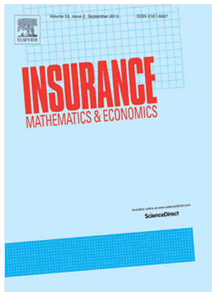
Accepted Manuscript



Bayesian credibility for GLMs

Oscar Alberto Quijano Xacur, José Garrido

PII: S0167-6687(17)30413-4
DOI: https://doi.org/10.1016/j.insmatheco.2018.05.001
Reference: INSUMA 2463

To appear in: *Insurance: Mathematics and Economics*

Received date : September 2017
Revised date : January 2018
Accepted date : 9 May 2018

Please cite this article as: Xacur O.A.Q., Garrido J., Bayesian credibility for GLMs. *Insurance: Mathematics and Economics* (2018), https://doi.org/10.1016/j.insmatheco.2018.05.001

Bayesian Credibility for GLMs

Oscar Alberto Quijano Xacur* and José Garrido
Concordia University, Montreal, Canada

January 20, 2018

Dedicated to the memory of Bent Jørgensen;
a true scholar, a gentlemen and a great person.

Abstract

We revisit the classical credibility results of Jewell [8] and Bühlmann [2] to obtain credibility premiums for a GLM using a modern Bayesian approach. Here the prior distribution can be chosen without restrictions to be conjugate to the response distribution. It can even come from out–of–sample information if the actuary prefers.

Then we use the relative entropy between the "true" and the estimated models as a loss function, without restricting credibility premiums to be linear. A numerical illustration on real data shows the feasibility of the approach, now that computing power is cheap, and simulations software readily available.

1 Introduction

The well known classical result from Jewell [8] gives exact linear credibility estimators for the exponential family. More precisely, it applies to random variables Y with distribution in some exponential dispersion family (EDF), i.e. with density

*Corresponding author: oscar.quijano@use.startmail.com. The authors gratefully acknowledge the partial financial support of NSERC grant 36860-2017.

$$f(y|\theta,\phi) = a(y,\phi)\exp\left(\frac{1}{\phi}\left\{y\theta - \kappa(\theta)\right\}\right), \qquad \theta \in \Theta, \phi \in \Phi. \tag{1}$$

Assuming that ϕ is known, Jewell uses the following prior density on θ:

$$\pi_{n_0,x_0}(\theta) \propto \exp\big(n_0[x_0\theta - \kappa(\theta)]\big), \tag{2}$$

for some hyper–parameters $n_0 > 0$ and x_0 in the support of Y. For a conditionally i.i.d. sample $y_1,\ldots,y_n$ of Y, given θ, Jewell showed that the marginal mean of Y given this sample is

$$\begin{aligned}\mathbb{E}[Y|y_1\ldots y_n] &= \frac{\phi n_0}{\phi n_0 + n}x_0 + \frac{n}{\phi n_0 + n}\bar{y} \\ &= (1-z)x_0 + z\bar{y},\end{aligned} \tag{3}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $z = \frac{n}{\phi n_0 + n}$ is known as the credibility factor. Note that (3) is the Bühlmann [2] Bayesian point estimator of the mean of Y that minimizes the expected square loss error and it is known as the linear credibility estimator.

The purpose of this article is to present a methodology for obtaining credibility estimators that generalize these ideas and allow them to be applicable to Generalized Linear Models (GLMs).

Hachemeister [7] and De Vylder [5] give classical credibility results for regression models. The main idea in these articles is similar to Bühlman's: they first impose linearity on the regression covariates and then find the optimal linear parameter estimators by minimizing a distance or error function.

Nelder and Verall [17] and Ohlsson [18] propose linear credibility estimators for GLMs. Although different in substance, these two articles have in common that they are both likelihood based and both rely on random effects in order to obtain credibility estimates. The resulting Generalized Linear Mixed Models (GLMMs) credibility premiums remain essentially linear, which differs fundamentally from the method proposed here.

In this article we use a modern Bayesian approach to obtain credibility estimators. We focus on extending Jewell's results to Generalized Linear Models (GLMs). Our estimator is not linear as we do not impose linearity constraints, nor force the prior to be conjugate. It also offers the advantage of considering the uncertainty on the dispersion parameter.

The exposition is divided as follows. In Section 2 we review the essential elements needed to introduce our estimator. Section 3 presents entropic estimators and some properties of the unit deviance that are important for the rest of the article. In Section 4 we discuss the un–feasibility of exact linear credibility for GLMs. Section 5 proves that it is possible to obtain entropic estimators for GLMs, which is our proposed credibility approach. Moreover, Proposition 5.1 suggests an algorithm to obtain the estimators.

Finally, in Section 6 we show that our method is applicable to real life datasets by fitting the model to a publicly available dataset. A commented version of the R code used for this can be found at `https://oquijano.net/articles/bayesian_credibility/`.

2 Preliminaries

2.1 Exponential Dispersion Families

In (1) θ and Θ are called the canonical parameter and canonical space, respectively and ϕ is known as the dispersion parameter. A neat property of reproductive exponential families is that for $\theta \in \operatorname{int}(\Theta)$ (here int stands for interior),

$$\mathbb{E}[Y] = \dot{\kappa}(\theta) \qquad \text{and} \qquad \mathbb{V}[Y] = \phi\ddot{\kappa}(\theta), \tag{4}$$

where $\dot{\kappa} = \kappa'$ and $\ddot{\kappa} = \dot{\kappa}'$. This motivates the following definitions.

Definition 2.1. *Given an exponential dispersion family, the mean domain of the family is defined as*

$$\Omega = \{\mu = \dot{\kappa}(\theta) : \theta \in int(\Theta)\}.$$

Another important property is that the support of the distribution only depends on ϕ (and not on θ). For a given family, let C_ϕ be the convex support of any member of the family with dispersion parameter ϕ. We define the convex support of the family as

$$C_\Phi = \bigcup_{\phi\in\Phi} C_\phi.$$

Definition 2.2. *The unit deviance function of an exponential dispersion family is defined as* $d : C_\Phi \times \Omega \to [0, \infty)$ *with*

$$d(y,\mu) = 2\left[\sup_{\theta\in\Theta}\{\theta y - \kappa(\theta)\} - y\dot{\kappa}^{-1}(\mu) + \kappa\big(\dot{\kappa}^{-1}(\mu)\big)\right]. \tag{5}$$

The unit deviance function plays an important role in the theory of GLMs. The model assessment of a GLM is through hypothesis tests that are based on the asymptotic behaviour of the unit deviance function. It also allows to re-parametrize (1) as

$$f(y|\mu,\phi) = c(y,\phi)\exp\left(-\frac{1}{2\phi}d(y,\mu)\right). \tag{6}$$

This is known as the mean–value parameterization and it is the one used in this article.

When the canonical space Θ is open, it is said that the EDF is regular. In this case $C_\Phi = \Omega$ and (5) is equivalent to

$$d\,(y,\mu) = 2\left[y\{\dot\kappa^{-1}(y) - \dot\kappa^{-1}(\mu)\} - \kappa\big(\dot\kappa^{-1}(y)\big) + \kappa\big(\dot\kappa^{-1}(\mu)\big)\right]. \tag{7}$$

In the rest of the article we work with regular EDFs. Notice that all the usual distributions used for GLMs are regular (e.g. all the distributions in the Tweedie family are regular).

There is a useful property of reproductive exponential dispersion families that allows for data aggregation. Jørgensen's notation (from [11]) is very convenient to express this property: given a fixed exponential family, if Y has mean μ and density given by (6), we say that it is $ED(\mu,\phi/w)$ distributed. The property is then as follows: if $Y_1, Y_2, \cdots, Y_n$ are independent, and $Y_i \sim ED(\mu,\phi/w_i)$, then

$$\bar Y = \frac{w_1Y_1 + \cdots + w_nY_n}{w_+} \sim ED(\mu,\phi/w_+), \qquad w_+ = \sum_{i=1}^{n} w_i. \tag{8}$$

2.1.1 A Note on Aggregating Discrete Exponential Dispersion Models

There are two usual parametrizations of exponential dispersion families. (1) gives the density of *reproductive* EDFs and it is used for continuous distributions. Discrete distributions are usually parametrized as *additive* EDFs, whose densities have the form

$$f(y|\theta,\phi) = a(y,\phi)\exp\left(y\theta - \frac{1}{\phi}\kappa(\theta)\right), \qquad \theta\in\Theta, \phi\in\Phi. \tag{9}$$

Both parametrizations are defined and discussed in Jørgensen [11] and Jørgensen [10]. GLMs assume the reproductive parameterization (see Nelder and Wedderburn [16]). Now, for many discrete EDFs, the dispersion parameter has a known value. Specifically, for the Poisson, Bernoulli and negative binomial distributions $\phi = 1$. This makes (1) and (9) the same parametrization and allows such distributions to enter the GLMs framework. Nevertheless, it is important to be aware that for the discrete case, one cannot aggregate data using (8). The properties of the Poisson distribution allow to use an offset for data aggregation (see for example Section 9.5 of Kaas, Goovaerts et al. [12]). Other discrete distributions can be aggregated using quasi-likelihood.

2.2 Relative Entropy

Let m_i be probability measures with $dm_i(\boldsymbol{y}) = f_i(\boldsymbol{y})ds(\boldsymbol{y})$ for some density functions f_1, f_2 and some probability measure s, with $m_1 \equiv m_2 \equiv s$ (that is m_1, m_2 and s are absolutely continuous with respect to each other).

Definition 2.3. *The relative entropy of m_2 from m_1 is defined as*

$$D(m_1 \parallel m_2) = \mathbb{E}_{m_1}\left[\log\left(\frac{f_1(\boldsymbol{Y})}{f_2(\boldsymbol{Y})}\right)\right] = \int \log\left(\frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})}\right) dm_1(\boldsymbol{y}).$$

This definition was introduced by Kullback and Leibler in [14]. $D(\cdot \parallel \cdot)$ is often called the Kullback–Leibler divergence, nevertheless we prefer the term relative entropy since what they called divergence between m_1 and m_2 was $D(m_1 \parallel m_2) + D(m_2 \parallel m_1)$.

The relative entropy is a measure of information. As such, it satisfies the invariance property. Intuitively this means that bijective transformations do not increase or decrease the information. The following paragraph expresses this formally.

Let $(\Omega_1, \mathcal{F}, m_i)$ and $(\Omega_2, \mathcal{G}, \nu_i)$, for $i = 1, 2$, be probability spaces and $T : \Omega_1 \to \Omega_2$ a measurable transformation such that $\nu_i(G) = m_i(T^{-1}(G))$, for $G \in \mathcal{G}$. Define also $\gamma(G) = s(T^{-1}(G))$. Since $m_1 \equiv m_2 \equiv s$, then $\nu_1 \equiv \nu_2 \equiv \gamma$. This implies, by Radon–Nykodim's theorem that there exist g_1 and g_2 such that

$$\nu_i(G) = \int_G g_i(\boldsymbol{y})d\gamma(\boldsymbol{y}), \qquad G \in \mathcal{G}.$$

With these definitions in mind, the following theorem asserts the invariance property of the relative entropy. Its proof can be found in Chapter 2 of Kullback [13].

Theorem 2.1. *$D(m_1 \parallel m_2) = D(\nu_1 \parallel \nu_2)$ if and only if T is a bijective transformation.*

2.3 GLMs

In a GLM the response variable is assumed to follow a EDF with density

$$f(y|\theta, \phi) = a(y, \phi) \exp\left(\frac{w}{\phi}\{y\theta - \kappa(\theta)\}\right), \tag{10}$$

Note that ϕ in (1) corresponds to ϕ/w in (10) which implies that the mean and variance can be expressed as $\mu = \kappa'(\theta)$ and $\sigma^2 = \phi\kappa''(\theta)/w$, respectively. Here $w \geq 0$ is know as the weight. In applications w is known usually and ϕ needs to be estimated. It is further assumed that there is a vector of explanatory variables, also known as covariates, $\boldsymbol{x} = (x_1 \cdots x_p)^T$, a vector of coefficients $\boldsymbol{\beta} = (\beta_0\ \beta_1 \cdots \beta_p)^T$ and a function g known as the link function such that

$$g(\mu) = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p. \tag{11}$$

It is useful for further developments to express the canonical parameter θ in terms of the coefficients. Since $\mu = \kappa'(\theta) \equiv \dot{\kappa}(\theta)$ then:

$$\begin{aligned} (g \circ \dot{\kappa})(\theta) &= \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p \\ \theta &= (g \circ \dot{\kappa})^{-1}(\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p). \end{aligned} \tag{12}$$

The population can be divided into different classes according to the values of the explanatory variables. Thus, given a sample, we can group together all the observations that share the same values of the explanatory variables and aggregate them using (8). It is important to mention that with this grouping there is no loss of information for estimating the mean since $\bar{Y}$ is a sufficient statistic for θ (but not for ϕ, thus some information is lost for the estimation of ϕ).

Possibly after aggregating, let m be the number of classes and $\boldsymbol{\theta} \in \Theta^m$, where $\Theta^m = \left\{(\theta_1 \cdots \theta_m)^T : \theta_1, \ldots, \theta_m \in \Theta\right\}$. The density of the sample can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \phi) = A(\boldsymbol{y}, \phi) \exp\left(\frac{\boldsymbol{y}^T W \boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta})}{\phi}\right), \qquad \boldsymbol{y} \in \mathbb{R}^m, \tag{13}$$

where $\boldsymbol{\kappa}(\boldsymbol{\theta}) = \left(\kappa(\theta_1) \cdots \kappa(\theta_m)\right)^T$, $W = \text{diag}(w_1, \cdots, w_m)$, with w_i being the sum of all the weights in the i-th class, $\mathbf{1} = (1 \cdots 1)^T$ and

$A(\boldsymbol{y}, \phi) = \prod_{i=1}^{m} \left(a(y_i, \frac{w_i}{\phi}) \right)$. In order to express $\boldsymbol{\theta}$ in terms of $\boldsymbol{\beta}$, we define the following maps

$$\boldsymbol{\mu} = \dot{\boldsymbol{\kappa}}(\boldsymbol{\theta}) = \begin{pmatrix} \dot{\kappa}(\theta_1) \\ \vdots \\ \dot{\kappa}(\theta_m) \end{pmatrix}, \quad G(\boldsymbol{\mu}) = G \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_m) \end{pmatrix},$$

and the design matrix

$$X = \begin{pmatrix} 1 & \boldsymbol{x_1}^T \\ & \vdots \\ 1 & \boldsymbol{x_m}^T \end{pmatrix},$$

where $\boldsymbol{x}_i$ is the vector of explanatory variables for the i-th class. Along this article we assume that X has full rank and that the model is not saturated (i.e. $p + 1 < m$). With all these definitions, we have that

$$\begin{aligned} G(\boldsymbol{\mu}) &= X\boldsymbol{\beta}, \\ (G \circ \dot{\boldsymbol{\kappa}})(\boldsymbol{\theta}) &= X\boldsymbol{\beta}, \\ \boldsymbol{\theta} &= (G \circ \dot{\boldsymbol{\kappa}})^{-1}(X\boldsymbol{\beta}). \end{aligned} \tag{14}$$

It is useful to reparameterize (13) in terms of the mean vector $\boldsymbol{\mu}$ instead of $\boldsymbol{\theta}$. Using the mean value parameterization (this is (6) but substituting ϕ for ϕ/w), (13) can be reparameterized as

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \phi) = C(\boldsymbol{y}, \phi) \exp\left(-\frac{1}{2\phi} D(\boldsymbol{y}, \boldsymbol{\mu}) \right), \tag{15}$$

where $C(\boldsymbol{y}, \phi) = \prod_{i=1}^{m} c(y_i, \frac{\phi}{w_i})$, and $D : \Omega^m \times \Omega^m \to [0, \infty)$ with

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{m} w_i d(y_i, \mu_i), \tag{16}$$

where $\Omega^m = \left\{ (\mu_1 \cdots \mu_m)^T : \mu_1, \ldots, \mu_m \in \Omega \right\}$. D is called the deviance of the model. We give here some of its properties:

- Given a sample, finding the mle of $\boldsymbol{\theta}$ is equivalent to finding the value of $\boldsymbol{\beta}$ that minimizes the deviance.
- D can be used to estimate the dispersion parameter (although it is not the only method). The deviance estimator of ϕ is given by

$$\hat{\phi} = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{m - p}.$$

- The asymptotic distribution of D plays an important role in model assessment and variable selection.

For further details about the use and properties of the deviance we recommend Jørgensen [10].

3 Entropic Estimator

A posterior distribution is more informative than a point estimation since it reflects our uncertainty about the true parameter. Now, in insurance, it is necessary to charge a premium, which is a point estimate. In this section we define the point estimators that we propose as credibility premiums.

Consider a parametric family of distributions with a parameter $\boldsymbol{\theta}$ to be estimated. Assume that $\boldsymbol{\theta}_0$ is the "true" parameter. In Bayesian point estimation one first chooses a loss function $\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ that represents the cost of estimating $\boldsymbol{\theta}$ to be $\boldsymbol{\theta}_1$ instead of $\boldsymbol{\theta}_0$. Now, since $\boldsymbol{\theta}_0$ is not known, we define a risk function as

$$\mathcal{R}(\theta) = \mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\right],$$

where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$. Then the point estimator $\hat{\boldsymbol{\theta}}$ of θ_0 is the value of θ that minimizes $\mathcal{R}$. That is:

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{R}(\boldsymbol{\theta}).$$

The entropic estimator is defined as the Bayesian point estimator when the loss function $\mathcal{L}$ is the relative entropy of the distribution with the real parameter $\boldsymbol{\theta_0}$ over the estimated one.

More precisely, assume that the density of a random vector $\boldsymbol{Y}$ depends on a parameter $\boldsymbol{\theta}$. Denote with $f(\boldsymbol{Y}|\boldsymbol{\theta}_1)$ the density of $\boldsymbol{Y}$ when $\boldsymbol{\theta}$ takes some value $\boldsymbol{\theta}_1$. The loss function is defined as

$$\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right].$$

The corresponding risk function is then defined as

$$\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right] := \mathbb{E}_{\pi}\left[\mathbb{E}_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right]\right],$$

where $\mathbb{E}_{\pi}$ is the expectation taken with respect to the posterior distribution of $\boldsymbol{\theta}$.

Definition 3.1. *The entropic estimator is defined as*

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right],$$

where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$.

Consider $\boldsymbol{\theta}$ to be the parameter of some model and let $\boldsymbol{\beta} = g(\boldsymbol{\theta})$ be a bijective transformation. Entropic estimators have the appealing invariance property that if $\boldsymbol{\theta}^*$ is the entropic estimator of $\boldsymbol{\theta}$, then $\boldsymbol{\beta}^* = g(\boldsymbol{\theta}^*)$ is the entropic estimator of $\boldsymbol{\beta}$. This property is called *invariance* (this terminology is consistent with Bernardo [1]) .

Not all estimators have this property. For example the estimator that minimizes the square loss error, i.e. the posterior mean, is not invariant. On the other hand, it is well known that maximum likelihood estimators are invariant.

3.1 Entropic Estimators for univariate EDFs

A key part of this article is to show how to find entropic estimators for GLMs. This is done in Section 4. In the rest of this section, we focus on entropic estimators for univariate EDFs and their relation to linear credibility. The main result of the section is Proposition 3.1, which is preceded by two technical lemmas that show properties of the unit deviance that are fundamental for finding entropic estimators of exponential families and GLMs (see the appendix for the proofs of these lemmas).

Lemma 3.1. *Let* d *be the unit deviance of a univariate EDF in* (5). *Then, there exist functions* d_1 *and* d_2 *such that for* $(y, \mu) \in \Omega \times \Omega$, *we have the following decomposition:*

$$d(y, \mu) = d_1(y) + d_2(y, \mu). \tag{17}$$

Moreover, d_2 *has the property that if* Y *is a random variable with support in* Ω, *and* μ *is fixed, then*

$$\mathbb{E}\big[d_2(Y, \mu)\big] = d_2\big(\mathbb{E}[Y], \mu\big).$$

Lemma 3.2. *Let* y *be fixed, then*

1. *the value of* μ *that minimizes* $d_2(y, \mu)$ *is the same one that minimizes* $d(y, \mu)$.

2. $d_2(y,\mu)$ *is minimized when* $\mu = y$.

Table 1 shows d, d_1 and d_2 for the normal, Poisson and gamma distributions.

Distribution	$d(y,\mu)$	$d_1(y)$	$d_2(y,\mu)$
Normal	$(y-\mu)^2$	y^2	$\mu^2 - 2y\mu$
Poisson	$2\left\{y\log\left(\frac{y}{\mu}\right) - (y-\mu)\right\}$	$2y[\log(y)-1]$	$2\left[\mu - y\log(\mu)\right]$
Gamma	$2\left\{\log\left(\frac{\mu}{y}\right) + \frac{y}{\mu} - 1\right\}$	$2\left[\frac{y}{\mu} + \log(\mu)\right]$	$2\left[\frac{y}{\mu} + \log\left(\frac{\mu}{y}\right) - 1\right]$

Table 1: Deviance decomposition of some common EDF's

Proposition 3.1. *Let* Y *be a random variable whose density is given by* (6) *for some unknown values of* μ *and* ϕ, $\pi(\mu,\phi)$ *be a prior distribution for* (μ,ϕ), *the vector* $\boldsymbol{y} = (y_1 \dots y_n)^T$ *be a conditionally i.i.d. sample given* (μ,ϕ), *and* $\pi(\mu,\phi|\boldsymbol{y})$ *the corresponding posterior. The entropic estimator* $\hat{\mu}$ *of* μ *is then given by*

$$\hat{\mu} = \mathbb{E}[Y|\boldsymbol{y}] = \mathbb{E}_\pi\left[\mathbb{E}_{\mu,\phi}[Y]\right],$$

where $\mathbb{E}_\pi$ *represents the expectation with respect to the posterior distribution and* $\mathbb{E}_{\mu,\phi}$ *represent the expectations with respect to fixed values of* (μ,ϕ).

Proof. Let (μ_0,ϕ_0) be the true parameters. By Lemma 3.1, the entropic risk measure can be expressed as

$$\begin{aligned}
\mathcal{R}(\mu,\phi) &= \mathbb{E}\left[\log\left(\frac{f(Y|\mu_0,\phi_0)}{f(Y|\mu,\phi)}\right)\right] \\
&= \mathbb{E}\left[\log\left(\frac{c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)}{c(Y,\phi)\exp\left(-\frac{1}{2\phi}d(Y,\mu)\right)}\right)\right] \\
&= \mathbb{E}\left[\log\left(c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)\right)\right] \\
&\quad - \mathbb{E}[\log(c(Y,\phi))] + \frac{1}{2\phi}\mathbb{E}[d(Y,\mu)] \\
&= \mathbb{E}\left[\log\left(c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)\right)\right] \\
&\quad - \mathbb{E}[\log(c(Y,\phi))] + \frac{1}{2\phi}\mathbb{E}[d_1(Y)] + \frac{1}{2\phi}d_2(\mathbb{E}[Y],\mu).
\end{aligned}$$

Note that, regardless of the value of ϕ, the value of μ that minimizes the expression above is the same one that minimizes the simpler function

$$\mathcal{R}_1(\mu) = d_2\big(\mathbb{E}[Y], \mu\big).$$

Then, by Part 2 of Lemma 3.2, the entropic estimator of μ is given by $\hat{\mu} = \mathbb{E}[Y]$. □

This result shows that for univariate EDFs the posterior mean not only minimizes the expected square error risk, but also the posterior entropic risk. In the following sections we will see that this property does not generalize to GLMs due to the difference of dimension between the response vector and the regression coefficients. For now, realize that a direct consequence of Proposition 3.1 is that Jewell's estimator in [8] is an entropic estimator.

Corollary 3.1. *The linear credibility estimator* (3) *is the entropic estimator when ϕ is assumed known and* (2) *is used as prior for θ.*

4 Linear Credibility for GLMs

This section discusses whether it is possible to extend Jewell's result to GLMs. In other words we address the question: is there a prior for the regression coefficients $\boldsymbol{\beta}$ for which the posterior mean is a weighted mean between an out–of–sample estimate and the sample mean of a GLM?

There are two ways in which the question above can be interpreted. One could think of it as all m dimensions having the same credibility factor, that is, the credibility premium $\hat{\boldsymbol{\mu}}_c$ is given by

$$\hat{\boldsymbol{\mu}}_c = z\bar{\boldsymbol{y}} + (1-z)\boldsymbol{M}, \tag{18}$$

where $\bar{\boldsymbol{y}}$ is the GLM observed sample mean (i.e. a vector for which (8) applies to each coordinate), $\boldsymbol{M}$ is a vector of out–of–sample "manual" premiums as coordinates and $z \in (0,1)$ is the credibility factor. We call this interpretation *Linear Credibility of Type 1.*

The other interpretation is to give a different credibility factor to each coordinate. This is

$$\hat{\boldsymbol{\mu}}_c = Z\bar{\boldsymbol{y}} + (I-Z)\boldsymbol{M}, \tag{19}$$

where $\hat{\boldsymbol{\mu}}_c$, $\bar{\boldsymbol{y}}$ and $\boldsymbol{M}$ are as in (18), but $Z = \operatorname{diag}(z_1, \ldots, z_m)$, where z_i is the credibility factor of the i-th class and I is the identity matrix.

We call this interpretation *Linear Credibility of Type 2*. Note that linear credibility of Type 1 is a special case of linear credibility of Type 2.

4.1 Linear Credibility of Type 1 is Impossible

Jewell's prior in (2) is a conjugate prior to (1) that gives linear credibility premiums based on the posterior mean. Diaconis and Ylvisaker [6] generalized Jewell's result to multivariate exponential dispersion models. Since a GLM assumes that the response vector follows such a distribution, it could be conjectured that this automatically implies linear credibility for GLMs. In what follows we show that this is not the case.

After adapting the conjugate prior discussed in [6] to correspond to (13) so as to consider weights we get

$$\pi_{n_0,\boldsymbol{x_0}}(\boldsymbol{\theta}) \propto \exp\left(n_0\{\boldsymbol{x_0}^T W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{k}(\boldsymbol{\theta})\}\right)\mathbb{I}_{\Theta^m}(\boldsymbol{\theta}), \qquad (20)$$

where $n_0 > 0$, and $\boldsymbol{x_0} \in \Omega^m$ are the parameters of the prior distribution, $\Theta^m = \left\{(\theta_1 \dots \theta_m)^T : \theta_1, \dots, \theta_m \in \Theta\right\}$, $\mathbb{I}_{\Theta^m}$ is an indicator function and $\Omega^m = \left\{(\mu_1 \dots \mu_m)^T : \mu_1, \dots, \mu_m \in \Omega\right\}$. Theorem 3 of [6] proves that if the support of (13) contains an interval then (20) is the only prior that gives linear credibility. This implies that for any continuous response (and also for any Tweedie distribution), (20) is the only prior that gives linear credibility. In the paper it is also proven that (20) is the unique prior that gives linear credibility for the binomial distribution and in Johnson [9] the same is proven for the Poisson distribution.

As shown in (14), $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are related by $\boldsymbol{\theta} = ((G \circ \dot{\boldsymbol{\kappa}})^{-1} \circ X)(\boldsymbol{\beta})$. Thus, when a prior for $\boldsymbol{\beta}$ is chosen, a distribution is induced on $\boldsymbol{\theta}$. In what follows we refer to this distribution as the *induced prior* on $\boldsymbol{\theta}$. We have then that for continuous and Tweedie distribution and for the Poisson and negative binomial, a prior on the betas gives linear credibility if and only if the induced prior on $\boldsymbol{\theta}$ is (20).

Our strategy to prove the impossibility of linear credibility of type 1 is to show that no prior of $\boldsymbol{\beta}$ induces a prior on $\boldsymbol{\theta}$ that has density (20). We can see that this is the case by focusing on the support of the induced distribution. Since the support of (20) is Θ^m, then it is enough to prove that the support of every induced prior of $\boldsymbol{\theta}$ is different than Θ^m.

Proposition 4.1. *For any prior of $\boldsymbol{\beta}$ in a non-saturated GLM, the support of the induced prior of $\boldsymbol{\theta}$ is a proper subset of Θ^m.*

Proof. Since there is no restriction for the value of $\boldsymbol{\beta}$, it can take any value on $\mathbb{R}^{p+1}$. This is represented on the left rectangle of Figure 1.

$X\boldsymbol{\beta}$ can take values in $R(X)$, where $R(X)$ is the range of X. Since $\dim(R(X)) = p+1 < m$, then $R(X) \subsetneq \mathbb{R}^m$. This is represented in the middle rectangle of Figure 1.

Let S be the support of the induced prior on $\boldsymbol{\theta}$. Then $S \subset (G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X)) := \left\{(G \circ \dot{\boldsymbol{\kappa}})^{-1}(X\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^{p+1}\right\}$ (subset but not equality since some values of $\mathbb{R}^{p+1}$ may not be in the support of $\boldsymbol{\beta}$). Now, $(G \circ \dot{\boldsymbol{\kappa}})^{-1}$ is a bijective function. Let $\boldsymbol{p}$ be a point in $\mathbb{R}^m$ that is not in $R(X)$ and let $\boldsymbol{q} = (G \circ \dot{\boldsymbol{\kappa}})^{-1}(\boldsymbol{p})$. Then $\boldsymbol{q} \in \Theta^m$ but $\boldsymbol{q} \in S$ because otherwise $(G \circ \dot{\boldsymbol{\kappa}})^{-1}$ would not be one-to-one. This proves that $(G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X)) \subsetneq \Theta^m$ and therefore also that $S \subsetneq \Theta^m$. This is represented in the right rectangle of Figure 1.

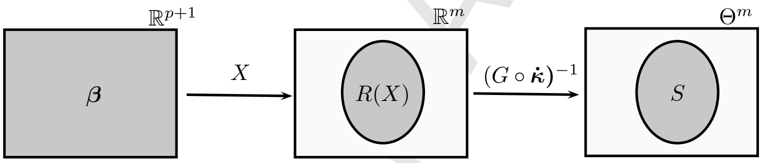


Figure 1: From left to right the grey zone represents the values that $\boldsymbol{\beta}$, $R(X)$ and S can take, respectively.

□

Now, on a different but related note, it is possible to generalize (20) in a way that allows to obtain conjugate priors that are suitable for GLMs. Define

$$\pi_1(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta}) \exp\left(n_0\{\boldsymbol{x_0}^T W\boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{k}(\boldsymbol{\theta})\}\right)\mathbb{I}_{\Theta^m}(\boldsymbol{\theta}),$$

where h is some integrable function for which the integral on the right hand–side above is finite and denote this distribution by $D_{conj}(n_0, \boldsymbol{x_0})$.

Proposition 4.2. *π_1 is a conjugate prior to* (13) *with posterior distribution* $D_{conj}\left(n_0 + \frac{1}{\phi}, \frac{1}{n_0\phi+1}\bar{\boldsymbol{y}} + \frac{\phi n_0}{\phi n_0+1}\boldsymbol{x_0}\right)$.

Proof. Let $\pi_1(\cdot|\boldsymbol{y})$ denote the posterior of π_1. Then, by definition of the posterior:

$\pi_1(\cdot|\boldsymbol{y}) \propto \pi_1(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}, \phi)$

$$
\begin{aligned}
&\propto h(\boldsymbol{\theta}) \exp\left(n_0 \boldsymbol{x}_{\mathbf{0}}^T W\boldsymbol{\theta} + \frac{\boldsymbol{y}^T W\boldsymbol{\theta}}{\phi} - n_0 \mathbf{1}^T W \boldsymbol{k}(\boldsymbol{\theta}) - \frac{\mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta})}{\phi}\right) \\
&= h(\boldsymbol{\theta}) \exp\left((n_0 \boldsymbol{x}_{\mathbf{0}}^T + \frac{\boldsymbol{y}^T}{\phi}) W\boldsymbol{\theta} - \left(n_0 + \frac{1}{\phi}\right) W \boldsymbol{\kappa}(\boldsymbol{\theta})\right) \\
&= h(\boldsymbol{\theta}) \exp\left(\left(n_0 + \frac{1}{\phi}\right)\left\{\frac{n_0 \boldsymbol{x}_{\mathbf{0}}^T + \frac{\boldsymbol{y}^T}{\phi}}{n_0 + \frac{1}{\phi}} W\boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta})\right\}\right) \\
&= h(\boldsymbol{\theta}) \exp\left(\left(n_0 + \frac{1}{\phi}\right)\left\{\frac{\phi n_0 \boldsymbol{x}_{\mathbf{0}}^T + \boldsymbol{y}^T}{\phi n_0 + 1} W\boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta})\right\}\right),
\end{aligned}
$$

which proves the result. □

Now, in order for π_1 to overcome the problems that do not allow π in (20) to be used as a prior for GLMs, it is only necessary to chose h such that π_1 is outside of $(G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X))$ with probability zero. This way π_1 has the "right" support and there is a distribution of $\boldsymbol{\beta}$ that gives this distribution when transformed with $((G \circ \dot{\boldsymbol{\kappa}})^{-1} \circ X)$.

Two important remarks about π_1:

1. It does not give linear credibility (since this is impossible as has been shown above).
2. It is not easy to find an analytic expression for $\boldsymbol{\mu}$ (although this might be possible for some choices of π). Thus most likely one has to use some numerical method or MCMC in order to find the posterior means, but this defeats the purpose of using a conjugate prior.

4.2 Linear Credibility of Type 2 is Sometimes Feasible

Since the model is a GLM, there should be a $\hat{\boldsymbol{\beta}}_c$ such that $\hat{\boldsymbol{\mu}}_c = G^{-1}(X\hat{\boldsymbol{\beta}}_c)$. Thus, (19) becomes

$$G^{-1}(X\hat{\boldsymbol{\beta}}_c) = Z\bar{\boldsymbol{y}} + (I - Z)\boldsymbol{M}. \tag{21}$$

It turns out that for non saturated models (i.e. $\dim(\boldsymbol{\beta}) < \dim(\boldsymbol{\mu})$), the existence of some $\hat{\boldsymbol{\beta}}_c$ for which (21) can be satisfied depends on the observed sample. We demonstrate why this is the case with a simple example in dimension 2.

Consider a situation in which you divide your population in only 2 segments using a binary covariate with no intercept (otherwise we would have a saturated model). The design matrix in this case would be

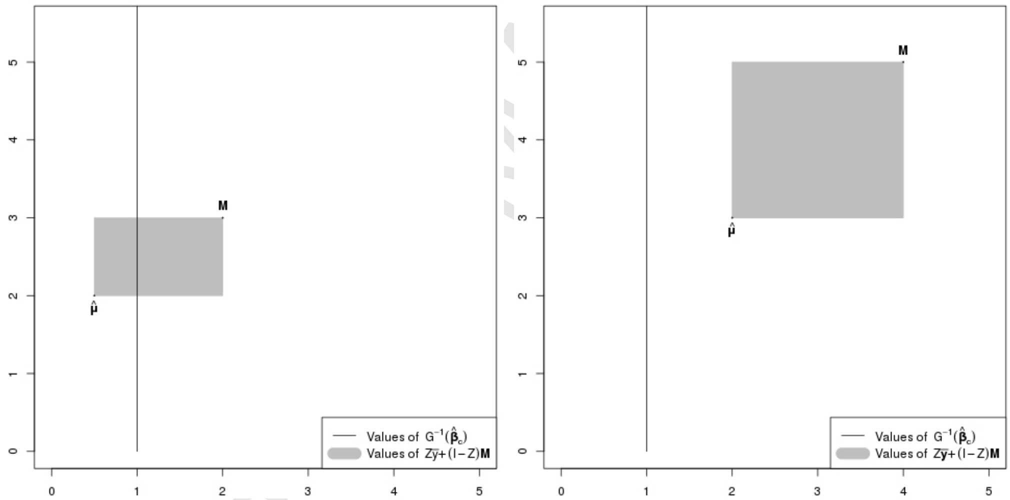
$$X = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \hat{\beta}_c \in \mathbb{R}.$$

Then, assuming a log–link function, the left hand side of (21) can be expressed as

$$\hat{\boldsymbol{\mu}}_c = G^{-1}(X\hat{\beta}_c) = G^{-1}\begin{pmatrix} 0 \\ \hat{\beta}_c \end{pmatrix} = \begin{pmatrix} \exp(0) \\ \exp(\hat{\beta}_c) \end{pmatrix} = \begin{pmatrix} 1 \\ \exp(\hat{\beta}_c) \end{pmatrix}.$$

If we graphed it we would see that the left hand side of (21) takes values only on the half upper side of the vertical line $x = 1$.

Imagine now two scenarios. In Scenario 1, $\bar{\boldsymbol{y}} = (0.5, 2)$ and $\boldsymbol{M} = (2, 3)$, while in Scenario 2, $\bar{\boldsymbol{y}} = (2, 3)$ and $\boldsymbol{M} = (4, 5)$.



(a) Scenario 1 (b) Scenario 2

Figure 2: Values of the left and right hand side of (21) in both scenarios

As the values of the elements of Z vary, the right hand side of (19) can take the values of the rectangle defined by $\bar{\boldsymbol{y}}$ and $\boldsymbol{M}$. Figure 2 shows graphs with the possible values of the left and right hand side of (19) for each scenario.

In both graphs, the vertical line represents the values of $\hat{\boldsymbol{\mu}}_c$. The rectangle represents all the possible values that $Z\bar{\boldsymbol{y}} + (I - Z)\boldsymbol{M}$ can take as the entries in the diagonal of Z vary from 0 to 1. In order to have exact linear credibility of type 2, it is necessary for the line and the rectangle to intersect. This is because the points of intersection, correspond to combinations of values of $\boldsymbol{\beta}_c$ and Z for which (21) holds. If there is no intersection it is not possible to have linear credibility of type 2.

The graph of Scenario 1 shows that (19) is satisfied for some values of Z, while in the graph for Scenario 2 it is impossible to satisfy (19).

The results of this section show that Jewell's result cannot be generalized to GLM's. That is, no prior for the parameters of a GLM guarantee linear credibility for all observed samples. In the next section we propose credibility estimators for GLMs that are not linear.

5 Entropic Credibility for GLMs

The Bayesian models proposed by Bühlmann and Jewell, using conjugate priors that ensured linear credibility formulas, made great sense in the 1960's and 70's, when computational issues ruled out more general (non–linear) Bayesian solutions. However, computing power is no longer scarce nor expensive. In this section we propose a modern, computational Bayesian approach to credibility.

5.1 Estimation of the Mean

We propose an entropic estimator of the mean vector of a GLM as the credibility premium. This section focuses on how to find such an estimator. We start by enunciating the following technical lemma, which is an extension of Lemma 3.1 to greater dimensions. A proof can be found in the Appendix.

Lemma 5.1. *Let D be the deviance of a GLM (see (16)). Then there exist functions D_1 and D_2 such that for $(\boldsymbol{y}, \boldsymbol{\mu}) \in \Omega^m \times \Omega^m$*

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = D_1(\boldsymbol{y}) + D_2(\boldsymbol{y}, \boldsymbol{\mu}). \tag{22}$$

Moreover, D_2 has the property that if $\boldsymbol{Y}$ is a random vector with support in Ω^m and $\boldsymbol{\mu} \in \Omega^m$ is fixed, then

$$\mathbb{E}[D_2(\boldsymbol{Y}, \boldsymbol{\mu})] = D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \tag{23}$$

In what follows, an arbitrary prior π is assumed (not necessarily conjugate) with posterior

$$\pi(\boldsymbol{\beta}, \phi|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \phi)\pi(\boldsymbol{\beta}, \phi), \tag{24}$$

where f is as in (13) or, equivalently (15), depending on the chosen parameterization. As stated, $\mathbb{E}_\pi[\cdot]$ denotes expectation with respect to the posterior measure. Whenever the expectation symbol is used without a subindex, it means expectation with respect to the predictive posterior distribution, i.e. $\mathbb{E}[\cdot] = \mathbb{E}_\pi[\mathbb{E}_{\boldsymbol{\beta},\phi}(\cdot)]$, where $\mathbb{E}_{\boldsymbol{\beta},\phi}(\cdot)$ means expectation with respect to the density in (13) with fixed coefficients vector $\boldsymbol{\beta}$ and fixed dispersion parameter ϕ.

Proposition 5.1. *The entropic estimator $\boldsymbol{\beta}^*$ of the coefficients of a Bayesian GLM are equal to the maximum likelihood estimator of a frequentist GLM with the same covariates, response distribution and weights, but with an observed response vector equal to $\mathbb{E}[\boldsymbol{Y}]$.*

Proof. Let $(\boldsymbol{\beta}_0, \phi_0)$ be the real parameters and $(\boldsymbol{\beta}, \phi)$ some fixed values. We use here for f the mean value parameterization in (15). Then, the risk function is given by

$$\mathcal{R}(\boldsymbol{\beta}, \phi) = \mathbb{E}_\pi\{\mathcal{L}[(\boldsymbol{\beta}_0, \phi_0), (\boldsymbol{\beta}, \phi)]\} = \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\mu}_0, \phi_0)}{f(\boldsymbol{Y}|\boldsymbol{\mu}, \phi)}\right)\right],$$

where $\boldsymbol{\mu} = G^{-1}(X\boldsymbol{\beta})$ and $\boldsymbol{\mu}_0 = G^{-1}(X\boldsymbol{\beta_0})$. Then, by Lemma 5.1 and (15) the expression above becomes

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\beta}, \phi) &= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0}D(\boldsymbol{Y}, \boldsymbol{\mu}_0)\right] - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] \\
&\quad + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0}D(\boldsymbol{Y}, \boldsymbol{\mu}_0)\right] \\
&\quad - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi}\mathbb{E}[D_1(\boldsymbol{Y}) + D_2(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0}D(\boldsymbol{Y}, \boldsymbol{\mu}_0) + \frac{1}{2\phi}D_1(\boldsymbol{Y})\right] \\
&\quad - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi}\mathbb{E}[D_2(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0}D(\boldsymbol{Y}, \boldsymbol{\mu}_0) + \frac{1}{2\phi}D_1(\boldsymbol{Y})\right]
\end{aligned}$$

$$- \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi} D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \tag{25}$$

The Bayesian point–estimator of $(\boldsymbol{\beta}_0, \phi_0)$ is given by the vector $(\boldsymbol{\beta}^*, \phi^*)$ that minimizes $\mathcal{R}$. Let us first focus on finding $\boldsymbol{\beta}^*$. Note that this is equivalent to minimizing

$$\mathcal{R}_1(\boldsymbol{\beta}) = D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \tag{26}$$

We first need to compute $\mathbb{E}[\boldsymbol{Y}]$, which can be expressed as

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}_\pi \left[\mathbb{E}_{\boldsymbol{\beta}_0, \phi_0}[\boldsymbol{Y}] \right] = \mathbb{E}_\pi \left[G^{-1}(X\boldsymbol{\beta}) \right].$$

Now, (24) gives $\pi(\cdot|\boldsymbol{y})$ up to a normalizing constant, thus the expectation above can be calculated with MCMC methods. In this way we consider the problem of computing $\mathbb{E}[\boldsymbol{Y}]$ solved.

Compare now the minimization of $\mathcal{R}_1(\boldsymbol{\beta})$ with a different optimization problem for which the solution method is well known. Consider a frequentist (non–Bayesian) GLM with the same response distribution, explanatory variables and weights. Imagine a sample under this model in which the observed response vector is equal to $\mathbb{E}[\boldsymbol{Y}]$. Using the mean value parameterization and Lemma 5.1, the log–likelihood function based on such a sample is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}, \phi) &= \log(C(\mathbb{E}[\boldsymbol{Y}], \phi)) - \frac{1}{2\phi} D(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}) \\ &= \log(C(\mathbb{E}[\boldsymbol{Y}], \phi)) - \frac{1}{2\phi} D_1(\mathbb{E}[\boldsymbol{Y}]) - \frac{1}{2\phi} D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}), \end{aligned}$$

where $\boldsymbol{\mu} = G^{-1}(X\boldsymbol{\beta})$. Since the only term that depends on $\boldsymbol{\beta}$ is the third one, then maximizing $\ell(\boldsymbol{\beta}, \phi)$ is equivalent to minimizing $D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu})$, i.e. the same as minimizing $\mathcal{R}_1(\boldsymbol{\beta})$. Hence, by obtaining the mle of the regression coefficients of this hypothetical frequentist GLM, we obtain $\boldsymbol{\beta}^*$ (or conclude that there is no solution, whenever this is the case). □

Once $\boldsymbol{\beta}^*$ has been found, the invariance property of the relative entropy allows to find the entropic premium straightforwardly.

Corollary 5.1. *If $\boldsymbol{\beta}^*$ is the entropic estimator of the coefficients of a Bayesian GLM, then the entropic premium is given by*

$$\boldsymbol{\mu}^* = G^{-1}(X\boldsymbol{\beta}^*). \tag{27}$$

Remark 5.1. *For a saturated model, i.e. when the dimension of $\boldsymbol{\beta}$ is equal to the dimension of $\boldsymbol{Y}$ (in other words $m = p$), the entropic premium is equal to $\mathbb{E}[\boldsymbol{Y}]$. This is because in a saturated model, the predicted mean is equal to the observed response mean.*

5.2 Estimation of the Dispersion Parameter

It is important to remark that the credibility estimator from the previous section takes into consideration the uncertainty of the dispersion parameter. This is the case because the posterior distribution of $\boldsymbol{\beta}$ depends on the posterior of ϕ.

This differs from classical credibility results where the dispersion parameter is considered known (e.g. Jewell [8] and Diaconis and Ylvisaker [6]). To the best of our knowledge there is only one article that considers a prior distribution for the dispersion parameter, Landsman and Makov [15], about which we have the following remarks:

1. The exponential distribution for the index parameter is justified using the principle of maximum entropy. The authors maximize the continuous entropy (that is entropy for continuous random variables), and use it for the index parameter assuming a known mean. Now, the continuous entropy does not have good properties as a measure of information. For instance it is not invariant under bijective transformations, which implies that one can loose or gain information by just transforming a random variable. Thus, the principle of maximum entropy is not a valid justification for the exponential distribution. Nevertheless, it is a valid prior and one can use it in those cases where it reflects properly the out of sample information.
2. A more serious problem exists with the result in Theorem 2; the integrals in (7) are carried out assuming that λ is exponential with mean λ_0. In other words, these are computed assuming the prior distribution for λ. This is erroneous since it is the posterior distribution that should be used in this integral. This would be justified if the prior for λ were natural conjugate. In this way the posterior of λ would also be exponential, but the parameter of the posterior would be different than the parameter of the prior, in this case.

We have not found a general procedure for obtaining the entropic estimator of the dispersion parameter. We discuss here the cases for which it can be found and present the difficulties in obtaining a general solution. Notice that a point–estimator for ϕ is not necessary to obtain the credibility premium (as seen in Section 5.1) or its uncertainty (which is measured by the posterior distribution).

Suppose that the credibility premium $\boldsymbol{\mu}^*$ has been obtained. From (25), one can see that finding the entropic estimator ϕ^* of the dispersion parameter is equivalent to minimizing

$$\mathcal{R}_2(\phi) = -\mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi))\right] + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)], \tag{28}$$

where $\boldsymbol{\mu}^* = G^{-1}(X\boldsymbol{\beta}^*)$. There are standard methods for minimizing univariate functions, but $\mathcal{R}_2$ is more difficult because the first expectation in (28) depends on ϕ. We consider first a special case where this minimization is rather straightforward. This is when there exists a function $H : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ such that

$$-\mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] = H(\mathbb{E}[\boldsymbol{Y}], \phi), \tag{29}$$

for every ϕ. In this case the problem simplifies considerably because once $\mathbb{E}[\boldsymbol{Y}]$ and $\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)]$ have been found (most likely by simulations), then it is possible to use standard methods to find ϕ^*, since (28) becomes

$$\mathcal{R}_2(\phi) = H(\mathbb{E}[\boldsymbol{Y}], \phi) + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)], \tag{30}$$

which is simple to evaluate.

A case worth mentioning when (29) occurs is when the response distribution is a proper dispersion model (see Jørgensen [11, Chap. 5]), i.e. when c in (6) can be decomposed as

$$c(y, \phi) = d(y)e(\phi), \tag{31}$$

for some functions d and e. Then, the first term on the right hand side of (28) becomes

$$\begin{aligned} -\mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] &= -\mathbb{E}\left[\prod_{i=1}^{m} \log\left(c\left(Y_i, \frac{\phi}{w_i}\right)\right)\right] \\ &= -\mathbb{E}\left[\prod_{i=1}^{m} \log\left(d(Y_i)e\left(\frac{\phi}{w_i}\right)\right)\right] \end{aligned}$$

$$= -\sum_{i=1}^{m} \mathbb{E}[\log(d(Y_i))] - \sum_{i=1}^{m} \log\left(e\left(\frac{\phi}{w_i}\right)\right).$$

Since $\sum_{i=1}^{m} \mathbb{E}[\log(d(Y_i))]$ does not depend on ϕ, the problem reduces to minimizing

$$\mathcal{R}_3(\phi) = -\sum_{i=1}^{m} \log\left(e\left(\frac{\phi}{w_i}\right)\right) + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)],$$

which can be done using standard optimization methods. Now, it is known that there are only three exponential dispersion models for which the factorization in (31) holds: the gamma, inverse Gaussian and normal distributions (this result is commented in Jørgensen [11, Chap. 5] and proven in Daniels [3]). Table 2 gives e is for these three models.

Distribution	Normal	Gamma	Inverse Gaussian
$e(\phi)$	$\phi^{-1/2}$	$\dfrac{e^{-1/\phi}}{\Gamma(\frac{1}{\phi})\phi^{1/\phi}}$	$\phi^{-1/2}$

Table 2: $e(\phi)$ for the three proper exponential dispersion families

Let us now consider the general case where (30) does not hold. Again MCMC methods can be helpful. Let $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^N$ be N simulations of $\boldsymbol{Y}$ from the posterior predictive distribution (superscripts are used since Y_i was already defined to be the i-th entry of $\boldsymbol{Y}$). Now define

$$\tilde{\mathcal{R}}_N(\phi) = -\frac{1}{N}\sum_{i=1}^{N} \log(C(\boldsymbol{Y}^i, \phi)) + \frac{1}{2\phi N}\sum_{i=1}^{N} D(Y^i, \boldsymbol{\mu}^*),$$

then, for every fixed ϕ

$$\lim_{N\to\infty} \tilde{\mathcal{R}}_N(\phi) = \mathcal{R}_2(\phi) \qquad \text{a.s.}. \tag{32}$$

Let $\tilde{\phi}_N = \operatorname{argmin} \tilde{\mathcal{R}}_N(\phi)$. Since $\tilde{\mathcal{R}}_N$ is simple to evaluate with a computer, standard univariate optimization methods can be used to find $\tilde{\phi}_N$. The question now is whether $\tilde{\phi}_N$ converges to ϕ^* as $N \to \infty$? We have not found easy–to–check sufficient conditions that guarantee convergence, although the following theorem might be useful in some cases.

Proposition 5.2. *If the convergence in* (32) *is uniform almost surely w.r.t. ϕ, then*

$$\mathcal{R}_2(\phi^*) = \lim_{N\to\infty} \tilde{\mathcal{R}}_N(\tilde{\phi}_N) \qquad a.s.$$

Proof. On the one hand we have that for every $n \in \mathbb{N}$

$$\begin{aligned} & \mathcal{R}_2(\phi^*) \leq \mathcal{R}_2(\tilde{\phi}_n) \\ \therefore \qquad & \mathcal{R}_2(\phi^*) \leq \liminf_n \mathcal{R}_2(\tilde{\phi}_n). \end{aligned}$$

On the other hand, let $\epsilon > 0$, since $\tilde{\mathcal{R}}_N \to \mathcal{R}_2$ uniformly a.s., then with probability one there exists $M > 0$, such that for every $n \geq M$,

$$|\tilde{\mathcal{R}}_n(\tilde{\phi}_n) - \mathcal{R}(\tilde{\phi}_n)| < \epsilon.$$

By the definition of $\tilde{\phi}_n$,

$$\tilde{\mathcal{R}}_n(\tilde{\phi}_n) \leq \mathcal{R}_n(\phi^*), \qquad \text{for all } n \in \mathbb{N}.$$

Then for every $n \geq N$,

$$\mathcal{R}_2(\tilde{\phi}_n) - \epsilon < \mathcal{R}_n(\phi^*),$$

thus

$$\begin{aligned} & \limsup_n \mathcal{R}_2(\tilde{\phi}_n) - \epsilon \leq \limsup \mathcal{R}_n(\phi^*) \\ \therefore \qquad & \limsup_n \mathcal{R}_2(\tilde{\phi}_n) - \epsilon \leq \mathcal{R}_2(\phi^*) \qquad a.s. \end{aligned}$$

Since this is true for $\epsilon > 0$, this implies that

$$\limsup_n \mathcal{R}_2(\tilde{\phi}_n) \leq \mathcal{R}_2(\phi^*),$$

and therefore

$$\mathcal{R}_2(\phi^*) = \lim_{n\to\infty} \mathcal{R}(\tilde{\phi}_n) \qquad a.s.$$

□

6 On the Applicability of the Entropic Premium

The previous section showed how one can find the entropic premium of a GLM, theoretically. In this section we address its practicability. In other words, we address the following question: is entropic credibility feasible for real–life datasets?

From Proposition 5.1 and Corollary 5.1, we have that once the response distribution, explanatory variables and prior have been chosen, the following steps give the entropic premium:

1. Find $\mathbb{E}[\boldsymbol{Y}]$ (see the paragraph preceding Proposition 5.1 for the definition of $\mathbb{E}[\boldsymbol{Y}]$).
2. Fit a frequentist GLM with the same covariates, response distribution and weights, but with observed response vector equal to $\mathbb{E}[\boldsymbol{Y}]$. This gives $\boldsymbol{\beta}^*$, the entropic estimator of the coefficients.
3. Find the entropic mean using (27).

Steps 2 and 3 are simple: one can perform these computations in R (see [19]) without major problems. The difficult part is Step 1. To the best of our knowledge, the simplest way to solve this problem is using Markov Chain Monte Carlo (MCMC). In the paragraphs that follow we give some recommendations on how to use this method.

It is important to consider that the greater m (the dimension of $\mathbb{E}[\boldsymbol{Y}]$), the more demanding the computations (both in terms of memory and CPU). Thus, it is very useful to first aggregate the data as in (8). This can drastically reduce m and turn an infeasible computation into something manageable.

A continuous variable can make data aggregation useless, especially when this happens with several different variables in the dataset. In such cases one should consider converting the support of these variables into intervals, hence transforming them into categorical variables.

Using Bayesian methods for variable selection can be time consuming. This is because one would need to run MCMC simulations for each combination of variables. A pragmatic approach to deal with this is to choose the variables using a frequentist GLM (which is much faster to fit). The resulting combination of variables can be used to build a starting model in the Bayesian case.

6.1 Illustrative Example

In this section we use the R interface to STAN (see [20]) to find an entropic credibility estimate of a severity model for a publicly available dataset. The main purpose of this example is to show that it is feasible to obtain entropic credibility premiums. We leave out the discussion about the convergence of the MCMC and the goodness of fit of the model. The interested reader can find the commented R code used for this example at `https://oquijano.net/articles/bayesian_credibility/`.

Variable name	Description
`veh_value`	vehicle value, in $10,000s
`exposure`	0-1
`clm`	occurrence of claim (0 = no, 1 = yes)
`numclaims`	number of claims
`claimcst0`	claim amount (0 if no claim)
`veh_body`	vehicle body, coded as BUS CONVT = convertible COUPE HBACK = hatchback HDTOP = hardtop MCARA = motorized caravan MIBUS = minibus PANVN = panel van RDSTR = roadster SEDAN STNWG = station wagon TRUCK UTE - utility
`veh_age`	age of vehicle: 1 (youngest), 2, 3, 4
`gender`	gender of driver: M, F
`area`	driver's area of residence: A, B, C, D, E, F
`agecat`	driver's age category: 1 (youngest), 2, 3, 4, 5, 6

Table 3: Vehicle insurance variables

The dataset appears in de Jong and Heller [4]. It is based on

67,856 one–year policies from 2004 or 2005. It can be downloaded from the companion site of the book: `http://www.acst.mq.edu.au/GLMsforInsuranceData`, as the dataset called Car. Table 3 shows the description of the variables as provided at the website.

We use a GLM with gamma responses to model the severity (variable `claimcst0`). We modified two explanatory variables, dividing the support of the continuous variable `veh_value` into three intervals $[0, 1.2)$,$[1.2, 1.86)$ and $[1.86, \infty)$, which we label as `P1`, `P2` and `P3`, respectively. The areas of residence A,B,C and D were also grouped together, thus the variable `area` included in our model takes three values: ABCD, E or F.

Model information		MCMC information	
Response distribution	gamma	**No. of chains**	3
Weight variable	`numclaims`	**Warmup period**	$2,000$
Covariates	`agecat(1)`		
	`gender(F)`	**Simulations kept (per chain)**	$28,000$
	`area(ABCD)`		
	`veh_value(P1)`		
Prior	betas are i.i.d. $U(-20, 20)$ and $\phi \sim U(0, 1000)$.		

Table 4: Severity model.

There is no out–of–sample information for this example and therefore we use non–informative priors for all the parameters (see Remark 6.1): The beta regression coefficients are assumed to be i.i.d. and to follow a uniform distribution on the interval $(-20, 20)$. The dispersion parameter is assumed to follow a uniform distribution on $(0, 1000)$, independently from the betas.

After aggregating the data (using (8)), the number of observations are reduced from $67, 856$ policies to 101 classes. Table 4 shows the information used for the Bayesian severity GLM. The value between parenthesis on the right of each explanatory variable corresponds to the reference category used in the model.

As shown in Proposition 5.1, in order to find the entropic estimator for the betas, it is first necessary to find the posterior mean of $\mathbb{E}[\boldsymbol{Y}]$. This is where MCMC is needed. For this example the simulations on STAN took around fifty seconds, counting compilation time, on three

2.67GHz processors. After, the entropic betas are found by fitting a frequentist GLM with the posterior mean as response vector. Table 5 shows the entropic coefficients obtained for this example.

Coefficient	Value	Coefficient	Value
(Intercept)	7.784	genderM	0.183
agecat2	-0.207	areaE	0.152
agecat3	-0.303	areaF	0.377
agecat4	-0.301	veh_valueP2	-0.117
agecat5	-0.403	veh_valueP3	-0.156
agecat6	-0.331		

Table 5: Entropic coefficients

Corollary 5.1 can be used to obtain the entropic premium for each homogeneous class. We do not show here a full table of premium values since there are 101 classes and it would take too much space. Nevertheless, a full table can be found at this article's website in the section "Classes Table". Figure 3 shows a graph of the entropic premiums in increasing order for all the classes in this example and compares it to premiums obtained using a frequentist GLM without any credibility considerations. We see that for some risk classes the GLM premiums are fully credible while for others the entropic credibility premiums are larger and more conservative.

Remark 6.1. *The use of the uniform prior here is arbitrary and for illustrative purposes only. The goal here is not to seek the best Bayesian analysis to this particular dataset, but rather to illustrate the feasibility of our method. That is, to show that it can be used with real–life datasets and get the results in a reasonable amount of time. It is up to the user of the method to choose a prior based on the criterion of their preference.*

Conclusion

As a Bayesian model, linear credibility (described at the beginning of the introduction) is rather artificial: the adequacy of Jewell's prior in any given situation is never discussed and the main focus is to ensure a credibility premium that is easy to compute. This convenience was crucial when Bühlman and Jewell originally published their work

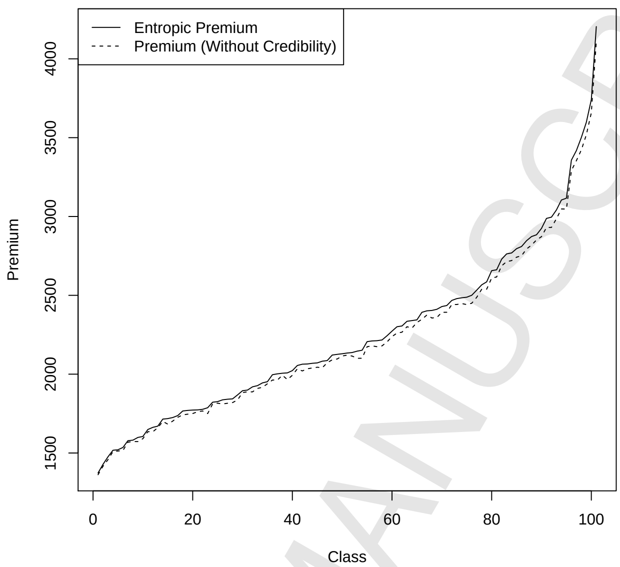


Figure 3: Entropic credibility premiums in increasing order

since computing power was scarce and expensive. Nowadays, not only computing power is cheap, but also sophisticated simulation software is available to anyone on the internet.

We propose then a modern Bayesian approach to credibility. In this way one can choose a prior based on out–of–sample information rather than on ease of computation. The limitation of possible priors is now set by the convergence of MCMC simulations. We use the relative entropy between the "true" model and the estimated one as the loss function.

Our proposal, when compared to classical credibility results for the exponential families, has the additional advantage of considering the uncertainty on the dispersion parameter. Finally, applying our method to a publicly available dataset shows that although substantial computations are required to obtain the credibility estimates, it is possible to apply this method to real–life datasets.

Acknowledgments

The authors are sincerely grateful to the anonymous reviewers for their constructive comments. We dedicate this paper to the memory of Bent Jørgensen whose contribution to the theory of exponential dispersion families opened the way for future developments.

References

[1] José M. Bernardo. Intrinsic credible regions: An objective bayesian approach to interval estimation. *Test*, 14(2):317–384, 2005.

[2] H. Bühlmann. Experience rating and credibility. *ASTIN Bulletin*, 4(3):99–207, 1967.

[3] H. E. Daniels. Exact saddlepoint approximations. *Biometrika*, 67(1):59–63, 1980. URL `http://biomet.oxfordjournals.org/content/67/1/59.abstract`.

[4] P. de Jong and G.Z. Heller. *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008. ISBN 9780511755408. URL `http://dx.doi.org/10.1017/CBO9780511755408`. Cambridge Books Online.

[5] F.E. De Vylder. Non-linear regression in credibility theory. *Insurance: Mathematics and Economics*, 4(3):163–172, 1985. URL `http://EconPapers.repec.org/RePEc:eee:insuma:v:4:y:1985:i:3:p:163-172`.

[6] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979. URL `http://dx.doi.org/10.1214/aos/1176344611`.

[7] C.A. Hachemeister. Credibility for regression models with application to trend. *Proc. of the Berkeley Actuarial Research Conference on Credibility*, pages 129–163, 1975.

[8] W.S. Jewell. Credible means are exact Bayesian for exponential families. *ASTIN Bulletin*, 8(1):77–90, 1974.

[9] N.L. Johnson. Uniqueness of a result in the theory of accident proneness. *Biometrika*, 44:530–531, 1957.

[10] B. Jørgensen. *The Theory of Exponential Dispersion Models and Analysis of Deviance*. Instituto de Matemática Pura e Aplicada, (IMPA), Brazil, 1992.

[11] B. Jørgensen. *The Theory of Dispersion Models*. Chapman & Hall, London, 1997.

[12] R. Kaas, M Goovaerts, J Dhaene, and M. Denuit. *Modern Actuarial Risk Theory.* Springer-Verlag Berlin Heidelberg, 2008.

[13] S. Kullback. *Information Theory and Statistics.* Dover Publications, New York, 1968.

[14] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. URL http://dx.doi.org/10.1214/aoms/1177729694.

[15] Z.M. Landsman and U.E. Makov. Exponential dispersion models and credibility. *Scandinavian Actuarial Journal*, 1998(1):89–96, 1998. URL http://dx.doi.org/10.1080/03461238.1998.10413995.

[16] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972.

[17] J.A. Nelder and R.J. Verall. Credibility theory and generalized linear models. *ASTIN Bulletin*, 27(1):71–82, 1997.

[18] E. Ohlsson. Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, 2008(4):301–314, 2008.

[19] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[20] Stan Development Team. RStan: the R interface to Stan, 2016. URL http://mc-stan.org/. R package version 2.14.1.

Appendix: Proofs of technical lemmas

Proof of Lemma 3.1
Let d be the unit deviance function of the response distribution and $y, \mu \in \Omega$. By (7), we have that

$$\begin{aligned} d(y,\mu) &= 2\left[y\left\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\right\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu))\right] \\ &= 2\left[y\dot{\kappa}^{-1}(y) - \kappa(\dot{\kappa}^{-1}(y))\right] + 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - y\dot{\kappa}^{-1}(\mu)\right] \end{aligned}$$

$$= d_1(y) + d_2(y, \mu),$$

where $d_1(y) = 2\left[y\dot{\kappa}^{-1}(y) - \kappa(\dot{\kappa}^{-1}(y))\right]$ and $d_2(y,\mu) = 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - y\dot{\kappa}^{-1}(\mu)\right]$. Now, let Y be a random variable with support in Ω and $\mu \in \Omega$ be fixed, then

$$\begin{aligned}
\mathbb{E}[d_2(Y,\mu)] &= \mathbb{E}\left[2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - Y\dot{\kappa}^{-1}(\mu)\right]\right] \\
&= 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - \mathbb{E}[Y]\dot{\kappa}^{-1}(\mu)\right] \\
&= d_2(\mathbb{E}[Y], \mu).
\end{aligned}$$

□

Proof of Lemma 3.2
The first part is a direct consequence of (17). Since y is fixed, minimizing the right hand side is equivalent to minimizing d_2 and therefore the claim is true.

The unit deviance d is such that (see Chapter 1 of Jørgensen [11]) $d(y,\mu) > 0$ if $y \neq \mu$ and $d(y,\mu) = 0$ when $y = \mu$. Thus $d(y,\mu)$ is minimized when $y = \mu$ and by Part 1 above the same applies to $d_2(y,\mu)$.

□

Proof of Lemma 5.1
From the definition of D and Lemma 3.1, we have that

$$\begin{aligned}
D(\boldsymbol{y},\boldsymbol{\mu}) &= \sum_{i=1}^{m} w_i d(y_i,\mu_i) \\
&= \sum_{i=1}^{m} w_i (d_1(y_i) + d_2(y_i,\mu_i)) \\
&= \sum_{i=1}^{m} w_i d_1(y_i) + \sum_{i=1}^{m} w_i d_2(y_i,\mu_i) \\
&= D_1(\boldsymbol{y}) + D_2(\boldsymbol{y},\boldsymbol{\mu}),
\end{aligned}$$

where $D_1(\boldsymbol{y}) = \sum_{i=1}^{m} w_i d_1(y_i)$ and $D_2(\boldsymbol{y},\boldsymbol{\mu}) = \sum_{i=1}^{m} w_i d_2(y_i,\mu_i)$. This proves (22). Let $\boldsymbol{Y} = (Y_1,\ldots,Y_m)$ be a random vector with support on Ω^m, and $\boldsymbol{\mu} \in \Omega^m$ be fixed. Then

$$\mathbb{E}[D_2(\boldsymbol{Y},\boldsymbol{\mu})] = \mathbb{E}[\sum_{i=1}^{m} w_i d_2(Y_i,\mu_i)] = \sum_{i=1}^{m} w_i d_2(\mathbb{E}[Y_i],\mu_i) = D_2(\mathbb{E}[\boldsymbol{Y}],\boldsymbol{\mu}),$$

which proves (23). □