Using the Softplus Function to Construct Alternative Link Functions in Generalized Linear Models and Beyond

Paul F.V. Wiemann*[1,2], Thomas Kneib[2], and Julien Hambuckers[3]

[1]Department of Statistics, University of Wisconsin–Madison, WI, USA
[2]Chair of Statistics, University of Göttingen, Germany
[3]HEC Liège, Department of Finance, University of Liège, Belgium

Response functions that link regression predictors to properties of the response distribution are fundamental components in many statistical models. However, the choice of these functions is typically based on the domain of the modeled quantities and is usually not further scrutinized. For example, the exponential response function is often assumed for parameters restricted to be positive, although it implies a multiplicative model, which is not necessarily desirable or adequate. Consequently, applied researchers might face misleading results when relying on such defaults. For parameters restricted to be positive, we propose to construct alternative response functions based on the softplus function. These response functions are differentiable and correspond closely to the identity function for positive values of the regression predictor implying a quasi-additive model. Consequently, the proposed response functions allow for an additive interpretation of the estimated effects by practitioners and can be a better fit in certain data situations. We study the properties of the newly constructed response functions and demonstrate the applicability in the context of count data regression and Bayesian distributional regression. We contrast our approach to the commonly used exponential response function.

Keywords— generalized linear model; link function; regression; response function; softplus;

1. Introduction

Regression analysis is an essential tool for understanding relationships between variables in many fields, including economics, social sciences, and engineering. Response functions and their inverse, known as link functions, play a pivotal role in modern regression methods as they relate distribution parameters to predictors. While default response functions such as the logistic and exponential functions are widely used, they may not always provide the best fit for the specific problem at hand. The selection of an appropriate response function impacts the model in two ways: First, unsuitable response functions lead to poor model fit, potentially

*corresponding author: pwiemann@uni-goettingen.de; Humboldtallee 3, 37073 Göttingen, Germany

violating model assumptions. Second, the choice of the response function significantly impacts the interpretation of covariate effects in a regression model.

The exponential function is popular for strictly positive parameters due to the interpretability of effects as multiplicative. However, this assumption of multiplicativity can be restrictive, as additive effects are often desired in statistical modeling. Traditionally, researchers resort to not using a response function if they desire additivity of effects. However, this can be problematic when the parameter modeled is strictly positive. Certain covariate combinations might lead to a negative value and, thus, invalidate the model.

In this paper, we propose constructing novel types of response functions for strictly positive parameters based on the softplus function $\text{softplus}(x) = \log(1 + \exp(x))$. These response functions may prove valuable when seeking a model with an additive interpretation of effects since they allow for a quasi-additive interpretation over a certain part of the range of predictor values while guaranteeing that the function's value complies with the positivity restriction. Furthermore, it is a strictly increasing bijective function mapping the real values to its positive subset. Therefore, it is an eligible response function for positively restricted distribution parameters. It can be used instead of the exponential response function (providing a quasi-additive model) or the identity response function (avoiding the restriction of regression coefficients).

In addition to the quasi-additive interpretation, the softplus function enables the design of response functions with interesting properties. Augmented with an additional parameter, it yields further flexibility to model the data given, (i) it avoids exponential growth, which can be an issue under certain covariate combinations, and (ii) it enables the construction of an exponential-like function that avoids potential numerical overflow when evaluating it for large positive predictor values.

For the choice of the response function, most researchers rely on default choices such as the logistic response function for parameters restricted to the unit interval (e.g., probabilities) or the exponential response function for strictly positive parameters. In generalized linear models (GLMs, McCullagh and Nelder, 1989), these defaults can often be justified by their characterization as natural link functions arising in the context of exponential families. In other cases, the default response functions are chosen to entail specific modes of interpretation, e.g., multiplicative effects on odds in the case of the logistic response function or multiplicative effects on the parameter of interest in the case of the exponential response function. The straightforward interpretability is also the reason why Fahrmeir et al. (2013) recommend using the exponential response function for gamma-distributed responses in the GLM framework instead of the canonical link function. Additionally, special situations require the use of rather exotic response functions. For example, the square root link for Poisson distributed data helps to mimic a least squares estimation on the square root transformed data with a likelihood approach within the GLM framework.

To determine the correct response function in a set of candidate functions, Ntzoufras et al.

(2003) propose employing a reversible jump algorithm instead of resorting to model selection criteria. An alternative to pre-chosen response functions is to estimate the response function flexibly from the data. The most well-known example of this approach is the single-index model introduced by Ichimura (1993). The kernel-based single-index models share the disadvantage that the estimated response function is often too flexible. To counter this characteristic, Yu and Ruppert (2002) and Yu et al. (2017) introduced penalization to single-index models. Recently, Spiegel et al. (2019) presented an approach that combines the single-index models based on penalized splines with the flexibility of generalized additive models (Hastie and Tibshirani, 1986).

One practical challenge when employing flexible link functions lies in interpreting the resulting model since restrictions must be assigned to the regression predictor to render the response function estimate identifiable. In contrast, simple, fixed response functions considerably facilitate interpretation. Having easily interpretable effects may be why the exponential response function is still the most common approach for positively bounded parameters.

Regardless of how well default choices can be justified, no single response function can fit all situations. Pregibon (1980) points out that a misspecification of the response function is systematic model misspecification. Moreover, domain-specific knowledge about the application can invalidate a multiplicative model entirely and, e.g., suggest an additive model. Therefore, investigating alternative response functions is a worthwhile and relevant endeavor.

The remainder of this paper is structured as follows: Section 2 introduces the softplus response function, justifies the quasi-additive interpretation, and gives a guideline for its proper use. Furthermore, Section 2 describes statistical inference when employing the softplus response function. Section 3 investigates the softplus response function in simulation studies. The practical applicability of softplus-based regression specifications is demonstrated in Section 4. The final Section 5 summarizes our findings and discusses future research directions. The code associated with this manuscript can be found on GitHub[1].

2. The Softplus Function

2.1. Definition and Properties

The softplus function (Dugas et al., 2001) has primarily been employed in deep neural networks (Zheng et al., 2015) as a smooth and differentiable approximation of the rectifier function. In the statistics literature, the rectifier function is commonly known as a linear spline, defined as $x_+ = \max\{0, x\}$. The softplus function maps real numbers to their positive counterparts, such that the output is always in the positive subset $\mathbb{R}_+$. We use a generalized version of the softplus function (similar to Liu and Furber, 2016) that incorporates an extra parameter $a > 0$ and

[1] `https://github.com/wiep/softplus-response-function`

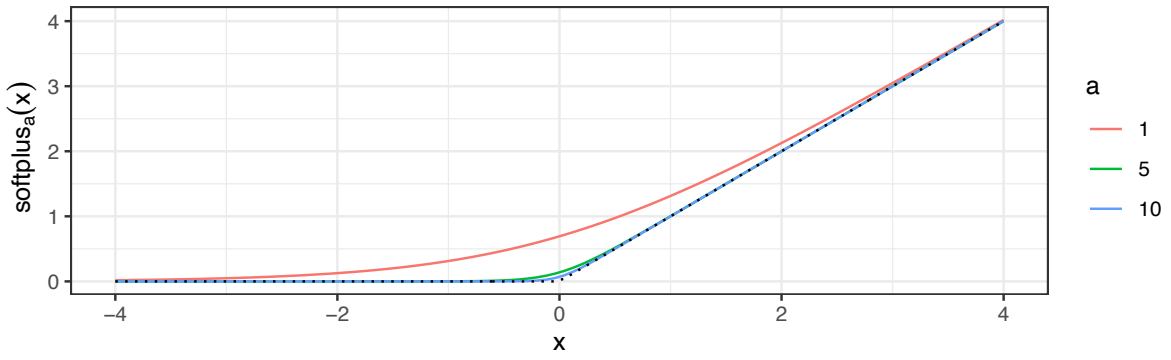

Figure 1: Plot of the softplus function (left) for different values of softplus parameters a. The approximated linear spline is shown as the dotted line.

can be defined using the equation

$$\text{softplus}_a(x) = \frac{\log\left(1 + \exp(ax)\right)}{a}. \tag{1}$$

When $a = 1$, the function reduces to its simple form. Figure 1 illustrates the softplus response function for different values of a. By introducing the softplus parameter a, we can control the approximation error with respect to the linear spline. It can be shown that for every $\varepsilon > 0$, exists some $a > 0$ such that

$$0 < \text{softplus}_a(x) - \max\{0, x\} \leq \log(2)/a < \varepsilon$$

holds for all $x \in \mathbb{R}$. The largest approximation error is at $x = 0$ as visually indicated by Figure 1 and follows from the reformulation for numerical stability discussed in the Appendix (see Equation (3) in Section A.1). Besides, one can observe in Figure 1 that the softplus function follows the identity function very closely in the positive domain and rapidly approaches zero in the negative domain for $x \to -\infty$. This behavior can be further accentuated by increasing the parameter a. Therefore, the softplus parameter a can be used to control how long the quasi-linear relationship should be maintained when approaching zero and consequently, how fast the boundary of the linked distribution parameter is approached.

The adoption of the softplus function as a response function features two major advantages:

- It translates the additivity of effects on the predictor level to the parameter space for a majority of the relevant distribution parameter space while always guaranteeing the positivity of the distribution parameter. This is achieved by being quasi-linear in its argument as long as the predictor is large enough for a given value of a.

- The softplus function allows for a straightforward interpretation of the covariate effects. When the predictor value is large enough, the effects can be interpreted directly on the

parameter. Consider the linear effect of a covariate x with the corresponding regression coefficient β. A difference in x by one unit is associated with a difference of β in the predictor and, when using the softplus function, also with a difference of almost β in the distribution parameter or expressed as a formula $\text{softplus}(\beta(x+1)) \approx \beta + \beta x$.

Clearly, the quasi-additive interpretation is no longer valid once the argument of the softplus function is not within the approximately linear part of the softplus function, which we define in detail in Section 2.2. However, by choosing a sufficiently large a, the linear part covers almost the entire positive domain. In the negative domain and for a sufficiently large a, a slight change of the covariate usually does not cause a relevant change in the parameter value since the softplus function outputs values very close to zero. To ensure the validity of this interpretation, it is necessary to check the range of values of the linear predictor for the observations in the data set. Most of them should be located within the linear part of the softplus function.

The quasi-additive interpretation contrasts the usual multiplicative interpretation for positively constrained parameters that arise from using the exponential response function. For example, $\exp(\beta(x+1)) = \exp(\beta)\exp(\beta x)$ leads to the interpretation that a change of one unit in the covariate is associated with a multiplicative change of $\exp(\beta)$ units on the parameter.

In contrast to large values for a, which enable the quasi-additive interpretation, with the choice of sufficiently small values for a, the softplus function resembles the exponential function with a scaled and shifted argument. This becomes more obvious when taking into account that $\log(x+1)$ is almost linear in x, for $|x| \ll 1$, and thus $\log(1+\exp(ax))/a \approx \exp(ax - \log(a))$ for $\exp(ax) \ll 1$. Consequently, the choice of the softplus parameter a allows to continuously vary between an identity-like response function (for $a \to \infty$) and the exponential response function (with scaled and shifted argument for $a \to 0$). This property facilitates the construction of the *softplus exponential* response function, approximating the exponential function for small arguments but with a limiting gradient (see the Supplementary Material for details). The function can be a viable alternative if one desires an exponential-like response function, but unbounded growth is an (e.g., numerical) issue.

2.2. Linear Part of the Softplus Function

A crucial consideration in employing the softplus function as a response function in regression modeling is identifying the region where it behaves approximately linearly. Only within this section the interpretation of regression effects is quasi-additive. However, due to the nonlinearity of the softplus function across its entire domain, it is necessary to specify the conditions under which this interpretation is valid.

Consider a change in the regression predictor of $\gamma \in \mathbb{R}$ starting from the value $\eta \in \mathbb{R}$. Using the identity function as the response function, the change in the parameter modeled is equally γ. Thus, a regression effect of size γ can be interpreted as influencing the parameter by γ.

When working with the softplus function, it is important to note that the same change in the argument can lead to different changes depending on the location of η. We need to identify the subset of the domain in which the function value change approximately matches the argument's change. This requires an assessment of the error induced by the softplus function and establishing an acceptable error threshold. The error induced by using the softplus function is

$$\text{error}_a(\eta, \gamma) = \gamma - (\text{softplus}_a(\eta + \gamma) - \text{softplus}_a(\eta)) \tag{2}$$

while the relative error is

$$\text{rerr}_a(\eta, \gamma) = \frac{\text{error}_a(\eta, \gamma)}{\gamma} = 1 - \frac{\text{softplus}_a(\eta + \gamma) - \text{softplus}_a(\eta)}{\gamma}.$$

We say that interpreting a regression effect γ directly on the parameter is valid if, for some pre-specified acceptable relative error α, the predictor η is in the interval $[T, \infty) \subseteq \mathbb{R}$ for which $\text{rerr}_a(T, \gamma) < \alpha$ holds. The acceptable relative error, of course, depends on the application and should be chosen carefully. In this work, we consider a relative error of 5% acceptable.

2.3. Choosing the Softplus Parameter

There are two approaches to selecting a value for the softplus parameter a: based on the model fit or to achieve the desired interpretability. In cases where interpretability is of higher priority, we recommend selecting a value for a that permits the quasi-additive interpretation with an relative error margin of no more than five percent. More precisely, after fitting the model, it should be confirmed that the estimated predictors for (almost) all observations are within the interval $[T, \infty)$ when interpreting effects of size γ. However, since this assessment can only be done after fitting the model, testing multiple values of a might be required. Additionally, we want to stress that it is at least of equal importance to check that the model assumptions are met.

2.4. Inference

Replacing the standard exponential response function with the softplus response functions introduced in this paper does not cause major difficulties as long as the parameter a is fixed. Since the softplus-based response functions are continuously differentiable, standard maximum likelihood inference can be used in GLM-type setting where only the derivative of the link function in the definition of the working weights and the working observations of iteratively weighted least squares (IWLS) optimization have to be replaced (see for example Fahrmeir et al., 2013, for details on the IWLS algorithm).

In our simulations and applications, we rely on the Bayesian paradigm for statistical infer-

ence, since this allows us to apply the softplus-based response functions also beyond GLMs, for example in generalized structured additive regression models with complex additive predictor (Brezger and Lang, 2006) or in structured additive distributional regression models (Klein et al., 2015). For the pre-specified response function with parameter a, we rely on an MCMC simulation scheme where we update the parameter vector block-wise with a Metropolis-Hastings (MH) step in conjunction with IWLS proposals (Gamerman, 1997; Klein et al., 2015). IWLS proposals automatically adapt the proposal distribution to the full conditional distribution and therefore avoid manual tuning which is, for example, required in random walk proposals. This is achieved by approximating the full conditional distribution with a multivariate normal distribution whose expectation and covariance matrix match mode and curvature of the full conditional distribution at the current state of the chain which can be determined based on the IWLS algorithm of frequentist maximum likelihood estimation without requiring the normalizing constant of the full posterior. More precisely, the parameters of the proposal distribution are determined by executing one Fisher-Scoring step and using the new position as the mean for the multivariate normal distribution while the covariance of the normal distribution is set to be the inverted observed Fisher-Information at the old position. More formally, let $\boldsymbol{\theta}$ be the vector of parameters that should be updated within a MH-block and let $\mathcal{L}(\boldsymbol{\theta})$ be the unnormalized full conditional posterior log density with respect to the parameter vector $\boldsymbol{\theta}$. The proposal distribution is Normal with mean $\boldsymbol{\mu} = \boldsymbol{\theta} + \boldsymbol{g}\boldsymbol{F}^{-1}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{F}^{-1}$ where $\boldsymbol{g}$ denotes the gradient of $\mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{F}$ denotes the Hessian of $-\mathcal{L}(\boldsymbol{\theta})$ each with respect to $\boldsymbol{\theta}$. This sampling scheme has proven effective in various regression models (Lang and Brezger, 2004; Klein et al., 2015; Klein and Kneib, 2016). Our implementation relies on an extension of the R-package `bamlss` (Umlauf et al., 2018) which implements methodology described above.

3. Simulations

With our simulations, we

- conduct a proof of concept evaluation that investigates how reliable models with the softplus response function can be estimated and whether the resulting credible intervals are well calibrated,
- study the ability of model selection criteria to distinguish between data-generating processes involving either the softplus or the exponential response function, and

For all simulations, estimation is conducted within the Bayesian paradigm and carried out in `R` (R Core Team, 2022) with the package `bamlss` (Umlauf et al., 2018). We use a similar data-generating process varying only the sample size and the response function. In particular, we employ data generated from a Poisson distribution with expectation $\mathrm{E}(y_i) = \lambda_i = \mathrm{h}(\eta_i)$ where h denotes the response function. For a single observation, we choose the predictor structure

$\eta = 1.0 + 0.5x_1 + 1.0x_2 + 2.0x_3$ with x_1, x_2, x_3 being independent and identically uniform distributed on the interval from -1 to 1. All observations are stochastically independent. Throughout this section, we assume flat priors for all regression coefficients.

3.1. Point Estimates and Credible Intervals

In the first segment of the simulation studies, we present the results that demonstrate the reliability of using the softplus function as a response function with well-calibrated posterior means and credible intervals. The simulation scenarios feature the sample sizes $n \in \{50, 100, 200, 500, 1000, 5000\}$ and the softplus parameter a is set to a value from $\{1, 5, 10\}$. Within each scenario, we carry out 6150 replications, a number that is determined by considering the coverage of the true parameter as a Bernoulli experiment and imposing that the normal approximation of the 95% confidence interval for a coverage rate of 0.8 does not surpass 0.02.

In the first part of the simulation studies, we show that the softplus function can be reliably used as a response function and that posterior means and credible intervals are well-calibrated. The simulation scenarios feature the sample sizes $n \in \{50, 100, 200, 500, 1000, 5000\}$ and the softplus parameter a was set to a value from $\{1, 5, 10\}$. Within each scenario, we simulated 6150 replications. The number 6150 is determined by considering the coverage of the true parameter as a Bernoulli experiment and requiring that the normal approximation of the 95% confidence interval for a coverage rate of 0.8 is smaller than 0.02. We run one MCMC chain with 12000 iterations, of which the first 2000 iterations are deemed the burn-in phase.

The results are visualized using box plots of the posterior mean estimates in Figure 2 and coverage ranges of 80% and 95% credible intervals in Figure 3. In summary, we draw the following conclusions:

Bias For most simulation settings, the bias is negligibly small. The only exception is a small sample size in conjunction with a rather large softplus parameter a, where we can observe a slight bias, especially for the intercept. However, one must keep in mind that a large parameter a implies an almost linear link function such that there is considerably less variability (and therefore information) in data sets with the softplus response function compared to the exponential response function with the same linear predictor value. Furthermore, the softplus function maps even small negative values to a positive value close to zero and thus close to the boundary of the parameter space. The bias quickly diminishes as the sample size increases.

Coverage rates Figure 3 supports that our Bayesian approach provides accurate 'credible intervals' for sufficiently sized samples. However, for smaller samples, the coverage rates suffer from the bias that arises due to the use of higher values for the softplus parameter.

In short, the results obtained with the softplus response function are reliable. Especially for larger sample sizes, no biases are observed and the coverage rates behave as expected. Results

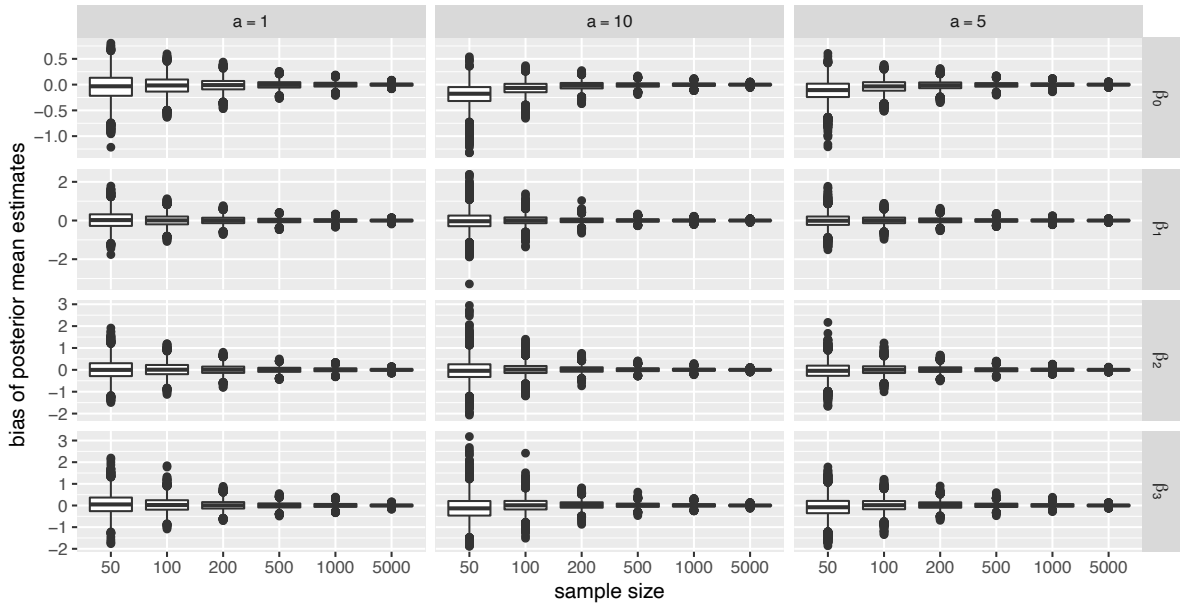

Figure 2: Box plots of deviations from posterior mean estimates of regression coefficients to the true value for different sample sizes and different softplus parameters a. Replications that include an absolute deviation larger than five for one coefficient have been excluded from plotting for better visualization. This applies to one replication with $a = 5$ and to ten replications with $a = 10$ each with a sample size of 50.

of this simulation exercise obtained with maximum likelihood inference are virtually identical and are omitted for the sake of brevity.

3.2. Model Selection Based on DIC

In this simulation setting, we study how successfully the well-established deviance information criterion (DIC, Spiegelhalter et al., 2002) can be used to discriminate between data generated by either the softplus or the exponential response function. As in the last subsection, we vary the sample size and use $a = 1$ or $a = 5$ for the softplus parameter. Each scenario is replicated 500 times, and as before, we run one MCMC chain with a burn-in phase of 2000 iterations and 10000 sampling iterations.

In Figure 4, we present the results summarized as percentages of correct model selection. In addition, we consider a more conservative model decision rule where a minimum difference in DIC has to be achieved and use 1, 10, and 100 as threshold values.

In all settings, a larger sample size leads to the correct model being recognized more frequently. Furthermore, the correct model for the same sample size is better identified if the data are generated with the exponential response function. As described above, this can be attributed to the fact that the information per observation (quantified by the expected Fisher information) is larger when generated with the exponential response function than with the

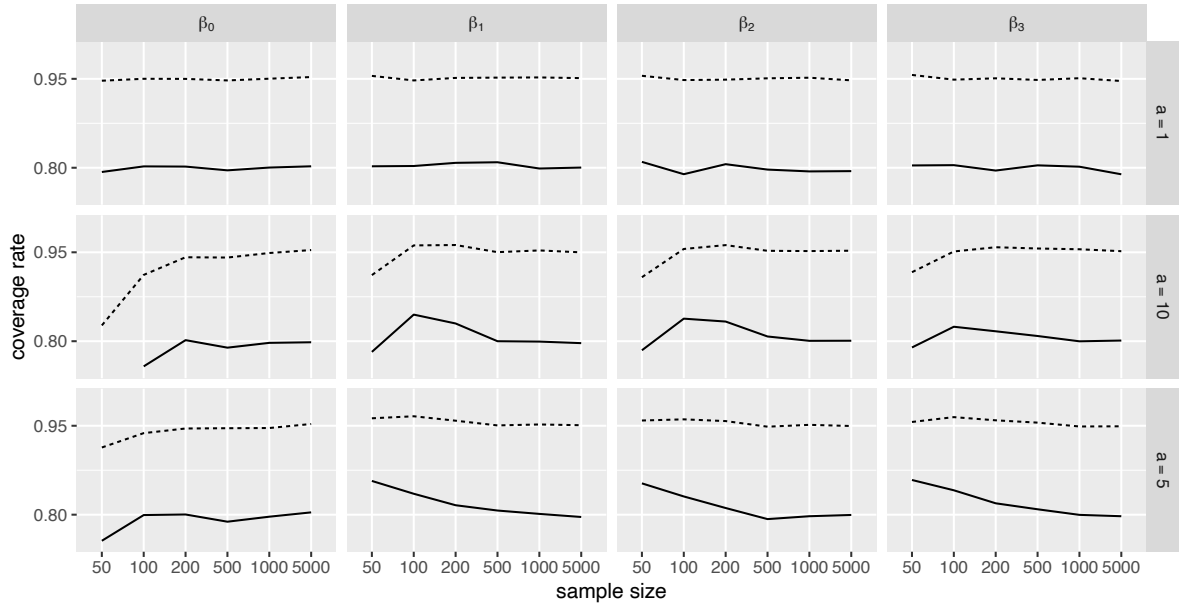

Figure 3: Coverage probability for 80% (solid line) 95% (dotted line) credible intervals for different sample sizes and different softplus parameters a.

softplus response function with $a = 5$. Thus, a larger sample size is needed to have the same probability of selecting the correct model. Yet, our simulations show that the DIC is a reliable metric for differentiating between the softplus response function and the exponential response function.

4. Applications

We present four applications to demonstrate how the softplus response function can be used in practice. We contrast our novel approach to the commonly used exponential response function. First, we employ a well-known data set from ethology about horseshoe crab mating behavior as an illustrative example for count data regression with the softplus response function (Section 4.1). Then, we illustrate the usefulness of the softplus response function in a distributional regression model with smooth effects (Section 4.2). For that, we fit a model to data from a bike-sharing service in Washington, D.C., where the softplus function can be used as a response function for the variance parameter of a normally distributed outcome.

In an application to operational loss data (Section 4.3), we demonstrate the usefulness of the softplus response function apart from the quasi-additive interpretation.

In the supplementary material, we revisit the horseshoe crab data estimating the limiting gradient of the softplus exponential function, which suggests using a linear response function. Furthermore, using data from the Munich rent index, we demonstrate similarities between results obtained from quantile regression and a location-scale model with the softplus response

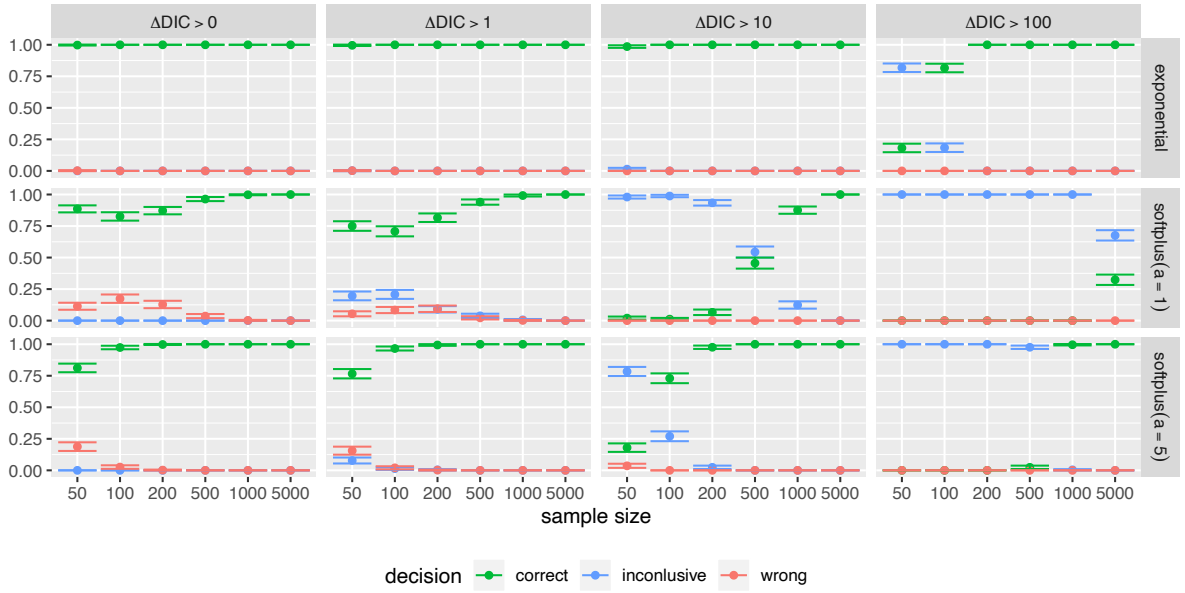

Figure 4: Percentages of correct model selections based on DIC differences with thresholds 0, 1, 10 and 100 when data are generated with the exponential response function (top row) and the softplus response function (second and third row).

function.

4.1. Horseshoe Crabs

Brockmann (1996) investigates horseshoe crab mating behavior. Horseshoe crabs have a strongly male-biased sex ratio which becomes particularly apparent in spring when male and female horseshoe crabs arrive in pairs at the beach, ready to spawn. Unattached males also come to the beach, gather around females and try to gain fertilization at the expense of the attached males. Brockmann (1996) shows that the number of unattached males, so-called satellites, gathering around a couple depends mainly on the properties of the female crab and, to a lesser extent, on environmental factors.

Agresti (2013)[2] and Kleiber and Zeileis (2016) reanalyze these data using count data regression techniques to model the number of satellite males. Agresti (2013) assumes the response value to be Poisson or negative binomial distributed, and for each response distribution, he compares the exponential response function and the identity response function. He finds that the negative binomial regression model with identity response function fits the data best among these four models. Kleiber and Zeileis (2016) extend this approach by using hurdle models to allow excess zeros. The authors favor the negative binomial hurdle model with an exponential response function. However, they omit results for the identity response function since they

[2]The accompanying website makes the data available; see `https://users.stat.ufl.edu/~aa/cda/cda.html`.

Table 1: The table displays the DIC values broken down by response function and response distribution for each model fitted. ZA indicates the zero-adjusted response distribution.

	negbin (ZA)	negbin
exp	716	740
softplus	<u>715</u>	738

claim that the negative binomial hurdle model is superior with respect to predictions compared to the negative binomial model with identity response function favored by Agresti (2013). Their argumentation includes that, in contrast to the identity response function, the exponential response function avoids negative predictions for small carapace widths. The softplus function prevents negative predictions as well.

To illustrate how the softplus function can be used to model the bounded expectation of a count data model, we extend the analyses mentioned. We use the softplus function with $a = 5$ as a response function in negative binomial regression models with and without accounting for excess zeros. Following Kleiber and Zeileis (2016), the carapace width and a numeric coding of the color variable are used as regressors in all models. All models are fitted with `bamlss` using uninformative priors on all coefficients.

We use the DIC to compare the eight models' relative performances (see Table 1). Similar to Kleiber and Zeileis (2016), we find that the negative binomial hurdle models fit best, and the DIC slightly favors the softplus response function. The slight difference in fit between the response functions is not surprising since Kleiber and Zeileis (2016) already point out that, given at least one satellite, neither carapace width nor color seems to have a significant contribution. Note that an intercept-only model does not depend on the response function used since the intercept parameter can adapt to the response function yielding the same distribution parameter. Consequently, the limited impact of the response function in the zero-adjusted model is expected.

Nonetheless, the application gives insight into the usefulness of the softplus function. In Figure 5, we display the expected number of satellites predicted as a function of carapace width with color set to the mean value. When considering the negative binomial regression, one can clearly observe the differently shaped curves reflecting the response function employed. A visual examination suggests that the exponential response function might not decay fast enough for small values of carapace width while increasing too fast for large values. On the contrary, the softplus response function seems to fit better when compared to the pattern arising from the model with zero-adjusted negative binomial response distribution (i.e., the hurdle model).

In particular, when considering the probabilities of observing zero satellites ($\mathrm{P}(y = 0)$; these are represented as dashed lines in Figure 5), the model based on the softplus function is closer the output from the zero-adjusted response distribution. This is especially true for small width values of the carapace. This is due to the fact that the softplus function with $a = 5$ approaches

zero much faster than the exponential response function does. Furthermore, quantile-quantile plots (QQ-plots) of the randomized quantile residuals (RQRs, Dunn and Smyth, 1996) indicate a decent fit to the data for all models with a preference for the hurdle model (see Figure 6 for one realization).

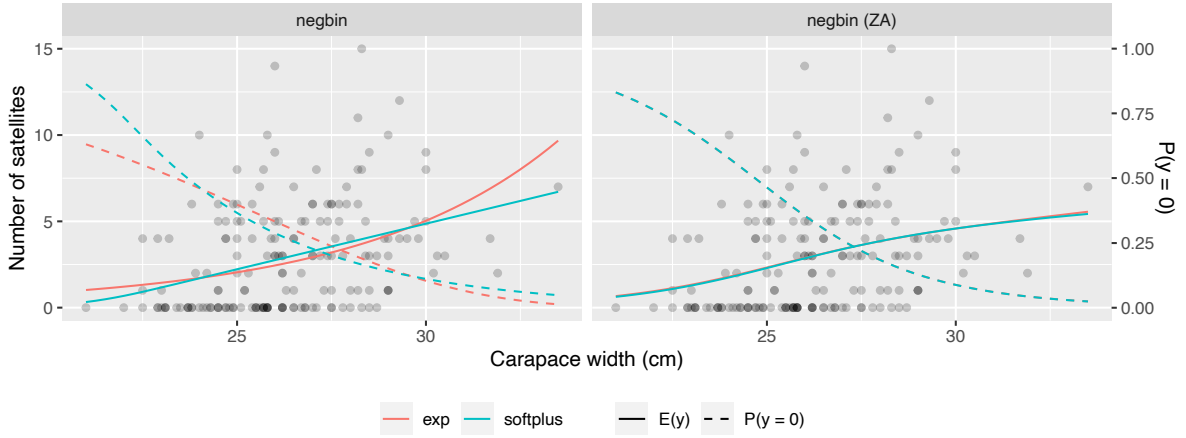

Figure 5: Plots of the expected response given carapace width and mean color for different response distributions broken down by response function. ZA indicates the the zero-adjusted response distribution. In addition, the dashed line indicates the probability of 0 satellites. The points show the observed data.

When removing the zero-adjusted model from our considerations, the DIC suggests that the model with the softplus response function has an advantage. This finding aligns with Agresti (2013) and his claim of a better fit using the identity response function compared to the exponential response function. By fitting Poisson models with softplus and exponential response functions, we can confirm the results from Agresti (2013), i.e., the quasi-linear response function fits the data better in terms of DIC. However, we omit the results here because Kleiber and Zeileis (2016) have already pointed out that the Poisson response distribution can not appropriately model the data.

To illustrate the difference in the interpretation of softplus and exponential response functions, we focus on the model assuming a negative binomial distributed response without adjusting for zeros since the impact of the different link functions becomes almost indistinguishable when adjusting for zeros. Posterior means of the parameters are displayed in Table 2 together with the corresponding 95% credible interval (equal-tailed). For a change of 0.53, the linear threshold, as defined in Section 2.2, is 0.37, while for a change of -0.54, its value is 0.91. Notice that more than 98% and 94% of the posterior means of the linear predictor are larger than these linear thresholds. Thus, we consider the linear interpretation of the covariate effects of width and color valid for almost all observations. In particular, a change by one unit in carapace size or color would increase the expected number of satellites by 0.53 or -0.54, respectively. This is in contrast to the interpretation of the exponential response function, where the same changes

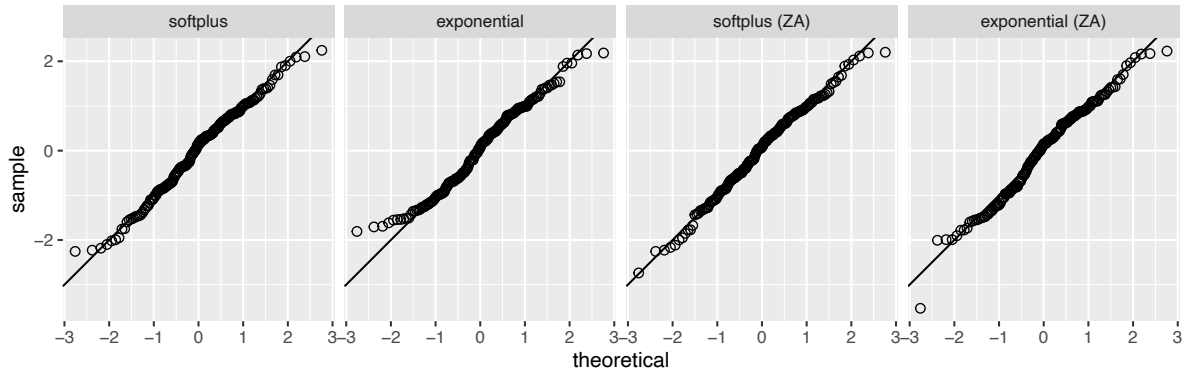

Figure 6: QQ-plots of one realization of RQRs for negative binomial distributed responses without and with zero-adjustment (indicated by ZA) employing the softplus or the exponential response function.

would lead to a multiplicative change of 1.20 and 0.77, respectively. The 95% credible interval of the effect of color includes 0 for the softplus response function but not for the exponential response function. In both cases, however, the null effect is very close to the credible interval's boundary.

4.2. CapitalBikeshare

In this section, we demonstrate the applicability of the softplus function as a response function in a Bayesian distributional regression model with flexible covariate effects. We employ data from CapitalBikeshare, a bicycle-sharing service located in Washington D.C., to analyze the mean rental duration in minutes within each hour in the years 2016 – 2017[3]. The operator might want to predict the number of trips and their expected duration in order to know how

[3]the raw data can be found at https://www.capitalbikeshare.com/system-data

Table 2: Posterior estimates of the regression coefficients on the expected value together with their 95% credible intervals (equal-tailed). The last column shows the posterior mean of the exponential function applied to the regression coefficient. Besides, σ denotes the dispersion parameter.

	softplus			exponential			
	Mean	2.5%	97.5%	Mean	2.5%	97.5%	exp(·)
(Intercept)	-9.65	-15.86	-3.59	-3.12	-5.66	-0.86	0.08
width	0.53	0.32	0.75	0.18	0.11	0.27	1.20
color	-0.54	-1.12	0.01	-0.27	-0.52	-0.04	0.77
log(σ)	1.13	0.74	1.53	1.14	0.76	1.51	-

many bikes have to be stocked. However, the variance of the average journey time is evenly essential, as it can prevent bottlenecks caused by unforeseen fluctuations.

The data have been preprocessed by the following rules:

- Trips taken by staff for service and inspection of the system have been removed, as well as trips toward test stations.
- Trips taken by non-members have been removed.
- All trips with a duration of fewer than 60 seconds have been removed since they most likely indicate a false start or users ensuring that the bike is secure by redocking it.
- Trips longer or equal to 60 minutes have been removed. This amounts to roughly 0.5% of the eligible trips. We consider them outliers since the financial incentive system of CapitalBikeshare strongly encourages users to return bikes within the first hour.

The mean rental duration per hour is, on average, based on 308.24 trips. A raw descriptive analysis of this quantity gives an average of 10.9 minutes with a standard deviation of 1.88 minutes.

The framework of structured additive distributional regression models (Rigby and Stasinopoulos, 2005; Umlauf et al., 2018) extends the generalized additive models such that multiple parameters of a response distribution can be modeled with structured additive predictors and suitable response functions. For our analysis, we assume the mean rental duration to be conditionally independent and normally distributed. We model both distributional parameters (mean and standard deviation) with structured additive predictors. In particular, the mean rental duration within each hour y_i is assumed to be independently and normally distributed with mean μ_i and standard deviation σ_i. The parameters are linked to predictors $(\eta_i^\mu, \eta_i^\sigma)$ via response functions h^μ and h^σ.

We use the same structure for both predictors and drop in the following superscript index. The predictor is specified as

$$\eta_i = f_1(\texttt{yday}_i) + f_2(\texttt{dhour}_i) + \boldsymbol{x}_i'\boldsymbol{\beta},$$

where `yday` denotes the day of the year, `dhour` denotes the hour of the day and the last term contains the intercept and additional linear effects. As linear effects, we consider a dummy variable for the year 2017 and a binary variable that encodes if the trip took place on a weekend. The smooth functions are represented by cyclic P-splines (Eilers and Marx, 1996; Hofner et al., 2016) with second-order random walk penalty (Lang and Brezger, 2004).

To illustrate the difference in interpretation between the softplus response function and the popular exponential response function, we estimate the model for both response functions, i.e., $\mathrm{h}^\sigma = \exp$ or $\mathrm{h}^\sigma = \mathrm{softplus}_{10}$. The DIC favors the softplus response function (exponential:

58152, softplus: 57943). The softplus parameter was not chosen on the basis of an information criterion but rather to enable the quasi-additive interpretation.

Detailed results concerning the mean predictor and its components are omitted since both models employ the same response functions and the results are very similar (see the Supplementary Material for a full description).

We focus on the effect of the response function concerning the standard deviation σ_i and, in particular, on the smooth effect of `dhour` and the linear effect of `weekend`. Figure 7 shows the estimated effect of the time of the day on the predictor of the standard deviation. We find that both models yield similar patterns. The standard deviation is much larger in the early hours of the day with a peak around 3 am, then drops steeply, crosses the zero line shortly after 5 am and is comparatively low in the morning. Over the course of the morning, the standard deviation increases slightly until lunchtime, then decreases over the early afternoon before starting to increase again around 4 pm. At first slightly and then very steep until it reaches its peak again in the early morning hours.

The direct interpretation of these effects is difficult, especially when using the exponential response function. For the softplus model, the estimated values of the linear predictor are larger than 0.42 and, in conjunction with a softplus parameter of 10, covariate effects can be interpreted as quasi-additive effects on the parameter (the relative error for a change of 0.0001 at predictor value 0.42 is smaller than 2%). In the following, consider the difference between the initial peak at 2.5 am and the second peak at lunchtime. In the model with the softplus response function, we observe that the predictor decreases by about 2.7 units and, consequently, the standard deviation likewise. In contrast, for the competing model, the exponential function must be applied to the predictor, and the outcome can subsequently be interpreted multiplicatively. The exponential model outputs an additive change of -1.25 on the predictor η^σ, which is reflected in a multiplicative change of the standard deviation by 3.5^{-1}.

When considering the variable `weekend` (Table 3), its effect is similar in both models: the mean rental duration exhibits more variance during the weekend, and both 95% credible intervals exclude zero. However, the interpretation of the exponential model is not straightforward: the posterior mean of the regression coefficient related to `weekend` is 0.39. In order to assess the multiplicative effect of weekend on the standard deviation, one needs to consider the posterior mean of the transformed parameter, that is $\overline{\exp(\beta_{\text{weekend}})} = 1.48$. We conclude that on a weekend, the standard deviation is 1.48 times larger than on weekdays.

The softplus model directly outputs the additive effect of weekend. We expect that the standard deviation of the mean rental duration is 0.56 minutes larger on weekends than on working days.

In both models, the interpretation of regression effects w.r.t. the predictor η^σ is straightforward, and the nature of the effects w.r.t. the standard deviation is known (i.e., additive or multiplicative). Despite this, the combination of effects and their interaction with the response

Table 3: Posterior estimates of the linear effects on the predictor of the standard deviation together with their 95% credible intervals (equal-tailed).

	exponential				softplus		
	Mean	2.5%	97.5%	exp(·)	Mean	2.5%	97.5%
(Intercept)	0.151	0.134	0.168	1.163	1.325	1.303	1.348
weekend	0.392	0.364	0.419	1.480	0.566	0.525	0.604
year2017	-0.016	-0.037	0.005	0.984	-0.009	-0.032	0.014

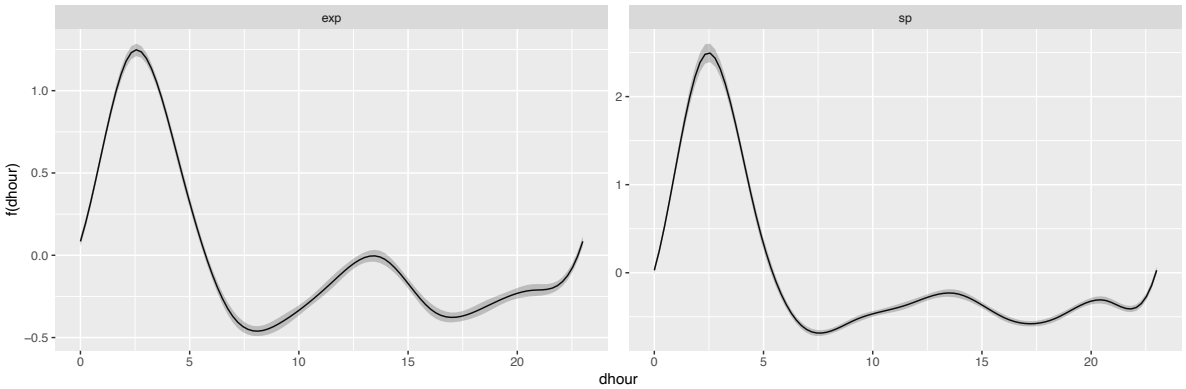

Figure 7: Posterior mean estimates on the predictor of the standard deviation together with 95% point-wise credible intervals (equal-tailed) for both response functions.

function makes assessing the absolute effects on the distribution parameter difficult. However, it becomes more apparent when considering plots of the predicted parameter values. In Figure 8, we show the predicted values (using the posterior mean of the estimated parameters) for σ over the course of two selected days of the year 2016 (these are the 1st of January and the 1st of July). We further add the effect of `weekend` and display the predicted values for both models. We observe that both models output similar values for a weekday on the 1st of July. Even the spike in the early morning appears similar. In the winter or on a weekend, the standard deviation is larger in both models (exponential model: 1.48 times larger on weekends, 1.31 times larger on the 1st of January; softplus model: 0.56 minutes larger on weekends, 0.23 minutes larger on the 1st of January). The difference between the models is most apparent at the 3 am peak, where it is now about one minute. The exponential function's almost explosive behavior becomes apparent when considering the combined effect (left panel in Figure 7 B). The difference between the peak at noon and in the morning is almost 5 minutes with the exponential function (that is a 3.5 fold increase) and just 2.75 minutes with the softplus function. Again, for the remaining time of the day, both models output relatively minor differences. Compared to the right panel in A, we find $\exp(0.27 + 0.39 + 1.25) = 6.75$ fold increase between noon and morning peak with the exponential function. Due to the additive nature of the softplus function, the effects are not multiplied and the difference is just 4.53 fold (4.49 minutes compared

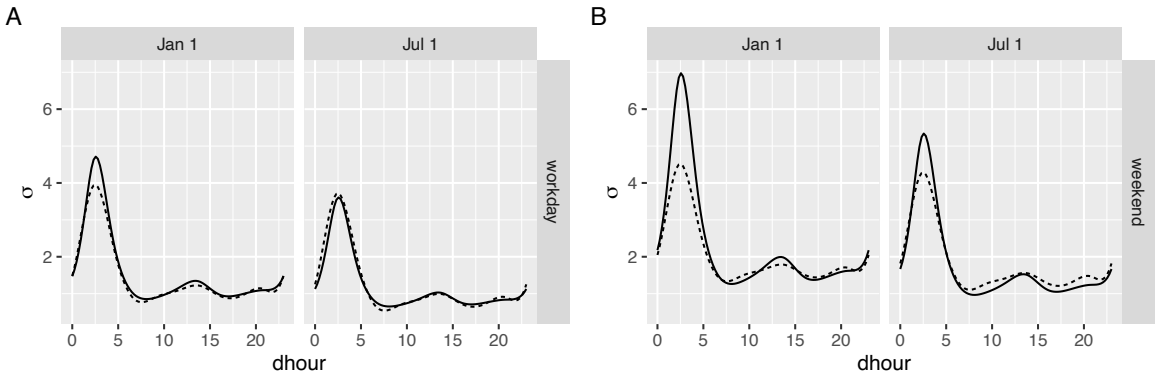

Figure 8: Predicted standard deviation over the course of a day. Exemplary for the first day of the year 2016 and July 1st on a working day (Panel A) and on a weekend (Panel B). The solid line refers to predictions for the model with exponential response function while the dashed line refers to the model with softplus response function.

to 0.99 minutes).

We can draw the following conclusions from this application: the softplus function can serve as an alternative response function in distributional regression for a bounded parameter and result in improved model fit. Additionally, the limited growth rate of the softplus function can prevent the explosive behavior of the exponential function while also providing a quasi-additive interpretation of regression effects. Even when both functions fit equally well, the softplus function remains a viable alternative, and the choice between the two functions ultimately rests with the practitioner. A final remark: In the distributional regression framework, choosing an appropriate response distribution is essential. In our case, the mean rental duration is modeled, and hence the normal distribution appears to be a natural choice. However, the quantile-quantile plots (shown in the Supplementary Material) of the estimated residuals indicate that potentially skewed distributions with heavier tails may be more suitable and warrant further investigation in future research.

4.3. Operational Losses at UniCredit

In this section, we demonstrate the usefulness of the softplus function in the context of a distributional regression model. To do so, we employ the data[4] used in Hambuckers et al. (2018a) where the authors model the size distribution of extreme operational losses in a large Italian bank (i.e., losses stemming from fraud, computer crashes or human errors) given a set of economic factors (e.g., interest rates, market volatility or unemployment rate). This conditional distribution is then used to estimate a quantile at the 99.9% level, a quantity needed to establish the regulatory capital held by the bank, with large quantile values requesting more capital, and

[4]The data can be accessed through: `http://qed.econ.queensu.ca/jae/datasets/hambuckers001/`.

to monitor operational risk exposure in various economic situations, such as a financial crisis or economic expansion periods.

Since operational loss data are heavy-tailed and the focus is on extreme value dynamics, distributional regression techniques are needed to properly reflect the effect of the covariates on extreme quantiles. Following Chavez-Demoulin et al. (2016), an approach based on extreme value theory is traditionally used: a high threshold τ is defined by the statistician, and only losses larger than this threshold are kept for the analysis. Then, we assume that the distribution of the exceedances above the threshold is well approximated by a Generalized Pareto distribution (GPD). In the context of extreme value regression, the parameters of the GPD are additionally modeled as functions of covariates, defining a Generalized Pareto (GP) regression model. Estimated parameters of this model are used to derive the quantile of interest given values of the covariates.

For mathematical and conceptual reasons, both parameters of the GPD are restricted to strictly positive values: the scale parameter $\sigma(x)$ is strictly larger than 0, whereas the shape parameter $\gamma(x)$ is restricted to positive values to guarantee the consistency of the maximum likelihood estimator and to reflect the tail-heaviness of the loss distribution. Thus, an exponential response function is commonly used for computational simplicity, although no theoretical support for a multiplicative model exists (see, e.g., Umlauf and Kneib (2018); Hambuckers et al. (2018b); Bee et al. (2019) and Groll et al. (2019)). However, this choice for $\gamma(x)$ might quickly generate explosive quantile estimates for some combinations of the covariates, making the model economically unexploitable to derive capital requirements. In addition, it can have a similar undesired effect on uncertainty quantification: the width of the confidence interval on the quantile increases exponentially with the estimated quantile itself. Consequently, it is in times of high estimated risk exposure (i.e., large values of the 99.9% quantile) that risk managers face the highest model uncertainty to take decisions.

To illustrate how the softplus function helps mitigate these issues, we reanalyze the UniCredit loss data for three categories of operational losses, namely the categories *execution, delivery and process management* (`EDPM`), *clients, products, and business practices* (`CPBP`) and *external fraud* (`EFRAUD`). The data were collected over the period January 2004 – June 2014. As in Hambuckers et al. (2018a), we work with the 25% largest losses in each category. Descriptive statistics and histograms of the data are provided in Table 4 and Figure 9. They both highlight the presence of extreme values that need to be accounted for. For each loss registered during a given month, we associate the values taken by a set of economic covariates observed the month before that were found susceptible to influence the loss distribution by Hambuckers et al. (2018a) (the complete list can be found in the Supplementary Material). Denoting by $y_i = z_i - \tau$ the exceedance of a loss z_i above the threshold τ, and by $\mathbf{x}_{\gamma,i}$ and $\mathbf{x}_{\sigma,i}$ the corresponding vectors

of covariates for both γ and σ, our model can be written in generic form as

$$\begin{aligned} y_i &\sim G(\gamma(\mathbf{x}_i), \sigma(\mathbf{x}_i)), \\ \gamma(\mathbf{x}_i) &= h^{\gamma}(\mathbf{x}_{\gamma,i}^{'}\boldsymbol{\beta}_{\gamma}), \\ \sigma(\mathbf{x}_i) &= h^{\sigma}(\mathbf{x}_{\sigma,i}^{'}\boldsymbol{\beta}_{\sigma}), \end{aligned}$$

with $G(\cdot)$ denoting the cumulative distribution function of the GPD, and $\boldsymbol{\beta}_{\gamma}$ and $\boldsymbol{\beta}_{\sigma}$ being the vectors of regression parameters for γ and σ, respectively.

We fit separate GP regression models to each sample with `bamlss` using 24000 MCMC iterations, treating the first 4000 iterations as burn-in and applying a thinning factor of 20. We compare the results obtained with various response functions h^{γ} (we keep the exponential function for h^{σ}). Estimated regression parameters can be found in the Supplementary Material. We report the DIC in Table 5, whereas Figure 10 displays the QQ-plots of the RQRs. They both indicate that the overall goodness-of-fit is satisfactory and similar across models, with a slight preference for the softplus models for `EFRAUD`, and an advantage of the exponential model for the other categories.

Table 4: Descriptive statistics of the exceedance. *iqr* denotes the inter-quantile range.

Category	n	mean	median	skewness	kurtosis	iqr
`CPBP`	4034	255879	29453	24	674	68361
`EDPM`	3302	133468	15571	20	539	43689
`EFRAUD`	1598	64027	15277	37	1412	31169

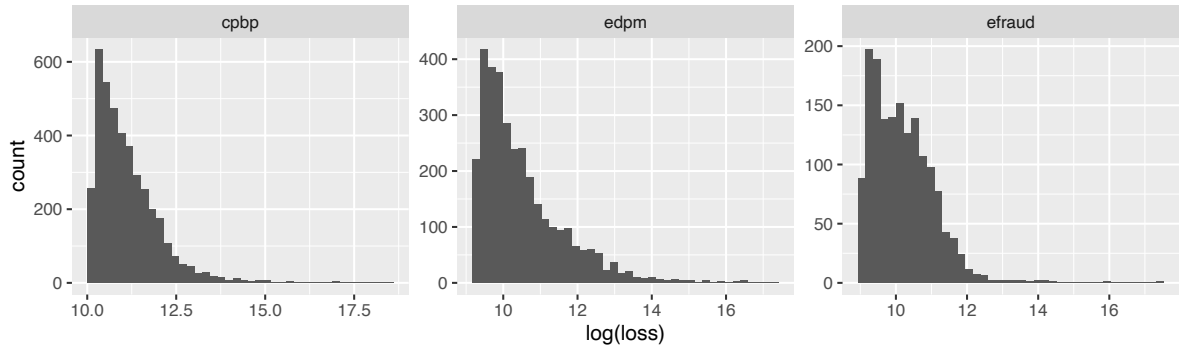

Figure 9: Histograms of the log-losses larger than the 75% quantile per category, for the three event types.

However, looking first at the predicted values of the 99.9% quantile of the conditional distributions, models based on the softplus functions generate fewer outliers than the exponential function (Figure 11): the largest predicted quantiles are between 1.5 and 3471 times smaller with the softplus models than with the exponential model. Whereas UniCredit is exposed to

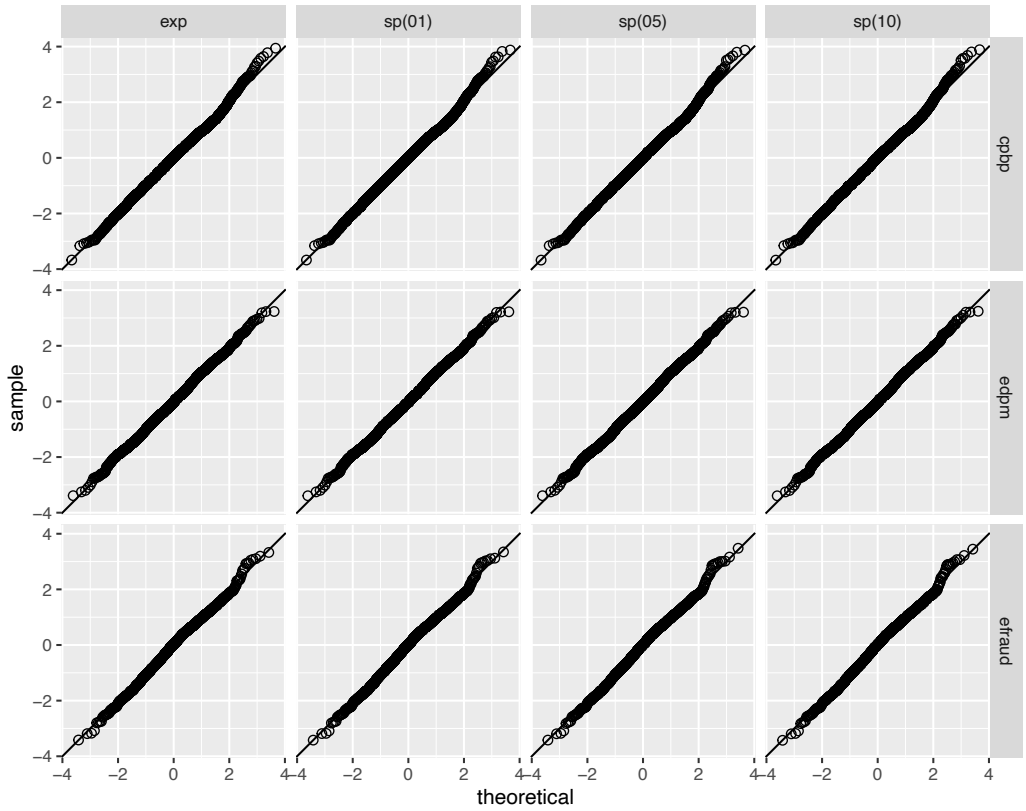

Figure 10: QQ-plots of the RQRs for the different models. The number in parentheses refers to the employed goodness of approximation parameter in the softplus function.

extremely high (and unrealistic) capital requirements if it uses the exponential model, this issue is well mitigated with the softplus model. This effect is particularly strong for `EFRAUD`. Second, looking at the size of the confidence intervals for the 99.9% quantiles, we observe a clear trend: in Figure 12, we show the ratio between the size of the confidence intervals obtained with the softplus functions and those obtained with the exponential function, with values smaller than 1 indicating an advantage for the softplus functions. For large values of the estimated quantile, we obtain much narrower confidence intervals with the softplus functions, with most ratios below 1. This result implies that, in times of financial stress characterized by high values of the quantiles, the softplus models deliver more informative estimations.

Table 5: Deviance information criterion for the different models. *Null model* refers to the exponential model with no covariates.

Category	Exp.	sofptplus 1	softplus 5	softplus 10	Null model
`CPBP`	23,593.24	23,594.99	23,598.48	23,603.37	23,626.18
`EDPM`	16,532.26	16,532.77	16,535.42	16,540.51	16,542.68
`EFRAUD`	6,547.11	6,545.59	6,545.89	6,546.31	6,567.46

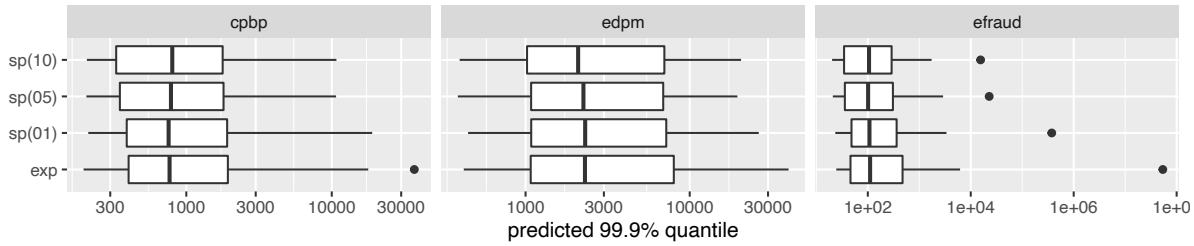

Figure 11: Box plots of the estimated 99.9% quantiles of the conditional loss size distribution (x-axis is in log-scale).

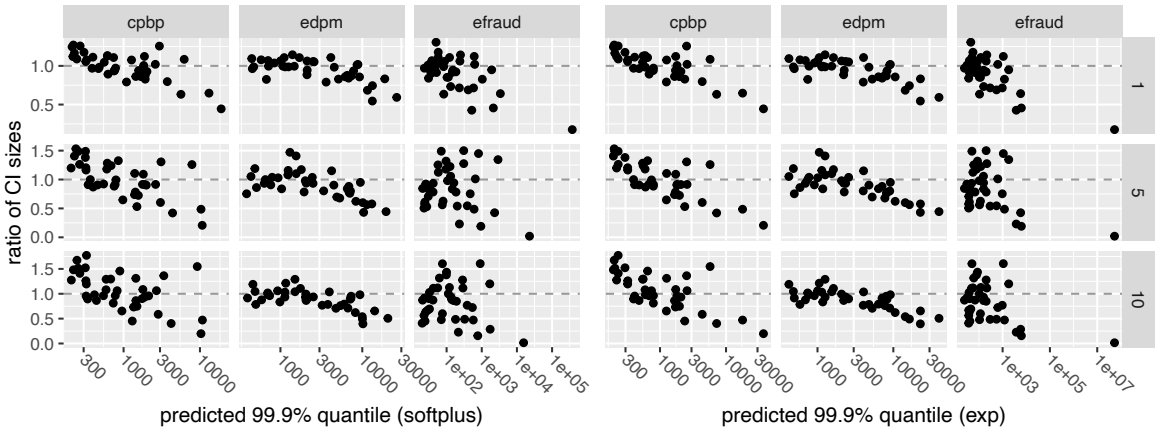

Figure 12: Ratio of the sizes of the confidence intervals (sp / exp). Left panel: ratio expressed as a function of the softplus posterior mean estimate. Right panel: ratio expressed as a function of the exponential posterior mean estimate. Each row displays the results with the softplus parameter set to the value indicated on the right ($a = 1, 5, 10$).

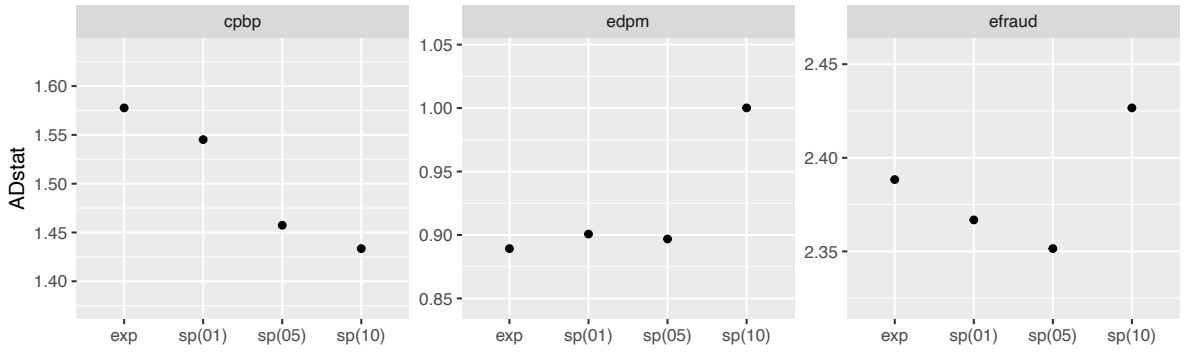

Figure 13: Anderson-Darling statistics obtained from the RQRs for the three categories.

Finally, we investigate if these results also imply a better fit of the models based on the softplus function for the observations far in the tail. To do so, we report the Anderson-Darling (AD) statistics (Stephens, 1974) computed on the RQRs (Figure 13). Compared to AIC or DIC, the AD statistic gives more weight to extreme residuals and is therefore routinely used to assess the goodness-of-fit of extreme value regression models (Choulakian and Stephens, 2001; Bader et al., 2018). On this latter criterion, we observe a better fit of the softplus functions for CPBP and EFRAUD. The fit is rather similar for EDPM, although slightly better for the exponential model.

Overall, this application demonstrates the usefulness of the softplus function to prevent outliers among estimated quantities of interest (in the present case, a quantile far in the tail) when there are no justifications for a multiplicative model. In addition, it shows that the softplus models provide similar global goodness-of-fit levels but dramatically reduce the uncertainty around large estimated quantities of interest, a desirable feature for end-users.

Finally, notice that, once the softplus response function has been chosen to avoid outliers and to decrease estimation uncertainty, one may be interested in selecting an appropriate softplus parameter a, as discussed in Subsection 2.3. In the present case, this selection can be conducted using either an information criterion like DIC, or a goodness-of-fit statistic focusing on the tail observations such as the AD statistic, since this is a primary concern in our application.

5. Summary and Conclusion

This paper introduces the softplus response function and showcases its applicability in a broad range of statistical models. The novel response function ensures the positivity of the associated distribution parameter while allowing for a quasi-additive interpretation of regression effects for the majority of the relevant predictor space. We highlight the interesting theoretical properties of the softplus response function, justify the quasi-additive interpretation and give a guideline to assess the validity of this interpretation.

Particular emphasis is placed on demonstrating the straightforward quasi-linear and quasi-additive interpretation of covariate effects with several applications. Furthermore, we highlight that the limited growth rate of the softplus response function can prevent outliers in predictions and, thus, can reduce prediction uncertainty. Thereby, we show that the new response function is applicable to a great variety of model classes and data situations.

Our simulation studies demonstrate that the softplus function behaves well as a response function with no noticeable shortcomings. Estimates are consistent and our Bayesian approach yields reliable credible intervals. Furthermore, we show that information criteria can be used to distinguish between data generated by the exponential and the softplus response function.

We do not claim that the softplus function is generally a better response function than the exponential function, nor that the softplus response function is always the best choice for a quasi-additive interpretation. Indeed, other response functions to approximate the linear spline can be easily constructed. For example, based on the work of Bacon and Watts (1971) one can construct $\lim_{0<\epsilon\to 0} 0.5x + 0.5\sqrt{x^2+\epsilon} = x_+$. However, since the implementation of the softplus response function is straightforward, it provides an easy quasi-additive alternative response function for empirical verification. Thus, we are optimistic that the softplus response function will be available in more software for regression modeling, as it is already included in the R-packages `brms` (Bürkner, 2017) and `bamlss` (Umlauf et al., 2018), and will allow researchers to benefit from it. The work of Weiß et al. (2021), evaluating the softplus response function in the context of INGARCH models, is a first indication for interest of the statistical community in the novel response function.

Acknowledgements

Financial support from the German Research Foundation (DFG) within the research project KN 922/9-1 is gratefully acknowledged. Julien Hambuckers acknowledges the financial support of the National Bank of Belgium, project REFEX. We thank an anonymous reviewer and an associate editor for their valuable feedback on which we improved the initial manuscript.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions*. Number 55 in National Bereau of Standards: Applied Mathematics. U.S. Government Printing Office, Washington, D.C., tenth edition, 1972.

Agresti, A. *Categorical Data Analysis*. Number 792 in Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, third edition, 2013. ISBN 978-0-470-46363-5.

Bacon, D. W. and Watts, D. G. Estimating the transition between two intersecting straight lines. *Biometrika*, 58(3):525–534, 1971.

Bader, B., Yan, J., and Zhang, X. Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Annals of Applied Statistics*, 12(1):310–329, 2018.

Bee, M., Dupuis, D. J., and Trapin, L. Realized peaks over threshold: a time-varying extreme value approach with high-frequency-based measures. *Journal of Financial Econometrics*, 17(2):254–283, 2019.

Brezger, A. and Lang, S. Generalized structured additive regression based on Bayesian P-Splines. *Computational Statistics & Data Analysis*, 50(4):967–991, 2006. doi:10.1016/j.csda.2004.10.011.

Brockmann, H. J. Satellite male groups in horseshoe crabs, limulus polyphemus. *Ethology*, 102(1):1–21, 1996. doi:10.1111/j.1439-0310.1996.tb01099.x.

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

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2016.

Choulakian, V. and Stephens, M. A. Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4):478–484, 2001.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating second-order functional knowledge for better option pricing. *Advances in Neural Information Processing Systems 13*, pages 451–457, 2001.

Dunn, P. K. and Smyth, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996. doi:10.1080/10618600.1996.10474708.

Eilers, P. H. C. and Marx, B. D. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102, 1996.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. *Regression*. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-34332-2 978-3-642-34333-9. doi:10.1007/978-3-642-34333-9.

Gamerman, D. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7:57–68, 1997.

Groll, A., Hambuckers, J., Kneib, T., and Umlauf, N. Lasso-type penalization in the framework of generalized additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 140:59 – 73, 2019.

Hambuckers, J., Groll, A., and Kneib, T. Understanding the economic determinants of the severity of operational losses: a regularized generalized pareto regression approach. *Journal of Applied Econometrics*, 33(6):898–935, 2018a.

Hambuckers, J., Kneib, T., Langrock, R., and Silbersdorff, A. A Markov-switching generalized additive model for compound Poisson processes, with applications to operational loss models. *Quantitative Finance*, 18(10):1679–1698, 2018b.

Hastie, T. and Tibshirani, R. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986. doi:10.1214/ss/1177013604.

Hofner, B., Kneib, T., and Hothorn, T. A unified framework of constrained regression. *Statistics and Computing*, 26(1-2):1–14, 2016. doi:10.1007/s11222-014-9520-y.

Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993. doi:10.1016/0304-4076(93)90114-K.

Kleiber, C. and Zeileis, A. Visualizing count data regressions using rootograms. *The American Statistician*, 70(3):296–303, 2016. doi:10.1080/00031305.2016.1173590.

Klein, N. and Kneib, T. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4):841–860, 2016. doi:10.1007/s11222-015-9573-6.

Klein, N., Kneib, T., and Lang, S. Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509):405–419, 2015. doi:10.1080/01621459.2014.912955.

Lang, S. and Brezger, A. Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004. doi:10.1198/1061860043010.

Liu, Q. and Furber, S. Noisy softplus: a biology inspired activation function. In A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, and D. Liu, editors, *Neural Information Processing (ICONIP)*, volume 9950 of *Lecture Notes in Computer Science*, pages 405–412. Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-46681-1_49.

McCullagh, P. and Nelder, J. *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, second edition, 1989. ISBN 978-0-203-75373-6.

Nielsen, F. and Sun, K. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise Log-Sum-Exp inequalities. *Entropy*, 18(12):442–467, 2016. doi: 10.3390/e18120442.

Ntzoufras, I., Dellaportas, P., and Forster, J. J. Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111(1-2):165–180, 2003. doi:10.1016/S0378-3758(02)00298-7.

Pregibon, D. Goodness of link tests for generalized linear models. *Applied Statistics*, 29(1):15, 1980. doi:10.2307/2346405.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

Rigby, R. A. and Stasinopoulos, D. M. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005. doi:10.1111/j.1467-9876.2005.00510.x.

Spiegel, E., Kneib, T., and Otto-Sobotka, F. Generalized additive models with flexible response functions. *Statistics and Computing*, 29(1):123–138, 2019. doi:10.1007/s11222-017-9799-6.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi:10.1111/1467-9868.00353.

Stephens, M. A. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974. doi:10.1080/01621459.1974.10480196.

Umlauf, N., Klein, N., and Zeileis, A. BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627, 2018. doi:10.1080/10618600.2017.1407325.

Umlauf, N. and Kneib, T. A primer on Bayesian distributional regression. *Statistical Modelling*, 18(3-4):219–247, 2018.

Weiß, C. H., Zhu, F., and Hoshiyar, A. Softplus INGARCH model. *Statistica Sinica*, 2021. doi:10.5705/ss.202020.0353.

Yu, Y. and Ruppert, D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002. doi:10.1198/016214502388618861.

Yu, Y., Wu, C., and Zhang, Y. Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing*, 27(2):571–582, 2017. doi:10.1007/s11222-016-9639-0.

Zheng, H., Yang, Z., Liu, W., Liang, J., and Li, Y. Improving deep neural networks using softplus units. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE, Killarney, Ireland, 2015. doi:10.1109/IJCNN.2015.7280459.

Zuras, D., Cowlishaw, M., Aiken, A., Applegate, M., Bailey, D., Bass, S., Bhandarkar, D., Bhat, M., Bindel, D., Boldo, S., et al. IEEE standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, 2008.

A. On Implementation and Properties of the Softplus Function

A.1. Numerical Stability

A naive implementation of the softplus function derived from Equation (1) can easily lead to numerical issues. The value of the exponential function for relatively small inputs is infinity on common computer hardware. To give some intuition, according to the IEEE 754 standard (Zuras et al., 2008) the largest 32-bit and 64-bit floating point numbers are roughly $3.4028 \cdot 10^{38}$ and $1.7977 \cdot 10^{308}$, respectively. Consequently, calculating exp(89) and exp(710), respectively, yields infinity.

This is of special concern for the implementation of softplus function since the argument to the softplus function is multiplied with the softplus parameter a before the exponential function is applied. Consider a Poisson regression model with softplus response function and $a = 10$. The predictor $\eta = 9$, targeting an expected value of 9, would already yield infinity on a 32-bit system using the naive implementation although the correct result is between 9 and $9 + 10^{-40}$. Albeit 64-bit CPUs are common nowadays, one should still consider 32-bit floating point arithmetic since it is often used in high-performance computing or when the computation is carried out on graphical processing units (GPUs) or tensor processing units (TPUs).

Despite the difficulties described, the softplus function becomes numerically stable by using the equality

$$\text{softplus}_a(x) = \max\{0, x\} + \frac{\log(1 + \exp(-|ax|))}{a} \tag{3}$$

in conjunction with the `log1p` procedure. `log1p` evaluates $\log(1 + x)$ very precisely even for $|x| \ll 1$ (Abramowitz and Stegun, 1972, p. 68) and is available in most programming languages. In this formulation, the exponential function must be evaluated only for arguments less than 0 which can be done accurately. Besides numerical stability, Equation (3) also implies that the softplus function has its largest approximation error with respect to the linear spline at $x = 0$ with $\log(2)/a$.

The correctness of the numerical stable formulation is easily verified by expressing the softplus function in terms of the log-sum-exp (LSE) function and exploiting its translation property. The LSE function takes l real valued arguments $x_1, \ldots, x_l$. Its value is given by

$$\text{LSE}(x_1, \ldots, x_l) = \log\left(\sum_{i=1}^{l} \exp(x_i)\right).$$

The translational property (Nielsen and Sun, 2016) states that for $c \in \mathbb{R}$

$$\mathrm{LSE}(x_1, \ldots, x_l) = c + \log\left(\sum_{i=1}^{l} \exp(x_i - c)\right)$$

holds. Consequently, we have

$$\begin{aligned}
\mathrm{softplus}_a(x) &= \frac{\log(1 + \exp(ax))}{a} = \frac{\mathrm{LSE}(0, ax)}{a} \\
&= \max\{0, x\} + \frac{\log(\exp(0 - \max\{0, ax\}) + \exp(ax - \max\{0, ax\}))}{a} \\
&= \max\{0, x\} + \frac{\log(1 + \exp(-|ax|))}{a}
\end{aligned}$$

where the second line arises by setting $c = \max\{0, ax\}$ and the last line follows from the observation that $-|ax| = ax$ for $x < 0$.

A.2. On the inverse function

The inverse of the softplus is easily derived as

$$\mathrm{softplus}_a^{-1}(y) = \frac{\log(\exp(ay) - 1)}{a}.$$

However, potential numerical issues have to be taken into consideration when implementing. We suggest to implement

$$\mathrm{softplus}_a^{-1}(y) = \begin{cases} \frac{\log(\exp(ay) - 1)}{a} & , \text{ if } y < \frac{\log(2)}{a} \\ y + \frac{\log(1 - \exp(-ay))}{a} & , \text{ otherwise} \end{cases}$$

and use the procedure `expm1(x)` to calculate $\exp(x) - 1$. The first derivative is given by

$$\frac{\mathrm{d}}{\mathrm{d}y}\,\mathrm{softplus}_a^{-1}(y) = \frac{1}{1 - \exp(-ay)}.$$

A.3. Further properties

The softplus function shares a number of its properties with the exponential function. Both functions are smooth and bijective mappings from the real numbers to the positive half-axis. The first derivative of the softplus function is always positive and is given by

$$0 < \frac{\mathrm{d}}{\mathrm{d}x}\,\mathrm{softplus}_a(x) = \frac{1}{1 + \exp(-ax)} < 1.$$

The second derivative is likewise strictly positive

$$0 < \frac{\mathrm{d}^2}{\mathrm{d}x^2} \operatorname{softplus}_a(x) = \frac{a \exp(ax)}{(1 + \exp(ax))^2}.$$

Therefore, the softplus function is strictly monotonically increasing and strictly convex.