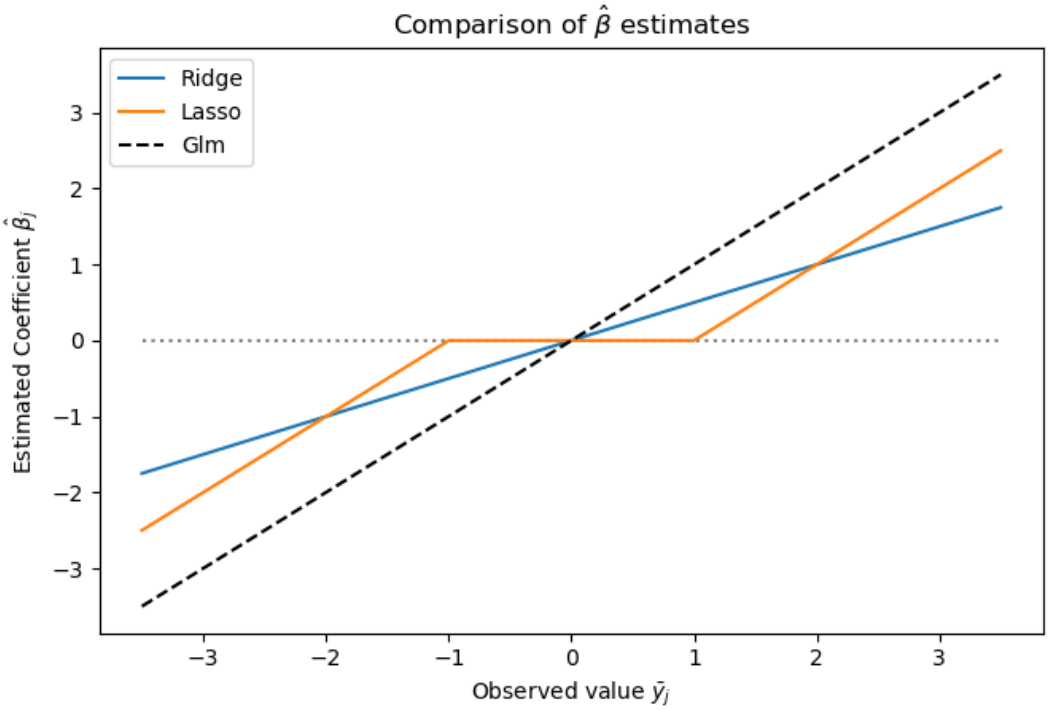

Figure 6: Plot of the correspondence between observed value $\bar{y}_j$ with estimates $\hat{\beta}_j$ for GLM and Penalized regression. For GLMs, the dashed line represents the identity function. For the Ridge, the relationship is linear, by a factor of $\frac{n_j}{n_j+2\lambda}$. For the Lasso, the relationship is piecewise linear, as described in 17.

Once these are set, the parameters are estimated by maximizing the likelihood over the data. This maximization will try to replicate the observed data as closely as possible, giving 100% credibility to the data.

From a statistical perspective, this happens because the frequentist approach to modeling assumes that there exists a fixed set of true coefficients: the observed values of the target are assumed to have been generated assuming these fixed coefficients and the hypothesis assumed above. Our best guess is thus the set of coefficients maximizing the probability of observing the actual values of the target, motivating a maximum likelihood approach.

In a Bayesian perspective, on top of the standard hypothesis done over the observations, a distributional assumption is made on the coefficients of the model themselves (called **the prior distribution**). This assumption describes our a priori knowledge and uncertainty over the value of the coefficients (hence the name), which are now random variables, and allows for the inclusion of some additional structure on the coefficients' estimates.

Buhlmann Credibility can be considered either from a Bayesian or a Frequentist point of view: as proven by Jewell [Jewell, 1974], for all data generating distributions used in GLMs, one can find an appropriate prior distribution such that the resulting Bayesian estimation

coincides with the estimator obtained via Buhlmann credibility.

The result by Jewell allows to put Buhlmann Credibility in the much more generic framework of Bayesian statistics. Under the Bayesian perspective, the connection with Penalized regression will be evident.

To see how this connection works in practice, it is helpful to apply Jewell's result to the workers' compensation use case. The goal is to show that a Bayesian model with proper priors does returns the exact same estimates than the Buhlmann Credibility formula.

As a reminder, the Buhlmann Credibility estimator is given by

$$\hat{\beta}_j = \frac{n_j}{n_j + k}\, \bar{y}_j \tag{18}$$

where $k = \frac{\sigma^2}{\tau^2}$ is the ratio between the within and between class variance (see Table 1 and Appendix).

Jewell's result shows that, for a given statistical hypothesis on the target, there exists a certain prior distribution for the coefficient β (called **conjugate**) that will return the Buhlmann estimator.

Statistical assumption	Conjugate distribution
$y \sim$ Gaussian	$\beta \sim$ Gaussian
$y \sim$ Poisson	$\beta \sim$ Gamma
$y \sim$ Gamma(α)	$\beta \sim$ Gamma
$y \sim$ Binomial	$\beta \sim$ Beta

Table 2: Table of couples Statistical assumptions / Conjugate Distribution for commonly used GLMs. Gamma(α) refers to a Gamma distribution with known parameter α.

The initial assumption of the workers' compensation use case was that the loss deviations y are normally distributed around the estimations, that is $y_i \sim \mathcal{N}(X_i\beta, \sigma^2)$. The conjugate of the normal distribution with known variance is the normal distribution itself. For this reason we'll suppose that *a priori* β follows itself a normal distribution with constant variance τ^2 and mean zero: $\beta \sim \mathcal{N}(0, \tau^2)$.

The choice of the normal **a priori** on $\beta \sim \mathcal{N}(0, \tau^2)$ can be motivated as well by pragmatic considerations: deviations of class code losses from the grand average should be centered around zero. Furthermore, large deviations from the grand average should be considered a priori as less likely than minor deviations. In this sense, **the priors allow to formalize common sense intuitions in a robust mathematical framework.**

To define a Bayesian estimator for this model it's necessary to derive the so-called *posterior* distribution, that updates through the likelihood (through the data) our *a priori* belief

about the unknown parameter (the β should not deviate too much from zero, the mean). This update (by Bayes' theorem) takes the form of

$$p_{posterior}(\beta|y,X) = \frac{1}{K} \times p(y|\hat{y}(X,\beta)) \times p_{prior}(\beta) \tag{19}$$

where K is a constant that doesn't depend on β.

Specifically, the Bayesian estimator can be given by the solution of [4]:

$$\begin{aligned} \hat{\beta} &= \underset{\beta}{\text{argmax}} p(y|\hat{y}(X),\beta) \times p_{prior}(\beta) \\ &= \underset{\beta}{\text{argmin}} \text{NLL}(y,X,\beta) - \log(p_{prior}(\beta)) \end{aligned} \tag{20}$$

The first summand is, for the workers' compensation use case, given by the same formula optimized by a GLM (2), that is

$$\text{NLL}(y,X,\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 \tag{21}$$

The second summand, since $\beta \sim \mathcal{N}(0,\tau^2)$ is equal to

$$-\log(p_{prior}(\beta)) = \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 + C \tag{22}$$

where C is a constant that does not depend on β , and can be removed from the optimization problem, which becomes

$$\begin{aligned} \hat{\beta} &= \underset{\beta}{\text{argmin}} \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 \\ &= \underset{\beta}{\text{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \frac{\sigma^2}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 \end{aligned} \tag{23}$$

Formula 23 shows as well that the Bayesian estimates' optimization formula is equal to that of Ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. Furthermore, by the Jewell theorem, we already know that the optimal solution is equal to the Buhlmann estimates, with $k = \frac{\tau^2}{\sigma^2}$.

The reasoning above proves that, under the hypothesis of the workers' compensation use case, **Ridge regression and Buhlmann credibility are equivalent** with $k = \frac{1}{\lambda}$.

The formulas highlight a strong connection between Bayesian and Penalized regression modeling:

[4] The posterior distribution in this case can be proven to be once again a Normal distribution with parameters obtained through a combination of the prior parameters and the maximum likelihood estimate; a typical estimator would then be the posterior mean. Outside of a limited number of convenient combinations of likelihoods and relative conjugate priors though, the posterior distribution might not be known analytically and inference becomes more cumbersome. For this reason the Maximum A Posteriori (MAP) estimator shown above can be a practical, efficient and reasonable choice. Similarly to GLMs and their *maximum likelihood* approach, we want to choose our estimate as the parameters that have the most support from the data, all the while respecting the additional structure imposed by the prior.

1. Bayesian modeling offers a generic framework to model the uncertainty of the estimation β when the number of observations is limited. This translates into a choice of a **prior** hypothesis on the coefficients β, that is, a certain distribution that the estimates β are assumed to follow.
2. The estimator of the Bayesian model can be found by maximizing the posteriori log-likelihood $p_{posterior}(\beta|y, X)$. When taking the logarithm its structure decomposes naturally in two terms: the log-likelihood (equal to the GLM formula) and the log-probability of the prior (which acts as a penalty).

In the example above, with a normal hypothesis on the target, and a normal prior, the choice of a prior distribution translates to a penalization term in the log-likelihood of the posterior.

This shows once again that Penalized regression can be seen under a Bayesian lens and vice versa; the Ridge penalty can be interpreted as the log-probability (up to some constants) of a normal prior distribution with zero mean and $1/\lambda$ common variance.

The Lasso is no different in principle, but with respect to a different prior distribution.

The Laplace distribution with mean $\mu = 0$ and scale γ , has density function

$$f_{\text{Laplace}(0,\gamma)}(x) = \frac{1}{2\gamma} \exp(-\frac{|x|}{\gamma}) \tag{24}$$

Under the assumptions that $\beta \sim Laplace(0, \gamma)$, then

$$-\log(p_{prior}(\beta)) = \frac{1}{2\gamma} \sum_{j=1}^{p} |\beta_j| + C \tag{25}$$

where C is a constant independent on β , which can be ignored when computing a MAP estimator. The log-prior is the same as the Lasso penalization with $\lambda = 1/2\gamma$.

The equivalence between Credibility and Penalized regression can be extended for all possible Penalized regressions, as proved in [Fry, 2015].

4.2 The credibility factor of Ridge and Lasso

The connection between Credibility and Penalized regression is motivated by

1. Buhlmann Credibility is the solution of a Bayesian model whose prior depends on the likelihood hypothesis for the target
2. Penalized regression is the solution of a Bayesian model with normal (for Ridge) or Laplace (for Lasso) priors
3. In the workers' compensation use case, Credibility and Ridge coincide (since they share the same prior hypothesis).

Since Ridge and Lasso can be seen as a Credibility methodology, it is helpful to compare them with other Credibility methods. A meaningful comparison can be given by showing how

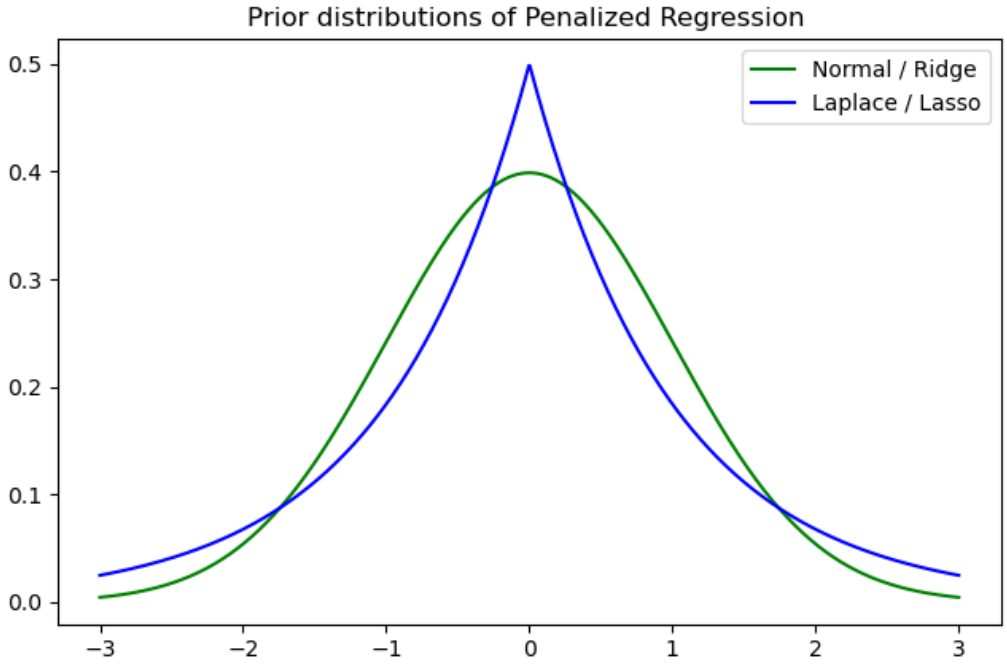

Figure 7: Comparison of densities of both the Normal / Ridge prior and the Laplace / Lasso prior.

the credibility estimates evolve as the underlying amount of exposures increases. Figure 8 displays the evolution of the estimates under the workers' compensation use case hypothesis. The value $\bar{y}$ for a specific code is kept fixed, and the underlying number of observations increased.

The behaviour of the Lasso is different from the others, as it exhibits a minimum amount of observations (here $\lambda_{\text{Lasso}}/\bar{y}$) for which the recorded experience does not influence the estimates $\hat{\beta}$ and the predictions are equal to the complement of credibility. This can be seen as equivalent to considering a specific level of significance to include a variable in a GLM model. The choice of including or excluding a specific level in a GLM model may be decided upon the result of a statistic (p-value) which will depend, amongst others, on the number of observations. If the statistics is below a certain threshold (e.g. 5% for p-values), then the factor will be included in the modeling, giving an underlying 100% credibility. The Lasso regression leads to similar results, but allows to interpolate between 0 and full credibility instead of a binary split (a yes or no decision).

When the number of observations is above the Lasso's threshold, experience gains weight onto the final estimate much faster than in the Ridge/Buhlmann estimates.

The visualization above shows, in a univariate example, how the signal is interpolated from the complement of credibility to the observed data by Credibility and Penalized regression.

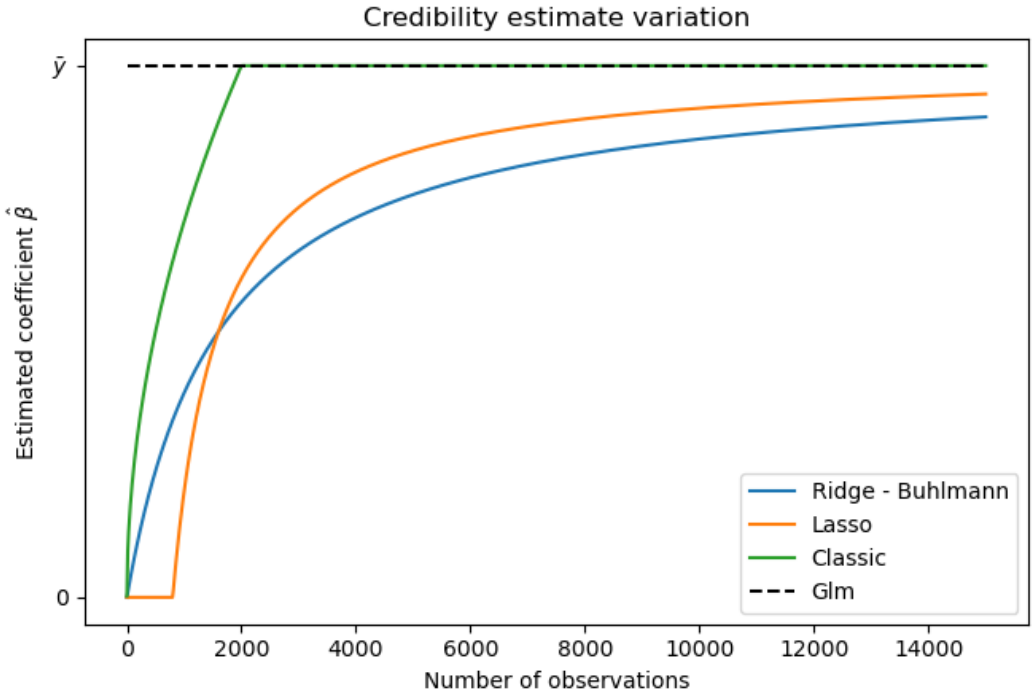

Figure 8: Plot of the estimates $\hat{\beta}$ with $\bar{y} = 5$ and the number of observations varying. Parameters for each model were: Classical credibility $N_{full} = 2000$, Buhlmann $k = 1600$, Ridge $\lambda_{\text{Ridge}} = 1/k$, Lasso $\lambda_{\text{Lasso}} = 4000$. The formulas used to compute the estimates for the use case can be found in the relative section in the paper. The parameters were chosen arbitrarily to best display the differences of trend. Depending on the parameters, the curve would more or less be similar.

It also shows how both frameworks can be derived from a Bayesian prior hypothesis on the coefficients distribution, demonstrating how the Penalized regression approach extends the GLMs by integrating credibility in a multivariate context.

4.3 Lasso fits signal up to a threshold

In an univariate setting it is possible to show how Penalized regression blends Credibility via explicit formulas. In a multivariate setting, while the Bayesian interpretation holds true, explicit weighing formulas are not available. This is similar to GLM estimates, where an explicit formula for the optimal solution is not available for most of the statistical assumptions, and coefficients are computed via iterative algorithms.

It is however possible, by looking at the structure of the gradient of the GLM formula, to see how the estimates adapt to the data with or without credibility assumptions.

This is because the gradient of a GLM (with canonical link) can be seen as the difference, for each level of a variable, of the total observations and the total estimates included in the

GLM model [Ohlsson and Johansson, 2010]:

$$\nabla NLL(y, X, \beta)_j = \frac{\partial NLL(y, X, \beta)}{\partial \beta_j} = \sum_{i=1}^{n} (\mu_i - y_i) x_{ij} \tag{26}$$

where $\mu_i = \text{Link}^{-1}(X_i\beta)$ is the prediction for a given β. For all coordinates j whose covariates X_j have binary encoding ($x_{ij} = 1$ if $i \in J$, else 0) , the gradient is the difference of the total observations $\sum_{i \in J} y_i$ and the total predictions $\sum_{i \in J} \mu_i$. At the GLM solution $\hat{\beta}$, the gradient is null. In particular, for all levels j we have

$$\sum_{i \in J} (\mu_i - y_i) = 0 \tag{27}$$

This is consistent with the results described in Section 2. GLMs give 100% credibility to the data, and average predictions will coincide with the average of the observations for each level j included in the model (regardless of the number of exposures).

If one wants to leave some room for the complement of credibility and shrink the estimates (as seen in the previous chapter the two concepts are tightly linked), instead of a strict matching of observations and predictions we could consider adding some "slack", so that the model can match the data only up to a certain threshold. For example $|\sum_{i \in J}(y_i \mu_i)| \leq \varepsilon$ for a certain constant ε. This is exactly how the Lasso guarantees optimality.

In the Appendix, we prove that the optimality condition for

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \text{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{28}$$

is equivalent to (where $\mu_i = X_i\beta$)

$$\begin{cases} |\sum_{i \in J}(\mu_i - y_i)| \leq \lambda & \leftrightarrow \ \hat{\beta}_j = 0 \\ \sum_{i \in J}(\mu_i - y_i) = \lambda \, \text{sign}(\hat{\beta}_i) & \leftrightarrow \ \hat{\beta}_j > 0 \\ \sum_{i \in J}(\mu_i - y_i) = -\lambda \, \text{sign}(\hat{\beta}_i) & \leftrightarrow \ \hat{\beta}_j \leq 0 \end{cases} \tag{29}$$

Any final estimate of the Lasso, regardless of fitting procedure, will satisfy the above optimality condition system 29. We can clearly see the "slack" discussed above when matching observed and predicted averages: a coefficient will be deemed non-relevant (and thus set to zero) if its contribution to the likelihood via the gradient falls below the threshold λ; when the effect is instead considered relevant the coefficient will moved to capture it but just until the error tolerance threshold λ is hit, rather than going all the way like it would on a GLM.

Once again the parallel with Credibility is clear, but note that this time the analysis is completely general and potentially multivariate, providing an intuitive yet sound credibility result outside of the univariate setting.

5 Conclusions

Albeit misrepresented in the actuarial literature, Penalized regression methods are effective tools for actuarial modeling that belongs to the toolkit of all pricing expert.

Throughout this whitepaper a detailed comparison of GLM, Credibility and Penalized regression was explored, exemplified through a univariate modeling scenario. We've seen that in order to blend desirable credibility assumptions, a modeling framework cannot only rely on optimizing likelihood of data, but needs additional structure that will allow to "shrink" the estimates, moving them toward some complement of credibility. Penalized regression methods such as Ridge or Lasso satisfy this condition. A strong link between Penalized regression and Buhlmann credibility in particular has been established via their Bayesian interpretation, which allowed us to describe how Penalized regression also naturally extends the univariate credibility approach to a multivariate, real-world setting.

We hope this manuscript helped shed some light on some lesser known results about Penalized regression both from a theoretical and a practical perspective, highlighting them as valid alternative to ad-hoc adjustments and allowing a more widespread adoption of such techniques.

A Optimality condition of the Lasso solution

Lasso regression is known for its ability to achieve sparse solutions, where some of its estimated coefficients will be exactly equal to zero. The mathematical reasons on why sparsity happens are however either not discussed at all or are hidden in very technical explanations that require a certain level of familiarity with advanced convex optimization concepts. This need not be the case. Furthermore, as we have previously established a strong connection between Lasso regression and Bayesian statistics, understanding the impact of the λ parameter on the coefficients allows the practitioner to better understand the output of Lasso regression.

First, by considering a simple example, we will show how sparsity naturally arises from the non-differentiability of the absolute value $|\beta|$ contained in the Lasso penalty. Then, we will introduce the least amount of concepts from convex optimization necessary to provide the optimality guarantees for the Lasso problem.

A.1 Simplified proof of the Lasso problem

Consider the simplest possible Lasso regression expressed as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}(y - \beta)^2 + \lambda |\beta| \tag{30}$$

where we have a one dimensional parameter β aiming to approximate a single observation y.

Computing the solution by setting the gradient to zero is not possible as the absolute value is non-differentiable at zero. Instead, one can write the function as a piecewise parabolic

function:

$$\frac{1}{2}(y-\beta)^2 + \lambda|\beta| = \begin{cases} \frac{1}{2}(y-\beta)^2 + \lambda\beta \text{ if } \beta \geq 0 \\ \frac{1}{2}(y-\beta)^2 - \lambda\beta \text{ if } \beta < 0 \end{cases} \tag{31}$$

For every value of y and λ, this function is convex: this means that there is one and only one global minimum. Figure 9 highlights all the possible cases, depending on the value of y. It is clear that the optimum $\hat{\beta}$ of the piecewise function either lies at the global minimum of the parabola, or at $\beta = 0$, where the two parabolas intersect.

We can then deduce that if the minimum of the function lies in the right interval $\beta > 0$, then the optimum will be $\hat{\beta} = y - \lambda$ (which is the global minimum of $\frac{1}{2}(y-\beta)^2 + \lambda\beta$). Equivalently, if it lies on the left part of the parabola ($\beta < 0$), then the optimum will be $\hat{\beta} = y + \lambda$. Combining the inequalities, we proved that the optimum of 30 is

$$\hat{\beta} = \begin{cases} y - \lambda \text{ if } y > \lambda \\ y + \lambda \text{ if } y < -\lambda \\ 0 \qquad \text{otherwise} \end{cases} \tag{32}$$

The plot of the optimum $\hat{\beta}$ as a function of the value of y is provided in Figure 9

The example above highlights how the discontinuity in the Lasso penalty $|\beta|$ achieves a sparse solution. The Lasso estimate will be exactly equal to zero in correspondence of values of y smaller than λ in absolute value and will be equal to the value or y shrunken by a constant of size λ otherwise. It is thanks to its non-differentiable nature that the Lasso problem allows us to obtain sparse solutions.

We will now review how to demonstrate the Lasso solution in a general way using simple tools from convex optimisation.

A.2 General proof of the Lasso problem

When required to find a minimum of a function analytically, the practitioner would naturally compute the gradient of that function and find the parameter β that solves the equality of the gradient to zero.

In the case of the Lasso problem, we saw how this is not possible due to the non-differentiability of the penalty at $\beta = 0$. As a matter of fact, it is still possible to compute a minimum of the Lasso regression by setting the gradient to zero: we just need to generalize the definition of the gradient.

The gradient is defined as the slope of the tangent to the graph of a function. When there is a discontinuity, there may be multiple slopes that are tangent to the graph of the function. The gradient loses its uniqueness property and it is hence said that the function is "not differentiable".

A generalization of the gradient, the subgradient, is defined as the *set* of possible slopes that are tangent to a graph. Formally, given a convex function $f \in \mathbb{R}^p$, the subgradient is

$$\partial f(\beta_0) = \{u \in \mathbb{R}^p \mid f(\beta) - f(\beta_0) \geq u(\beta - \beta_0)\} \tag{33}$$

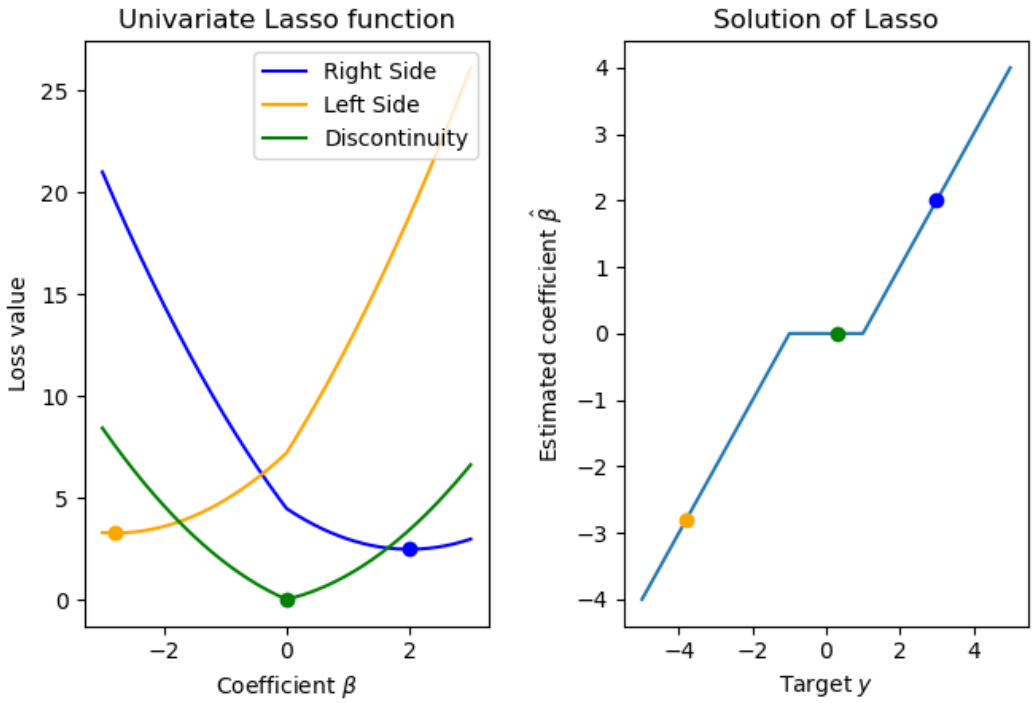

Figure 9: The left graph represents the plot of three different Lasso formulas 31 with different fixed values of y and $\lambda = 1$. Right side represents $\beta \to 1/2(3-\beta)^2 + |\beta|$, Left side is $\beta \to 1/2(-3.8-\beta)^2 + |\beta|$ and Discontinuity is $\beta \to 1/2(-3.8-\beta)^2 + |\beta|$. The dots represent the optimum $\hat{\beta}$ for each function. The right graph represent the evolution of the optimum $\hat{\beta}$ as a function of the values y. The colored points represent the couples $(y, \hat{\beta})$ of the three functions of the left sided graph. The function is also called soft-thresholding in the literature.

In the case of the absolute value function, since the subgradient is equal to the gradient when the function is differentiable, for all values strictly different than zero the gradient will be equal to the sign function *i.e.* 1 for all positive values and -1 for all negative values. In the discontinuity point at 0, it will take all possible values between -1 and 1.

$$\partial|\beta| = \begin{cases} -1 & \text{if } \beta < 0 \\ (-1, 1] & \text{if } \beta = 0 \\ 1 & \text{if } \beta > 0 \end{cases} \tag{34}$$

Generalizing the gradient to the subgradient allows us to compute the minimum for the Lasso. It is established that if f is differentiable and convex, then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} f(\beta) \quad \Longleftrightarrow \quad 0 = \nabla f(\hat{\beta}) \tag{35}$$

If f is not differentiable (but still convex), then the optimality condition becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} f(\beta) \quad \Longleftrightarrow \quad 0 \in \partial f(\hat{\beta}) \tag{36}$$

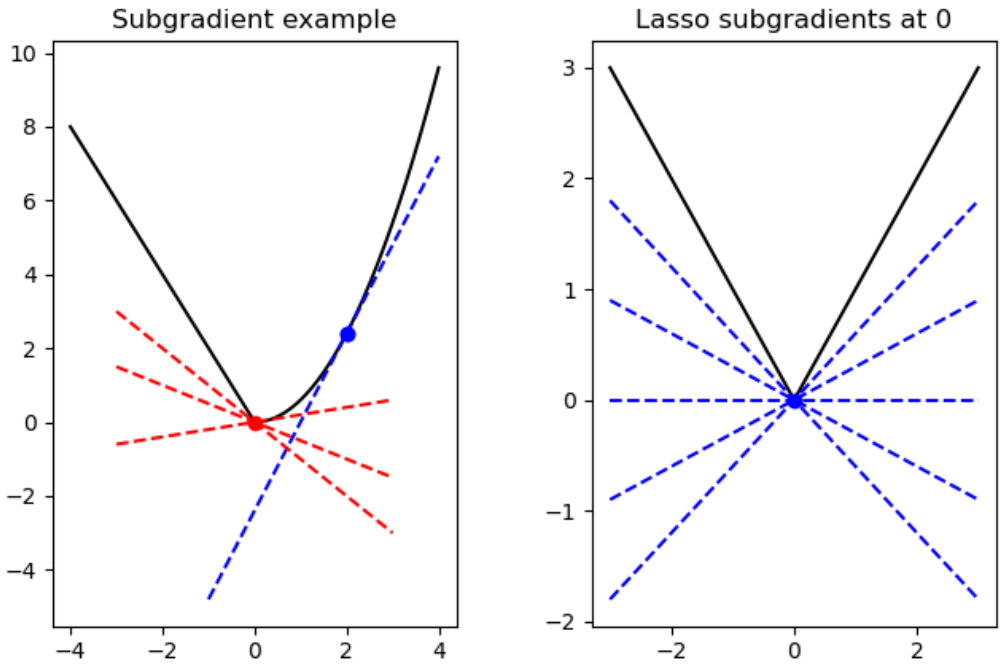

Figure 10: The left figure illustrates the subgradient for a piecewise function. In blue, the subgradient at value 2 is displayed. As the function is differentiable, there exists only one subgradient, and it is equal to the gradient. Since the function at 0 is not differentiable, there is more than one tangent line to the graph. The subgradient is drawn in red. The right figure represents some possible tangent lines for the Lasso function $\beta \to |\beta|$. This provides a visual intuition of why $\partial|\beta|\Big|_{\beta=0} = [-1, 1]$.

by the subgradient optimality condition (see [Boyd et al., 2004]).

Since the subgradient of a sum is the sum of the subgradients and the subgradient of a differentiable function is the gradient, in the case of the (simplified) Lasso regression we can write the condition 36 as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}(y-\beta)^2 + \lambda|\beta| \quad \iff \quad 0 \in (\hat{\beta} - y) + \lambda\partial|\hat{\beta}| \tag{37}$$

where $(\hat{\beta} - y)$ is the derivative of $\frac{1}{2}(y-\beta)^2$.

Since the subgradient of $\beta \to |\beta|$ is given by 34 , we can prove the optimality of 32: if the optimum is $\hat{\beta} = 0$, then there exists a number $|u| \leq 1$ such that $0 = -y + \lambda u$. This happens only when $|y| \leq \lambda$. For the other cases ($\hat{\beta} > 0$, $\hat{\beta} < 0$) the Lasso penalty is differentiable and by standard arguments one can verify the optimality of 32.

The subgradient definition provides the optimality conditions of the Lasso regression in all its generality, both in a multivariate setting and using a generic negative log-likelihood. It

provides as well the tools to understand the optimality conditions 29. To see this, consider the general Lasso problem.

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \text{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{38}$$

To compute the optimal solution, first the **sub**gradient of the negative log-likelihood is required

$$\frac{\partial}{\partial \beta_j}[\text{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|] = \nabla NLL(y, X, \beta)_j + \lambda \partial |\beta_j| \tag{39}$$

At optimum $\hat{\beta}$ zero must belong to the subgradient, which means that depending on the sign of $\hat{\beta}_j$ we have that

$$\begin{cases} |\nabla NLL(y, X, \beta)_j| \leq \lambda & \leftrightarrow \ \hat{\beta}_j = 0 \\ \nabla NLL(y, X, \beta)_j = \lambda \, \text{sign}(\hat{\beta}_i) & \leftrightarrow \ \hat{\beta}_j > 0 \\ \nabla NLL(y, X, \beta)_j = -\lambda \, \text{sign}(\hat{\beta}_i) & \leftrightarrow \ \hat{\beta}_j \leq 0 \end{cases} \tag{40}$$

which proves 29. The results of this section combined provide as well all required tools to prove the optimal solution formula 17 for the workers' compensation use case.

B Classical and Buhlmann Credibility

The whitepaper often refers to the Classical and Buhlmann Credibility methodologies without explaining the rationale behind these techniques and the underlying hypothesis which lead to different computations of the credibility factor Z. This section aims to provide a simplified summary of the key characteristics of Buhlmann and Classical credibility and their comparison.

B.1 Classical Credibility

The objective of Classical Credibility is to build estimates that are robust to the impact of random fluctuations in the data.

First, this requires to compute the minimum amount of data N_{full} to be confident by a probability P that the future observations will be within a threshold K of the estimates.

In the case of the workers' compensation example, we may observe a deviation of $\bar{y}_j = 100$ for a specific class j. N_{full} can be the amount of data required so that the future estimates will fall in a threshold of $K = 5\%$ (that is, within the [95, 105] interval) by a probability $P = 95\%$.

By knowing the nature of the risk, whether it be Pure Premium or Claim Severity for instance, and using the Central Limit theorem, the modeler can resort to an explicit formula to compute N_{full} as a function of K and P [Tse, 2009].

Assume for instance we're measuring Gaussian realizations x with mean μ and variance σ^2; denote the observed average over n observations $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$. We obtain full credibility if

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\bar{x}-\mu}{\mu}\right| < K\right) \geq P \\
&\mathbb{P}\left(|\bar{x}-\mu| < K\mu\right) \geq P \\
&\mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\right| < \frac{\sqrt{n}K\mu}{\sigma}\right) \geq P \\
&\mathbb{P}\left(-\frac{\sqrt{n}K\mu}{\sigma} < \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} < \frac{\sqrt{n}K\mu}{\sigma}\right) \geq P \\
&\mathbb{P}\left(-\frac{\sqrt{n}K\mu}{\sigma} < \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} < \frac{\sqrt{n}K\mu}{\sigma}\right) \geq P \\
&z_{\frac{1+P}{2}} \leq \frac{\sqrt{n}K\mu}{\sigma} \Rightarrow n \geq \left(\frac{z_{\frac{1+P}{2}}\ \sigma}{K\mu}\right)^2 = N_{full}
\end{aligned} \tag{41}$$

where we used the fact that $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$ and thus we can recover easily z_q, the q^{th} quantile of a standard normal. Since μ and σ are not generally known one can susbstitute empirical estimates for them in the above expression.

If the amount of exposures j is not high enough to reach full credibility, one can prove that, under several assumptions, that the credibility factor Z_j can be computed via the square root formula $\sqrt{\frac{n_j}{N_{full}}}$ described above.

For a more comprehensive and detailed review of Classical Credibility methodology, we refer the interested reader to [Tse, 2009]

B.2 Buhlmann Credibility

Buhlmann credibility provides a more statistically sound framework to blend experience and complement of credibility, by using a different definition to what a credible estimate is. We'll present it here as deriving from a hierarchical model using a Bayesian perspective but, as we've seen in the main text, multiple interpretations exist, for example as a random effect model.

The assumption of Buhlmann Credibility is that the observation of loss for class j is the outcome of a random variable with distribution p, which depends on a parameter j. We'll assume that each j is different but they're all related via a hierarchical prior distribution q. Under these hypotheses, the data generating process can be expressed as : $\beta\_j \sim q$; $Y_j \sim p_{\beta_j}$.

It is possible to prove that the variance of the observation Y_j can be exactly decomposed as the sum of the two steps, *i.e.* the sum of the variance within the class j and the variance between classes (coming from the prior q on j).

If we want to find the best estimator for Y_j in a Mean Squared Error sense (or, equivalently, if we want to use the posterior mean as our point-estimate), the Buhlmann Credibility estimator comes out, as a linear combination of the observations with the grand average $\bar{Y}$ As proved in [Tse, 2009], the Buhlmann Credibility factor Z_j for class j is given by

$$Z_j = \frac{n_j}{n_j + k} \tag{42}$$

Where the parameter k is the ratio between the variance within a class (often called the Expected Process Variance - EPV) and variance across the classes (also called the Variance of Hypothetical Means - VHM). Since these quantities are in general not known, one can resort to estimating them, most commonly via unbiased estimators (making this an empirical Bayes estimator). One can see how the higher the k, the higher the amount of exposure required to reach a given amount of credibility.

B.3 Comparison and considerations

Both Classical and Buhlmann Credibility provide formulas to compute the credibility factor Z as a function of exposure and some additional parameters.

In Classical Credibility the main parameters P and K that determine the strictness of the credibility constraint are to be chosen by the modeler. The computation of the exposure needed for full credibility N_{full} may require, depending on the nature of the risk, other quantities that can be robustly and easily estimated from the data.

The simplicity of the formulas used by Classical Credibility have contributed to the success of such methodology. It must be noted however that:

> *While the classical credibility theory addresses the important problem of combining claim experience and prior information to update the prediction for loss, it does not provide a very satisfactory solution. The method is based on arbitrary selection of the coverage probability and the accuracy parameter. Furthermore, for tractability some restrictive assumptions about the loss distribution have to be imposed.*
>
> Tse, Yiu-Kuen. *Nonlife actuarial models: theory, methods and evaluation* [Tse, 2009]

Buhlmann Credibility provides a more robust statistical framework for the definition of credibility but the estimation of its parameters can be convoluted, especially for the unobserved variance between classes. Furthermore, the higher the number of distinct classes, the less accurate the estimates of the variance within each single class can be. In a sense, the variance estimation formulas can be affected by the same data credibility problems that the Buhlmann approach tries to tackle.

References

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Casualty Actuarial Society. *Foundations of Casualty Actuarial Science, Fourth Edition*. Casualty Actuarial Society (CAS), fourth edition, 2001.

Taylor Fry. A discussion on credibility and penalised regression, with implications for actuarial work. *Actuaries Institute*, 2015.

Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller. *Generalized linear models for insurance rating*, volume 5. Casualty Actuarial Society, 2016.

Trevor Hastie and Junyang Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

William S Jewell. Credible means are exact bayesian for exponential families. *ASTIN Bulletin: The Journal of the IAA*, 8(1):77–90, 1974.

Fred Klinker. Generalized linear mixed models for ratemaking: a means of introducing credibility into a generalized linear model setting. In *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2*, 2010.

NAIC. *Whitepaper of Regulatory Review of Predictive Models (NAIC)*. National Association of Insurance Commissioners, 2020.

Esbjörn Ohlsson and Björn Johansson. *Non-life insurance pricing with generalized linear models*, volume 174. Springer, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Yiu-Kuen Tse. *Nonlife actuarial models: theory, methods and evaluation*. Cambridge University Press, 2009.

Gary Venter. Bayesian regularization for class rates. ., 2018.