Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Calibration bands for mean estimates within the exponential dispersion family

Łukasz Delong, Selim Gatti & Mario V. Wüthrich

To cite this article: Łukasz Delong, Selim Gatti & Mario V. Wüthrich (04 Feb 2026): Calibration bands for mean estimates within the exponential dispersion family, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2620835

To link to this article: https://doi.org/10.1080/24754269.2026.2620835

Published online: 04 Feb 2026.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2620835

Taylor & Francis
Taylor & Francis Group

Calibration bands for mean estimates within the exponential dispersion family

Łukasz Delong[a], Selim Gatti [b] and Mario V. Wüthrich[b]

[a]Department of Statistics and Econometrics, Faculty of Economic Sciences, University of Warsaw, Poland;
[b]RiskLab, Department of Mathematics, ETH Zurich, Zürich, Switzerland

ABSTRACT

A statistical model is said to be calibrated if the resulting mean estimates perfectly match the true means of the underlying responses. Aiming for calibration is often not achievable in practice as one has to deal with finite samples of noisy observations. A weaker notion of calibration is auto-calibration. An auto-calibrated model satisfies that the expected value of the responses being given the same mean estimate matches this estimate. Testing for auto-calibration has only been considered recently in the literature, and we propose a new approach based on calibration bands. Calibration bands denote a set of lower and upper bounds such that the probability that the true means lie simultaneously inside those bounds exceeds some given confidence level. Such bands were constructed by Yang and Barber ((2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, *13*(1), 646–677. https://doi.org/10.1214/18-EJS1520) for sub-Gaussian distributions. Dimitriadis et al. ((2023). Honest calibration assessment for binary outcome predictions. *Biometrika*, *110*(3), 663–680. https://doi.org/10.1093/biomet/asac068) then introduced narrower bands for the Bernoulli distribution. We use the same idea in order to extend the construction to the entire exponential dispersion family that contains, for example, the binomial, Poisson, negative binomial, gamma, and normal distributions. Moreover, we show that the obtained calibration bands allow us to construct various tests for calibration and auto-calibration, respectively. As the construction of the bands does not rely on asymptotic results, we emphasize that our tests can be used for any sample size.

ARTICLE HISTORY
Received 25 March 2025
Revised 8 October 2025
Accepted 11 January 2026

KEYWORDS
Auto-calibration; calibration; calibration bands; exponential dispersion family; binomial distribution; Poisson distribution; gamma distribution; normal distribution

1. Introduction

Various statistical methods can be used to derive mean estimates from available observations, and it is important to understand whether these mean estimates are reliable for decision making. A statistical model is said to be *calibrated* if the resulting mean estimates perfectly match the true means of the underlying responses. In practice, calibration is often not achievable, as estimates are obtained from finite samples of noisy observations. A desirable property for a statistical model is *auto-calibration*, which is a related and weaker notion of calibration; see Krüger and Ziegel (2021) and Gneiting and Resin (2023). This property means

CONTACT Selim Gatti selim.gatti@math.ethz.ch RiskLab, Department of Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zürich, Switzerland

that when the responses are partitioned according to their mean estimates, i.e., responses with equal mean estimates are grouped, the estimated mean within each of these groups should match the expected value of the responses of that group. Pohle (2020), Gneiting and Resin (2023), Krüger and Ziegel (2021), Denuit et al. (2021), Fissler et al. (2022) as well as Wüthrich and Merz (2023) emphasized the importance of auto-calibration when assessing a fitted model, especially for insurance pricing, because an auto-calibrated pricing system avoids systematic cross-subsidy.

Testing for auto-calibration has only been considered recently in the literature. Denuit et al. (2024) proposed a test using Lorenz and concentration curves that requires the evaluation of a non-explicit asymptotic distribution using Monte-Carlo simulations. Simpler versions of this test were provided in Wüthrich (2024) for discrete and finite regression functions. Additionally, Delong and Wüthrich (2025) considered the use of bootstrap techniques to assess the auto-calibration of a model. In the special case of binary observations, Hosmer and Lemeshow (1980) derived a χ^2-test by binning observations over disjoint intervals, whereas Gneiting and Resin (2023) proposed a bootstrap approach to test for auto-calibration in this binary setup.

We take a different approach in this paper. Our goal is to construct *calibration bands* for mean estimates within the exponential dispersion family (EDF). A calibration band denotes a set of lower and upper bounds for each mean estimate such that the probability of having simultaneously all the true means lying inside these bounds exceeds a given confidence level. This allows us to assess the calibration of a model by evaluating whether the mean estimates fall inside these bounds, and as the construction of the band does not rely on any asymptotical results, one can use it to construct statistical tests for calibration and auto-calibration for all sample sizes, in contrast to the approaches mentioned above.

Calibration bands were first constructed by Yang and Barber (2019) for mean estimates of sub-Gaussian distributions, which are distributions having similar or lighter tails than a Gaussian distribution as, for example, the binomial and the normal distributions. Dimitriadis et al. (2023) then provided another construction in the binary case and showed that the resulting calibration bands are narrower than Yang–Barber's bands for the same case. Our construction of the calibration bands is similar to the construction of Dimitriadis et al. (2023). We extend their results to the entire EDF by exploiting stochastic ordering results and convolution formulas within the EDF. The EDF is a broad class of distributions commonly used in statistical modelling and, particularly, in generalized linear models (GLMs); we refer the reader to McCullagh and Nelder (1983), Jørgensen (1986, 1997) and Barndorff-Nielsen (2014). We provide a general construction of the calibration bands for the EDF, and we show that these bands can be expressed explicitly for the binomial, Poisson, negative binomial, gamma and normal cases. Our bands are identical to the ones derived by Dimitriadis et al. (2023) in the binary case and we show that they are narrower than the calibration bands of Yang–Barber in the normal case.

We then extend the above construction to regression modelling, where the mean estimation task consists in approximating the conditional mean of a response given an observed set of features. In this framework, we construct two opposite statistical tests for calibration using calibration bands, i.e., statistical tests where the calibration property once lies in the null-hypothesis and once in the alternative. Moreover, we show that one can construct two opposite tests for auto-calibration.

This paper is organized as follows. In the next section, we introduce the EDF and state the framework under which we aim at constructing calibration bands on the mean. In Section 3,

we outline the necessary assumptions and derive some stochastic ordering results within the EDF that allow us, along with some convolution formulas, to derive the bands. Then, we exploit these results by using a union bound argument in order to construct the calibration bands in Section 4. In Section 5, we show that these bands can actually be expressed explicitly for the binomial, Poisson, negative binomial, gamma, and normal distributions. In Section 6, we extend the construction of the calibration bands to regression modelling and in Section 7, we introduce the auto-calibration property and provide conditions under which this property is equivalent to calibration. Moreover, we derive statistical tests in the same section that enable us to test for calibration and auto-calibration of a given regression function, and we emphasize the importance of using a suitable dispersion estimate to perform these tests. Finally, in Section 8, we highlight the impact of various factors on the resulting calibration bands through a series of numerical examples. The last section concludes this work. All mathematical proofs are provided in the appendix.

2. Calibration bands within the exponential dispersion family

A random variable Y belongs to the EDF if its density can be written as

$$f_Y(y;\theta,v,\varphi,\kappa(\cdot)) = \exp\left\{\frac{y\theta-\kappa(\theta)}{\varphi/v} + a(y;v/\varphi)\right\}, \tag{1}$$

where $\theta \in \Theta$ is called the canonical parameter, Θ is the effective domain, $\kappa : \Theta \to \mathbb{R}$ is the cumulant function, $v > 0$ is the volume, $\varphi > 0$ is the dispersion parameter, and $a(y; v/\varphi)$ is a normalizing function depending only on y and v/φ such that the density integrates to one. We write $Y \sim \text{EDF}(\theta, v, \varphi, \kappa(\cdot))$ to denote a member of the EDF and emphasize that the density in (1) is understood w.r.t. a σ-finite measure ν on $\mathbb{R}$ that is independent of the specific choice of the canonical parameter $\theta \in \Theta$. In particular, the random variable Y can, for example, be absolutely continuous or discrete.

In this paper, we construct a calibration band on the mean of responses belonging to the EDF. To this end, we consider n independent responses $Y_i \sim \text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ for a fixed and known cumulant function κ and a given dispersion parameter $\varphi > 0$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and denote the mean of each response by $\mu_i = \mathbb{E}[Y_i]$. Under the assumption that the responses are ordered such that their means are non-decreasing, i.e.,

$$\mu_i \le \mu_j \text{ whenever } i \le j, \tag{2}$$

we construct random variables $L^{\alpha}_{\boldsymbol{Y},i}$ and $U^{\alpha}_{\boldsymbol{Y},i}$ such that

$$\mathbb{P}\left(L^{\alpha}_{\boldsymbol{Y},i} \le \mu_i \le U^{\alpha}_{\boldsymbol{Y},i} \text{ for all } i \in \{1,\ldots,n\}\right) \ge 1-\alpha, \tag{3}$$

for any given confidence level $1-\alpha \in (0,1)$. The resulting calibration band

$$\left(L^{\alpha}_{\boldsymbol{Y},i}, U^{\alpha}_{\boldsymbol{Y},i}\right)_{i=1}^{n}$$

is *data-dependent* as it depends on the realizations of the random vector $\boldsymbol{Y}$. Its construction relies on the ordering assumed in (2) as well as on stochastic ordering properties and convolution formulas of the EDF that are discussed in the next section.

3. Stochastic orders within the exponential dispersion family

The construction of the calibration bands on the means within the EDF is motivated by the same idea used by Dimitriadis et al. (2023) in order to construct calibration bands for independent binary random variables. Binary random variables have the nice property that we can aggregate them to binomial random variables, which satisfy some stochastic ordering properties. The calibration bands constructed by Dimitriadis et al. (2023) are based on these aggregations and stochastic orderings. In this section, we outline the assumptions and properties needed to generalize these ideas to the entire EDF by extracting similar stochastic orders. For this, we start with an assumption on the effective domain Θ and the σ-finite measures ν_i that define the supports of the responses Y_i.

Assumption 3.1: We assume that the effective domain has a non-empty interior $\overset{\circ}{\Theta}$ and that the σ-finite measures ν_i are not a single point mass.

This assumption excludes any trivial case of the EDF and it implies that the effective domain Θ is a (possibly infinite) interval in $\mathbb{R}$; we refer to Jørgensen (1986, 1997). Moreover, under Assumption 3.1, the cumulant function κ is smooth and strictly convex on the interior of the effective domain $\overset{\circ}{\Theta}$, which implies that the derivative of the cumulant function κ' can be inverted on this range. There is thus a one-to-one correspondence between the canonical parameter $\theta \in \overset{\circ}{\Theta}$ and the mean of the random variable $Y \sim \text{EDF}(\theta, \nu, \varphi, \kappa(\cdot))$ that is given by

$$\mathbb{E}[Y] = \kappa'(\theta).$$

This correspondence can be expressed as

$$h(\mathbb{E}[Y]) = \theta, \tag{4}$$

where $h = (\kappa')^{-1} : \kappa'(\overset{\circ}{\Theta}) \to \overset{\circ}{\Theta}$ is a strictly increasing function called the *canonical link* of the chosen distribution within the EDF. In order to exploit this bijective map between means and canonical parameters, we make the following assumption and call the set $\kappa'(\overset{\circ}{\Theta})$ the *mean parameter space*.

Assumption 3.2: We assume that the canonical parameters of all the considered random variables lie in the interior of the effective domain, i.e., $\theta_i \in \overset{\circ}{\Theta}$ for all $1 \leq i \leq n$.

Under Assumptions 3.1 and 3.2, one can now derive various stochastic ordering relations that will be used to construct the calibration bands on the means in Section 4. To do so, we introduce the usual stochastic order and the likelihood ratio order as in Shaked and Shanthikumar (2007).

Definition 3.3: The usual stochastic order and the likelihood ratio order are defined as follows.

- A random variable X is said to be smaller than a random variable Y in the usual stochastic order, written as $X \leq_{\text{st}} Y$, if

$$\mathbb{P}(X \leq x) \geq \mathbb{P}(Y \leq x), \qquad \text{for all } x \in \mathbb{R}.$$

- A random variable X (with density f) is said to be smaller than a random variable Y (with density g) in the likelihood ratio order, written as $X \leq_{\text{lr}} Y$, if

$$t \mapsto \frac{g(t)}{f(t)} \tag{5}$$

is a non-decreasing function in t, for t being in the union of the supports of f and g, and where $a/0$ is taken to be ∞, whenever $a > 0$.

Note that the densities in (5) are understood with respect to σ-finite measures on $\mathbb{R}$, which means that the random variables X and Y can, for example, be absolutely continuous or discrete. Theorem 1.C.1 of Shaked and Shanthikumar (2007) states that the likelihood ratio order is weaker than the usual stochastic order. That is, for any two random variables X and Y satisfying $X \leq_{\text{lr}} Y$, we have $X \leq_{\text{st}} Y$. This implication leads to a first stochastic ordering result within the EDF; all proofs are provided in the appendix.

Proposition 3.4: Suppose that Assumptions 3.1 and 3.2 hold and let $\mu_1 \leq \mu_2$ be in the mean parameter space $\kappa'(\overset{\circ}{\Theta})$. Then, for any volume $\nu > 0$, dispersion parameter $\varphi > 0$ and cumulant function κ, the random variables $Y_1 \sim \text{EDF}(h(\mu_1), \nu, \varphi, \kappa(\cdot))$ and $Y_2 \sim \text{EDF}(h(\mu_2), \nu, \varphi, \kappa(\cdot))$ satisfy $Y_1 \leq_{\text{st}} Y_2$.

Denote the distribution of a random variable $Y \sim \text{EDF}(h(\mu), \nu, \varphi, \kappa(\cdot))$ for $\mu \in \kappa'(\overset{\circ}{\Theta})$ by

$$F(y; h(\mu), \nu, \varphi, \kappa(\cdot)) = \mathbb{P}(Y \leq y),$$

and the left-continuous, right-limit distribution of this random variable by

$$F^*(y; h(\mu), \nu, \varphi, \kappa(\cdot)) = \mathbb{P}(Y < y).$$

For fixed $y \in \mathbb{R}$, $\nu > 0$, $\varphi > 0$ and cumulant function κ, the stochastic ordering result in Proposition 3.4 implies that the functions

$$\mu \in \kappa'(\overset{\circ}{\Theta}) \mapsto F(y; h(\mu), \nu, \varphi, \kappa(\cdot)),$$

and

$$\mu \in \kappa'(\overset{\circ}{\Theta}) \mapsto F^*(y; h(\mu), \nu, \varphi, \kappa(\cdot)),$$

are non-increasing in μ. This observation leads to the construction of the bounds on the mean in the next proposition.

Proposition 3.5: Suppose that Assumptions 3.1 and 3.2 hold. Let $Y \sim \text{EDF}(\theta, \nu, \varphi, \kappa(\cdot))$ for $\theta \in \overset{\circ}{\Theta}$, $\delta \in (0, 1)$, and define the random variables

$$l^\delta(Y, \nu, \varphi, \kappa(\cdot)) = \inf \left\{ \mu \in \kappa'(\overset{\circ}{\Theta}) \;\middle|\; F^*\left(Y; h(\mu), \nu, \varphi, \kappa(\cdot)\right) \leq 1 - \delta \right\}, \tag{6}$$

and

$$u^\delta(Y, \nu, \varphi, \kappa(\cdot)) = \sup \left\{ \mu \in \kappa'(\overset{\circ}{\Theta}) \;\middle|\; F\left(Y; h(\mu), \nu, \varphi, \kappa(\cdot)\right) \geq \delta \right\}. \tag{7}$$

Then,

$$\mathbb{P}\left(\mathbb{E}[Y] \geq l^\delta(Y, \nu, \varphi, \kappa(\cdot))\right) \geq 1 - \delta \quad \text{and} \quad \mathbb{P}\left(\mathbb{E}[Y] \leq u^\delta(Y, \nu, \varphi, \kappa(\cdot))\right) \geq 1 - \delta.$$

Proposition 3.5 provides lower and upper bounds holding for the case of a single response $Y \sim \text{EDF}(\theta, v, \varphi, \kappa(\cdot))$. These bounds directly depend on the value of $\delta \in (0,1)$ as for any $y \in \mathbb{R}$, the interval

$$\left[l^{\delta}(y, v, \varphi, \kappa(\cdot)), u^{\delta}(y, v, \varphi, \kappa(\cdot))\right]$$

is wide for small values of $\delta \in (0, 1/2]$ and narrow for large values of $\delta \in (0, 1/2]$. Additionally, note that for $\delta \geq 1/2$, we might even have that the lower bound exceeds the upper bound, i.e.,

$$l^{\delta}(y, v, \varphi, \kappa(\cdot)) \geq u^{\delta}(y, v, \varphi, \kappa(\cdot)).$$

To lift this result to the case of n independent responses being ordered such that their canonical parameters (or means) are increasing, referring to (2), we aim at using Proposition 3.5 to derive lower and upper bounds on the weighted partial sums

$$Z_{j:k} = \frac{1}{v_{j:k}} \sum_{i=j}^{k} v_i Y_i, \tag{8}$$

for any $1 \leq j \leq k \leq n$, and where we define the aggregated volumes as

$$v_{j:k} = \sum_{i=j}^{k} v_i. \tag{9}$$

For this, we use the following stochastic bounds on the random variables $Z_{j:k}$.

Lemma 3.6: Let $Y_j, \ldots, Y_k$ be independent $\text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ distributed random variables for given volumes $v_i > 0$ and indices $j \leq i \leq k$ such that $\theta_j \leq \cdots \leq \theta_k$. Under Assumptions 3.1 and 3.2, the weighted sum $Z_{j:k}$ in (8) satisfies

$$Z_{j:k}^{-} \leq_{\text{st}} Z_{j:k} \leq_{\text{st}} Z_{j:k}^{+},$$

for

$$Z_{j:k}^{-} \sim \text{EDF}(\theta_j, v_{j:k}, \varphi, \kappa(\cdot)) \quad \text{and} \quad Z_{j:k}^{+} \sim \text{EDF}(\theta_k, v_{j:k}, \varphi, \kappa(\cdot)). \tag{10}$$

Using these stochastic bounds and Proposition 3.5, we can now derive bounds on the means of the weighted partial sums $Z_{j:k}$.

Proposition 3.7: Suppose that Assumptions 3.1 and 3.2 hold. Let $Y_j, \ldots, Y_k$ be independent $\text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ distributed random variables for given volumes $v_i > 0$ and indices $j \leq i \leq k$ such that $\theta_j \leq \cdots \leq \theta_k$. Moreover, let $\delta \in (0,1)$, define $v_{j:k}$ and $Z_{j:k}$ as in (8)–(9) and denote by μ_j and μ_k the means of Y_j and Y_k, respectively. Then, we have

$$\mathbb{P}\left(\mu_j \leq u^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\right) \geq 1 - \delta \quad \text{and} \quad \mathbb{P}\left(\mu_k \geq l^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\right) \geq 1 - \delta,$$

for the random variables $l^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ and $u^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ defined in (6)–(7).

As we will outline in the next section, Proposition 3.7 is in fact at the core of the construction of the calibration bands on the mean holding for n independent responses. because for any indices $j \leq k$, it provides upper and lower bounds on the true means μ_j and μ_k that only depend on the realizations of the responses $Y_j, \ldots, Y_k$, the given volumes $v_j, \ldots, v_k$ as well as the dispersion parameter φ and cumulant function κ.

4. Construction of the calibration bands

4.1. Main result

The aim of this section is to construct calibration bands on the means, as defined in (3), for independent responses $(Y_i)_{i=1}^n \sim \text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ that are ordered such that their canonical parameters fulfill $\theta_1 \leq \cdots \leq \theta_n$. This construction makes use of sets of ordered pairs $\mathcal{J} \subseteq \{1, \ldots, n\}^2$ that we define as sets satisfying

$$(j,k) \in \mathcal{J} \implies j \leq k.$$

By using a union bound argument, a corollary of Proposition 3.7 is that

$$\mathbb{P}\left(\mu_j \leq u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ and } \mu_k \geq l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\right) \geq 1 - 2\delta,$$

for any pair $(j,k) \in \mathcal{J}$, because the complement of the above event is nested in the event where at least one of the true means fails to lie above or below the constructed lower and upper bounds, respectively. Similarly, we have

$$\begin{aligned}\mathbb{P}\left(\mu_j \leq u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ and } \mu_k \geq l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ for all } (j,k) \in \mathcal{J}\right) \\ \geq 1 - 2|\mathcal{J}|\delta,\end{aligned} \tag{11}$$

where the bounds on the means μ_j and μ_k now hold simultaneously for all pairs $(j,k) \in \mathcal{J}$. This last inequality allows us to construct calibration bands on the means of n independent responses, as stated in the next theorem.

Theorem 4.1: Suppose that Assumptions 3.1 and 3.2 hold. Let $Y_1, \ldots, Y_n$ be independent $\text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ distributed random variables for given volumes $v_i > 0$ and indices $1 \leq i \leq n$ such that $\theta_1 \leq \cdots \leq \theta_n$. Moreover, let $\mathcal{J} \subseteq \{1, \ldots, n\}^2$ be any set of ordered pairs. By writing $\mu_i = \mathbb{E}[Y_i]$ for $1 \leq i \leq n$, we have for any given confidence level $1 - \alpha \in (0,1)$ that

$$\mathbb{P}\left(L^\alpha_{\boldsymbol{Y},i} \leq \mu_i \leq U^\alpha_{\boldsymbol{Y},i} \text{ for all } i \in \{1, \ldots, n\}\right) \geq 1 - \alpha, \tag{12}$$

with

$$L^\alpha_{\boldsymbol{Y},i} = \sup_{(j,k) \in \mathcal{J} : \theta_i \geq \theta_k} l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)), \tag{13}$$

and

$$U^\alpha_{\boldsymbol{Y},i} = \inf_{(j,k) \in \mathcal{J} : \theta_i \leq \theta_j} u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)), \tag{14}$$

for $\delta = \alpha/(2|\mathcal{J}|)$.

We emphasize that the construction of the calibration band in Theorem 4.1 only depends on the realizations of the random variables $(Y_i)_{i=1}^n$, the volumes $(v_i)_{i=1}^n$, the dispersion parameter φ and the cumulant function κ. In particular, it does not depend on the means $(\mu_i)_{i=1}^n$, but only on their rankings. Moreover, it holds for any sample size, i.e., it does not rely on asymptotic sample size considerations.

4.2. Choice of the set of ordered pairs and binning of the observations

The statement in Theorem 4.1 holds for any set of ordered pairs $\mathcal{J} \subseteq \{1, \dots, n\}^2$ and, in fact, the resulting calibration band directly depends on the choice of this set. Moreover, note that in principle, $\mathcal{J}$ might only contain a few pairs, which could lead to take the supremum and the infimum of empty sets in (13) and (14). In such cases, we adopt the convention

$$\inf \emptyset = \sup_{\theta \in \mathring{\Theta}} \kappa'(\theta) \quad \text{and} \quad \sup \emptyset = \inf_{\theta \in \mathring{\Theta}} \kappa'(\theta).$$

That is, the underlying lower and upper bounds are equal to the infimum and the supremum of the mean parameter space, respectively. An intuitive choice for $\mathcal{J}$ is the set of all possible combinations of ordered pairs that we denote by

$$\mathcal{J}^{\,\text{full}} = \{(j,k) \in \{1, \dots, n\}^2 \,|\, j \leq k\}.$$

In this case, we call the constructed band the *full calibration band*. Many other choices are possible and we want to discuss two contrasting factors that create a trade-off situation. On the one hand, for a fixed $\delta \in (0,1)$, the lower and upper bounds in (13) and (14) lead to a wider band than the full calibration band for any smaller set of ordered pairs $\mathcal{J} \subseteq \mathcal{J}^{\,\text{full}}$. This suggests that the set $\mathcal{J}$ should be large. On the other hand, the map $\delta \mapsto l^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ is non-decreasing, whereas the map $\delta \mapsto u^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ is non-increasing. Consequently, for a fixed set of ordered pairs $\mathcal{J}$, the resulting band becomes narrower as δ increases. However, since the value of δ is directly determined by the number of elements in the set $\mathcal{J}$ via the relation $\delta = \alpha/(2|\mathcal{J}|)$ in Theorem 4.1, a large set of ordered pairs leads to a low value for δ and vice versa. This creates a trade-off and there is thus no optimal choice for the set of ordered pairs in general.

In their construction of calibration bands for the binary case, Dimitriadis et al. (2023) suggest to use a slightly modified version of $\mathcal{J}^{\,\text{full}}$ that we call

$$\mathcal{J}^{\,\text{distinct}} = \{(j,k) \in \mathcal{D}^2 \,|\, j \leq k\},$$

where $\mathcal{D}$ is any largest subset of $\{1, \dots, n\}$ such that there are no ties in the canonical parameters, i.e., $\theta_i \neq \theta_j$ for all $i \neq j \in \mathcal{D}$. Note that using the convolution property of the EDF, one can always merge observations associated to the same canonical parameter and appropriately adapt the volumes v_i before constructing the calibration band using $\mathcal{J}^{\,\text{distinct}}$; we refer to Corollary 2.15 of Wüthrich and Merz (2023).

Another consideration is the computational time required for constructing the band, which corresponds to $\mathcal{O}(|\mathcal{J}|)$. Indeed, using $\mathcal{J}^{\,\text{full}}$ as the set of ordered pairs leads to a computational time of $\mathcal{O}(n^2)$, which might be problematic for large datasets. One way to improve the run time is to reduce the amount of pairs in the set $\mathcal{J}$. Another way is to reduce the number of observations by binning them even if there are no ties in the canonical parameters. We come back to those methods through the numerical examples in Section 8.

4.3. Crossings inside the calibration bands

Although we call the simultaneous lower and upper bounds on the means $(L^{\alpha}_{\boldsymbol{Y},i}, U^{\alpha}_{\boldsymbol{Y},i})_{i=1}^{n}$ derived in Theorem 4.1 a calibration band, we point out that, in general, we might have $U^{\alpha}_{\boldsymbol{Y},i} < L^{\alpha}_{\boldsymbol{Y},i}$ for some indices $1 \leq i \leq n$. This phenomenon was already observed by Dimitriadis et al. (2023) in the binary case and these authors argue that this typically happens

when the ranking of the responses violates (2), i.e., when a ranking obtained from empirical data is not fully accurate. In order to construct calibration bands that do not exhibit any crossings, Dimitriadis et al. (2023) propose to take the pointwise minimum (maximum) of the lower (upper) band with the isotonic regressor of the responses that is defined by

$$\widehat{\boldsymbol{\mu}}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v}) = \underset{\boldsymbol{\mu}\in\mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} v_i (Y_i - \mu_i)^2 : \mu_1 \leq \cdots \leq \mu_n \right\},$$
$$\text{for } \boldsymbol{v} = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n. \tag{15}$$

The vector $\widehat{\boldsymbol{\mu}}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})$ provides a non-parametric estimator of the means of $\boldsymbol{Y}$ that satisfies the ordering in (2) by construction. In the case where the calibration band shows some crossings, we follow the proposition of Dimitriadis et al. (2023), and we suggest to use the following modified bands

$$\widetilde{L}^{\alpha}_{\boldsymbol{Y},i} = \min\left(L^{\alpha}_{\boldsymbol{Y},i}, \widehat{\mu}_i^{\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})\right) \quad \text{and} \quad \widetilde{U}^{\alpha}_{\boldsymbol{Y},i} = \max\left(U^{\alpha}_{\boldsymbol{Y},i}, \widehat{\mu}_i^{\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})\right), \tag{16}$$

for $1 \leq i \leq n$. Of course, the main result of this section still holds because this makes the interval wider. That is, for any given confidence level $1 - \alpha \in (0, 1)$, we still have

$$\mathbb{P}\left(\widetilde{L}^{\alpha}_{\boldsymbol{Y},i} \leq \mu_i \leq \widetilde{U}^{\alpha}_{\boldsymbol{Y},i} \text{ for all } i \in \{1, \ldots, n\}\right) \geq 1 - \alpha.$$

5. Explicit calibration bands for selected distributions

The calibration band derived in Theorem 4.1 can be constructed for any member of the EDF under Assumptions 3.1 and 3.2. To do so, the evaluation of the lower and upper bounds $l^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ and $u^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))$ requires the use of a root-finding algorithm. For some members of the EDF, these bounds can be calculated in closed form. We give the resulting expressions for the binomial, Poisson, negative binomial, gamma and normal cases in this section. These expressions are derived using closed form characterizations for the quantiles of the above distributions.

We point out that the explicit calibration bands presented in this section could also be derived using fiducial distributions. Fiducial distributions were introduced by Fisher (1935, 1973) in the 1930s, who aimed at providing a framework for constructing probability distributions of unknown parameters based on available observations. The use of fiducial distributions has been shown to lead to some contradictory results; see, for example, Chapter 5.4 in Sprott (2000). Therefore, such distributions may only be used under specific assumptions that are discussed by Pedersen (1978). Veronese and Mellili (2015) compute the fiducial distributions of selected members of the EDF, including the above mentioned examples, and they show that these fiducial distributions satisfy those assumptions. As a consequence, all the explicit calibration bands derived in this section could actually also be derived using fiducial distributions.

5.1. Discrete distributions

The binomial, Poisson, and negative binomial distributions are members of the EDF since any random variable N belonging to one of these distributions can be written as

$$N = \frac{vY}{\varphi}, \tag{17}$$

for $Y \sim \text{EDF}(\theta, v, \varphi, \kappa(\cdot))$ and for a carefully chosen effective domain, volume, dispersion parameter and cumulant function; we refer to Section 3.3 of Jørgensen (1997) and Section 2.2 of Wüthrich and Merz (2023). Note that the transformation (17) is called the duality transformation as it provides a duality between the random variables

$$Y \sim \text{EDF}(\theta, v, \varphi, \kappa(\cdot)) \quad \text{and} \quad N = \frac{vY}{\varphi}.$$

The former random variable Y, whose density is given in (1), is said to be in the *reproductive* form of the EDF, whereas the latter random variable N is said to be in the *additive* form of the EDF; see Section 3.1 of Jørgensen (1997). Using the duality transformation, Theorem 4.1 can be used to derive calibration bands for members of the additive form of the EDF under Assumptions 3.1 and 3.2. We show in the next result that the resulting bands for the binomial, Poisson and negative binomial cases can be given in closed form using explicit expressions for the quantiles of those distributions. Note that the calibration bands derived by Dimitriadis et al. (2023) for the binary case are contained in the binomial case, below.

Proposition 5.1: Suppose that Assumptions 3.1 and 3.2 hold. Let $N_1, \ldots, N_n$ be independent members of the EDF in the additive form, i.e., there exist canonical parameters $(\theta_i)_{i=1}^n$, volumes $(v_i)_{i=1}^n$, a dispersion parameter $\varphi > 0$ and a cumulant function κ such that $Y_i = \varphi N_i / v_i$, for $Y_i \sim \text{EDF}(\theta_i, v_i, \varphi, \kappa(\cdot))$ and $1 \le i \le n$. By writing $\mu_i = \mathbb{E}[Y_i]$ and assuming that $\mu_1 \le \cdots \le \mu_n$, we have for any set of ordered pairs $\mathcal{J} \subseteq \{1, \ldots, n\}^2$ and any confidence level $1 - \alpha \in (0, 1)$ that

$$\mathbb{P}\left(L^{\alpha}_{\boldsymbol{Y},i} \le \mu_i \le U^{\alpha}_{\boldsymbol{Y},i} \text{ for all } i \in \{1, \ldots, n\}\right) \ge 1 - \alpha,$$

where the lower and upper bounds $L^{\alpha}_{\boldsymbol{Y},i}$ and $U^{\alpha}_{\boldsymbol{Y},i}$ are defined in (13)–(14). These bounds can be explicitly expressed in the following three cases using the weighted partial sums $Z_{j:k}$ and aggregated volumes $v_{j:k}$ in (8)–(9).

- *Binomial case*. The lower and upper bounds are given by

$$L^{\alpha}_{\boldsymbol{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \ge \mu_k} q_B(\delta; v_{j:k} Z_{j:k}/\varphi, 1 + v_{j:k}/\varphi - v_{j:k} Z_{j:k}/\varphi)\, \mathbb{1}_{\{Z_{j:k}>0\}}, \tag{18}$$

  and

$$U^{\alpha}_{\boldsymbol{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i \le \mu_j} q_B(1-\delta; 1 + v_{j:k} Z_{j:k}/\varphi, v_{j:k}/\varphi - v_{j:k} Z_{j:k}/\varphi)\, \mathbb{1}_{\{Z_{j:k}<1\}} + \mathbb{1}_{\{Z_{j:k}=1\}}, \tag{19}$$

  for $\delta = \alpha/(2|\mathcal{J}|)$, and where $q_B(\delta; \alpha, \beta)$ denotes the δ-quantile of a beta distribution with parameters $\alpha, \beta > 0$.
- *Poisson case*. The lower and upper bounds are given by

$$L^{\alpha}_{\boldsymbol{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \ge \mu_k} \frac{\varphi\, q_\Gamma(\delta; v_{j:k} Z_{j:k}/\varphi, 1)}{v_{j:k}}\, \mathbb{1}_{\{Z_{j:k}>0\}}, \tag{20}$$

  and

$$U^{\alpha}_{\boldsymbol{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i \le \mu_j} \frac{\varphi\, q_\Gamma(1-\delta; 1 + v_{j:k} Z_{j:k}/\varphi, 1)}{v_{j:k}}, \tag{21}$$

for $\delta = \alpha/(2|\mathcal{J}|)$, and where $q_\Gamma(\delta;\gamma,c)$ denotes the δ-quantile of a gamma distribution with shape parameter $\gamma > 0$ and scale parameter $c > 0$.

- *Negative binomial case*. The lower and upper bounds are given by

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i\geq\mu_k} \frac{q_B(\delta; v_{j:k}Z_{j:k}/\varphi, v_{j:k}/\varphi)}{1-q_B(\delta; v_{j:k}Z_{j:k}/\varphi, v_{j:k}/\varphi)}\,\mathbb{1}_{\{Z_{j:k}>0\}}, \tag{22}$$

and

$$U^{\alpha}_{\mathbf{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i\leq\mu_j} \frac{q_B(1-\delta; 1+v_{j:k}Z_{j:k}/\varphi, v_{j:k}/\varphi)}{1-q_B(1-\delta; 1+v_{j:k}Z_{j:k}/\varphi, v_{j:k}/\varphi)}, \tag{23}$$

for $\delta = \alpha/(2|\mathcal{J}|)$.

5.2. Continuous distributions

The gamma and normal distributions are also members of the EDF. This time, note that any random variable Y belonging to one of these distributions can be directly written as an EDF$(\theta, v, \varphi, \kappa(\cdot))$ random variable for a carefully chosen effective domain, volume, dispersion parameter and cumulant function; we refer to Section 3.3 of Jørgensen (1997) or Section 2.2 of Wüthrich and Merz (2023). That is, the normal and the gamma distributions can be directly expressed in the reproductive of the EDF. Moreover, these distributions satisfy Assumptions 3.1 and 3.2, which allows us to derive calibration bands on the mean of gamma or normal responses. As above, a closed form expression for the calibration bands can be obtained using the quantiles of these distributions.

Proposition 5.2: Let $\mathcal{J} \subseteq \{1,\dots,n\}^2$ be any set of ordered pairs. For any confidence level $1-\alpha \in (0,1)$, the calibration band in Theorem *4.1* can be explicitly expressed in the following two cases using the weighted partial sums $Z_{j:k}$ and aggregated volumes $v_{j:k}$ in (8)–(9).

- *Gamma case*. The lower and upper bounds in (13)–(14) are given by

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i\geq\mu_k} \frac{v_{j:k}/\varphi}{q_\Gamma(1-\delta; v_{j:k}/\varphi, Z_{j:k})}\,\mathbb{1}_{\{Z_{j:k}>0\}}, \tag{24}$$

and

$$U^{\alpha}_{\mathbf{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i\leq\mu_j} \frac{v_{j:k}/\varphi}{q_\Gamma(\delta; v_{j:k}/\varphi, Z_{j:k})}\,\mathbb{1}_{\{Z_{j:k}>0\}}, \tag{25}$$

for $\delta = \alpha/(2|\mathcal{J}|)$.

- *Normal case*. The lower and upper bounds in (13)–(14) are given by

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i\geq\mu_k} Z_{j:k} - \frac{\Phi^{-1}(1-\delta)}{\sqrt{v_{j:k}/\varphi}}, \tag{26}$$

and

$$U^{\alpha}_{\mathbf{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i\leq\mu_j} Z_{j:k} - \frac{\Phi^{-1}(\delta)}{\sqrt{v_{j:k}/\varphi}}, \tag{27}$$

for $\delta = \alpha/(2|\mathcal{J}|)$, and where $\Phi^{-1}(\delta)$ denotes the δ-quantile of the standard normal distribution.

Remark 5.1: Let $Y_1, \dots, Y_n$ be independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables for known standard deviations $\sigma_i > 0$ and indices $1 \le i \le n$ such that $\mu_1 \le \cdots \le \mu_n$. The underlying aggregated volumes and weighted partial sums are given by

$$v_{j:k} = \sum_{i=j}^{k} \frac{1}{\sigma_i^2} \quad \text{and} \quad Z_{j:k} = \frac{1}{v_{j:k}} \sum_{i=j}^{k} v_i Y_i,$$

where we set $\varphi = 1$, because the distribution of an $\mathrm{EDF}(\theta, v, \varphi, \kappa(\cdot))$ distributed random variable only depends on the ratio v/φ. These sums correspond to weighted sums of normal random variables and these weights are determined by Theorem 4.1 for the general EDF case. Note that large weights are given to responses Y_i that are associated to a small variance and vice versa. The resulting weighted partial sums $Z_{j:k}$ are thus scaled sums of $\mathcal{N}(\mu_i/\sigma_i^2, 1/\sigma_i^2)$ random variables. Since the normal distribution has the nice property that any weighted sum of independent normal responses is again normal, other weights could in principle be chosen as, for example,

$$\tilde{v}_{j:k} = \sum_{i=j}^{k} \frac{1}{\sigma_i} \quad \text{and} \quad \widetilde{Z}_{j:k} = \frac{1}{\tilde{v}_{j:k}} \sum_{i=j}^{k} \frac{Y_i}{\sigma_i},$$

which results in weighted partial sums $\widetilde{Z}_{j:k}$ being scaled sums of normal random variables that all have variance 1. The resulting calibration band can be expressed as

$$\widetilde{L}^{\alpha}_{\boldsymbol{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \ge \mu_k} \widetilde{Z}_{j:k} - \frac{\Phi^{-1}(1-\delta)\sqrt{k-j+1}}{\tilde{v}_{j:k}},$$

and

$$\widetilde{U}^{\alpha}_{\boldsymbol{Y},i} = \inf_{(j,k)\in\mathcal{J}\,:\,\mu_i \le \mu_j} \widetilde{Z}_{j:k} - \frac{\Phi^{-1}(\delta)\sqrt{k-j+1}}{\tilde{v}_{j:k}}.$$

This new calibration band is in general different from the one derived in Proposition 5.2. However, both calibration bands coincide in the case of a constant variance for the responses, i.e., when $\sigma_i = \sigma$ for $1 \le i \le n$.

5.3. Comparison with Yang–Barber's calibration bands

In the literature, calibration bands on the mean have first been constructed by Yang and Barber (2019), under the assumption that the responses $(Y_i)_{i=1}^n$ satisfy the additive relation

$$Y_i = \mu_i + \sigma\varepsilon_i,$$

for means $\mu_1 \le \cdots \le \mu_n$, for some fixed and known $\sigma > 0$ and for independent and zero-mean random variables ε_i that are sub-Gaussian, i.e.,

$$\mathbb{P}(|\varepsilon_i| > t) \le 2\,\mathrm{e}^{-t^2/2}, \qquad \text{for all } t > 0 \text{ and for all } 1 \le i \le n.$$

The construction of their calibration bands makes use of the isotonic regressor of $\boldsymbol{Y}$ defined in (15) in order to introduce the empirical partial sums

$$Z_{j:k}^{\mathrm{Iso}} = \frac{1}{k-j+1} \sum_{i=j}^{k} \widehat{\mu}_i^{\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1}),$$

for $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$. Yang and Barber (2019) show that for the set of ordered pairs

$$\mathcal{J}^{\,\mathrm{full}} = \{(j,k) \in \{1, \ldots, n\}^2 \,|\, j \leq k\},$$

and for any confidence level $1 - \alpha \in (0,1)$, we have

$$\mathbb{P}\left(L_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}} \leq \mu_i \leq U_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}} \text{ for all } i \in \{1, \ldots, n\}\right) \geq 1 - \alpha,$$

with

$$L_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}} = \sup_{(j,k) \in \mathcal{J}^{\,\mathrm{full}} \,:\, \mu_i \geq \mu_k} Z_{j:k}^{\mathrm{Iso}} - \frac{\sqrt{2\sigma^2 \log(1/\delta)}}{\sqrt{k-j+1}}, \tag{28}$$

and

$$U_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}} = \inf_{(j,k) \in \mathcal{J}^{\,\mathrm{full}} \,:\, \mu_i \leq \mu_j} Z_{j:k}^{\mathrm{Iso}} + \frac{\sqrt{2\sigma^2 \log(1/\delta)}}{\sqrt{k-j+1}}, \tag{29}$$

for $\delta = \alpha / (2|\mathcal{J}^{\,\mathrm{full}}|) = \alpha/(n^2+n)$.

A particular case of the framework used by Yang and Barber (2019) arises by taking i.i.d. Gaussian random variables $\varepsilon_i \sim \mathcal{N}(0,1)$. In this case, we obtain independent responses $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for $1 \leq i \leq n$. Similar to Dimitriadis et al. (2023) for the binary case, we show that our calibration band is narrower than Yang–Barber's band in this setting.

Theorem 5.3: Let $Y_1, \ldots, Y_n$ be independent $\mathcal{N}(\mu_i, \sigma^2)$ random variables for a known standard deviation $\sigma > 0$ and indices $1 \leq i \leq n$ such that $\mu_1 \leq \cdots \leq \mu_n$. The bands derived in Proposition *5.2* using the set of ordered pairs $\mathcal{J}^{\,\mathrm{full}}$ satisfy

$$L_{\boldsymbol{Y},i}^{\alpha} \geq L_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}} \quad \text{and} \quad U_{\boldsymbol{Y},i}^{\alpha} \leq U_{\boldsymbol{Y},i}^{\alpha,\,\mathrm{YB}},$$

for all $1 \leq i \leq n$ and for any confidence level $1 - \alpha \in (0,1)$.

The proof of this theorem relies on Proposition B1 of Dimitriadis et al. (2023) that characterizes the pairs $(i,j) \in \mathcal{J}^{\,\mathrm{full}}$ for which the maximum in (28) and the minimum in (29) are attained. It is provided in the appendix.

6. Extension to regression modelling

6.1. Regression modelling within the exponential dispersion family

The calibration bands on the means derived in Theorem 4.1 can be extended to regression modelling. To this end, let $(\boldsymbol{\Omega}, \mathcal{F}, \mathbb{P})$ be the underlying probability space and consider an

independent sample $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ with responses Y_i and i.i.d. features $\boldsymbol{X}_i$ satisfying

$$Y_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \;\sim\; \mathrm{EDF}(\theta(\boldsymbol{x}_i), v_i, \varphi, \kappa(\cdot)),$$

for given volumes $v_i > 0$, as well as a dispersion parameter $\varphi > 0$ and a cumulant function κ that do not depend on i. We denote the support of the features $\boldsymbol{X}_i$ by $\mathcal{X}$ and call it the *feature space*. The goal of a regression on the mean is to estimate the true mean function

$$\mu^* : \mathcal{X} \to \kappa'(\overset{\circ}{\Theta}), \quad \boldsymbol{x} \mapsto \kappa'(\theta(\boldsymbol{x})), \tag{30}$$

where the map $\theta : \mathcal{X} \to \overset{\circ}{\Theta}$ is unknown. Note that this true mean function is a strictly increasing map of the canonical parameter $\boldsymbol{x} \mapsto \theta(\boldsymbol{x})$ due to the strict convexity of the cumulant function κ under Assumptions 3.1 and 3.2; see (30). In particular, the true mean function does not depend on the volume and the dispersion parameter. Therefore, one can write

$$\mu^*(\boldsymbol{X}) = \mathbb{E}[Y \,|\, \boldsymbol{X}], \quad \mathbb{P}\text{-a.s.}, \tag{31}$$

for any pair $(Y, \boldsymbol{X})$ satisfying

$$Y \,|\, \boldsymbol{X} = \boldsymbol{x} \;\sim\; \mathrm{EDF}(\theta(\boldsymbol{x}), v, \varphi, \kappa(\cdot)),$$

regardless of the volume $v > 0$ and the dispersion parameter $\varphi > 0$; we refer to Section 3. That is, the true mean function $\mu^* : \mathcal{X} \to \kappa'(\overset{\circ}{\Theta})$ maps each feature $\boldsymbol{x} \in \mathcal{X}$ to the conditional expectation of the response Y, given this feature is observed.

6.2. Construction of the calibration bands

In regression modelling, a calibration band on the mean denotes a set of lower and upper bounds such that the probability that the true mean function in (30) lies simultaneously inside these bounds for almost every (a.e.) feature $\boldsymbol{x} \in \mathcal{X}$ exceeds a given confidence level. As in Section 4, where we required the responses to be ordered such that their canonical parameters are increasing, the construction of this band is based on the assumption of knowing a ranking function that indicates the ordering of the true mean function for a.e. $\boldsymbol{x} \in \mathcal{X}$.

Assumption 6.1: There exists a measurable ranking function $\pi : \mathcal{X} \to \mathbb{R}$ and a version of the true mean function $\mu_\pi^* : \mathcal{X} \to \mathbb{R}$, i.e., a regression function satisfying

$$\mu_\pi^*(\boldsymbol{X}) = \mu^*(\boldsymbol{X}), \quad \mathbb{P}\text{-a.s.},$$

such that for any two features $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$\pi(\boldsymbol{x}) \le \pi(\boldsymbol{x}') \implies \mu_\pi^*(\boldsymbol{x}) \le \mu_\pi^*(\boldsymbol{x}'). \tag{32}$$

Because the conditional mean in (31) is only given $\mathbb{P}$-a.s., we only require μ_π^* to align with μ^* $\mathbb{P}$-a.s. In other words, the ranking function (32) can be chosen such that it complies with the ranking of the true mean function μ^* for a.e. feature $\boldsymbol{x} \in \mathcal{X}$.

The existence of such a ranking function is clear as $\pi(\boldsymbol{x}) = \mu^*(\boldsymbol{x})$ fulfills (32). In fact, there are infinitely many ranking functions since, for example, any positive affine transformation of a ranking function is again a ranking function. The crucial point is that we assume to know at

least one of these functions. Moreover, note that the above assumption is actually equivalent to saying that there exists a non-decreasing map $G : \mathbb{R} \to \mathbb{R}$ such that

$$G(\pi(\boldsymbol{x})) = \mu^*_\pi(\boldsymbol{x}),$$

for every $\boldsymbol{x} \in \mathcal{X}$. Under Assumption 6.1, we aim at constructing a data-dependent calibration band

$$\left(L^\alpha_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}), U^\alpha_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x})\right)_{\boldsymbol{x}\in\mathcal{X}},$$

i.e., a band depending on the realizations of the responses $(Y_i)_{i=1}^n$ and the features $(\boldsymbol{X}_i)_{i=1}^n$ such that

$$\mathbb{P}\left(L^\alpha_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \le \mu^*_\pi(\boldsymbol{x}) \le U^\alpha_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X}\right) \ge 1-\alpha,$$

for any given confidence level $1-\alpha \in (0,1)$. To do so, we further make the following assumption.

Assumption 6.2: The map

$$\mathbb{Q} : \Omega \times \mathcal{F} \to [0,1], \quad \mathbb{Q}(\omega, A) = \mathbb{E}[\mathbb{1}_A \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n](\omega),$$

is a regular conditional probability of $\mathbb{P}$, given the features $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

This assumption fails to hold in general, and we refer to Section 3.2 of Rao and Swift (2006) for necessary conditions ensuring the existence of this regular conditional probability. Moreover, we emphasize that Assumption 6.2 means that the map

$$A \in \mathcal{F} \mapsto \mathbb{Q}_{(\boldsymbol{x}_i)_{i=1}^n}(A) = \mathbb{P}(A \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n),$$

is a probability measure on $(\Omega, \mathcal{F})$ for any realization $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the features $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In particular, we have for any $A \in \mathcal{F}$,

$$\begin{aligned}\mathbb{P}(A) &= \int \mathbb{P}(A \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n)\, \mathrm{d}\mathbb{P}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\ &= \int \mathbb{Q}_{(\boldsymbol{x}_i)_{i=1}^n}(A)\, \mathrm{d}\mathbb{P}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n). \end{aligned} \tag{33}$$

Denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the observed features. The calibration band constructed in Theorem 4.1 can now be extended to regression modelling under the probability measure $\mathbb{Q}_{(\boldsymbol{x}_i)_{i=1}^n}$. For this, let $\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n} : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be any permutation on the indices such that for the given ranking function π, we have

$$\pi(\boldsymbol{x}_{\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(1)}) \le \cdots \le \pi(\boldsymbol{x}_{\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(n)}). \tag{34}$$

We point out that such a permutation always exists and depends on the given ranking function π. However, as for the ordering assumed in (2), the map $\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$ is in general not unique. We use it in order to rank the responses according to their conditional means and define the following weighted partial sums that depend on the responses $(Y_i)_{i=1}^n$, the features $(\boldsymbol{X}_i)_{i=1}^n$

and the ranking function π through the permutation $\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$. We drop the subscript of the permutation function for convenience and write

$$Z_{j:k} = \frac{1}{v_{j:k}} \sum_{i=j}^{k} v_{\tau(i)} Y_{\tau(i)},$$

for $1 \leq j \leq k \leq n$, with aggregated volumes

$$v_{j:k} = \sum_{i=j}^{k} v_{\tau(i)}.$$

Theorem 6.3: Suppose that Assumptions 3.1, 3.2, 6.1 and 6.2 hold. Let $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ be independent random variables such that $Y_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \sim \mathrm{EDF}(\theta(\boldsymbol{x}_i), v_i, \varphi, \kappa(\cdot))$ for i.i.d. features $(\boldsymbol{X}_i)_{i=1}^n$ and given volumes $v_i > 0$. Moreover, let $\mathcal{J} \subseteq \{1,\ldots,n\}^2$ be any set of ordered pairs. Then, for a.e. realization of the features $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ and any given confidence level $1-\alpha \in (0,1)$, we have

$$\mathbb{Q}_{(\boldsymbol{x}_i)_{i=1}^n}\left(L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \leq \mu^*_{\pi}(\boldsymbol{x}) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X}\right) \geq 1-\alpha, \tag{35}$$

where

$$L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) = \sup_{(j,k)\in\mathcal{J}} \left[I^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\, \mathbb{1}_{\{\pi(\boldsymbol{x}) \geq \pi(\boldsymbol{X}_{\tau(k)})\}} + \inf_{\theta\in\overset{\circ}{\Theta}} \kappa'(\theta)\, \mathbb{1}_{\{\pi(\boldsymbol{x}) < \pi(\boldsymbol{X}_{\tau(k)})\}} \right],$$

and

$$U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) = \inf_{(j,k)\in\mathcal{J}} \left[u^{\delta}(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\, \mathbb{1}_{\{\pi(\boldsymbol{x}) \leq \pi(\boldsymbol{X}_{\tau(j)})\}} + \sup_{\theta\in\overset{\circ}{\Theta}} \kappa'(\theta)\, \mathbb{1}_{\{\pi(\boldsymbol{x}) > \pi(\boldsymbol{X}_{\tau(j)})\}} \right],$$

for $\boldsymbol{x} \in \mathcal{X}$ and $\delta = \alpha/(2|\mathcal{J}|)$.

We emphasize again that the construction of this calibration band on the mean is similar to Theorem 4.1 and relies on the ranking function $\pi : \mathcal{X} \to \mathbb{R}$. Moreover, the statement in (35) can be rewritten as

$$\mathbb{P}\left(L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \leq \mu^*_{\pi}(\boldsymbol{x}) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X} \,\middle|\, \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \geq 1-\alpha.$$

That is, given a.e. realization of the features, the probability that the realizations of the underlying responses lead to a calibration band being able to fully bound the mean function $\mu^*_{\pi} : \mathcal{X} \to \mathbb{R}$ for all $\boldsymbol{x} \in \mathcal{X}$ exceeds $1-\alpha$. Due to (33), a corollary of Theorem 6.3 is that

$$\mathbb{P}\left(L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \leq \mu^*_{\pi}(\boldsymbol{x}) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X}\right) \geq 1-\alpha, \tag{36}$$

for any given confidence level $1-\alpha \in (0,1)$. We emphasize, however, that the conditional probability bound in Theorem 6.3 is stronger than this inequality as it holds for a.e. fixed and

known realization of the features $(\boldsymbol{X}_i)_{i=1}^n$, i.e., when only the responses $(Y_i)_{i=1}^n$ are random. As the mean function $\mu_\pi^* : \mathcal{X} \to \mathbb{R}$ was assumed to be a version of the true mean function $\mu^* : \mathcal{X} \to \kappa'(\overset{\circ}{\Theta})$, note that the calibration band constructed in this section provides a bound on the true mean function for a.e. feature $\boldsymbol{x} \in \mathcal{X}$ with a probability exceeding the confidence level $1 - \alpha$. This leads to the construction of the statistical tests being introduced in the next section.

7. Statistical testing of calibration and auto-calibration

7.1. The auto-calibration property

In regression modelling, the true mean function in (30) is approximated by a regression function that we denote by $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$. This regression function is said to be calibrated if it matches the true mean function for a.e. realization of the features $\boldsymbol{x} \in \mathcal{X}$, i.e.,

$$\widehat{\mu}(\boldsymbol{x}) = \mu^*(\boldsymbol{x}), \quad \text{for a.e. } \boldsymbol{x} \in \mathcal{X}. \tag{37}$$

As the true mean function often exhibits a complex behaviour w.r.t. the features $\boldsymbol{x} \in \mathcal{X}$ and the mean estimation task is performed over a finite sample of (noisy) observations, it is in general impossible to aim for a calibrated regression function in practice. A related notion was introduced in the literature under the name of *auto-calibration*. It is defined as follows; see e.g., Krüger and Ziegel (2021).

Definition 7.1: A regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ is auto-calibrated for $(Y, \boldsymbol{X})$ if

$$\widehat{\mu}(\boldsymbol{X}) = \mathbb{E}[Y \mid \widehat{\mu}(\boldsymbol{X})], \quad \mathbb{P}\text{-a.s.}$$

Note that for any calibrated regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}[Y \mid \widehat{\mu}(\boldsymbol{X})] = \mathbb{E}[\mathbb{E}[Y \mid \boldsymbol{X}] \mid \widehat{\mu}(\boldsymbol{X})] = \mathbb{E}[\mu^*(\boldsymbol{X}) \mid \widehat{\mu}(\boldsymbol{X})] = \widehat{\mu}(\boldsymbol{X}), \quad \mathbb{P}\text{-a.s.}, \tag{38}$$

where in the first equality, we use that $\sigma(\widehat{\mu}(\boldsymbol{X})) \subseteq \sigma(\boldsymbol{X})$ and in the last equality, that the regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ is calibrated. The auto-calibration property means that the expected value of responses, for which the associated features are mapped to the same estimate under the regression function, matches this estimate. An auto-calibrated regression function thus guarantees that if we partition the feature space $\mathcal{X}$ into subsets according to the estimated regression values $(\widehat{\mu}(\boldsymbol{x}))_{\boldsymbol{x} \in \mathcal{X}}$, the mean of all the responses within such a subset matches the estimated mean for this subset, leading to locally unbiased estimates. While this notion is weaker than calibration, it is of particular interest in several applications, where mean estimates within given specific groups have to be unbiased. For example, this is the case in insurance pricing, where the auto-calibration of a regression function is a minimal requirement, as it ensures that each cohort of individuals paying a certain price remains self-financing; we refer to Wüthrich and Ziegel (2024).

7.2. Statistical tests for calibration

The calibration band derived in Theorem 6.3 can be used to construct a statistical test for calibration with confidence level $1 - \alpha \in (0, 1)$ as for any ranking function $\pi : \mathcal{X} \to \mathbb{R}$, (37)

is equivalent to

$$\widehat{\mu}(\boldsymbol{x}) = \mu^*_\pi(\boldsymbol{x}), \qquad \text{for a.e. } \boldsymbol{x} \in \mathcal{X}.$$

Moreover, the set

$$\left\{ L^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \le \mu^*_\pi(\boldsymbol{x}) \le U^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X} \right\}$$

is included in the set

$$\left\{ L^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \le \mu^*_\pi(\boldsymbol{x}) \le U^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \mathcal{X} \right\},$$

when $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ is calibrated and the former holds with probability greater or equal than $1-\alpha$; see (36). Therefore, by writing $(y_i, \boldsymbol{x}_i)_{i=1}^n$ for the observed realizations of the responses and features and

$$\left(L^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}), U^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \right)_{\boldsymbol{x}\in\mathcal{X}},$$

for the resulting calibration band, we reject the null-hypothesis

$$\mathbb{H}_0 : \widehat{\mu}(\boldsymbol{x}) = \mu^*(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \mathcal{X}, \tag{39}$$

with confidence level $1-\alpha$ whenever the set

$$\mathcal{X}^{\text{ out}} = \left\{ \boldsymbol{x} \in \mathcal{X} \;\middle|\; \widehat{\mu}(\boldsymbol{x}) \notin \left[L^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}), U^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}) \right] \right\} \tag{40}$$

satisfies $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}^{\text{ out}}) > 0$. That is, we reject the null-hypothesis of calibration of a model whenever the set of features for which the mean estimates fall outside the band has non-zero probability under the distribution of the features $\boldsymbol{X} \in \mathcal{X}$.

We propose a procedure in order to graphically determine the decision induced by this statistical test. First, the calibration band can be plotted against the ranking function, which results in two non-decreasing step functions delimiting the band. Then, the pairs $(\pi(\boldsymbol{x}), \widehat{\mu}(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$ can be drawn in the same plot for all features $\boldsymbol{x} \in \mathcal{X}$, and we reject the null-hypothesis $\mathbb{H}_0$ whenever the set of pairs falling outside the calibration band is large enough, i.e., whenever it corresponds to a set of features being larger than a nullset. We call such a plot a *calibration plot*, and emphasize that as the distribution of $\boldsymbol{X}$ is unknown in practice, the decision of rejecting or not the above null-hypothesis requires some assumptions on the support of the random variable $\widehat{\mu}(\boldsymbol{X})$; we come back to this in Section 8, below.

We also point out that the decision of the statistical test depends on the ranking function through the constructed calibration band. In practice, there are a few cases where a ranking function is known, and in such cases, statistical techniques under order restrictions could be used for mean estimation; we refer to Barlow et al. (1972) and Robertson et al. (1988). Note that one of these techniques is isotonic regression, which has the nice property to lead to empirically auto-calibrated regression functions. Most of the time, however, we do not have access to any ranking function giving the ordering of the true mean function over the feature space $\mathcal{X}$. In such cases, the ranking function needs to be approximated. We start our discussion from a related work, where Wüthrich and Ziegel (2024) propose a method to restore the auto-calibration of a given regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$. For this, they perform an isotonic regression by using the regression function itself as a ranking function and they

call their method the *isotonic recalibration* step because in a first step, a regression function is estimated and under the assumption that it provides the correct ranking, this ranking is lifted to be (empirically) auto-calibrated in a second (isotonic recalibration) step. We make the same choice here and take $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ as a ranking function.

Assumption 7.2: The regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ satisfies that there exists a version of the true mean function $\mu^*_{\widehat{\mu}} : \mathcal{X} \to \mathbb{R}$, i.e., a regression function satisfying

$$\mu^*_{\widehat{\mu}}(\boldsymbol{X}) = \mu^*(\boldsymbol{X}), \quad \mathbb{P}\text{-a.s.},$$

such that for any two features $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$\widehat{\mu}(\boldsymbol{x}) \leq \widehat{\mu}(\boldsymbol{x}') \implies \mu^*_{\widehat{\mu}}(\boldsymbol{x}) \leq \mu^*_{\widehat{\mu}}(\boldsymbol{x}').$$

This choice was also implicitly made by Dimitriadis et al. (2023) for the binary case and, actually, we emphasize that Assumption 7.2 is nested in the null-hypothesis (39), i.e., it holds under $\mathbb{H}_0$. This implies, in particular, that the above statistical test can always be applied by taking the regression function itself as a ranking function.

Finally, as pointed out by Dimitriadis et al. (2023), an opposite statistical test can be constructed where the calibration property now lies in the alternative. This test allows one to quantify deviations from calibration as its null-hypothesis reads as

$$\mathbb{H}_0 : \left|\widehat{\mu}(\boldsymbol{x}) - \mu^*(\boldsymbol{x})\right| > \varepsilon \text{ for a.e. } \boldsymbol{x} \in \mathcal{X}, \tag{41}$$

for some given $\varepsilon > 0$. This hypothesis is rejected with confidence level $1 - \alpha$, whenever the set

$$\mathcal{X}_\varepsilon^{\text{in}} = \left\{\boldsymbol{x} \in \mathcal{X} : \left[L^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x}), U^\alpha_{\pi,(y_i,\boldsymbol{x}_i)_{i=1}^n}(\boldsymbol{x})\right] \subseteq [\widehat{\mu}(\boldsymbol{x}) - \varepsilon, \widehat{\mu}(\boldsymbol{x}) + \varepsilon]\right\} \tag{42}$$

satisfies $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}_\varepsilon^{\text{in}}) > 0$. As above, note that the evaluation of this test can be done graphically by plotting the calibration band and the estimated means against the ranking function. Moreover, the previous discussion about the choice of the ranking function also equivalently applies here.

7.3. Statistical tests for auto-calibration

In order to construct statistical tests for the auto-calibration property of a given regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, we assume that Assumption 7.2 holds. That is, the regression function manages to correctly provide the ordering of the true mean function. Interestingly, we can show that under this assumption, calibration is equivalent to auto-calibration; see Proposition 5.1 in Denuit and Trufin (2023). This means that any auto-calibrated regression function managing to correctly express the ranking of the means over the feature space is actually equal to the true mean function for a.e. feature $\boldsymbol{x} \in \mathcal{X}$. Thus, the tests (39) and (41) do not only provide a test for calibration but also for auto-calibration under Assumption 7.2. The first test consists in rejecting the auto-calibration of a regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ with

null-hypothesis

$$\mathbb{H}_0 : \mathbb{E}[Y \mid \widehat{\mu}(\boldsymbol{X}) = \widehat{\mu}(\boldsymbol{x})] = \widehat{\mu}(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \mathcal{X},$$

whenever the set $\mathcal{X}^{\text{ out}}$ in (40) satisfies $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}^{\text{ out}}) > 0$. The second test consists in rejecting the null-hypothesis

$$\mathbb{H}_0 : |\mathbb{E}[Y \mid \widehat{\mu}(\boldsymbol{X}) = \widehat{\mu}(\boldsymbol{x})] - \widehat{\mu}(\boldsymbol{x})| > \varepsilon \text{ for a.e. } \boldsymbol{x} \in \mathcal{X},$$

whenever the set $\mathcal{X}_{\varepsilon}^{\text{ in}}$ in (42) satisfies $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}_{\varepsilon}^{\text{ in}}) > 0$. Finally, we conclude this section by highlighting that the derived statistical tests for calibration and auto-calibration also apply to the framework of Section 4. Indeed, by choosing as feature space $\mathcal{X} = \{1, \ldots, n\}$ and an appropriate ranking function $\pi : \mathcal{X} \to \{1, \ldots, n\}$ such that $\theta(\boldsymbol{x}_i) = \theta_{\pi(\boldsymbol{x}_i)}$ holds for all $1 \leq i \leq n$, the mean estimation task in Section 4 can be expressed as a regression modelling problem.

7.4. Impact of the dispersion parameter

The calibration band on the mean derived in Theorem 6.3 holds for EDF responses, for which the dispersion parameter φ and the cumulant function κ are known and fixed. The cumulant function uniquely determines the distribution of the responses, while the dispersion parameter characterizes their variances. Indeed, the variance of an EDF random variable $Y \sim \text{EDF}(\theta, v, \varphi, \kappa(\cdot))$ for $\theta \in \overset{\circ}{\Theta}$ is given by

$$\text{Var}(Y) = \frac{\varphi}{v} \kappa''(\theta),$$

and it can equivalently be expressed in terms of the mean $\mu = \mathbb{E}[Y]$ using the canonical link in (4) through

$$\text{Var}(Y) = \frac{\varphi}{v} \kappa''(h(\mu)) = \frac{\varphi}{v} V(\mu),$$

where $V = \kappa'' \circ h$ is the *variance function* of the chosen distribution within the EDF. The larger the dispersion parameter φ is, the larger the variance of the underlying response will be. Thus, the constructed calibration bands depend on the value of φ, and this dependence actually lies in the lower and upper bounds in (6) and (7).

Using a suitable dispersion estimate is thus crucial in order to construct the above statistical tests as, in practice, the dispersion parameter is often unknown. There are various methods for estimating this parameter; see Section 5.3.1 of Wüthrich and Merz (2023). We introduce one of these methods here, which consists in computing the *Pearson's estimate* that is given by

$$\widehat{\varphi}^{\text{P}} = \frac{1}{n-q} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}(\boldsymbol{X}_i))^2}{V(\widehat{\mu}(\boldsymbol{X}_i))/v_i},$$

where q denotes the number of unknown parameters that are estimated in order to derive the regression function $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$. Note that when this regression function is obtained from a well-specified generalized linear model (GLM), Pearson's estimate has the advantage of providing a consistent estimator for φ; see Section 8.3.6 of McCullagh and Nelder (1983).

8. Numerical examples

This section provides numerical examples where we construct calibration bands on the mean of given responses. Our goal is first to highlight the impact of different factors on the resulting calibration bands, as the choice of the confidence level and the set of ordered pairs, or the influence of binning observations. Then, we study a lime trees real dataset and construct a calibration band on the mean of the foliage biomass of the trees. We show that for this small dataset, the statistical test for auto-calibration introduced in Section 7.3 manages to detect violations of auto-calibration in contrast to the test of Denuit et al. (2024). After that, we discuss the impact of estimating the dispersion parameter on the resulting calibration band for simulated inverse Gaussian responses. We then consider a popular French motor third party liability real dataset and construct a calibration band on the claim frequency of the insured drivers. We show that for this large dataset of more than half a million insurance policies, the resulting band allows us to detect violations of calibration in this example and that the isotonic recalibration step proposed by Wüthrich and Ziegel (2024) addresses this issue. Finally, we discuss the power of the statistical test for calibration presented in Section 7.2 by considering the same simulated example as Wüthrich (2024).

8.1. Example 1: calibration bands on the mean of simulated normal responses

In this first example, we consider simulated normal responses and aim at assessing the calibration of mean estimates that are obtained using the isotonic estimator introduced in (15). For this, we sample $n = 2000$ independent normal random variables $Y_i \sim \mathcal{N}(\mu_i, \sigma_i)$, where the means are equally spaced over the range $[1500, 2500]$, i.e.,

$$\mu_i = \mathbb{E}[Y_i] = 1500 + \frac{i-1}{n-1} \cdot 1000, \qquad \text{for } 1 \le i \le n,$$

and where the standard deviations are chosen to satisfy $\sigma_i = 0.5\mu_i$. This choice of the parameters $(\mu_i, \sigma_i)_{i=1}^n$ leads to a coefficient of variation being constant for all responses as we have

$$\mathrm{Vco}(Y_i) = \frac{\sqrt{\mathrm{Var}(Y_i)}}{\mathbb{E}[Y_i]} = \frac{\sigma_i}{\mu_i} = \frac{1}{2}, \qquad \text{for } 1 \le i \le n.$$

Then, we estimate the means of the responses using the simulated responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ based on the isotonic mean estimator $\widehat{\boldsymbol{\mu}}^{\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})$, i.e., we set

$$\widehat{\mu}_i = \widehat{\mu}^{\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})_i, \qquad \text{for } 1 \le i \le n,$$

where $\boldsymbol{v} = (1/\sigma_1^2, \ldots, 1/\sigma_n^2)^\top$; we refer to Remark 5.1 for the choice of the volumes $\boldsymbol{v}$. In order to assess the calibration of the mean estimates $(\widehat{\mu}_i)_{i=1}^n$, we construct a full calibration band on the mean of the above responses using the ordering of their true means and a confidence level of $1 - \alpha = 0.95$. The resulting calibration plot is provided in Figure 1. As all the mean estimates (black dots) lie within the band (red lines), the conclusion of the test in (39) is not to reject the calibration assumption of the isotonic mean estimator $\widehat{\boldsymbol{\mu}}^{\mathrm{Iso}}(\boldsymbol{Y}, \boldsymbol{v})$ in this example. Note that the decision of the performed test depends on the confidence level and the set of ordered pairs used to construct the calibration band. We show the impact of these factors on the band below and we additionally look at the effect of binning observations.

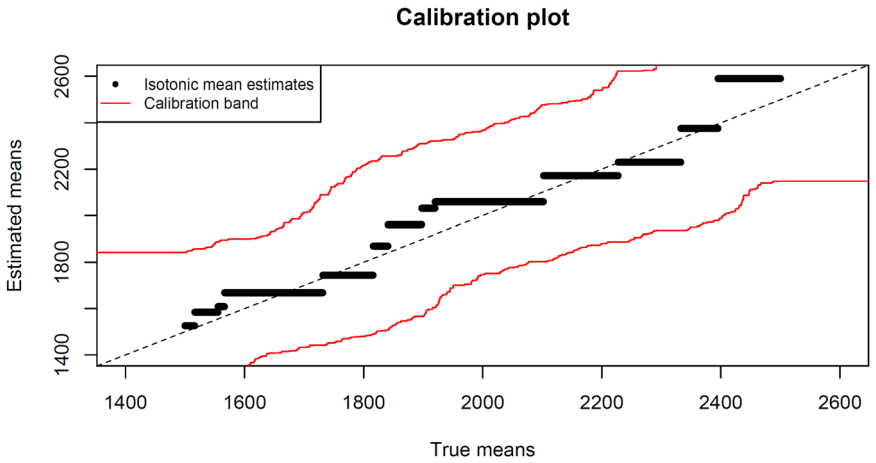

Figure 1. Calibration plot of the isotonic mean estimates of independent normal responses $(Y_i)_{i=1}^n$. The calibration band is plotted in red, whereas the mean estimates are drawn in black.

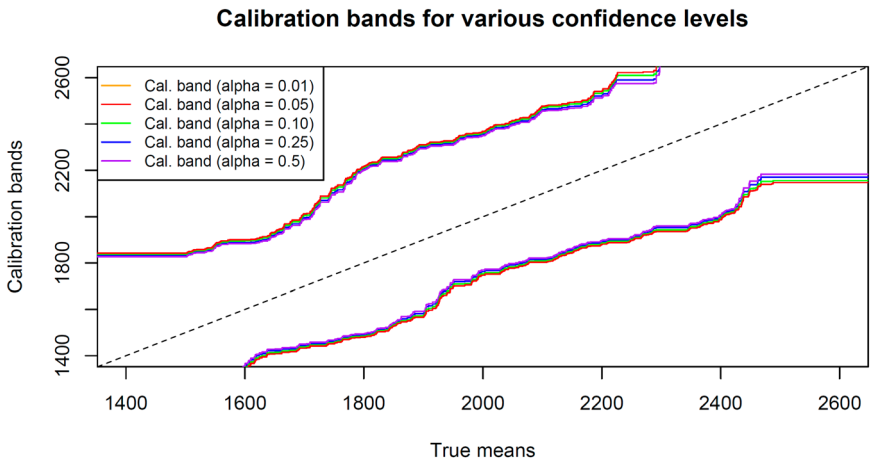

Figure 2. Calibration bands on the mean of independent normal responses $(Y_i)_{i=1}^n$ for various confidence levels.

8.1.1. Sensitivity with respect to the confidence level

The width of the calibration band depends on the chosen confidence level. We evaluate this impact in this example by providing full calibration bands on the mean of the above simulated responses for various confidence levels $1 - \alpha \in \{0.99, 0.95, 0.9, 0.75, 0.5\}$ in Figure 2.

As expected, the calibration bands get narrower as the value of α increases. However, we point out that the value of the confidence level seems not to lead to significant impacts on the width of the calibration band. That is, the constructed bands are not very sensitive to the confidence level, which indicates that the union bound inequality in (12) is not very sharp in this example.

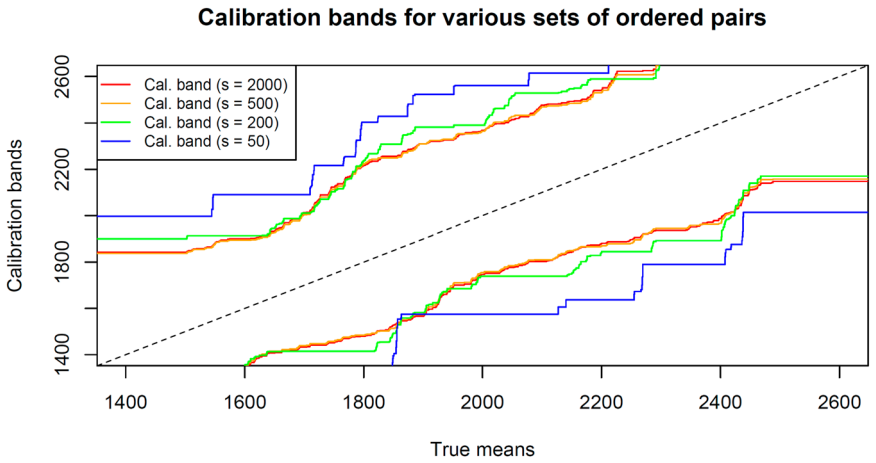

Figure 3. Calibration bands on the mean of independent normal responses $(Y_i)_{i=1}^n$ that are constructed using the set $\mathcal{J}_s^{\text{nbh}}$ for various sizes s.

Table 1. Time (seconds) required to construct the calibration bands on the mean of independent normal responses $(Y_i)_{i=1}^n$ using the set $\mathcal{J}_s^{\text{nbh}}$ for various sizes s.

s	2000	500	200	50
Time (s)	119.81	25.16	9.86	2.72

8.1.2. Impact of the chosen set of ordered pairs

We now study the impact of constructing calibration bands with sets of ordered pairs that are different from $\mathcal{J}^{\text{full}}$. For this, we use again the above simulated observations and fix the confidence level at $1-\alpha = 0.95$. As discussed in Section 4.2, note that the choice of a smaller set of ordered pairs enables us to reduce the computational time required to construct the band. We first consider a set of nearest neighbours (nbh) given by

$$\mathcal{J}_s^{\text{nbh}} = \left\{(j,k) \in \mathcal{J}^{\text{full}} \,\middle|\, k-j \leq s\right\},$$

for $s \in \mathbb{N}$. The use of such a set of ordered pairs can be justified by the intuition that the means of the weighted partial sums $Z_{j:k}$ are too far from μ_j and μ_k when the difference $k-j$ is large, implying the bounds in Proposition 3.7 not to be very sharp. We construct calibration bands using the set $\mathcal{J}_s^{\text{nbh}}$ for different sizes $s \in \{50, 200, 500, 2000\}$ and plot these bands in Figure 3. Moreover, the computational time required to construct the bands is shown in Table 1.

Interestingly, it seems that although small sizes s lead to small sets of ordered pairs $\mathcal{J}_s^{\text{nbh}}$, the constructed calibration bands seem close to each other for $s \in \{200, 500, 2000\}$, whereas the band is significantly wider for the case $s = 50$. This might be due to the underlying aggregated volumes $v_{j:k}$ that fail to be large enough in order to obtain a suitable band in the latter case.

Next, we consider another set of ordered pairs based on the distance (dist) between the available mean estimates $(\widehat{\mu}_i)_{i=1}^n$

$$\mathcal{J}_d^{\text{dist}} = \left\{(j,k) \in \mathcal{J}^{\text{full}} \,\middle|\, |\widehat{\mu}_i - \widehat{\mu}_j| \leq d\right\}, \tag{43}$$

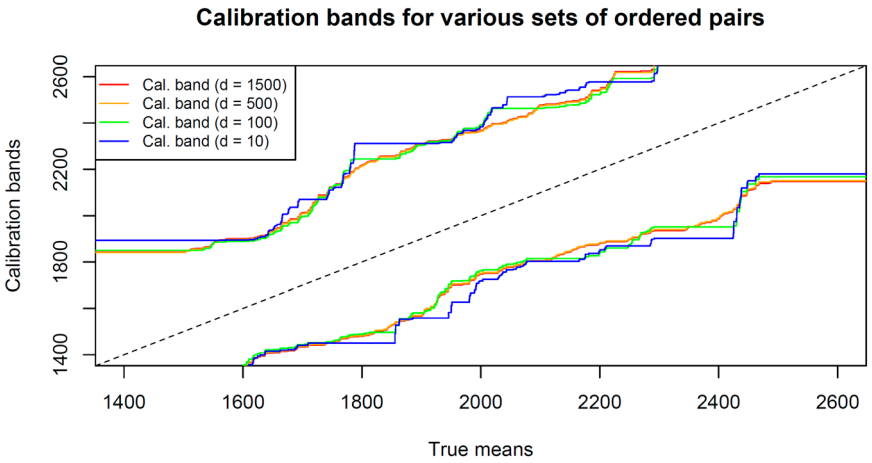

Figure 4. Calibration bands on the mean of independent normal responses $(Y_i)_{i=1}^n$ that are constructed using the set $\mathcal{J}_d^{\text{dist}}$ for various distances d.

Table 2. Time (seconds) required to construct the calibration bands on the mean of independent normal responses $(Y_i)_{i=1}^n$ using the set $\mathcal{J}_d^{\text{dist}}$ for various distances d.

d	1500	500	100	10
Time (s)	124.62	58.98	12.62	6.54

for $d \in \mathbb{R}$. The idea behind such a choice is to only take pairs into account, for which we believe that the means of the responses are close to each other. By using the above isotonic mean estimates in order to define the set $\mathcal{J}_d^{\text{dist}}$, i.e., by setting $\widehat{\mu}_i = \widehat{\mu}^{\text{Iso}}(\boldsymbol{Y}, \boldsymbol{\nu})_i$ for $1 \leq i \leq n$ in (43), we construct calibration bands for various distances $d \in \{10, 100, 500, 1500\}$ and plot them in Figure 4. The computational time required to construct these bands is provided in Table 2. This time, we notice that all the calibration bands have a similar width for almost all distances d, except for the case $d = 10$ where the band seems to be wider at some specific locations. That is, for both restricted sets of ordered pairs we consider in this example, our results show that constructing bands using smaller sets of ordered pairs than $\mathcal{J}^{\text{full}}$ is computationally more efficient, and these smaller sets lead to suitable calibration bands as long as their size is not too small.

8.1.3. *Impact of binning observations*

Another method for reducing the computational time needed to construct the calibration bands consists in binning observations; we refer to Section 4.2. To understand the impact of this method on the resulting bands in this example, we set the confidence level to $1 - \alpha = 0.95$, choose $\mathcal{J}^{\text{full}}$ as the set of ordered pairs, and use the same independent normal observations $(y_i)_{i=1}^n$ as above. Moreover, we create L equally sized bins in order to define new observations $(\tilde{y}_l)_{l=1}^L$ that satisfy

$$\tilde{y}_l = \sum_{k=n(l-1)/L+1}^{nl/L} \frac{\nu_k y_k}{\nu_{n(l-1)/L+1:nl/L}}, \tag{44}$$

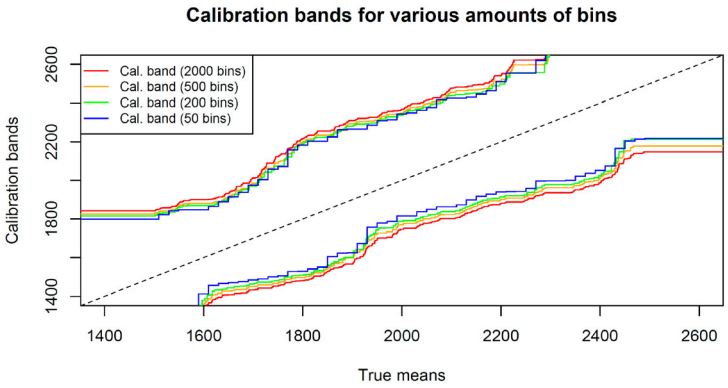

Figure 5. Calibration bands on the mean constructed by binning independent normal responses $(Y_i)_{i=1}^n$ for different amounts of bins.

Table 3. Time (seconds) required to construct the calibration bands on the mean of binned independent normal responses $(Y_i)_{i=1}^n$ for different amounts of bins.

l	2000	500	200	50
Time (s)	122.66	4.31	1.15	0.63

with

$$v_i = \frac{1}{\sigma_i^2}, \qquad \text{for } 1 \le i \le n.$$

That is, the new observations are weighted sums of the original ones, with weights being equal to the volumes of the original observations; see Remark 5.1. Under the assumption that the means of the responses within a given bin are equal, note that this weighting implies that the binned responses belong to the EDF due to the convolution formula in Corollary 2.15 of Wüthrich and Merz (2023). In general, we emphasize that the new observations are not realizations of EDF random variables as this assumption might be violated. In this example, however, this assumption is not needed as binned normal responses are always normally distributed, regardless of the chosen weights. Nonetheless, we still choose the volumes of the original observations as weights in (44) and construct calibration bands using as rankings the true means of the binned responses for $L \in \{50, 200, 500, 2000\}$. The resulting bands are provided in Figure 5 and the computational times required to construct them are given in Table 3.

We notice that all bin sizes lead to pretty similar calibration bands in this example and actually, we can even observe in Figure 5 that the bands get narrower the smaller the number of bins is. As discussed in Section 4.2, this might be a consequence of having a small number of observations, implying the set of ordered pairs $\mathcal{J}^{\text{full}}$ to be small. In fact, the ratio

$$\frac{\Phi^{-1}\left(\frac{0.05}{2000^2+2000}\right)}{\Phi^{-1}\left(\frac{0.05}{50^2+50}\right)} = 1.355$$

hints that the band constructed using 2000 bins should be approximately 1.355 wider than the band constructed using only 50 bins; see (26) and (27). However, this is not the case in

Table 4. AIC of the gamma and inverse Gaussian GLMs.

	Gamma GLM	Inverse Gaussian GLM
AIC	750.33	1089.50

Figure 5 due to the role played by the weighted partial sums $Z_{j:k}$ and the aggregated volumes $v_{j:k}$ that are used to construct the calibration bands.

Together with Section 8.1.2, this section shows that binning observations or choosing suitable sets of ordered pairs can be an interesting technique to reduce the computational time required to construct the calibration bands. In this example, it seems that the chosen sets of ordered pairs enable to reduce the running time of the construction at the cost of having slightly wider calibration bands, whereas binning observations leads to even narrower bands. More generally, we point out that the choice of the set of ordered pairs or the size of the bins has to be carefully made in practice as the true means of the responses are unknown and might not be evenly distributed over the range of interest. Moreover, note that while the inequality (12) holds for any chosen set of ordered pairs, it does not hold anymore when the observations are binned since the distribution of the underlying binned responses is unknown. The latter method has the advantage of leading to low computational times, while using all the observations and constructing the band with large aggregated volumes. On the contrary, by reducing the number of elements in the set of ordered pairs, the resulting band has to be constructed using small aggregated volumes, and this is actually the reason why the bands become too wide for small sets of ordered pairs in Figures 3 and 4. Therefore, we recommend to use the binning method for large datasets. We will follow this choice in Section 8.4 where we consider a portfolio of more than half a million insurance policies.

8.2. Example 2: calibration bands for a small sample-sized real dataset

We consider in this example a lime trees real dataset[1] available from the R (R Core Team, 2021) package `GLMsData` hosted by Dunn and Smyth (2022). This dataset contains measurements from $n = 385$ small-leaved lime trees growing in Russia such as the foliage biomass (in kg), the tree diameter (in cm) and the origin of the tree (`Coppice`, `Natural` and `Planted`). Our goal is to estimate the foliage biomass using the tree diameter and the origin of the tree. As in Examples 11.4 and 11.7 in Dunn and Smyth (2018), we fit a gamma and an inverse Gaussian GLM with a logarithmic link using a continuous feature providing the logarithm of the tree diameter and a categorical feature providing the origin of the tree. In both GLMs, we further include an interaction term between the two features and we denote the resulting regression functions by $\hat{\mu}^{\text{gamma}} : \mathcal{X} \to (0, \infty)$ and $\hat{\mu}^{\text{IG}} : \mathcal{X} \to (0, \infty)$, where

$$\mathcal{X} = (0, \infty) \times \{\texttt{Coppice}, \texttt{Natural}, \texttt{Planted}\}, \tag{45}$$

stands for the feature space. By construction, both models have the same number of parameters and the volumes are $\boldsymbol{v} = (1, \ldots, 1)^\top \in \mathbb{R}^n$, because we did not introduce any weights in the fitting procedure. Based on AIC, we clearly give preference to the gamma GLM; see Table 4.

In order to assess the auto-calibration of the above models, we construct full calibration bands on the mean of the foliage biomass by using the GLMs themselves as ranking functions

[1] The dataset can be downloaded by running `library(GLMsData); data(lime)` in R.

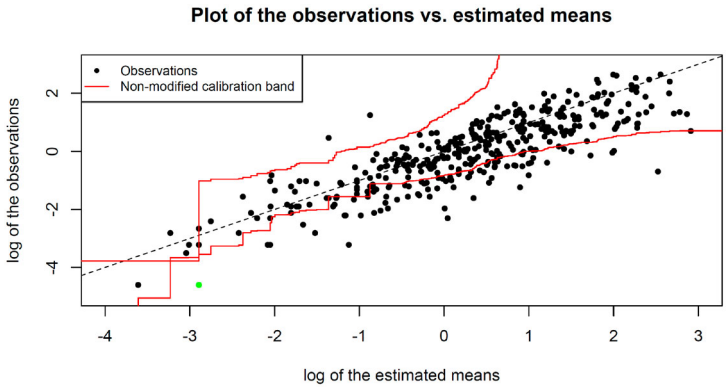

Figure 6. Plot of the observations versus the inverse Gaussian GLM mean estimates on the log scale. The non-modified calibration band is plotted in red, whereas the observations are drawn in black. The seventh observation is highlighted in green.

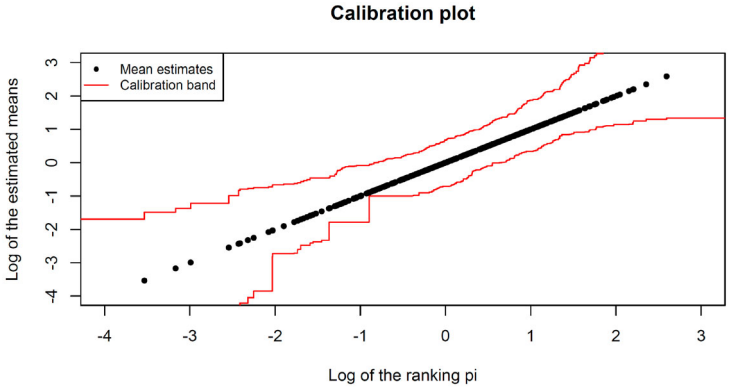

Figure 7. Calibration plot of the gamma GLM on the log scale. The calibration band is plotted in red, whereas the mean estimates are drawn in black.

and choosing the confidence level $1 - \alpha = 0.95$. The gamma and inverse Gaussian bands are shown in Figures 7 and 8. For the latter band, note that the constructed upper and lower bounds for small mean estimates are crossing; see Figure 6. We thus use the method in (15) in order to obtain a non-crossing band in Figure 8. Those crossings may happen because the ranking of the estimated means may not match the ranking of the smallest observations. The assumed variance function when using the inverse Gaussian GLM fulfills

$$V(\mu) \propto \mu^3,$$

which implies that the lower and upper bounds in (13) and (14) are extremely close to the weighted partial sums $Z_{j:k}$ in (8) for small values of the sums. Outliers are thus rare and we see in Figure 6 that in this example, the seventh observation, highlighted in green, is responsible for a low upper band on the left end of the interval.

The conclusion of the calibration plot in Figure 7 is not to reject the auto-calibration of the gamma GLM at a confidence level $1 - \alpha = 0.95$ as all the mean estimates fall within the constructed calibration band. However, the plot in Figure 8 leads to the rejection of the

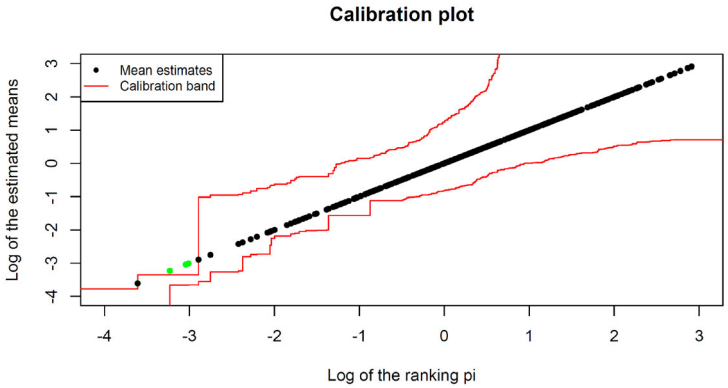

Figure 8. Calibration plot of the inverse Gaussian GLM on the log scale. The calibration band is plotted in red. The mean estimates $(\widehat{\mu}^{\,\mathrm{IG}}(\boldsymbol{x}_i))_{i=1}^{n}$ falling within the band are drawn in black, whereas those falling outside the bands are drawn in green.

Table 5. The p-values $\hat{p}_{\,\mathrm{auto}}$ for the gamma and inverse Gaussian GLMs.

toprule	gamma GLM	Inverse Gaussian GLM
$\hat{p}_{\,\mathrm{auto}}$	0.89	0.08

auto-calibration of the inverse Gaussian GLM because at the bottom left of the plot, some of the mean estimates fall outside the constructed band, those are plotted in green. In order to perform these tests, we assumed that $\hat{\mu}^{\,\mathrm{gamma}}(\boldsymbol{X})$ and $\hat{\mu}^{\,\mathrm{IG}}(\boldsymbol{X})$ are absolutely continuous random variables with strictly positive density over their supports.

Interestingly, the conclusion of the test derived by Denuit et al. (2024) is in this case different as the auto-calibration of both models is not rejected at the level $1-\alpha = 0.95$. To perform the latter test, non-parametric Monte Carlo methods have to be used to compute the p-value $\hat{p}_{\,\mathrm{auto}}$, and the null-hypothesis of auto-calibration is rejected at confidence level $1-\alpha$ whenever $\hat{p}_{\,\mathrm{auto}} < \alpha$. The p-values obtained for $\hat{p}_{\,\mathrm{auto}}$ for both GLMs are summarized in Table 5 by performing $B = 500$ Monte-Carlo simulations; we refer the reader to Section 4 of Denuit et al. (2024) for more technical details. While these values are both larger than $\alpha = 0.05$, note that the value for the gamma GLM is close to 1, whereas the value for the inverse Gaussian GLM is close to the chosen significance level. This indicates that the gamma GLM is indeed closer to auto-calibration than the inverse Gaussian GLM. For this small-sample sized real dataset, our test manages to detect violations of auto-calibration of the inverse Gaussian GLM in contrast to the test of Denuit et al. (2024). The reason might be that the latter test relies on asymptotic results and should thus only be used on large datasets, whereas our test adapts to the distribution of the responses and can be used for any sample size as the construction of the calibration band does not rely on any asymptotic result. This constitutes an advantage over the other methods when it comes to assess the auto-calibration of a small dataset.

8.3. Example 3: calibration bands for simulated inverse Gaussian responses

In Section 7.4 we emphasized the importance of using suitable dispersion estimates in order to construct calibration bands as all of our results are based on the assumption of a known and fixed dispersion estimate. Throughout this example, we construct calibration bands on

Table 6. Dispersion estimates for various sample sizes; the true value is $\varphi = 1.26$.

Sample sizes	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$
Pearson estimate	1.50	1.18	1.13	1.04
Deviance estimate	1.40	1.28	1.16	1.24
MLE estimate	1.32	1.27	1.15	1.24

the mean of simulated inverse Gaussian responses that rely on different dispersion estimates. Our goal is to show that although the dispersion parameter φ is unknown, and has to be estimated, its influence on the resulting bands is comparably small, allowing us to use the statistical tests of Section 7 in practice.

In order to obtain realizations of n independent inverse Gaussian responses, we first replicate the feature space of the above real dataset by simulating n feature components $\tilde{\boldsymbol{x}}_i$ lying in the set $\mathcal{X}$ given in (45). To this end, we assume that the tree diameters follow a gamma distribution, with scale and shape parameters chosen based on the empirical mean and standard deviation of the observed diameters in the above real dataset. In addition, the origin component of the simulated features is generated using the empirical distribution of origins from the same dataset. This leads to n new features $(\tilde{\boldsymbol{x}}_i)_{i=1}^n$. We then choose a dispersion parameter of $\varphi = 1.26$, corresponding to the Pearson estimate of the above dataset. Finally, we assume that the responses satisfy

$$\widetilde{Y}_i \sim \text{IG}(\hat{\mu}^{\text{IG}}(\tilde{\boldsymbol{x}}_i), \varphi), \quad 1 \leq i \leq n.$$

This provides us with an example where the true model is given by an inverse Gaussian GLM with an interaction term between the two feature components; see Section 8.2. As the inverse Gaussian distribution has the advantage that the MLE dispersion estimate can be expressed in closed form, we aim at constructing four different full calibration bands with confidence level $1 - \alpha = 0.95$ by using successively the true dispersion parameter, the Pearson dispersion estimate, the deviance dispersion estimate and the MLE dispersion estimate; we refer to Chapter 11.6 in Dunn and Smyth (2022) for more details about these estimates. Using the constructed bands, we then assess the calibration of a newly fitted inverse Gaussian GLM $\tilde{\mu}^{\text{IG}} : \mathcal{X} \to (0, \infty)$ on the new dataset $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)_{i=1}^n$ for $n \in \{100, 500, 1000, 2000\}$, and the results are provided in Figure 9. Although all the treated datasets are small, we notice that the bands nearly coincide for all sample sizes. This example shows that when the model is well-specified, i.e., when the data generating process matches the model's assumptions, using an estimate for the dispersion instead of the true dispersion parameter to construct calibration bands does not have a major impact on the resulting bands, even for small datasets. The values of the dispersion estimates are shown in Table 6, where we see that even a 20% difference with respect to the true parameter $\varphi = 1.26$ does not lead to any significant impact on the bands. Finally, note that while the statistical tests we derive can be applied for all sample sizes, we see in Figure 9 that the calibration band gets narrower the larger the sample size is, leading to more powerful statistical tests.

8.4. Example 4: calibration bands for a large sample-sized real dataset

After considering a small dataset in Section 8.2 we study in this example a French motor third party liability (MTPL) real dataset available from the R (R Core Team, 2021) package `CASdatasets` hosted by Dutang and Charpentier (2018). This dataset contains information on insurance policies and claim frequency of more than half a million French car drivers.

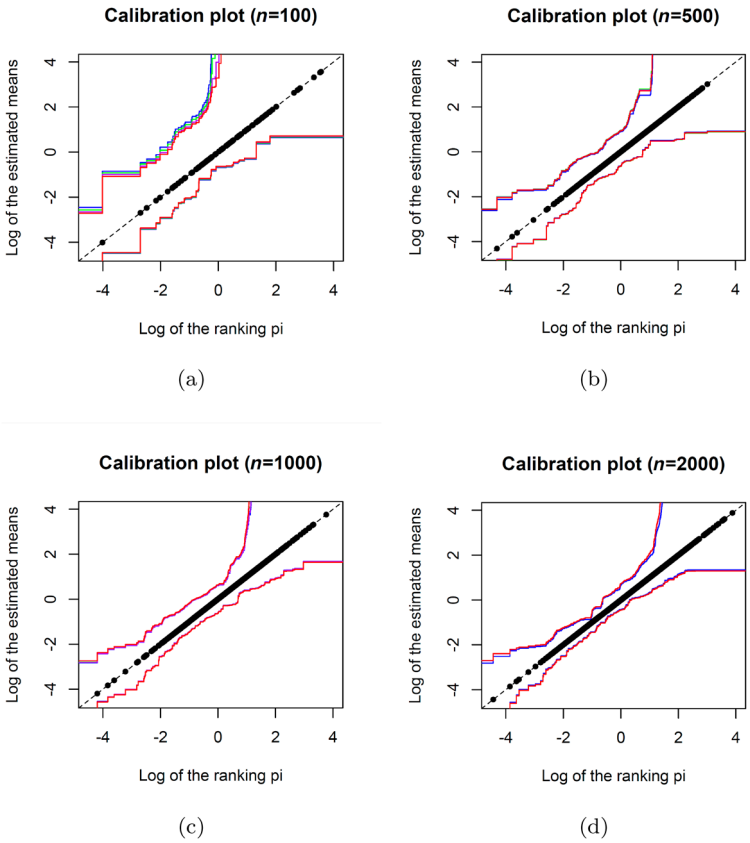

Figure 9. Calibration plots of the regression function $\tilde{\mu}^{\text{IG}}: \mathcal{X} \to (0, \infty)$ on the log scale for four different sample sizes. For each plot, four different calibration bands are constructed using $\pi(\cdot) = \tilde{\mu}^{\text{IG}}(\cdot)$ as a ranking function and the true dispersion parameter (red) as well as the Pearson estimate (blue), the deviance estimate (green) and the MLE estimate (purple) for the dispersion. The mean estimates are drawn in black.

We follow Listing 13.1 in Wüthrich and Merz (2023) in order to clean the data, leading to a portfolio of $n = 678{,}007$ insurance policies and 26,383 claims.[2] For each policy $1 \leq i \leq n$, the resulting dataset provides the number of claims $N_i \in \mathbb{N}$ that occurred during an exposure period $v_i \in (0, 1]$ (years-at-risk) and features containing information of the policyholder as, for example, the age of the driver, the brand and power of their car, or the region of residence. The total exposure at risk is equal to 358,359 years, indicating that some policyholders were covered for a period of less than one year, and as one might expect in motor liability insurance, most policies do not lead to any claim; see Table 7. We refer to Section 13.1 in Wüthrich and Merz (2023) for an extended description of the dataset.

In this section, we aim at modelling the claim frequency of each policyholder. For this, we follow Listings 5.1–5.2 in Wüthrich and Merz (2023) in order to pre-process the available features. Moreover, we consider a subset of the features by only keeping the information about the policyholder and not their car. That is, we use 3 continuous feature components and 2

[2] The cleaned dataset can be downloaded under https://people.math.ethz.ch/ wueth/Lecture/freMTPL2freq.rda.

Table 7. Number of policies and total exposure within the portfolio that is split with respect to the number of claims occurred for each policy.

Number of claims occurred for each policy	0	1	2	3	4	5
Number of policies	653,069	23,571	1298	62	5	2
Total exposure	341,090	16,315	909	42	2	1

categorical ones[3], leading to a feature space

$$\mathcal{X} \subset \mathbb{R}^3 \times \{0,1\}^6 \times \{0,1\}^{21}.$$

After pre-processing the categorical variables using dummy coding, we fit a Poisson GLM with the canonical link on the whole dataset in order to estimate the claim frequency of each policyholder with feature $\boldsymbol{x} \in \mathcal{X}$, given by the true mean function $\mu^* : \mathcal{X} \to (0,\infty)$. That is, we assume that

$$N_i \sim \text{Poi}(\mu^*(\boldsymbol{x}_i)v_i), \qquad \text{for } 1 \le i \le n,$$

where $\boldsymbol{x}_i \in \mathcal{X}$ is the considered feature of the policy i. We call the resulting estimated regression function $\widehat{\mu}^{\text{Poi}} : \mathcal{X} \to (0,\infty)$, and this function satisfies

$$\min_{1\le i\le n} \widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i) = 0.024 \quad \text{and} \quad \max_{1\le i\le n} \widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i) = 1.292.$$

This means that the model predicts that an accident occurs on average once every 40 years for some drivers, while for others, it predicts that more than one accident occurs each year on average. In order to assess whether the obtained regression function is calibrated, we construct a calibration band on the claim frequency. As the dataset is large, we bin the responses according to their estimated means for computational reasons. To do so, we first define $Y_i = N_i/v_i$ and use the convolution formula for the reproductive form of the EDF in order to derive $L = 5000$ new responses

$$\widetilde{Y}_l = \frac{\sum_{i=1}^n v_i Y_i \mathbb{1}_{\{\widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i)\in\mathcal{I}_l\}}}{\sum_{i=1}^n v_i \mathbb{1}_{\{\widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i)\in\mathcal{I}_l\}}}, \qquad \text{for } 1 \le l \le L,$$

where the intervals $\mathcal{I}_l = [a_{l-1}, a_l)$ are delimited by some partition $(a_l)_{l=0}^L$ of $[0.024, 1.292]$ such that the volumes of all the binned responses are approximately equal. This can be achieved by ranking the responses $(Y_i)_{i=1}^n$ according to their mean estimates and using a weighted quantile binning. This procedure leads to the new volumes

$$\tilde{v}_l = \sum_{i=1}^n v_i \mathbb{1}_{\{\widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i)\in\mathcal{I}_l\}} \in [70.7, 72.64], \qquad \text{for } 1 \le l \le L.$$

The same binning applied to the mean estimates $(\widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i))_{i=1}^n$ allows us to derive the mean estimates of the new responses $(\widehat{\mu}_l^{\text{Poi}})_{l=1}^L$. Using the realizations of the binned responses as observations and $\pi(\cdot) = \widehat{\mu}^{\text{Poi}}(\cdot)$ as a ranking function, we construct a full calibration band

[3] The used features are BonusMalusGLM, DensityGLM, AreaGLM, DrivAgeGLM, Region; see Section 5.2.4 in Wüthrich and Merz (2023).

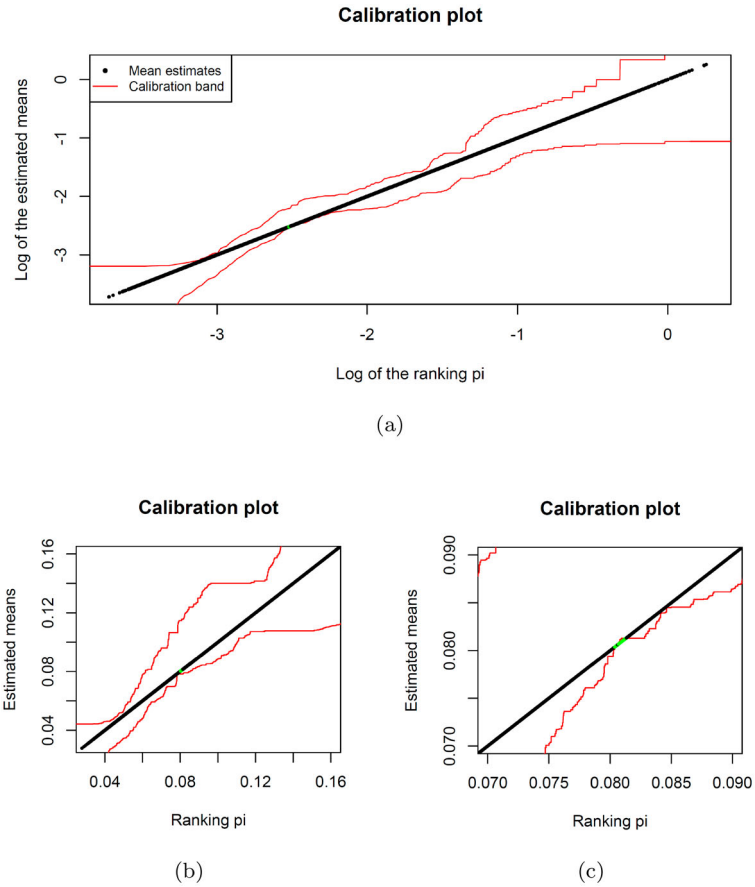

Figure 10. Calibration plot of the regression function $\widehat{\mu}^{\,\mathrm{Poi}} : \mathcal{X} \to (0, \infty)$ on the log scale (above). Zoomed versions of this plot are provided on the linear scale (below). The calibration band is constructed using $\pi(\cdot) = \widehat{\mu}^{\,\mathrm{Poi}}(\cdot)$ as a ranking function and is plotted in red. The mean estimates $(\widehat{\mu}^{\,\mathrm{Poi}}(\boldsymbol{x}_i))_{i=1}^n$ falling within the band are drawn in black, whereas those falling outside the bands are drawn in green.

for the confidence level $1 - \alpha = 0.95$. This band is drawn in Figure 10, where we additionally plot the mean estimates $(\widehat{\mu}^{\,\mathrm{Poi}}(\boldsymbol{x}_i))_{i=1}^n$ against the corresponding rankings. Note that the resulting points all lie on the diagonal due to our choice of the ranking function.

As we use the log scale for the upper plot in Figure 10, we notice that the band is narrow for small means and wide for large means. The reason for this is that the aggregated exposure of the policies for which the estimated mean is below 0.2 corresponds to 97.5% of the total exposure of the portfolio. In other words, the aggregated volumes used to compute the bounds fail to be large enough for mean estimates exceeding 0.2. The lower plots in Figure 10 show that some mean estimates fall outside the calibration band, which are plotted in green. By assuming that $\widehat{\mu}^{\,\mathrm{Poi}}(\boldsymbol{X})$ is an absolutely continuous random variable with strictly positive density over its support, the conclusion of the statistical test derived in Section 7.3 is thus to reject calibration at a confidence level of $1 - \alpha = 0.95$. This decision indicates that the regression function $\widehat{\mu}^{\,\mathrm{Poi}} : \mathcal{X} \to (0, \infty)$ is too far from the true mean function although the violation only happens for a small part of the support of $\widehat{\mu}^{\,\mathrm{Poi}}(\boldsymbol{X})$ in Figure 10. We emphasize, however, that as the inequality in (11) might not be very sharp in the construction of

the band, the lower left plot in Figure 10 hints that the regression function might not be sufficiently calibrated for other mean estimates too, which are close to the boundary of the band.

Our next goal is to improve the obtained regression function. For this, we assume that it provides the correct ordering of the true mean function, but the decision of the above statistical test indicates a violation of the auto-calibration of $\widehat{\mu}^{\text{Poi}} : \mathcal{X} \to (0, \infty)$; see Section 7.3. Therefore, we construct another regression function by applying the isotonic recalibration step proposed by Wüthrich and Ziegel (2024) to the Poisson GLM. We call the new resulting regression function $\widehat{\mu}^{\text{Poi}}_{\text{rec}} : \mathcal{X} \to (0, \infty)$ and point out that the isotonic recalibration step is performed using the ranking provided by $(\widehat{\mu}^{\text{Poi}}(\boldsymbol{x}_i))_{i=1}^n$ and the exposures $(v_i)_{i=1}^n$ as weights. The obtained regression function $\widehat{\mu}^{\text{Poi}}_{\text{rec}} : \mathcal{X} \to \mathbb{R}$ is provided in Figure 11, where we draw the same calibration band as above, i.e., we assume again the ranking function to be $\pi(\cdot) = \widehat{\mu}^{\text{Poi}}(\cdot)$ in order to construct the band.

This time, all the mean estimates lie at the middle of the constructed band, leading us not to reject the calibration of this model. Moreover, we point out that the regression function $\widehat{\mu}^{\text{Poi}}_{\text{rec}} : \mathcal{X} \to \mathbb{R}$ is empirically auto-calibrated; we refer to Wüthrich and Ziegel (2024). That is, the isotonic recalibration step provides an empirically auto-calibrated regression function, for which we do not reject the null-hypothesis of calibration. Note that one could alternatively construct a calibration band using $\pi(\cdot) = \widehat{\mu}^{\text{Poi}}_{\text{rec}}(\cdot)$ as a ranking function in order to assess the calibration of $\widehat{\mu}^{\text{Poi}}_{\text{rec}} : \mathcal{X} \to \mathbb{R}$. The corresponding calibration plot is given in Figure 12 and, again, we do not reject the null-hypothesis of calibration as all the mean estimates lie within the band.

8.5. Example 5: power of the statistical test for calibration

Finally, we study the power of the statistical test for calibration (39) derived in Section 7.2. To do so, we consider the example of Wüthrich (2024). That is, we first simulate n i.i.d. mean values μ_i from the law described by

$$\mu_i = \begin{cases} 10, & \text{with } p = 0.1, \\ 11, & \text{with } p = 0.15, \\ 12, & \text{with } p = 0.25, \\ 13, & \text{with } p = 0.25, \\ 14, & \text{with } p = 0.15, \\ 15, & \text{with } p = 0.1. \end{cases}$$

Then, we permute the indices of the sampled means so that (2) holds and simulate n independent responses Y_i by assuming

$$Y_i \sim \Gamma(3\mu_i, 3), \qquad \text{for } 1 \le i \le n.$$

The above permutation ensures that the calibration bands are constructed using responses that are correctly ranked according to their means. Our aim is to evaluate the power of the statistical test in (39). For this, we define the mean estimates

$$\widehat{\mu}_i = \mu_i, \qquad \text{for } 1 \le i \le n, \tag{46}$$

and assess their calibration when the realizations $(y_i)_{i=1}^n$ of the above simulated responses are shifted. We consider here two different types of contamination. First, we contaminate the

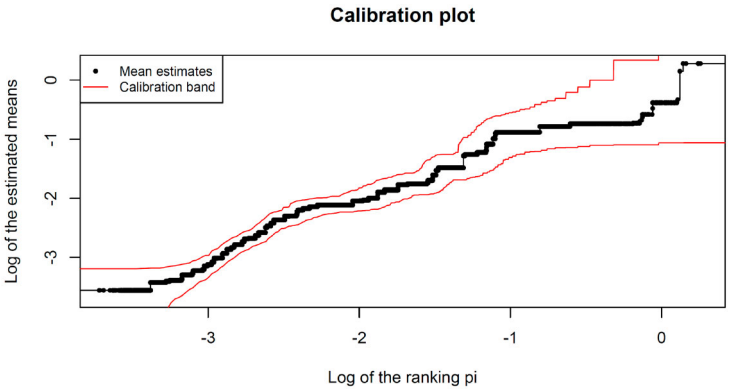

Figure 11. Calibration plot of the regression function $\widehat{\mu}^{\,\mathrm{Poi}}_{\mathrm{rec}} : \mathcal{X} \to (0, \infty)$ on the log scale. The calibration band is constructed using $\pi(\cdot) = \widehat{\mu}^{\,\mathrm{Poi}}(\cdot)$ as a ranking function and is plotted in red, whereas the mean estimates $(\widehat{\mu}^{\,\mathrm{Poi}}_{\mathrm{rec}}(\boldsymbol{x}_i))_{i=1}^n$ are drawn in black.

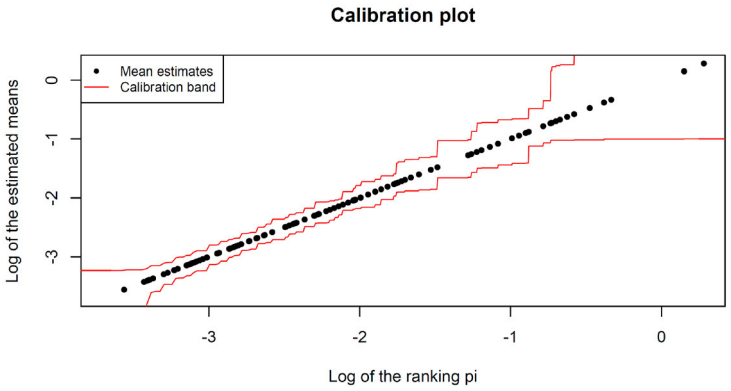

Figure 12. Calibration plot of the regression function $\widehat{\mu}^{\,\mathrm{Poi}}_{\mathrm{rec}} : \mathcal{X} \to (0, \infty)$ on the log scale. The calibration band is constructed using $\pi(\cdot) = \widehat{\mu}^{\,\mathrm{Poi}}_{\mathrm{rec}}(\cdot)$ as a ranking function and is plotted in red, whereas the mean estimates $(\widehat{\mu}^{\,\mathrm{Poi}}_{\mathrm{rec}}(\boldsymbol{x}_i))_{i=1}^n$ are drawn in black.

observations by a *global shift* $\delta \in \{0, 0.5, 1\}$, i.e.,

$$y_i^{\delta} = y_i + \delta, \qquad \text{for } 1 \le i \le n,$$

and we want to test for the calibration of $(\widehat{\mu}_i)_{i=1}^n$ for these new observations. Then, the same procedure is repeated but this time, only observations being associated to a given single mean are shifted, i.e., observations are transformed such that

$$y_i^{l,\delta} = y_i + \delta \mathbb{1}_{\{\mu_i = l\}},$$

for $l \in \{10, 13, 15\}$. We refer to this transformation as a *local shift of level l*.

For both of these shifts, the mean estimates $(\widehat{\mu}_i)_{i=1}^n$ are calibrated if and only if $\delta = 0$ and we want to understand whether the statistical test in (39) is able to detect these violations of calibration. The results, showing the number of rejections of the calibration of $(\widehat{\mu}_i)_{i=1}^n$ at a confidence level $1 - \alpha = 0.95$, are summarized in Table 8. They should be compared to

Table 8. Power of the performed statistical tests with confidence level $1 - \alpha = 0.95$.

Contamination δ	0	0.5	1
Global shift	5/1000	291/1000	1000/1000
Local shift of level $l = 10$	–	13/1000	612/1000
Local shift of level $l = 13$	–	128/1000	994/1000
Local shift of level $l = 15$	–	3/1000	270/1000

Table 9. Power of the performed statistical tests with confidence level $1 - \alpha = 0.95$ for binned responses.

Contamination δ	0	0.5	1
Global shift	16/1000	995/1000	1000/1000
Local shift of level $l = 10$	–	360/1000	986/1000
Local shift of level $l = 13$	–	769/1000	1000/1000
Local shift of level $l = 15$	–	203/1000	917/1000

the plots on Page 12 of Wüthrich (2024) as the mean estimates $(\widehat{\mu}_i)_{i=1}^n$ are calibrated if and only if they are auto-calibrated for the shifts considered in this example. Note that we only simulate 1000 different samples, each containing $n = 1000$ responses, while Wüthrich (2024) performs 10,000 simulations of size $n = 1000$. Additionally, we consider here only a limited set of contaminations $\delta \in \{0, 0.5, 1\}$.

When the shift factor δ is equal to 0, i.e., when the mean estimates $(\widehat{\mu}_i)_{i=1}^n$ are calibrated, we see in Table 8 that the rejection rate is equal to 5/1000, which is 10 times smaller than the significance level $\alpha = 0.05$. This is not surprising as the constructed calibration bands rely on the union bound inequality in (11), which implies that the power of the corresponding statistical test might be (much) lower than the significance level. Furthermore, we see in Table 8 that the constructed statistical test is not fully capable of detecting small deviations from calibration. However, the test seems to be more effective at identifying such deviations when they occur on a global scale or at the middle of the range of interest.

The same experiment is then repeated but this time, all the observations are binned according to their estimated means $(\widehat{\mu}_i)_{i=1}^n$ in (46), meaning that the calibration band is constructed using a full set of ordered pairs $\mathcal{J}^{\text{full}}$ and only six observations having large and different volumes due to aggregation. The corresponding rejection rates are provided in Table 9. As expected, the predictive power is now much better as the size of the set of ordered pairs heavily decreases, while we keep using the whole dataset to construct the calibration bands. Although the number of rejections is significantly higher, the same conclusions as for Table 8 hold. That is, the power of the test when the observations are not contaminated is still below the significance level $\alpha = 0.05$ and the test is more effective at identifying large deviations from calibration and contaminations that happen on a global scale or at the middle of the range of interest.

This example shows that the calibration bands we constructed in this paper can be in general wide, leading to statistical tests with lower power. We mention again that the reason for this lies in the union bound inequality (11) that might not be very sharp for large sets of ordered pairs. An interesting tool to reduce this size while using all the observations is to bin those observations. As a result, the bands get narrower and, thus, allow for more powerful statistical tests. This method assumes that the observations within a given bin have approximately the same mean and, by construction, this is the case in this example.

9. Conclusion

Using the stochastic ordering properties and the convolution formulas of the EDF, we extended the construction of the calibration bands on the means from the binary case of Dimitriadis et al. (2023) to the whole EDF. Our construction enables us to find closed form expressions for the calibration bands of independent binomial, Poisson, negative binomial, gamma and normal responses and the bands can be computed using a root-finding algorithm in the other cases of the EDF. Interestingly, we showed that our band is narrower than Yang–Barber's (Yang & Barber, 2019) band in the normal case.

As for the calibration bands derived by Dimitriadis et al. (2023) and Yang and Barber (2019), our construction relies on the assumption that the responses are ranked such that their true means are increasing. In a regression modelling context, we showed how this assumption can be extended by introducing a ranking function that provides the ordering of the true mean function for almost every feature in the feature space. In practice, such a ranking function is often unknown and in such cases, it has to be approximated by the regression function itself in order to construct the statistical tests for calibration or auto-calibration. Through numerical examples, we showed how these tests can be applied to detect violations of these properties, and we emphasized that in contrast to other approaches, our tests can be applied for any sample size as the construction of the calibration bands does not rely on asymptotic results. Moreover, we discussed some important factors that influence the shape of the calibration bands and proposed methods to construct them for large datasets. One of these methods consists in binning the available observations and we argued that it leads to suitable bands as it allows for using all the observations while remaining computationally efficient.

The decision of the statistical tests we propose depends on the underlying calibration band, which itself depends on the chosen ranking function. Going forward, it will be interesting to better understand the role of the ranking function on the resulting band. This is particularly true in cases where the ranking function cannot easily be inferred from the observations, e.g., when the signal-to-noise ratio of the available observations is low. Moreover, the construction of the band relies on the assumption of a known and given dispersion parameter. As this parameter is typically unknown in practice, it will be interesting to study the impact of misestimating this parameter. Another next step is to study alternative methods for binning the observations and other choices for the set of ordered pairs in order to understand the impact on the resulting band. Finally, the rate of convergence of the calibration bands for an increasing amount of observations is of interest, as well as asymptotic results.

Acknowledgments

We kindly thank the referees and the editor for their useful remarks that have helped us to improve this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Selim Gatti http://orcid.org/0009-0002-4970-1716

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley.

Barndorff-Nielsen, O. (2014). *Information and exponential families: In statistical theory*. Wiley.

Delong, Ł., & Wüthrich, M. V. (2025). Isotonic regression for variance estimation and its role in mean estimation and model validation. *North American Actuarial Journal*, *29*(3), 563–591. https://doi.org/10.1080/10920277.2024.2421221

Denuit, M., Charpentier, A., & Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, *101*(B), 485–497.

Denuit, M., Huyghe, J., Trufin, J., & Verdebout, T. (2024). Testing for auto-calibration with Lorenz and concentration curves. *Insurance: Mathematics and Economics*, *117*, 130–139.

Denuit, M., & Trufin, J. (2023). Model selection with Pearson's correlation, concentration and Lorenz curves under auto-calibration. *European Actuarial Journal*, *13*(2), 871–878. https://doi.org/10.1007/s13385-023-00353-5

Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., & Ziegel, J. (2023). Honest calibration assessment for binary outcome predictions. *Biometrika*, *110*(3), 663–680. https://doi.org/10.1093/biomet/asac068

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R*. Springer.

Dunn, P. K., & Smyth, G. K. (2022). *GLMsData R Package Vignette* (Reference manual. Version 1, packaged 2022-08-22) [Computer software manual].

Dutang, C., & Charpentier, A. (2018). *CASdatasets R Package Vignette* (Reference manual. Version 1.0-8, packaged 2018-05-20) [Computer software manual].

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*(4), 391–398. https://doi.org/10.1111/ahg.1935.6.issue-4

Fisher, R. A. (1973). *Statistical methods and scientific inference*. Hafner Press.

Fissler, T., Lorentzen, C., & Mayer, M. (2022). Model comparison and calibration assessment: User guide for consistent scoring functions in machine learning and actuarial practice. *Preprint*. arXiv:2202.12780 [stat.ML].

Gneiting, T., & Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, *17*(2), 3226–3286. https://doi.org/10.1214/23-EJS2180

Henzi, A., Mösching, A., & Dümbgen, L. (2022). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodology and Computing in Applied Probability*, *24*(4), 2633–2645. https://doi.org/10.1007/s11009-022-09937-2

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, *9*(10), 1043–1069. https://doi.org/10.1080/03610928008827941

Jørgensen, B. (1986). Some properties of exponential dispersion models. *Scandinavian Journal of Statistics*, *13*(3), 187–197.

Jørgensen, B. (1997). *The theory of dispersion Models*. Chapman and Hall.

Krüger, F., & Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, *39*(4), 972–983. https://doi.org/10.1080/07350015.2020.1741376

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. Chapman and Hall.

Pedersen, J. G. (1978). Fiducial inference. *International Statistical Review*, *46*(2), 147–170. https://doi.org/10.2307/1402811

Pohle, M. O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *Preprint*. arXiv:2005.01835 [stat.ME].

Rao, M. M., & Swift, R. J. (2006). *Probability theory with applications*. Springer.

R Core Team (2021). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Sprott, D. A. (2000). *Statistical inference in science*. Springer.

Veronese, P., & Mellili, E. (2015). Fiducial and confidence distributions for real exponential families. *Scandinavian Journal of Statistics*, *42*(2), 471–484. https://doi.org/10.1111/sjos.v42.2

Wüthrich, M. V. (2024). Auto-calibration tests for discrete finite regression functions. *Preprint*. arXiv:2408.05993 [math.ST].
Wüthrich, M. V., & Merz, M. (2023). *Statistical foundations of actuarial learning and its applications*. Springer.
Wüthrich, M. V., & Ziegel, J. (2024). Isotonic recalibration under a low signal-to-noise ratio. *Scandinavian Actuarial Journal*, *2024*(3), 279–299. https://doi.org/10.1080/03461238.2023.2246743
Yang, F., & Barber, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, *13*(1), 646–677. https://doi.org/10.1214/18-EJS1520

Appendix. Proofs

We prove all statements in this appendix.

Proof: Let t be in the support of the density f_{Y_1} (which is the same as the support of the density f_{Y_2}, see Section 2) and set $\theta_i = h(\mu_i)$ for $i \in \{1,2\}$. Since the canonical link h is strictly increasing on $\kappa'(\mathring{\Theta})$, we have $\theta_1 \le \theta_2$. Thus, if we divide the density of Y_2 by the density of Y_1, we obtain that the function

$$t \mapsto \frac{f_{Y_2}(t)}{f_{Y_1}(t)} = \frac{\exp\left\{\frac{t\theta_2-\kappa(\theta_2)}{\varphi/v} + a(t;v/\varphi)\right\}}{\exp\left\{\frac{t\theta_1-\kappa(\theta_1)}{\varphi/v} + a(t;v/\varphi)\right\}} = \exp\left\{\frac{t(\theta_2-\theta_1)-\kappa(\theta_2)+\kappa(\theta_1)}{\varphi/v}\right\}$$

is non-decreasing. This implies $Y_1 \le_{\mathrm{lr}} Y_2$ and using Theorem 1.C.1 in Shaked and Shanthikumar (2007), we conclude that $Y_1 \le_{\mathrm{st}} Y_2$. ■

Proof: The lower bound in (6) satisfies

$$\begin{aligned}\mathbb{P}\left(\mathbb{E}[Y] \ge l^\delta(Y,v,\varphi,\kappa(\cdot))\right) &= 1-\mathbb{P}\left(\mathbb{E}[Y] < l^\delta(Y,v,\varphi,\kappa(\cdot))\right)\\ &\ge 1-\mathbb{P}\left(F^*(Y;h(\mathbb{E}[Y]),v,\varphi,\kappa(\cdot)) > 1-\delta\right)\\ &\ge 1-\mathbb{P}(U>1-\delta) = 1-\delta,\end{aligned}$$

with $U \sim \mathrm{Unif}([0,1])$ and where we used in the first inequality that

$$\mathbb{E}[Y] < l^\delta(Y,v,\varphi,\kappa(\cdot)) \implies F^*(Y;h(\mathbb{E}[Y]),v,\varphi,\kappa(\cdot)) > 1-\delta.$$

Similarly, we have for the upper bound in (7) that

$$\begin{aligned}\mathbb{P}\left(\mathbb{E}[Y] \le u^\delta(Y,v,\varphi,\kappa(\cdot))\right) &= 1-\mathbb{P}\left(\mathbb{E}[Y] > u^\delta(Y,v,\varphi,\kappa(\cdot))\right)\\ &\ge 1-\mathbb{P}\left(F(Y;h(\mathbb{E}[Y]),v,\varphi,\kappa(\cdot)) < \delta\right)\\ &\ge 1-\mathbb{P}(U<\delta) = 1-\delta.\end{aligned}$$

■

Proof: The proof relies on Theorem 1.A.3 in Shaked and Shanthikumar (2007), which states that for any set of independent random variables $X_1,\dots,X_n$ and any another set of independent random variables $Y_1,\dots,Y_n$ satisfying $X_i \le_{\mathrm{st}} Y_i$ for all $i \in \{1,\dots,n\}$, we have

$$\psi(X_1,\dots,X_n) \le_{\mathrm{st}} \psi(Y_1,\dots,Y_n),$$

for any non-decreasing function $\psi : \mathbb{R}^n \to \mathbb{R}$. By choosing $\psi(x_1,\dots,x_n) = \sum_{i=1}^n v_i x_i / v_{1:n}$ and using Proposition 3.4, the stochastic ordering result holds for

$$Z^-_{j:k} = \frac{1}{v_{j:k}} \sum_{i=j}^{k} v_i Y_i^-, \quad \text{with} \quad Y_i^- \overset{\text{ind.}}{\sim} \mathrm{EDF}(\theta_j, v_i, \varphi, \kappa(\cdot)),$$

and

$$Z^+_{j:k} = \frac{1}{v_{j:k}} \sum_{i=j}^{k} v_i Y^+_i, \quad \text{with} \quad Y^+_i \overset{\text{ind.}}{\sim} \text{EDF}(\theta_k, v_i, \varphi, \kappa(\cdot)).$$

Note that these random variables satisfy (10) due to the convolution formula for the EDF in Corollary 2.15 of Wüthrich and Merz (2023). This completes the proof. ■

Proof: Letting $1 \leq j \leq k \leq n$, it follows from Proposition 3.5 that

$$\mathbb{P}\left(\mu_j \leq u^\delta(Z^-_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\right) \geq 1-\delta \quad \text{and} \quad \mathbb{P}\left(\mu_k \geq l^\delta(Z^+_{j:k}, v_{j:k}, \varphi, \kappa(\cdot))\right) \geq 1-\delta.$$

for the random variables $Z^-_{j:k}$ and $Z^+_{j:k}$ introduced in (10). Since $Z^-_{j:k} \leq_{\text{st}} Z_{j:k} \leq_{\text{st}} Z^+_{j:k}$ due to Lemma 3.6 and since the functions

$$y \in \mathbb{R} \mapsto u^\delta(y, v_{j:k}, \varphi, \kappa(\cdot)),$$

and

$$y \in \mathbb{R} \mapsto l^\delta(y, v_{j:k}, \varphi, \kappa(\cdot)),$$

are non-decreasing in y, we have by Theorem 1.A.3 of Shaked and Shanthikumar (2007) that

$$u^\delta(Z^-_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \leq_{\text{st}} u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)),$$

and

$$l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \leq_{\text{st}} l^\delta(Z^+_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)).$$

The claim then follows. ■

Proof: Let $\mathcal{J}$ be any set of ordered pairs. By Proposition 3.7, we deduce using a union bound argument that

$$\begin{aligned}&\mathbb{P}\left(\mu_j \leq u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ and } \mu_k \geq l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ for all } (j,k) \in \mathcal{J}\right)\\ &\quad \geq 1 - 2|\mathcal{J}|\delta.\end{aligned}$$

Due to the ordering assumed in (2), the above inequality can be rewritten as

$$\begin{aligned}&\mathbb{P}\Bigg(\sup_{(j,k)\in\mathcal{J}\,:\,\theta_i \geq \theta_k} l^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \leq \mu_i\\ &\qquad \leq \inf_{(j,k)\in\mathcal{J}\,:\,\theta_i \leq \theta_j} u^\delta(Z_{j:k}, v_{j:k}, \varphi, \kappa(\cdot)) \text{ for all } i \in \{1,\ldots,n\}\Bigg) \geq 1 - 2|\mathcal{J}|\delta.\end{aligned}$$

Choosing $\alpha = 2|\mathcal{J}|\delta$ provides the claim. ■

Proof: The binomial, Poisson and negative binomial distributions belong to the additive form of the EDF for carefully chosen canonical parameters, volumes, dispersion parameters and cumulant functions; we refer to Table 3.3 in Jørgensen (1997). The lower and upper bounds defined in (13)–(14) can be explicitly expressed for the following three cases using the weighted partial sums $Z_{j:k}$ and aggregated volumes $v_{j:k}$ in (8)–(9).

Binomial case. The lower bound in (18) can be expressed as

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i\geq\mu_k} l^{\delta}(Z_{j:k}, v_{j:k}, \varphi),$$

with

$$l^{\delta}(Z_{j:k}, v_{j:k}, \varphi) = \inf\left\{\mu \in (0,1) \,|\, F^*(v_{j:k}Z_{j:k}/\varphi; v_{j:k}/\varphi, \mu) \leq 1-\delta\right\},$$

and where $F^*(\cdot; m, \mu)$ denotes the left-continuous, right-limit distribution of a $\text{Bin}(m,\mu)$ random variable. Let $I_\mu(x,y)$ be the regularized incomplete beta function. For $l \in \{1,\dots,m\}$, we have

$$\begin{aligned} F^*(l; m, \mu) \leq 1-\delta &\iff 1 - I_\mu(1+(l-1), m-(l-1)) \leq 1-\delta, \\ &\iff 1 - G_B(\mu; l, 1+m-l) \leq 1-\delta, \\ &\iff \mu \geq q_B(\delta; l, 1+m-l), \end{aligned}$$

where $G_B(y;\alpha,\beta)$ and $q_B(\delta;\alpha,\beta)$ denote the distribution and the δ-quantile of a $\text{Beta}(\alpha,\beta)$ random variable, respectively. This shows the claim for the lower bound. The result for the upper bound in (19) follows similarly as

$$\begin{aligned} F(l; m, \mu) \geq \delta &\iff 1 - I_\mu(1+l, m-l) \geq \delta, \\ &\iff 1 - G_B(\mu; 1+l, m-l) \geq \delta, \\ &\iff \mu \leq q_B(1-\delta; 1+l, m-l), \end{aligned}$$

where $l \in \{0,\dots,m-1\}$. This completes the proof.

Poisson case. The lower bound in (20) can be expressed as

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i\geq\mu_k} l^{\delta}(Z_{j:k}, v_{j:k}, \varphi),$$

with

$$l^{\delta}(Z_{j:k}, v_{j:k}, \varphi) = \inf\left\{\mu \in (0,\infty) \,|\, F^*(v_{j:k}Z_{j:k}/\varphi; \mu v_{j:k}/\varphi) \leq 1-\delta\right\},$$

and where $F^*(\cdot; \mu v/\varphi)$ denotes the left-continuous, right-limit distribution of a $\text{Poi}(\mu v/\varphi)$ random variable. For $l \in \mathbb{N}$, we have

$$\begin{aligned} F^*(l; \mu v/\varphi) \leq 1-\delta &\iff \frac{\Gamma(l, \mu v/\varphi)}{(l-1)!} \leq 1-\delta, \\ &\iff \frac{\Gamma(l, \mu v/\varphi)}{\Gamma(l)} \leq 1-\delta, \\ &\iff 1 - G_\Gamma(\mu v/\varphi; l, 1) \leq 1-\delta, \\ &\iff \mu v/\varphi \geq q_\Gamma(\delta; l, 1), \\ &\iff \mu \geq \frac{\varphi q_\Gamma(\delta; l, 1)}{v}, \end{aligned}$$

where $G_\Gamma(y;\gamma,c)$ and $q_\Gamma(\delta;\gamma,c)$ denote the distribution and the δ-quantile of a $\Gamma(\gamma,c)$ random variable, respectively. This shows the claim for the lower bound. Similarly, the result for the upper bound in (21) follows from

$$\begin{aligned} F(l; \mu v/\varphi) \geq \delta &\iff \frac{\Gamma(1+l, \mu v/\varphi)}{l!} \geq \delta, \\ &\iff \frac{\Gamma(1+l, \mu v/\varphi)}{\Gamma(1+l)} \geq \delta, \\ &\iff 1 - G_\Gamma(\mu v/\varphi; 1+l, 1) \geq \delta, \\ &\iff \mu v/\varphi \leq q_\Gamma(1-\delta; 1+l, 1), \\ &\iff \mu \leq \frac{\varphi q_\Gamma(1-\delta; 1+l, 1)}{v}, \end{aligned}$$

where $l \in \mathbb{N}_0$.

Negative binomial case. The lower bound in (22) can be expressed as

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \geq \mu_k} l^{\delta}(Z_{j:k}, v_{j:k}, \varphi),$$

with

$$l^{\delta}(Z_{j:k}, v_{j:k}, \varphi) = \inf\left\{\mu \in (0,\infty) \,|\, F^*(v_{j:k}Z_{j:k}/\varphi; \mu, v_{j:k}/\varphi) \leq 1-\delta\right\},$$

and where $F^*(\cdot;\mu,\gamma)$ denotes the left-continuous, right-limit distribution of a $\text{NegBin}(\mu,\gamma)$ random variable with mean parameter $\mu > 0$ and shape parameter $\gamma > 0$. Define $p = \mu/(1+\mu)$ and let $I_p(x,y)$ be the regularized incomplete beta function. For $l \in \mathbb{N}$, we have

$$\begin{aligned}
F^*(l;\mu,v/\varphi) \leq 1-\delta &\iff 1 - I_p(1+(l-1), v/\varphi) \leq 1-\delta,\\
&\iff 1 - G_B(p; l, v/\varphi) \leq 1-\delta,\\
&\iff p \geq q_B(\delta; l, v/\varphi),\\
&\iff \frac{\mu}{1+\mu} \geq q_B(\delta; l, v/\varphi),\\
&\iff \mu \geq \frac{q_B(\delta; l, v/\varphi)}{1 - q_B(\delta; l, v/\varphi)},
\end{aligned}$$

where $G_B(y;\alpha,\beta)$ and $q_B(\delta;\alpha,\beta)$ denote the distribution and the δ-quantile of a $\text{Beta}(\alpha,\beta)$ random variable, respectively. This shows the claim for the lower bound. Similarly, the result for the upper bound in (23) follows from

$$\begin{aligned}
F(l;\mu,v/\varphi) \geq \delta &\iff 1 - I_p(1+l, v/\varphi) \geq \delta,\\
&\iff 1 - G_B(p; 1+l, v/\varphi) \geq \delta,\\
&\iff p \leq q_B(1-\delta; 1+l, v/\varphi),\\
&\iff \frac{\mu}{1+\mu} \leq q_B(1-\delta; 1+l, v/\varphi),\\
&\iff \mu \leq \frac{q_B(1-\delta; 1+l, v/\varphi)}{1 - q_B(1-\delta; 1+l, v/\varphi)},
\end{aligned}$$

where $l \in \mathbb{N}_0$. ■

Proof: The gamma and normal distributions belong to the EDF for carefully chosen canonical parameters, volumes, dispersion parameters and cumulant functions; we refer to Table 3.1 in Jørgensen (1997). The lower and upper bounds defined in (13)–(14) can be explicitly expressed for the following two cases using the weighted partial sums $Z_{j:k}$ and aggregated volumes $v_{j:k}$ in (8)–(9).

Gamma case. The lower bound in (24) can be expressed as

$$L^{\alpha}_{\mathbf{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \geq \mu_k} l^{\delta}(Z_{j:k}, v_{j:k}, \varphi),$$

with

$$l^{\delta}(Z_{j:k}, v_{j:k}, \varphi) = \inf\left\{\mu \in (0,\infty) \,|\, F^*(Z_{j:k}; v_{j:k}/\varphi, v_{j:k}/(\varphi\,\mu)) \leq 1-\delta\right\},$$

and where $F^*(\cdot; \nu/\varphi, \nu/(\varphi\mu))$ denotes the left-continuous, right-limit distribution of a $\Gamma(\nu/\varphi, \nu/(\varphi\mu))$ random variable. Let $\mathcal{G}(x, y)$ be the lower incomplete gamma function. For $l > 0$, we have

$$\begin{aligned} F^*(l; \nu/\varphi, \nu/(\varphi\mu)) \leq 1-\delta &\iff \frac{\mathcal{G}(\nu/\varphi, \nu l/(\varphi\mu))}{\Gamma(\nu)} \leq 1-\delta, \\ &\iff G_\Gamma(\nu/(\varphi\mu); \nu/\varphi, l) \leq 1-\delta, \\ &\iff \frac{\nu}{\varphi\mu} \leq q_\Gamma(1-\delta; \nu/\varphi, l), \\ &\iff \nu/\varphi \leq \mu \cdot q_\Gamma(1-\delta; \nu/\varphi, l), \\ &\iff \mu \geq \frac{\nu/\varphi}{q_\Gamma(1-\delta; \nu/\varphi, l)}, \end{aligned}$$

where $G_\Gamma(y; \alpha, \beta)$ and $q_\Gamma(\delta; \alpha, \beta)$ denote the distribution and the δ-quantile of a $\Gamma(\gamma, c)$ random variable, respectively. This shows the claim for the lower bound and as the left-continuous, right-limit distribution of a gamma random variable coincides with the distribution of this random variable, the result for the upper bound in (25) follows similarly.

Normal case. The lower bound in (26) can be expressed as

$$L^{\alpha}_{\boldsymbol{Y},i} = \sup_{(j,k)\in\mathcal{J}\,:\,\mu_i \geq \mu_k} l^{\delta}(Z_{j:k}, \nu_{j:k}, \varphi),$$

with

$$l^{\delta}(Z_{j:k}, \nu_{j:k}) = \inf\left\{\mu \in \mathbb{R} \,|\, F(Z_{j:k}; \mu, \nu_{j:k}/\varphi) \leq 1-\delta\right\},$$

and where $F(\cdot; \mu, \nu_{j:k}/\varphi)$ denotes the distribution of a $\mathcal{N}(\mu, \varphi/\nu_{j:k})$ random variable. This pointwise infimum is always attained as the map

$$\mu \mapsto \mathbb{P}(Y_{\mu,\nu_{j:k}/\varphi} \leq z), \text{ where } Y_{\mu,\nu_{j:k}/\varphi} \sim \mathcal{N}(\mu, \varphi/\nu_{j:k}),$$

is continuous for fixed $z \in \mathbb{R}$. The lower band $l = l^{\delta}(Z_{j:k}, \nu_{j:k}, \varphi)$ thus satisfies

$$\begin{aligned} F(Z_{j:k}; l, \nu_{j:k}/\varphi) = 1-\delta &\iff \Phi\left(\sqrt{\nu_{j:k}/\varphi}\,(Z_{j:k}-l)\right) = 1-\delta, \\ &\iff \sqrt{\nu_{j:k}/\varphi}\,(Z_{j:k}-l) = \Phi^{-1}(1-\delta), \\ &\iff l = Z_{j:k} - \frac{\Phi^{-1}(1-\delta)}{\sqrt{\nu_{j:k}/\varphi}}. \end{aligned}$$

A similar derivation provides the result for the upper bound in (27). ■

Proof: From Proposition B1 of Dimitriadis et al. (2023) and by choosing $\tau = \sqrt{2\sigma^2 \log(1/\delta)}$, we know that

$$U^{\alpha,\,\mathrm{YB}}_{\boldsymbol{Y},i} = Z^{\mathrm{Iso}}_{j:k} + \frac{\tau}{\sqrt{k-j+1}},$$

for some pair $(j,k) \in \mathcal{J}^{\,\mathrm{full}}$ with $j = i$ and such that either $\widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_k < \widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_{k+1}$ or $k = n$. Moreover, we have

$$Z_{j:k} \leq Z^{\mathrm{Iso}}_{j:k}, \text{ whenever } \widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_k < \widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_{k+1} \text{ or } k = n,$$

due to a property of isotonic regression (we refer, e.g., to Characterization II provided by Henzi et al., 2022). By defining the new set

$$\widetilde{\mathcal{J}}_i = \left\{(i,k) \in \mathcal{J}^{\,\mathrm{full}} \,\middle|\, \widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_k < \widehat{\mu}^{\,\mathrm{Iso}}(\boldsymbol{Y}, \mathbf{1})_{k+1} \text{ or } k = n\right\},$$

for $1 \leq i \leq n$, we thus obtain

$$\begin{aligned} U^{\alpha,\,\mathrm{YB}}_{\boldsymbol{Y},i} &= \min_{(j,k)\in\widetilde{\mathcal{J}}_i} Z^{\mathrm{Iso}}_{j:k} + \frac{\sqrt{2\sigma^2\log(1/\delta)}}{\sqrt{k-j+1}} \\ &\geq \min_{(j,k)\in\widetilde{\mathcal{J}}_i} Z_{j:k} + \frac{\sqrt{2\sigma^2\log(1/\delta)}}{\sqrt{k-j+1}} \\ &\geq \min_{(j,k)\in\widetilde{\mathcal{J}}_i} Z_{j:k} - \frac{\sigma\,\Phi^{-1}(\delta)}{\sqrt{k-j+1}} \\ &\geq \min_{(j,k)\in\mathcal{J}^{\,\mathrm{full}}:\,\mu_i\leq\mu_j} Z_{j:k} - \frac{\sigma\,\Phi^{-1}(\delta)}{\sqrt{k-j+1}} = U^{\alpha}_{\boldsymbol{Y},i}, \end{aligned}$$

where we used $\sqrt{2\log(1/\delta)} \geq -\Phi^{-1}(\delta)$ for all $\delta \in (0,1)$. A similar computation provides the result for the lower band. ■

Proof: Define the set

$$A = \left\{ L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{X}_l) \leq \mu^*_{\pi}(\boldsymbol{X}_l) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{X}_l) \text{ for all } l \in \{1,\ldots,n\} \right\}, \tag{A1}$$

that lies in $\mathcal{F}$ as all the random variables involved in its definition are measurable. Instead of proving the existence of the uniform calibration band in (35) holding simultaneously for all $\boldsymbol{x} \in \mathcal{X}$, we first prove that for a.e. realization $(\boldsymbol{x}_i)_{i=1}^n$ of the features $(\boldsymbol{X}_i)_{i=1}^n$, we have

$$\mathbb{Q}_{(\boldsymbol{x}_i)_{i=1}^n}\left(L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{X}_l) \leq \mu^*_{\pi}(\boldsymbol{X}_l) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{X}_l) \text{ for all } l \in \{1,\ldots,n\} \right) \geq 1-\alpha. \tag{A2}$$

Note that since the random variables $(Y_i,\boldsymbol{X}_i)_{i=1}^n$ are independent, they satisfy

$$Y_i \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1,\ldots,\boldsymbol{X}_n = \boldsymbol{x}_n \sim \mathrm{EDF}(\theta(\boldsymbol{x}_i), \nu_i, \varphi, \kappa(\cdot)), \quad \text{for } 1 \leq i \leq n.$$

By using the permutation function $\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$ introduced in (34), the indices of these random variables can be permuted such that

$$\theta(\boldsymbol{x}_{\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(1)}) \leq \cdots \leq \theta(\boldsymbol{x}_{\tau_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(n)}),$$

whenever the features $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ satisfy $\mu^*_{\pi}(\boldsymbol{x}_i) = \mu^*(\boldsymbol{x}_i)$ for all $i \in \{1,\ldots,n\}$. The latter happens for a.e. realization of the features $(\boldsymbol{X}_i)_{i=1}^n$ and in this case, the inequality in (A2) follows from Theorem 4.1. In order to show the inequality in (35), it suffices then to notice that under Assumption 6.1, the set A in (A1) is equal to the set B given by

$$B = \left\{ L^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \leq \mu^*_{\pi}(\boldsymbol{x}) \leq U^{\alpha}_{\pi,(Y_i,\boldsymbol{X}_i)_{i=1}^n}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X} \right\}.$$

This concludes the proof. ■