Statistica Sinica Preprint No: SS-2022-0412	
Title	Skewed Pivot-Blend Modeling with Applications to Semicontinuous Outcomes
Manuscript ID	SS-2022-0412
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0412
Complete List of Authors	Yiyuan She, Xiaoqiang Wu, Lizhu Tao and Debajyoti Sinha
Corresponding Authors	Yiyuan She
E-mails	yshe@stat.fsu.edu
Notice: Accepted version subject to English editing.	

Skewed Pivot-Blend Modeling with Applications to Semicontinuous Outcomes*

Yiyuan She*, Xiaoqiang Wu*, Lizhu Tao†, and Debajyoti Sinha*

* Department of Statistics, Florida State University

† College of Mathematics, Sichuan University

Abstract

Skewness is a common occurrence in statistical applications. In recent years, various distribution families have been proposed to model skewed data by introducing unequal scales based on the median or mode. However, we argue that the point at which unbalanced scales occur may be at any quantile and cannot be reparametrized as an ordinary shift parameter in the presence of skewness. In this paper, we introduce a novel skewed pivot-blend technique to create a skewed density family based on any continuous density, even those that are asymmetric and nonunimodal. Our framework enables the simultaneous estimation of scales, the pivotal point, and other location parameters, along with various extensions. We also introduce a skewed two-part model tailored for semicontinuous outcomes, which identifies relevant variables across the entire population and mitigates the additional skewness induced by commonly used transformations. Our theoretical analysis reveals the influence of skewness without assuming asymptotic conditions. Experiments on synthetic and real-life data demonstrate the excellent performance of the proposed method.

*The authors thank the editor, associate editor, and anonymous referees for their valuable suggestions to greatly improve this paper. Supplementary materials provide detailed proofs, additional theory and results, plus further experiments. This work was partially supported by the National Science Foundation.

Keywords: semicontinuous outcomes; skewed data; two-piece densities; two-part models; variable selection; composite models.

1 Introduction

Statisticians frequently encounter skewed data in biomedical, econometric, environmental, and social research. Commonly used models, such as linear regression, least absolute deviations, and robust regression, presume symmetric errors and are prone to significant distortions when confronted with skewness. To mitigate the issue, many researchers prefer transforming the data beforehand, with logarithmic-type transformations being among the most popular choices. Alternatively, some researchers use modal regression (Lee, 1989) or median-based methods, which are less sensitive to the assumption of symmetric errors. However, these approaches do not explicitly account for and describe skewness.

To comprehensively address this issue, adopting a "joint" modeling approach becomes essential and beneficial. This paper simultaneously estimates location, scale, and skewness parameters, thus avoiding the risk of either concealing true skewness (*masking*) or erroneously detecting spurious skewness (*swamping*). This risk is present when using a stepwise procedure, such as fitting a modal regression and then assessing skewness based on residuals (Boos, 1987). Our primary aim is not only to accommodate skewness, as many papers do, but to *explicitly* capture and characterize its effects.

Various distributions have been proposed in the literature for modeling skewed data. Azzalini (1985) proposed a skewed density family including the skewed normal density as an example. Fernández and Steel (1998) proposed a two-piece skewed distribution family that sets the mode at zero, including the skewed Student and Laplace distributions for Bayesian quantile regression (Arellano-Valle et al., 2005; Yu and Moyeed, 2001). Rubio and Steel (2015) extended the family by use of two scale parameters and additional shape parameters. Kottas and Gelfand (2001) described an alternative two-piece skewed distribution family

that keeps the median at zero, but the resulting density is discontinuous. For a historical account of two-piece distributions, interested readers may consult Rubio and Steel (2020).

The existing constructions rely on a symmetric and unimodal raw density, introducing asymmetric scales based on either the mode or the median of the raw density. However, in numerous real-life applications, these assumptions may not hold. Particularly, the point at which skewness is enforced, termed the "**pivotal point**" in this paper, could be situated at any position or quantile. Intriguingly, this pivotal point distinguishes itself from the commonly used shift parameter, as opposed to the prevailing assumption in the existing literature. To overcome these limitations, there is a demand for a novel skewed distribution family that offers flexibility, continuity, and adaptability to any pivotal point of interest.

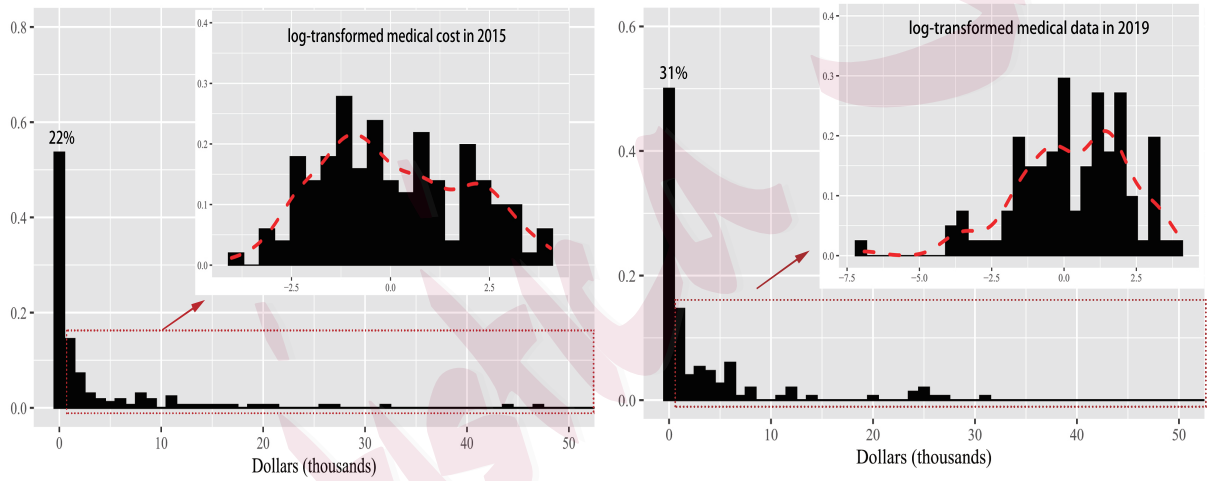


Figure 1: Some histograms of medical expenditures. Left: stratum ID 1098 (first PSU) of MEPS 2015, right: stratum ID 2109 (third PSU) of MEPS 2019. In each example, the main plot shows an excessive portion of zeros; the top right panel, excluding the zeros, plots the log-transformed positive values of response (with estimated density in red), which still exhibit asymmetry despite the transformation.

This study draws inspiration from the Medical Expenditure Panel Survey (MEPS) data, which is obtained from national surveys investigating the impact of various demographic variables on the medical expenses of patients in the United States. Notably, this dataset features a skewed response that includes a significant number of zeros, a phenomenon known as "**semicontinuous outcomes**" in the realms of economics and longitudinal studies (Olsen and Schafer, 2001). To provide a visual representation, we employed two datasets (Agency

for Healthcare Research and Quality, 2015, 2019), as depicted in Figure 1, for illustration.

According to Figure 1, more than 20% patients have zero medical expenditure, while the remaining exhibit highly skewed positive medical costs. Given that these zeros represent precisely zero medical expenses, rather than truncation, a *two-part* (or *hurdle*) model (Mullahy, 1998) is a more appropriate choice than the *Tobit* model (Tobin, 1958). In this approach, the binary part of the model captures zero-nonzero patterns, while the continuous part of the model addresses strictly positive outcomes. However, it is important to highlight that applying a standard log-normal two-part model may not yield sufficient power, owing to the asymmetry depicted in the upper-right panels of Figure 1. We have frequently observed that conventional transformations, such as logarithmic or power functions, not only fail to entirely eliminate skewness but also introduce nontrivial points around which asymmetric scales arise. Consequently, there may be a necessity to "reinforce" the transformed model to counteract the skewness effectively.

Another closely related challenge within the context of MEPS data analysis involves developing an interpretable two-part model. This entails the identification of a subset of medical cost-relevant predictors that apply to the *entire* population, serving as valuable guidance for policymakers. To the best of our knowledge, very few existing two-part models have considered the issue of joint variable selection, wherein each predictor can contribute to the response in a composite manner through the binary and continuous parts.

This paper attempts to address some aforementioned challenges for possibly skewed, semicontinuous outcomes. Our contributions are as follows.

1. We introduce a novel skewed pivotal-point adaptive family, designed to infuse skewness around an unknown pivotal point. The key "**skewed pivot-blend**" technique is versatile and can be applied to any raw density, regardless of its symmetry or unimodality. The resulting density remains continuous and accommodates many previous proposals.

2. We introduce the **SPEUS** framework (Skewed Pivot-Blend Estimation with Unsymmetric Scales) for simultaneous estimation of scales, pivotal point, and other location

parameters. This framework offers useful variants, especially for modeling semicontinuous outcomes with joint variable selection. The resulting two-part method is capable of identifying relevant variables across the entire population and concurrently addressing the excessive skewness introduced by imperfect transformations

3. We conduct nonasymptotic analysis for sparse skewed two-part models, utilizing a notion of effective noise and Orlicz norms to derive sharp statistical error bounds in the presence of skewness and heavy tails. Our work quantifies how skewness and tail decay impact regularization parameters, prediction and estimation errors.

Notations and symbols. Given two vectors $\alpha, \beta \in \mathbb{R}^n$, their inner product is $\langle \alpha, \beta \rangle = \alpha^T \beta$ and their elementwise product is denoted by the vector $\alpha \circ \beta$. Given a scalar function l and a vector a, $l(a) = [l(a_i)]_{i=1}^n$, i.e., l is applied componentwise. Throughout the paper, we use $1_A(x)$ to denote the indicator function of A, taking 1 if $x \in A$ and 0 otherwise. In particular, given any vector $a \in \mathbb{R}^n$, define two indicator vectors $1_+(a) = [1_{a_i>0}]_{i=1}^n, 1_-(a) = [1_{a_i<0}]_{i=1}^n$. Define $\mathbb{R}_+ = [0, \infty)$. Given a continuous density f (with respect to the Lebesgue measure μ), we use $f(\cdot|A)$ to denote the conditional density given A, or $f(\cdot|A) = \frac{f(\cdot)}{\int_A f \, d\mu}$. Given any matrix $A = [a_1, \ldots, a_p]^T \in \mathbb{R}^{p \times m}$, its spectral norm and Frobenius norm are denoted by $\|A\|_2$ and $\|A\|_F$, respectively. The (2,1)-norm of A is defined as $\|A\|_{2,1} = \sum_{j=1}^p \|a_j\|_2$. We use A_k to denote the kth column of A. Given $a, b \in \mathbb{R}$, we use the shorthand notation $a \vee b$ ($a \wedge b$) to denote the maximum (minimum) of a and b.

2 Skewed Pivotal-Blend Estimation

2.1 Skewed Pivot-Blend for Density Pasting

How to define a skewed distribution family from a unimodal, continuous, and symmetric density ϕ has attracted a lot of attention in the literature. Azzalini (1985) multiplied ϕ by a perturbation function to define a so-called "skewed symmetric distribution" family, one

well-known example being the skewed normal distribution. We refer the reader to Nadarajah and Kotz (2003), Wang et al. (2004), and Azzalini (2005) for variants and further extensions. On the other hand, the associated skewed distribution function often lacks an explicit form, and determining its mode can be a challenging task (Ma and Genton, 2004).

"Two-piece" skewed distributions are popularly used in recent years. Fernández and Steel (1998) introduced a two-piece transformation that rescales ϕ's negative and positive parts differently using an asymmetry parameter, allowing it to maintain the mode at zero. A reparametrization of the approach, following Arellano-Valle et al. (2005), includes the skewed Student and epsilon-skew-normal distributions (Fernández and Steel, 1998; Mudholkar and Hutson, 2000). Later, Rubio and Steel (2015) extended this idea to include two scale parameters (and additional shape parameters). The motivation for our work largely stems from their two-piece form, even though it assumes that the median of ϕ is zero. Another two-piece distribution family due to Kottas and Gelfand (2001) can guarantee a median at zero, but the resulting density is discontinuous, which may cause difficulties and instability in parameter estimation. Interested readers may refer to Jones (2014) for a systematic framework of how to construct skewed distributions from a given symmetric density.

Despite the research in this area, two issues have caught our particular attention and deserve further investigation. Firstly, the majority of existing works stipulate that ϕ should be unimodal and symmetric. However, situations can arise where skewness manifests when dealing with non-unimodal data. There might also be a need for additional reinforcement to counteract skewness, even when employing an asymmetric density. Another more critical concern is that in previous works, the transition point at which unequal scales are imposed, referred to as the **pivotal point** in this paper, is typically set at the mode or the median. Nevertheless, skewness can persist when the density deviates from the assumptions above or below *any* quantile, a common occurrence when using an imperfect transformation.

In the following, we introduce a process known as "*skewed pivot-blend*" (or sometimes pivot-blend for brevity) to characterize skewness as a combination of both first-order and

higher-order statistical effects. We define a versatile two-piece distribution framework with skewness, designed to (i) accommodate any asymmetric or non-unimodal ϕ, (ii) model skewness associated with any pivotal point of interest, and (iii) maintain continuity. See Figure 2 for an illustration.

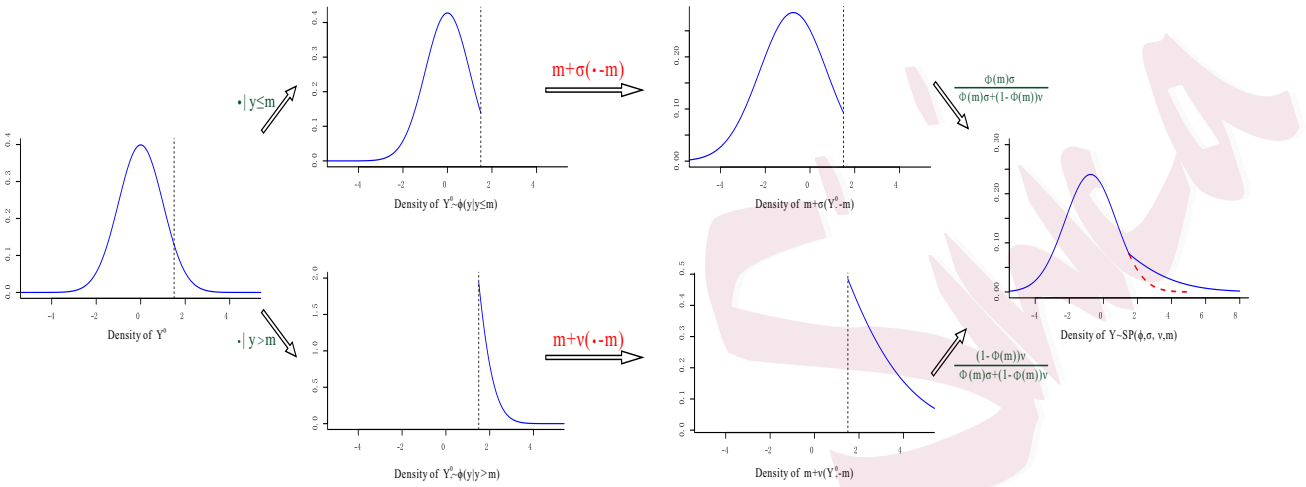


Figure 2: Diagram showing the process of "skewed pivot-blend" for constructing a skewed density: conditioning, affine transformations, and mixing. The two affine transformations ensure that the cut points remain fixed (alignment in the *x-direction*), and the mixing process guarantees the continuity of the resulting density (alignment in the *y-direction*).

We provide a step-by-step guide for constructing a new density function from an arbitrary continuous density (denoted as ϕ), during which skewness is imposed around a pivotal point m $(0 < \Phi(m) < 1)$ and is defined by the left and right scales, σ and $\nu > 0$.

a) *Pivotal-point conditioning*: The density ϕ is conditioned into two separate densities, one for $y \leq m$, and the other for $y > m$, resulting in $\frac{\phi(y)1_{y\leq m}}{\Phi(m)}$ and $\frac{\phi(y)1_{y>m}}{1-\Phi(m)}$.

b) *Affine transformation*: Apply two separate affine transformations to the random variables associated with the aforementioned densities, concerning the pivotal point m: $m+\sigma(\cdot-m)$ and $m+\nu(\cdot-m)$. The resulting densities are $\frac{\phi\left(\frac{y-m}{\sigma}+m\right)1_{y\leq m}}{\Phi(m)\sigma}$ and $\frac{\phi\left(\frac{y-m}{\nu}+m\right)1_{y>m}}{(1-\Phi(m))\nu}$. It is crucial to emphasize that these transformations are not simple scalings, but are designed to ensure that the cut points of the density functions remain aligned.

c) *Continuous mixing*: Probability masses p and $1-p$ are assigned to the two densities

obtained from the last step, resulting in a new density function:

$$f(y) = p \times \frac{1}{\Phi(m)\sigma}\phi\big(\frac{y-m}{\sigma}+m\big)1_{y\le m} + (1-p)\times\frac{1}{\{1-\Phi(m)\}\nu}\phi\big(\frac{y-m}{\nu}+m\big)1_{y>m}, \quad (1)$$

where Φ denotes the distribution function of ϕ throughout the paper unless otherwise specified. Given that $\phi(m) > 0$ typically holds, ensuring the continuity of f at m requires that $p/(\Phi(m)\sigma) = (1-p)/\{(1-\Phi(m)\nu\}$, which leads to a **unique** choice of p:

$$p = \frac{\Phi(m)\sigma}{\Phi(m)\sigma + \{1-\Phi(m)\}\nu}, \text{ or } \quad \frac{\mathbb{P}(Y\le m)}{\mathbb{P}(Y>m)} = \frac{\Phi(m)\sigma}{\{1-\Phi(m)\}\nu}. \quad (2)$$

We sometimes refer to the process as the "forward" pivot-blend transform (to contrast with the "backward" pivot-blend transform to be introduced in Remark 2). When $\sigma = \nu$ or m is not in the support of ϕ, pivot-blend operates as a location-scale transformation. Otherwise, it serves as a versatile tool for modeling skewed data, encompassing various existing skewed density functions. Notably, the incorporation of a single pivotal point parameter m substantially improves skewed data modeling in practical applications.

Definition 2.1 (Skewed pivot-blend (**SP**) family)**.** *Given a continuous density ϕ and a pivotal point m, we say that Y is a skewed random variable with m-associated left- and right-scale parameters σ and ν, i.e., $Y \sim SP^{(\phi)}(\sigma,\nu,m)$, if its density is given by*

$$f(y;m,\sigma,\nu) = \frac{\phi\big(\frac{y-m}{\sigma}+m\big)1_{y\le m} + \phi\big(\frac{y-m}{\nu}+m\big)1_{y>m}}{\Phi(m)\sigma + \{1-\Phi(m)\}\nu}. \quad (3)$$

We occasionally write $Y \sim \mathrm{SP}^{(\phi)}$ and omit the parameters when there is no ambiguity. Throughout the paper, we use the term *skewness* to refer to asymmetric scales ($\sigma \neq \nu$), regardless of the shape of ϕ.

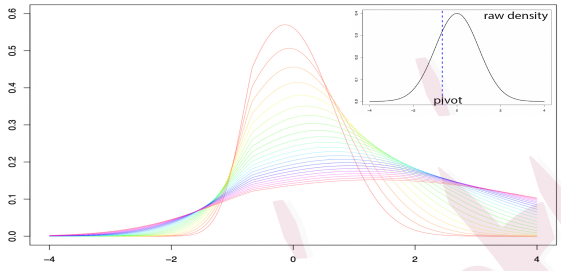
The pivotal location m can be translated to a *pivotal quantile* q. Let $q = \Phi(m)$, then an

equivalent form of (3) is

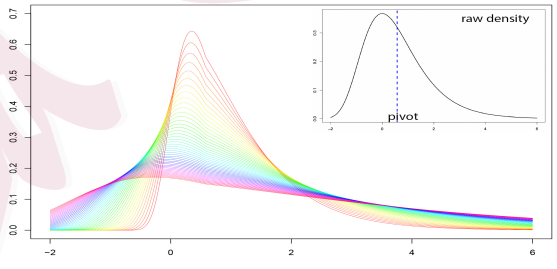
$$\frac{\phi\big(\frac{y-\Phi^{-1}(q)}{\sigma}+\Phi^{-1}(q)\big)1_{\Phi(y)\le q}+\phi\big(\frac{y-\Phi^{-1}(q)}{\nu}+\Phi^{-1}(q)\big)1_{\Phi(y)>q}}{q\sigma+(1-q)\nu}.$$

For the distribution function $F(y)=\frac{1}{\sigma\Phi(m)+\nu(1-\Phi(m))}\{\sigma\Phi(\frac{y-m}{\sigma}+m)1_{y\le m}+[\nu\Phi(\frac{y-m}{\nu}+m)+(\sigma-\nu)\Phi(m)]1_{y>m}\}$, the new quantile at m is related to the original quantile q by $F(m)=\frac{q\sigma}{q\sigma+(1-q)\nu}\lessgtr q$ when $\sigma\lessgtr\nu$.

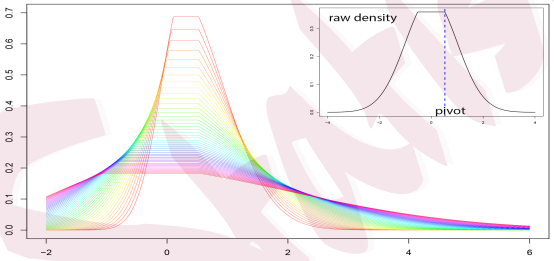
When working with the family described in Definition 2.1, it is a common practice to add a shift or intercept $\alpha\in\mathbb{R}$ and assume $Y-\alpha\sim \mathrm{SP}^{(\phi)}(\sigma,\nu,m)$; however, it is crucial to note that α and m are generally **not** redundant. This distinction arises because the operations of translation and asymmetric rescaling utilized in the skewed pivot-blend process do not commute (cf. Remark 1).



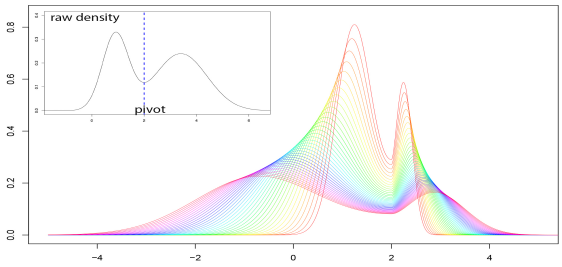
(a) $\phi\sim N(0,1)$, $m=\Phi^{-1}(0.25)$



(b) ϕ: Gumbel, m: mean



(c) ϕ: flat modal, m: (rightmost) mode



(d) $\phi\sim 0.4N(0.9,0.25)+0.6N(3.4,1)$, $m=2$

Figure 3: Illustration of some SP families with *varied* σ and ν (while maintaining a constant ratio). These plots demonstrate the *versatility* of skewed pivot-blend in generating a wide range of distributions for practical modeling, including asymmetry and diverse tails (which contrasts with traditional methods assuming symmetry in ϕ and median/mode in m).

Figure 3 provides visual examples of introducing skewness through pivot-blend around a nontrivial pivotal point, with variations in σ and ν to demonstrate different tail decay

behaviors.

Example 2.1 (Skewed double-gamma family)**.** *Skewed densities may involve heavy tails and multimodality. When applied to the deneralized double-Gamma density (GDG),* $p/\{2\gamma^d\Gamma(d/p)\}$ $|y|^{d-1}\exp\{-|y|^p/\gamma^p\}$*, an extension of Stacy (1962), skewed pivot-blend reveals*

$$\mathrm{SP}^{(\mathrm{GDG})}\colon\ \frac{p}{2[\Phi(m)\sigma+\{1-\Phi(m)\}\nu]\gamma^d\Gamma(d/p)}\Big[\Big|\frac{y-m}{\sigma}+m\Big|^{d-1}\exp\Big\{-\frac{|(y-m)/\sigma+m|^p}{\gamma^p}\Big\}1_{y\le m} + \Big|\frac{y-m}{\nu}+m\Big|^{d-1}\exp\Big\{-\frac{|(y-m)/\nu+m|^p}{\gamma^p}\Big\}1_{y>m}\Big],$$

where $\gamma, d, p > 0$ *are parameters. The skewed GDG family comprises bimodal types such as the skewed double gamma* ($p = 1, m = 0$) *and skewed double Weibull* ($d = p, m = 0$)*. The unimodal skewed exponential power distribution family (Zhu and Zinde-Walsh, 2009) is another instance* ($\gamma = 1$, $d = 1$)*, including the skewed Laplace distribution and skewed normal distributions (Arellano-Valle et al., 2005; Mudholkar and Hutson, 2000).*

2.2 SPEUS for Skewed Regression

Skewed pivot-blend is a valuable tool for statistical modeling of a skewed outcome $y \in \mathbb{R}^n$ associated with p predictors collected in the matrix $X \in \mathbb{R}^{n\times p}$. Given a density function ϕ, if we assume

$$y - X\beta^* \sim \mathrm{SP}^{(\phi)}(\sigma^*, \nu^*, m^*)$$

and define $\rho = -\log\phi$, the estimation of $\beta^*, \sigma^*, \nu^*, m^*$ can be formulated as a joint optimization problem

$$\min_{\beta,\sigma,\nu,m}\ n\log\big[\sigma\Phi(m)+\nu\{1-\Phi(m)\}\big] + \sum_{i=1}^{n}\Big\{\rho\big(\frac{r_i-m}{\sigma}+m\big)1_{r_i-m\le 0} + \rho\big(\frac{r_i-m}{\nu}+m\big)1_{r_i-m>0}\Big\},\quad \text{s.t.}\quad r = y - X\beta, \sigma > 0, \nu > 0, \tag{4}$$

where the first term arises from the so-called "normalizing constant" which is a joint function of m, σ, ν. Henceforth, we refer to the framework of (4) as the Skewed Pivot-blend

Estimation with Unsymmetric Scales (**SPEUS**). We always assume that ρ is constructed from a given density function unless otherwise specified. (4) is thus an instance of maximum likelihood estimation (MLE), and standard MLE asymptotic theory guarantees consistency and other properties. In practical implementation, the values of r_i are rarely equal to m and so conventional optimization algorithms like gradient descent, Newton's method, and quasi-Newton methods can be readily applied. Given the nonconvex nature, initialization impacts estimates, especially for small sample sizes. We usually start location parameters at 0, but using a preliminary estimate like Yang et al. (2019) tends to yield better performance. A Bayesian approach can be developed as well. It is also worth pointing out that skewed pivot-blend, like other skewness-introducing methods, operates on a given density with asymmetric scales to handle skewed data. We do not explore nonparametric approaches in this paper (but refer to Appendix B for potential ideas involving kernels and data ranks).

Remark 1 (**Pivotal Point vs. Intercept**). *Typically, an intercept α is included the model, and so $r = y - X\beta = y - X^\circ\beta^\circ - 1\alpha$, where $X = [1, X^\circ]$, $X^\circ = [\tilde{x}_1, \ldots, \tilde{x}_n]^T$, $\beta = [\alpha, (\beta^\circ)^T]^T$. Interestingly, when skewness is present, the pivotal point m diverges from the intercept α.*

Specifically, based on previous discussions, we have the following density form

$$\sum_{i=1}^{n} \frac{\phi\big(\frac{y_i - \tilde{x}_i^T\beta^\circ - \alpha - m}{\sigma} + m\big)1_{y_i - \tilde{x}_i^T\beta^\circ \le m+\alpha} + \phi\big(\frac{y_i - \tilde{x}_i^T\beta^\circ - \alpha - m}{\nu} + m\big)1_{y_i - \tilde{x}_i^T\beta^\circ > m+\alpha}}{\Phi(m)\sigma + \{1 - \Phi(m)\}\nu}.$$

It is evident that m plays a more intricate role compared to α. If $\sigma = \nu$, the expression within the sum can be rewritten in a location-scale form: $(1/\sigma)\phi((y_i - \tilde{x}_i^T\beta^\circ - \alpha')/\sigma)$, where $\alpha' = \alpha + (1 - \sigma)m$. In this special case, m can be absorbed into the combined intercept α', which is unique (ensuring the final model has no ambiguity). This also applies to $m \le \min r_i$ or $m \ge \max r_i$, regardless of scale differences. However, in situations beyond the simple unskewed case (e.g., when σ and ν are not exactly equal and m is within the support of r_i or $1/n < q < 1 - 1/n$), m cannot *be incorporated into the intercept or casually discarded.*

To the best of our knowledge, the distinct roles of pivotal point and intercept in the context

of skewness have received little attention in existing literature. Our proposal is one of the first attempts to introduce pivotal point estimation into the statistical modeling of skewed data.

Remark 2 (Backward Pivot-blend for Residual Diagnostics). *Let's start by rewriting the forward pivot-blend transform for generating a random variable following* $SP^{(\phi)}(\sigma, \nu, m)$*: With* $Y_-^0 \sim \phi(y \mid y \leq m), Y_+^0 \sim \phi(y \mid y > m)$ *and an independent Bernoulli variable* $U \sim Ber(\sigma\Phi(m)/[\Phi(m)\sigma + \{1 - \Phi(m)\}\nu])$*, we can construct*

$$Y = \begin{cases} m + \sigma(Y_-^0 - m) \text{ if } U = 1 \\ m + \nu(Y_+^0 - m) \text{ if } U = 0, \end{cases}$$

and guarantee that Y *follows* $SP^{(\phi)}(\sigma, \nu, m)$*. Conversely, given* f *representing* $SP^{(\phi)}(\sigma, \nu, m)$*, we can use* $Y_- \sim f(y \mid y \leq m)$*,* $Y_+ \sim f(y \mid y > m)$ *and an independent Bernoulli random variable* $V \sim Ber(\Phi(m))$ *to construct a random variable* $Y^0 \sim \phi$ *using the "*backward*" pivot-blend:*

$$Y^0 = \Big(\frac{Y_- - m}{\sigma} + m\Big)1_{V=1} + \Big(\frac{Y_+ - m}{\nu} + m\Big)1_{V=0}. \tag{5}$$

In addition to employing the inverse of the affine transformations in the forward process, the Bernoulli distribution here features a different probability. Thus, an $SP^{(\phi)}$ *sample can be transformed into a* **weighted** *sample that follows* ϕ*. Importantly, the functional form of the distribution* Φ *is not required to calculate the probability weights; instead, we can turn to the mixing probability formula* (2)

$$\mathbb{P}(V = 1) = \Phi(m) = \frac{\nu\,\mathbb{P}(Y \leq m)}{\nu\,\mathbb{P}(Y \leq m) + \sigma\,\mathbb{P}(Y > m)}, \mathbb{P}(V = 0) = \frac{\sigma\,\mathbb{P}(Y > m)}{\nu\,\mathbb{P}(Y \leq m) + \sigma\,\mathbb{P}(Y > m)} \tag{6}$$

and directly estimate these quantities from the data (also applicable to nonparametric skew estimation in Appendix B).

In the context of SPEUS, (5) *can be used to generate "back-transformed" residuals for*

model diagnostics: once the parameters β, σ, ν, m *are determined, a weighted sample can be created from the residual vector* $r = y - X\beta$*, which, if the model assumption holds, should adhere to* ϕ*. First, define* $\mathcal{L} = \{i : r_i \le m\}$*,* $L = |\mathcal{L}|$*,* $\mathcal{R} = \{i : r_i > m\}$*, and* $R = |\mathcal{R}|$*. As aforementioned,* $\Phi(m)$ *can be estimated by* $L\nu/\{L\nu + R\sigma\}$*. Next, define* $\tilde{r} = [\tilde{r}_i] \in \mathbb{R}^n$*:*

$$\tilde{r}_i = \begin{cases} (\frac{r_i - m}{\sigma} + m) & \text{if } r_i \le m \\ (\frac{r_i - m}{\nu} + m) & \text{if } r_i > m, \end{cases} \qquad \text{for } 1 \le i \le n.$$

Finally, assign two sets of nonuniform *probabilities* p_i *to* $\tilde{r}_i$ *based on* (6)*:*

$$p_i = \nu/(L\nu + R\sigma) \text{ for } i \in \mathcal{L}, \text{ and } \sigma/(L\nu + R\sigma) \text{ for } i \in \mathcal{R}.$$

Now we can use the R software to plot a weighted histogram *of* $\tilde{r}_i$ *and compare it to the hypothetical density* ϕ*. This is akin to standard OLS diagnostics for checking the goodness-of-fit of residuals under the Gaussian assumption. In practical data analysis, after fitting the SPEUS model, one can display the back-transformed residual plot to verify if skewness has been adequately addressed (cf. Figure 4).*

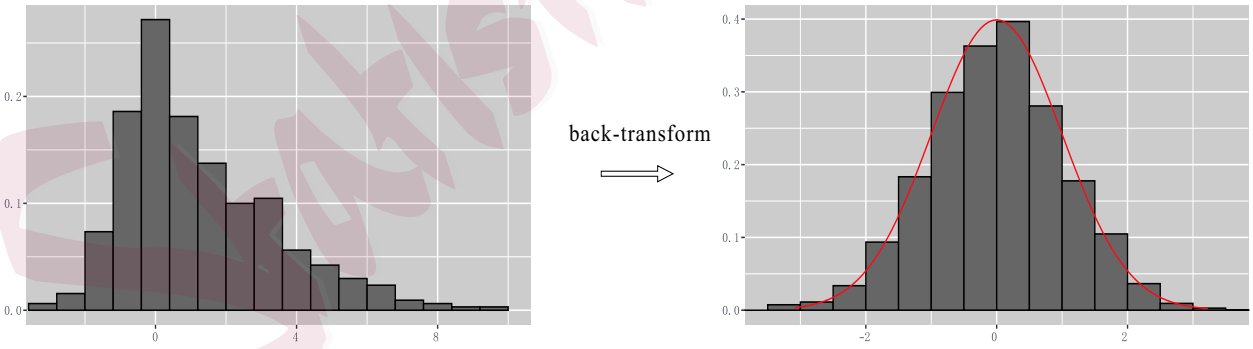


Figure 4: An illustration of the weighted histogram of back-transformed residuals. The left panel shows residuals from a skewed $\mathrm{SP}^{(\phi)}$ model with a symmetric ϕ, and the right panel displays backward pivot-blend residuals using estimated parameters. The model assumption is considered valid when the back-transformed residuals closely resemble ϕ (the red curve) and, importantly, exhibit *symmetry*, indicating effective handling of skewness.

2.3 Expansions of Skewed Pivot-Blend and Relevant Works

2.3.1 Extensions and Beyond

Our main focus is on applications where the loss function ρ is derived from a single density function ϕ. However, there are also variations of skewed pivot-blend that hold value across different applications and fields.

Skewed Pivot-blend for two densities. Skewed pivot-blend extends capabilities to seamlessly "paste" two distinct densities with varying scales, while ensuring continuity at the pivotal point. Consider ϕ and ψ as two continuous densities with respective distributions Φ and Ψ, and m an interior point within the support of both densities. The process of conditioning, affine transformations, and mixing, using two scales, σ and ν in relation to m, leads to

$$p\frac{\phi(\frac{y-m}{\sigma}+m)}{\Phi(m)\sigma}1_{y\le m}+(1-p)\frac{\psi(\frac{y-m}{\nu}+m)}{(1-\Psi(m))\nu}1_{y>m}. \tag{7}$$

Choosing

$$p=\frac{\psi(m)\Phi(m)\sigma}{\phi(m)(1-\Psi(m))\nu+\psi(m)\Phi(m)\sigma}$$

results in the following continuous density:

$$\frac{\psi(m)\phi(\frac{y-m}{\sigma}+m)1_{y\le m}+\phi(m)\psi(\frac{y-m}{\nu}+m)1_{y>m}}{\phi(m)(1-\Psi(m))\nu+\psi(m)\Phi(m)\sigma}. \tag{8}$$

This offers a means of fusing two distinct tail types with exceptional flexibility. Remarkably, even when $\sigma=\nu$, the pasted density in (8) does not conform to a location-scale form (in contrast to the single-density scenario, cf. Remark 1), and m cannot be simply interpreted as a location shift. Beyond the estimation of scales, an intriguing question is to determine the pivotal point at which the two densities coalesce. Furthermore, the concept of skewed pivot-blend can be iteratively applied to paste multiple densities with varying scales. In multidimensional spaces, the pivotal point can be extended to a *pivotal hyperplane* for combining

two densities, which is another intriguing topic for future exploration.

Skewed pivot-blend for bounded losses. In our discussions, we generally assume that ρ is a negative log-likelihood—for example, a convex ρ function like Huber's loss corresponds to a log-concave density. However, it is well established in robust statistics that bounded nonconvex losses are more effective in handling extreme outliers with high leverage. Two prominent examples are Tukey's bisquare loss and Hampel's three-part loss, both of which are bounded (or winsorized, preventing them from reaching $+\infty$) and are constructed using piecewise polynomials (Hampel et al., 2011). We can formulate a general objective for estimating the location parameters

$$\sum_{i=1}^{n} \big\{\rho(\frac{r_i - m}{\sigma} + m)1_{r_i \le m} + \rho(\frac{r_i - m}{\nu} + m)1_{r_i > m}\big\} + n\chi_0 \log\big(\frac{\sigma\Phi(m)/\{1-\Phi(m)\} + \nu}{1 + \Phi(m)/\{1-\Phi(m)\}}\big). \tag{9}$$

Here, $r = y - X\beta$, $0 \le \Phi(m) \le 1$, and χ_0 is for the purpose of calibration. The user can specify the particular forms of ρ and Φ (and in the convex-ρ case, $\Phi(m)$ may take $\lim_{M\to\infty} \int_{-M}^{m} \exp(-\rho(t))\,\mathrm{d}t / \int_{-M}^{M} \exp(-\rho(t))\,\mathrm{d}t$). In robust statistics, it is often recommended to first perform a separate ad-hoc robust scale estimation (Maronna et al., 2006), before proceeding to optimize (9) for the location parameters β, m. But various selections for σ and ν influence the structure of the resulting asymmetric loss. This practice prompts a theoretical inquiry: is it possible to set a finite-sample error bound for location estimation using data-dependent scales, regardless of scale construction or the data distribution of y? For a nonasymptotic analysis using statistical learning theory, see Theorem A.1 for insights on how skewness adds to problem complexity and increases "excess risk".

Finally, extensions of the skewed pivot-blend in nonparametric estimation, such as methods based on data ranks and kernels, can be found in Appendix B due to space constraints.

2.3.2 Other Related Works

Section 1 and Section 2.1 provide a list of relevant works on two-piece distributions. Moreover, as pointed out by a reviewer, our skewed pivot-blend idea shares similarities with the "composite models" in the fields of finance and actuarial sciences.

In such a context, researchers often aim to create a new size distribution by combining two distributions: one that is lighter-tailed on the left (e.g., lognormal or Weibull), and another that is heavier-tailed on the right (e.g., Pareto) (Cooray and Ananda, 2005). This can be expressed as $p \cdot \frac{\phi(y)}{\Phi(m)} \cdot 1_{y \le m} + (1-p) \cdot \frac{\psi(y)}{1-\Psi(m)} \cdot 1_{y>m}$ (Scollnik, 2007), with the choice of p to ensure the resulting density is smooth. An alternative proposal appeared in Bernardi and Bernardi (2018), which, however, does not guarantee continuity. For further discussions on composite models, we refer to Klugman et al. (2012) and Dominicy and Sinner (2017).

It is not difficult to see that the composite form described above corresponds to a specific instance of our skewed pivot-blend density (7) with $\sigma = 1, \nu = 1$. However, our research is driven by the need to tackle data skewness, where the values of σ and ν are typically unknown and can vary. Our primary goal is to estimate these potentially distinct scales while also identifying the central pivotal point in the context of skew data analysis. (7) or (8) is notably different from the composite distribution even when $\sigma = \nu$ but not equal to 1.

Additionally, the composite model may be rigid and restrictive due to limited choices for the mixing parameter p (Scollnik, 2007). In contrast, Figure 3 illustrates the flexibility of SP distributions, taking on various shapes through adjustments in σ and ν. The versatility of the skewed pivot-blend, capturing both asymmetry and varied tail behaviors, offers a variety of distributions for practical modelingg.

3 Skewed Two-Part Model with Joint Sparsity

As mentioned in Section 1, our work is driven by the study of "*semicontinuous outcomes*" (Olsen and Schafer, 2001), as defined by a significant proportion of values equaling 0, with

the remaining values following a continuous, often skewed, distribution. For example, the MEPS datasets have many patients showing no medical expenditure (including the sum of out-of-pocket payment, insurance, Medicaid, Medicare, and other payments), and the rest with positive, highly skewed and heavy-tailed medical costs (see Figure 1). Semicontinuous outcomes are frequently encountered in biomedical and economic applications, as well as rainfall levels and daily drinking records (Hyndman and Grunwald, 2000; Liu et al., 2008; Sarul and Sahin, 2015).

Because zero medical cost means no medical service, rather than an outcome resulting from truncation or sampling, commonly used biometric models like the Tobit model and zero-inflated models (Tobin, 1958; Lambert, 1992) are not suitable. Instead, the *two-part* model (Cragg, 1971; Mullahy, 1998), sometimes also referred to as a *hurdle* model, is more appropriate. This model can be expressed as:

$$\text{Two parts for semicontinuous } y\text{:} \quad \begin{cases} \mathbb{P}(y_i = 0) = \pi_i = 1/\{1 + \exp(-\tilde{x}_i^T b)\} \\ y_i \mid y_i > 0 \sim f(y_i; \tilde{x}_i^T \beta). \end{cases} \tag{10}$$

In the binary part, the probability of observing a zero response is typically modeled using logistic regression or a probit model; in the continuous part, the density function f represents a *positive* random variable with parameter $\tilde{x}_i^T \beta$. Without loss of generality, let's assume that for $1 \le i \le n$, $y_i > 0$, while for $n < i \le N$, $y_i = 0$, and so the response and the overall prediction matrix $\widetilde{X}$ can be partitioned as

$$y = \big[[y_1, \ldots, y_n]\ [0, \ldots, 0]\big]^T, \quad \widetilde{X} = \big[[\tilde{x}_1, \ldots, \tilde{x}_n]\ [\tilde{x}_{n+1}, \ldots, \tilde{x}_N]\big]^T = [X^T\ Z^T]^T. \tag{11}$$

We then derive the negative log-likelihood from the distribution defined in (10), with respect

to a combination of the Lebesgue measure on $\mathbb{R}_+$ and a counting measure at 0:

$$\begin{aligned} &-\sum_{i=1}^{N}\left[1_{y_i=0}\log\pi_i + 1_{y_i>0}\log\{(1-\pi_i)f(y_i;\tilde{x}_i^T\beta)\}\right] \\ =&\sum_{i=1}^{N}\left[-\tilde{x}_i^T b 1_{y_i=0} + \log\left\{1+\exp(\tilde{x}_i^T b)\right\}\right] + \sum_{i=1}^{n} -\log f(y_i;\tilde{x}_i^T\beta). \end{aligned} \tag{12}$$

Each predictor makes a composite contribution to the response through two parts, but (12) is separable with respect to b and β, making it amenable to optimization. Below, we will introduce two modifications to the classical two-part model to better address some challenges in modern applications: (a) mitigating skewness in the positive part through the use of pivot-blend, and (b) enhancing interpretability by incorporating joint variable selection across both model components.

First, specifying an ideal density function for the positive values of y_i can be challenging. As a result, many researchers opt to employ a transformation $T(\cdot):(0,\infty)\to\mathbb{R}$, and assume that the transformed response $T(y_i)$ $(1\le i\le n)$ follows a symmetric distribution, such as a normal or a Laplace. With ρ denoting the corresponding symmetric negative log-likelihood, the last term $\sum_{i=1}^{n} -\log f(y_i;\tilde{x}_i^T\beta)$ in (12) now takes the form

$$\sum_{i=1}^{n}\rho(r_i) \quad \text{with } r = T(y) - X\beta \tag{13}$$

where $r\in\mathbb{R}^n$ is the residual vector associated with the continuous component of the model. Nevertheless, asymmetry continues to manifest in the transformed data in various scenarios, as observed by Chai and Bailey (2008). Our experience shows that routine transformations may not only fail to completely rectify skewness but also introduce a nontrivial pivotal point around which asymmetric scales arise. The technique detailed in Section 2 offers an effective

remedy by replacing (13) with the following

$$n\log[\sigma\Phi(m)+\nu\{1-\Phi(m)\}]+\sum_{i=1}^{n}\big[\rho\big(\frac{r_i-m}{\sigma}+m\big)1_{r_i\le m}+\rho\big(\frac{r_i-m}{\nu}+m\big)1_{r_i>m}\big], \quad (14)$$

where σ, ν, m are all unknown.

Second, practitioners of two-part models encounter another pressing challenge—the abundance of predictors collected. Variable selection provides a valuable tool for enhancing model interpretation and prediction, but in the context of two-part models, it is crucial to identify predictors that are relevant to the entire population, rather than just focusing on the subpopulation with positive responses or the subpopulation with zero responses. In other words, a predictor can only be eliminated if it bears zero coefficients in *both* the binary and continuous parts of the model.

Combining both elements, the incorporation of regularization and skewed pivot-blend allows us to formulate a sparse skewed 2-part ($\mathbf{S}^2$) criterion for modeling semicontinuous outcomes with joint variable selection:

$$\begin{aligned}\mathbf{S}^2: \min_{b,\beta,\sigma,\nu,m}\ & n\log\big[\sigma\Phi(m)+\nu\{1-\Phi(m)\}\big]+\sum_{i=1}^{n}\big\{\rho\big(\frac{r_i-m}{\sigma}+m\big)1_{r_i-m\le 0}\\ &+\rho\big(\frac{r_i-m}{\nu}+m\big)1_{r_i-m>0}\big\}+\sum_{i=1}^{N}\big[-\tilde{x}_i^Tb1_{y_i=0}+\log\big\{1+\exp(\tilde{x}_i^Tb)\big\}\big]\\ &+\lambda\|B\|_{2,1}+P_2(\sigma,\nu,m,\beta;\tau)\ \text{ s.t. } r=T(y)-X\beta, B=[\sqrt{n}\beta,\sqrt{N}b],\sigma>0,\nu>0.\end{aligned} \quad (15)$$

Practically, it is common to include two intercepts, one for the binary part and one for the continuous part of the model, which are not subject to any penalty. The (2,1)-norm applied to matrix B enforces the desired row-wise sparsity for joint variable selection, but can be substituted with a row-wise nonconvex penalty like group SCAD or MCP. Incorporating the scaling factors in the construction of B is essential for the use of a single regularization parameter. The term P_2 represents an ℓ_2-penalty, akin to ridge regression, to account for significant noise and design collinearity. An example is adding $(\tau/2)(1/\sigma^2+1/\nu^2)$ (especially

when $p \geq n$), which from a Bayesian perspective amounts to an inverse gamma prior on σ^2 and ν^2. Empirical studies show that τ is not sensitive, and a small τ often suffices. Likewise, we suggest including an ℓ_2-penalty on m which translates to a Gaussian prior (or alternatively, a beta prior for the quantile parameter q), because in cases involving asymmetric scales $\sigma \neq \nu$, a moderate $|m|$ value (or q near 1/2) can exert a considerable influence on the model, warranting deeper exploration in applications. (This might contrast with outlier effects which exhibit more of a tail behavior inconsistent with most data and are complex to model with a single distribution due to heterogeneity.) Adding the ℓ_2 penalties also facilitates the theoretical analysis in the next section.

(15) involves the estimation of coefficients b, β, a pivotal point m, and two scales σ, ν and is one-sided directionally differentiable (She et al., 2021). In contrast to the conventional two-part (12), this criterion no longer shows separability in (b, β) and includes a non-differentiable penalty. Efficient computation of the estimates can be achieved through optimization techniques. In handling the nondifferentiable (2,1)-penalty, we can express $\|B\|_{2,1}$ as $(1/2)\sum_{j=1}^{p}(B_{j,1}^2 + B_{j,2}^2)/a_j + a_j$ with each $a_j > 0$. This yields a differentiable criterion, facilitating the use of standard optimization solvers such as Newton or quasi-Newton methods. Alternating optimization can also be used to improve scalability.

4 Analysis of Sparse Skewed Two Parts

In this section, we delve into the theoretical underpinnings of the sparse skewed two-part estimation (15) introduced earlier. Our investigation differs from classical asymptotics which assume a fixed number of predictors and an infinite sample size. Nonasymptotic theory remains relatively unexplored when considering the interplay of skewness, regularization, and heavy tails collectively. A significant challenge entails comprehending the impact of asymmetric scales, pivotal points, and sparsity on both statistical accuracy and the choice of the regularization parameter in finite sample sizes.

The key implications and contributions of our theoretical framework are as follows: (i) Applicability. Unlike conventional consistency studies, which frequently assume an i.i.d. data structure and require fixed p with $n \to +\infty$, our theory is applicable to any values of n and p, and does not require the design matrix to have i.i.d. rows. (ii) Effective Noise and Flexibility in Tails. Our investigation reveals that prediction and estimation errors, as well as the choice of the regularization parameter, are linked to the tail decay characteristics of the "effective noise". The concept diverges from raw noise and often exhibits *lighter* tails (cf. Remark 3). Moreover, we employ Orlicz ψ-norms to model various tail behaviors (cf. Lemmas A.1–A.4), providing flexibility in real applications. In essence, when the effective noise shows light tails, implying a finite Orlicz norm with a "large" ψ function, the presence of ψ^{-1} in the choice of the regularization parameter leads to a reduced error bound (cf. (25) or (27)). (iii) Misspecification Tolerance. The core theorems do not require a zero-mean effective noise, as demonstrated in Theorems 1, 2, A.2, A.3, A.4. Consequently, our analysis applies to misspecified models (where the risk function associated with the given loss does not necessarily vanish at the statistical truth).

4.1 Preliminaries: Reparametrization and Effective Noise

Recall the loss function in (15), which can be represented by

$$\begin{aligned} l_0(\beta, b, m, \sigma, \nu) = \sum_{i=1}^{n} \Big\{ & \rho\big(\frac{r_i - m}{\sigma} + m\big) 1_{r_i \le m} + \rho\big(\frac{r_i - m}{\nu} + m\big) 1_{r_i > m} \\ & + \log\big[\sigma\Phi(m) + \nu\{1 - \Phi(m)\}\big] \Big\} + \langle \mathfrak{L}(Zb; \mathcal{Y}), 1_N \rangle, \end{aligned} \tag{16}$$

where $r = y - X\beta$ denotes the plain residuals, ρ and $\mathfrak{L}$ are two differentiable losses that respectively operate on the continuous part and binary part. Here, the observed data are represented by $y, \mathcal{Y}, X, Z$, with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\mathcal{Y} \in \{0,1\}^N$ and $Z \in \mathbb{R}^{N \times p}$ (cf. (11)). For simplicity, the section assumes that the number of rows in X and Z is *each* bounded by cn, with n representing the order of the sample size and c a positive constant.

To ease theory and presentation, we introduce some concatenated symbols. First, β and b can be combined into $\bar{\beta}$ as the coefficient vector for an extended design matrix $\bar{X}$:

$$\bar{\beta} = \mathrm{vec}([\beta, b]) = \begin{bmatrix} \beta \\ b \end{bmatrix}, \ \bar{X} = \begin{bmatrix} X & 0 \\ 0 & Z \end{bmatrix}.$$

We denote $[\beta_k, b_k]^T$ by $\bar{\beta}_k$, the coefficients associated with the kth and $(p+k)$th columns of $\bar{X}$. Based on the problem structure in (16), we define

$$\bar{m} = m1_n, \quad \varsigma = \begin{bmatrix} (1/\sigma)1_n \\ (1/\nu)1_n \end{bmatrix},$$

where 1_n is to match the scale of the design matrices when considering prediction errors. Introduce ζ as the overall unknown vector, as well as γ and μ

$$\zeta = [\bar{\beta}^T, \bar{m}^T, \varsigma^T]^T, \quad \gamma = [\bar{m}^T, \varsigma^T]^T, \quad \mu = [\bar{\eta}^T, \bar{m}^T, \varsigma^T]^T, \quad \bar{\eta} = \bar{X}\bar{\beta}. \tag{17}$$

With the above notations, we can rewrite the general problem of interest as

$$l(\mu) + \|\varrho\bar{\beta}\|_{2,P} + \frac{\tau}{2}\|\gamma\|_2^2, \tag{18}$$

where l is the loss on μ, $\|\bar{\beta}\|_{2,P} := \sum_{k=1}^{p} P(\|\bar{\beta}_k\|_2; \lambda)$ and P is a sparsity-promoting penalty. Including ϱ in the penalty allows for a scale adjustment based on the size of the designs, enabling a universal choice of λ that is independent of the sample size in later theorems. In alternating optimization algorithms, ϱ can takes $\kappa_{2,\infty}$ which represents the maximum column ℓ_2-norm of $\bar{X}$, as a measure of the size of the design:

$$\kappa_{2,\infty} = \max_{1 \le k \le 2p} \|\bar{X}_k\|_2 = \max\left\{\|X_1\|_2, \ldots, \|X_p\|_2, \|Z_1\|_2, \ldots, \|Z_p\|_2\right\}.$$

This quantity is typically on the order of $\sqrt{n}$. When P is the ℓ_1-penalty, (18) reduces to the

previous $\mathbf{S}^2$-criterion (15) for two-part models with skew and sparsity,

$$l(\mu) + \lambda \|\varrho\bar{\beta}\|_{2,1} + \frac{\tau}{2}\|\gamma\|_2^2, \tag{19}$$

where $\|\bar{\beta}\|_{2,1}$ is short for $\sum_{k=1}^{p} \|\bar{\beta}_k\|_2$.

Next, we introduce the notion of "effective noise" to account for randomness, conditional on the design matrices X, Z. Given $l(\mu)$, define the effective noises associated with $\bar{\eta}^*, \gamma^*$ as

$$\epsilon_{\bar{\eta}} = -\nabla_{\bar{\eta}} l(\mu)\big|_{\mu=\mu^*}, \quad \epsilon_{\bar{m}} = -\nabla_{\bar{m}} l(\mu)\big|_{\mu=\mu^*}, \quad \epsilon_{\varsigma} = -\nabla_{\varsigma} l(\mu)\big|_{\mu=\mu^*}, \tag{20}$$

where l is assumed to be differentiable at the statistical truth μ^*.

When formulating statistical assumptions related to effective noises, it is important to account for different types of tail decay. We employ Orlicz ψ-norms, as well as some non-convex variants capable of handling significantly heavier tails (cf. Appendix A.1). In the context of the Orlicz ψ-norm for a random variable (or vector) X, represented as $\|X\|_\psi$, $\psi(\cdot)$ is consistently assumed to be a nondecreasing, nonzero function defined on $\mathbb{R}_+$ with $\psi(0) = 0$ (but not necessarily convex). For the Orlicz-norm of a random vector, please see (A.2). The inverse of ψ is defined as $\psi^{-1}(x) = \sup\{t \in \mathbb{R}_+ : \psi(t) \leq x\}$.

Some notable examples encompass the sub-Weibull ψ_q-norms, with ψ defined as

$$\psi_q(x) = \exp(x^q) - 1, \ x \in \mathbb{R}_+ \tag{21}$$

for $q > 0$. (21) covers sub-Gaussian ($q = 2$) and sub-Exponential ($q = 1$) random variables, but as $q < 1$, the sub-Weibull tails become much heavier (Götze et al., 2021). Another class is the L_q-norms, with $\psi(x) = x^q$ ($q \geq 1$). Orlicz norms provide a useful framework for analyzing skewed random variables (even when they lack a zero mean).

Remark 3 (Effective Noise vs. Raw Noise). *The effective noise, jointly determined by the data and the loss function, may differ from the plain "raw noise" defined by*

$$\epsilon^{raw} := y - \bar{m}^* - \eta^*, \tag{22}$$

where $\eta^* = X\beta^*$. *Comparing* (20) *with* (22), *one appealing aspect of* $\epsilon_{\bar{\eta}}$ *is that it tends to have* light *tails, even when* ϵ^{raw} *does not. Indeed, a straightforward derivative calculation based on* (16) *shows that*

$$\epsilon_{\bar{\eta},i} = \begin{cases} \frac{\rho'(\epsilon_i^-)}{\sigma^*} 1_-(\epsilon_i^{raw}) + \frac{\rho'(\epsilon_i^+)}{\nu^*} 1_+(\epsilon_i^{raw}), & 1 \le i \le n, \\ \mathfrak{L}'(Z_i^T b^*; \mathcal{Y}_i), & n < i \le N, \end{cases} \tag{23a}$$

$$\epsilon_{\bar{m},i} = (\frac{1}{\sigma^*} - 1)\rho'(\epsilon_i^-)1_-(\epsilon_i^{raw}) + (\frac{1}{\nu^*} - 1)\rho'(\epsilon_i^+)1_+(\epsilon_i^{raw}) + \frac{\Phi'(m^*)(\nu^* - \sigma^*)}{\sigma^*\Phi(m^*) + \nu^*(1 - \Phi(m^*))}, \tag{23b}$$

$$\epsilon_{\varsigma,i} = \begin{cases} -\epsilon_i^{raw}\rho'(\epsilon_i^-)1_-(\epsilon_i^{raw}) + \frac{\sigma^{*2}\Phi(m^*)}{\sigma^*\Phi(m^*)+\nu^*(1-\Phi(m^*))}, & 1 \le i \le n, \\ -\epsilon_i^{raw}\rho'(\epsilon_i^+)1_+(\epsilon_i^{raw}) + \frac{\nu^{*2}\{1-\Phi(m^*)\}}{\sigma^*\Phi(m^*)+\nu^*(1-\Phi(m^*))}, & n < i \le 2n, \end{cases} \tag{23c}$$

where $\epsilon_i^- = \epsilon_i^{raw}/\sigma^* + m^*, \epsilon_i^+ = \epsilon_i^{raw}/\nu^* + m^*$. *Therefore, if* $|\rho'| \le M, |\mathfrak{L}'| \le B$ *for some positive* M, B *(e.g., when using Huber's loss for* ρ *and logistic deviance for* $\mathfrak{L}$*), then*

$$|\epsilon_{\bar{\eta},i}| \le \frac{M}{\sigma^* \wedge \nu^*} + B, \tag{24a}$$

$$|\epsilon_{\bar{m},i}| \le (|1 - \frac{1}{\sigma^*}| \vee |1 - \frac{1}{\nu^*}|)M + \frac{|\sigma^* - \nu^*|}{\sigma^* \wedge \nu^*}, \tag{24b}$$

$$|\epsilon_{\varsigma,i}| \vee |\epsilon_{\varsigma,i+n}| \le M|\epsilon_i^{raw}| + (\sigma^* \vee \nu^*). \tag{24c}$$

It is evident that all components of $\epsilon_{\bar{\eta}}$ *and* $\epsilon_{\bar{m}}$ *are* bounded, *thereby possessing a finite* ψ_2*-norm regardless of heavy tails that* ϵ^{raw} *may exhibit. Finally, it is worth noting that our theorems below impose Orlicz-norm conditions on the entire random vectors in* (20), *which is more flexible than assuming that the vectors have independent components, each with a finite Orlicz norm and a mean of 0 (cf. Lemma A.1). Furthermore, one can employ generalized Bernstein-Orlicz norms for random vector marginals, as described in Kuchibhotla*

and Chakrabortty (2022) to develop sharper bounds under an additional minimum sample size constraint. We will not explore this further in the current paper.

4.2 Nonasymptotic Error Bounds

This part demonstrates some error bounds when using the $(2,1)$-penalty. Additional results can be found in the appendices, such as Theorem A.2 providing a universal form for λ applicable to a broad range of tails, Theorem A.3 presenting an elementwise error bound, and Theorem A.4 examining a general sparsity-inducing penalty.

In what follows, we denote the group support of $\bar{\beta}$ as $\mathcal{J}(\bar{\beta}) = \{k : \bar{\beta}_k = [\beta_k, b_k]^T \neq 0, 1 \leq k \leq p\}$ and $J(\bar{\beta})$ is the cardinality of $\mathcal{J}(\bar{\beta})$. Also, define $\mathcal{J}^* = \mathcal{J}(\bar{\beta}^*)$, $J^* = J(\bar{\beta}^*)$, and $\hat{\mathcal{J}} = \mathcal{J}(\hat{\bar{\beta}})$ for short, and let $\mathcal{J}^{*C} \subset \{1, \ldots, p\}$ denote the complement of $\mathcal{J}^*$. The generalized Bregman function $\boldsymbol{\Delta}_l$ is useful in defining an appropriate error measure and making regularity conditions: given a function l differentiable at η', $\boldsymbol{\Delta}_l(\eta, \eta') := l(\eta) - l(\eta') - \langle \nabla l(\eta'), \eta - \eta' \rangle$ and $\bar{\boldsymbol{\Delta}}_l(\eta, \eta') := \{\boldsymbol{\Delta}_l(\eta, \eta') + \boldsymbol{\Delta}_l(\eta', \eta)\}/2$. The differentiability can be replaced by directional differentiability (She et al., 2021). If l is also strictly convex, $\boldsymbol{\Delta}_l(\eta, \eta')$ becomes the standard Bregman divergence $\mathbf{D}_l(\eta, \eta')$ (Bregman, 1967). For the specific case of $l(\eta) = \|\eta - y\|_2^2/2$, $\boldsymbol{\Delta}_l(\eta, \eta') = \|\eta - \eta'\|_2^2/2$, or $\mathbf{D}_2(\eta, \eta')$ for short.

Theorem 1. *Assume that the effective noises $\epsilon_{\bar{\eta}}, \epsilon_{\bar{m}}$, and ϵ_{ς} are bounded in Orlicz norms: $\|\epsilon_{\bar{\eta}}\|_{\psi} \leq \omega_{\bar{\eta}}, \|\epsilon_{\bar{m}}\|_{\psi} \leq \omega_{\bar{m}}$, and $\|\epsilon_{\varsigma}\|_{\varphi} \leq \omega_{\varsigma}$, where ψ, φ satisfy: i) $\psi(x)$ is convex and $\psi(x)\psi(y) \leq c_1\psi(c_0 xy)$, $\forall x, y \geq c_2$, for some positive c_0, c_1, c_2 (dependent on ψ only), (ii) $\{\psi^{-1}(t)\}^2$ is concave or $\{\psi^{-1}(t)\}^2 \lesssim t$ on $\mathbb{R}_+$; iii) $\{\varphi^{-1}(t)\}^2$ is concave or $\{\varphi^{-1}(t)\}^2 \lesssim t$ on $\mathbb{R}_+$. Consider the estimator $\hat{\zeta} = [\hat{\bar{\beta}}^T, \hat{\gamma}^T]^T$ by minimizing (19) with $\varrho \geq \kappa_{2,\infty}$ and $\lambda = A\big\|\bar{X}^T \epsilon_{\bar{\eta}}\big\|_{\infty}/\varrho$, where A a large enough constant. Then*

$$\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat{\mu}, \mu^*) \vee \tau \mathbf{D}_2(\hat{\gamma}, \gamma^*)\Big\} \lesssim c_{\psi}\omega_{\bar{\eta}}\,\varrho\,\psi^{-1}(p)\|\bar{\beta}^*\|_{2,1} + \frac{1}{\tau}\{\psi^{-1}(1)\}^2\omega_{\bar{m}}^2 + \frac{1}{\tau}\{\varphi^{-1}(1)\}^2\omega_{\varsigma}^2 + \tau\|\gamma^*\|_2^2. \tag{25}$$

where $c_{\psi} = c_0(1 \vee c_1 \vee 2\psi(c_2))^2\psi^{-1}(1)$.

Theorem 1 provides a bound on prediction and estimator errors, measured using generalized Bregman functions. Notably, this bound does *not* necessitate any regularity conditions on the design matrices or signal strength.

The assumptions (i)–(iii) on effective noise tails are mild, and the functions φ and ψ can be applied to a wide range of cases. For example, it is straightforward to verify that $\|\cdot\|_\psi$ can represent a ψ_q-norm with $q \geq 1$ (van der Vaart and Wellner, 2013) where we can take $c_1 = 1, c_2 = 1, c_0 = 2^{1/q}$; $\|\cdot\|_\varphi$ can be sub-Weibull for some $q > 0$, or an L_q-norm ($q \geq 2$) with heavy polynomial tails.

The first term on the right-hand side of (25), $c_\psi \omega_{\bar\eta}\, \varrho\, \psi^{-1}(p) \big\|\bar\beta^*\big\|_{2,1}$, is the dominant term scaling with p. Remark 3 emphasizes that $\epsilon_{\bar\eta}$ can have considerably lighter tails, enabling the choice of a large ψ function. For instance, when $|\rho'| \leq M, |\mathfrak{L}'| \leq B$, we can take $\psi = \psi_2$, $\omega_{\bar\eta} = c\{M/(\sigma^* \wedge \nu^*) + B\}$. Such a substantial ψ function ensures that the error rate, which incorporates ψ^{-1}, stays well controlled, even when the raw noise (22) exhibits heavy tails.

Furthermore, with proper regularity conditions, another error bound that depends on $\bar\beta^*$ though its support J^* can be derived.

Theorem 2. *Assume that the tails of effective noises are bounded in Orlicz norms:* $\|\epsilon_\varsigma\|_\varphi \leq \omega_\varsigma$, $\|\epsilon_{\bar\eta}\|_{\psi_q} \leq \omega_{\bar\eta}, \|\epsilon_{\bar m}\|_{\psi_q} \leq \omega_{\bar m}$ *for some* $q > 0$. *Let* $\hat\zeta$ *denote the optimal solution for* (19) *with* $\varrho \geq \kappa_{2,\infty}$ *and*

$$\lambda = A\omega_{\bar\eta}(\log p)^{\frac{1}{q}}$$

for some large enough $A > 0$. *Suppose that there exist a large* $K > 0$ *and a constant* ϑ *such that for any* $\bar\beta, \gamma$

$$(1+\vartheta)\lambda\varrho\big\|(\bar\beta - \bar\beta^*)_{\mathcal{J}^*}\big\|_{2,1} \leq \mathbf{\Delta}_l(\mu, \mu^*) + \lambda\varrho\big\|(\bar\beta - \bar\beta^*)_{\mathcal{J}^{*C}}\big\|_{2,1} + K\lambda^2 J^*. \tag{26}$$

Then

$$\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)\vee\tau\mathbf{D}_2(\gamma,\gamma^*)\lesssim KA^2\omega_{\bar{\eta}}^2(\log p)^{\frac{2}{q}}J^*+A^2\frac{\omega_{\bar{m}}^2}{\tau}(\log p)^{\frac{2}{q}}+\frac{\omega_{\varsigma}^2}{\tau}\{\varphi^{-1}(p^{Aq})\}^2+\tau\|\gamma^*\|_2^2 \tag{27}$$

holds with probability at least $1-Cp^{-cA^q}$*, where* C,c *are positive constants.*

To obtain a sufficient condition for the regularity condition (26), we can confine $\xi=\bar{\beta}-\bar{\beta}^*$ within a cone: $(1+\vartheta)\|\xi_{\mathcal{J}^*}\|_{2,1}\geq\|\xi_{\mathcal{J}^{C*}}\|_{2,1}$, and require either $\varrho^2\|\xi_{\mathcal{J}^*}\|_{2,1}^2\leq\tilde{K}J^*\boldsymbol{\Delta}_l(\mu,\mu^*)$ or $\varrho^2\|\xi_{\mathcal{J}^*}\|_2^2\leq\tilde{K}\boldsymbol{\Delta}_l(\mu,\mu^*)$ for some large $\tilde{K}>0$. These conditions extend the compatibility and restricted eigenvalue conditions which are widely used in sparse regression (Bickel et al., 2009; van de Geer and Bühlmann, 2009). But (26) is less technically demanding.

In proving the theorem, we establish a more general result where ψ_q can be replaced with a general ψ and the appropriate choice for λ is of the order $\omega_{\bar{\eta}}\psi^{-1}\left(p\psi A\psi^{-1}(p)\right)$. For more, please refer to Appendix A.4.

To illustrate the bound (27), let's consider a scenario where $|\rho'|\vee|\mathfrak{L}'|\leq M$ for some $M>0$. Under the assumptions of independent centered effective noise components and $\|\epsilon_i^{\text{raw}}\|_{\psi_q}\leq\omega$ for some $q\in(0,2]$, we can deduce based on (24) in Remark 3 that the error bound in (27) is of the following order (treating K,A as constants):

$$\log p\cdot\Big[\big(1\vee\frac{1}{\sigma^{*2}}\vee\frac{1}{\nu^{*2}}\big)M^2J^*+\frac{1}{\tau}\Big\{[(1-\frac{1}{\sigma^*})^2\vee(1-\frac{1}{\nu^*})^2]M^2+\big(\frac{\sigma^*\vee\nu^*}{\sigma^*\wedge\nu^*}-1\big)^2\Big\}\Big]$$
$$+(\log p)^{2/q}\cdot\frac{1}{\tau}\{M\omega+\sigma^*\vee\nu^*\}^2+\tau\|\gamma^*\|_2^2.$$

The bound varies with the number of predictors logarithmically, and quantifies the impact of asymmetric scales in the context of sparse skewed two-part models.

Remark 4. *Under suitable regularity conditions similar to those in Theorem 2, the following* $(2,\infty)$*-norm bound holds (cf. Theorem A.3 in Appendix A.5):*

$$\|\hat{\bar{\beta}}-\bar{\beta}^*\|_{2,\infty}\leq C\frac{\sqrt{K\alpha}\vee\vartheta}{\alpha\sqrt{n}}A\Big\{\omega_{\bar{\eta}}+\frac{\omega_{\bar{m}}+\omega_{\varsigma}}{\sqrt{J^*}}\Big\}(\log p)^{\frac{1}{q}} \tag{28}$$

with probability at least $1 - Cp^{-(A\vartheta)^q} - 1/\varphi(cA\vartheta(\log p)^{1/q})$, *where* C, c *are positive constants and* A, K, α, ϑ *can often be treated as constants. Hence with a proper signal strength* $\min_{k\in\mathcal{J}(\bar{\beta}^*)} \|\bar{\beta}^*_k\|_2 > 2CA(\sqrt{K\alpha} \vee \vartheta)\{\omega_{\bar{\eta}} + (\omega_{\bar{m}} + \omega_\varsigma)/\sqrt{J^*}\}(\log p)^{\frac{1}{q}}/(\alpha\sqrt{n})$, (28) *guarantees faithful variable selection,* $\mathcal{J}^* \subset \hat{\mathcal{J}}$, *with high probability.*

Finally, our analysis can be extended to a general P *beyond the* ℓ_1*-type penalty. See Appendix A.6 for more details.*

5 Experiments

We conducted a variety of synthetic and real data experiments to evaluate the performance of the proposed method. Due to limited space, we only present a selection of our data analyses in the following subsections. Interested readers may refer to Appendix C for more experiment results.

5.1 Simulations

In this part, we conduct simulation experiments to compare our proposed methods with some popularly used approaches for skewed estimation. The predictor matrix $X = [X_1, \ldots, X_n]^T \in \mathbb{R}^{n\times p}$ is generated by $X_i \overset{i.i.d.}{\sim} N(0, \Sigma)$, where $\Sigma = [\kappa^{|i-j|}]$ has a Toeplitz structure. The response vector is generated according to $y = X\beta^* + 1\alpha^* + \epsilon$, where $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{SP}^{(\phi)}(\sigma^*, \nu^*, m^*)$. In the setups to be introduced, we set $\alpha^* = 0$ and $\beta^* = [12, 13, 14]^T$. The pivotal point will be chosen as the mean, median, mode, quartiles, and more.

The following methods are included for comparison: quantile regression (QR) (Koenker and Bassett Jr, 1978), Bayesian quantile regression (BQR) (Yu and Moyeed, 2001), Z-estimation quantile regression (ZQR) (Bera et al., 2016), adaptive M-estimation (AME) (Yang et al., 2019), epsilon-skew-normal regression (ESN) (Mudholkar and Hutson, 2000), in addition to SPEUS. The first two methods require the user to specify a quantile parameter (Koenker, 2009; Benoit and Van den Poel, 2017); we employ the *oracle quantile*

$(\Phi(m^*)\sigma^*/[\Phi(m^*)\sigma^* + \{1-\Phi(m^*)\nu^*\}]$ (computed using the truth) in all experiments, and so denote the methods by QR* and BQR*, respectively. In BQR*, the posterior mean estimate is obtained with 4000 MCMC draws after 1000 burn-in samples. For SPEUS, the scale parameters σ and ν, as well as the pivotal point m, are all considered unknown and are estimated from the data. Given each setup, we repeat the experiment for 50 times and evaluate the performance of each method based on Err(β), Err(σ) and Err(ν). Err(β) is the (absolute) root-mean-square error on β, and Err(σ) and Err(ν) denote the (relative) root-mean-square errors on σ and ν, i.e., $\text{Err}(\sigma) = \{\sum_{t=1}^{N}(\hat{\sigma}_t/\sigma^* - 1)^2/N\}^{1/2}$, where $\hat{\sigma}_t$ is the estimate on the tth simulation dataset.

Ex 1. (Skewed Gaussian and skewed Laplace, $m^* = 0$): Let ϕ be the standard normal or Laplace density, $n = 300$, $\kappa = 0.5$, $\sigma^* = 0.2$ and $\nu^* = 0.4, 0.6, 1.2$.

Ex 2. (Skewed Gaussian and skewed Laplace, $m^* \neq 0$): Let ϕ be the standard normal or Laplace density, $n = 300$, $\kappa = 0.2$, $m^* = \Phi^{-1}(0.75)$ (third quartile), $\sigma^* = 0.3$ and $\nu^* = 0.5, 0.7, 0.9$. (We also tried $m = \Phi^{-1}(0.25)$ but the results are similar and omitted.)

	Skewed Gaussian								
	$\nu^*/\sigma^* = 2$			$\nu^*/\sigma^* = 3$			$\nu^*/\sigma^* = 6$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	0.05	—	—	0.06	—	—	0.08	—	—
BQR*	0.04	0.20	0.20	0.05	0.20	0.20	0.07	0.20	0.20
AME	0.04	0.44	0.44	0.04	0.44	0.44	0.06	0.44	0.44
ESN	0.04	0.42	0.42	0.04	0.42	0.42	0.06	0.42	0.42
SPEUS	0.04	0.13	0.08	0.05	0.18	0.06	0.06	0.29	0.05
	Skewed Laplace								
	$\nu^*/\sigma^* = 2$			$\nu^*/\sigma^* = 3$			$\nu^*/\sigma^* = 6$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	0.04	—	—	0.05	—	—	0.17	—	—
BQR*	0.04	0.05	0.05	0.05	0.05	0.05	0.17	0.06	0.06
AME	0.04	0.23	0.20	0.05	0.25	0.21	0.17	0.24	0.20
ZQR	0.04	0.08	0.06	0.05	0.09	0.06	0.17	0.12	0.06
SPEUS	0.04	0.10	0.07	0.05	0.12	0.07	0.17	0.18	0.06

Table 1: Skewed normal and skewed Laplace with pivotal point at zero (Ex 1)

Tables 1 and 2 provide a detailed comparison of the performance of various methods. According to Table 1, when $m^* = 0$, the β-errors do not differ significantly. In this setup, BQR*, ZQR and SPEUS all perform well. In the setup of Ex 2 with a nontrivial pivotal

	Skewed Gaussian								
	$\nu^*/\sigma^* = 1.7$			$\nu^*/\sigma^* = 2.3$			$\nu^*/\sigma^* = 3$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	0.16	—	—	0.17	—	—	0.15	—	—
BQR*	0.16	1.04	0.30	0.16	0.67	0.43	0.14	0.46	0.50
AME	0.06	0.43	1.34	0.07	0.47	0.93	0.08	0.51	0.71
ESN	0.06	0.56	1.05	0.06	0.56	0.73	0.07	0.57	0.58
SPEUS	0.04	0.10	0.08	0.04	0.14	0.07	0.05	0.17	0.06
	Skewed Laplace								
	$\nu^*/\sigma^* = 1.7$			$\nu^*/\sigma^* = 2.3$			$\nu^*/\sigma^* = 3$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	0.17	—	—	0.18	—	—	0.17	—	—
BQR*	0.17	1.07	0.31	0.17	0.69	0.43	0.16	0.48	0.50
AME	0.09	0.19	1.32	0.11	0.26	0.90	0.11	0.31	0.66
ZQR	0.10	0.52	0.55	0.11	0.52	0.34	0.12	0.52	0.22
SPEUS	0.04	0.10	0.08	0.05	0.10	0.07	0.06	0.14	0.07

Table 2: Performance comparison for skewed normal and skewed Laplace with a nontrivial pivotal point (Ex 2)

point, as shown in Table 2, SPEUS significantly outperforms the other methods in both Gaussian and Laplace cases. We also tried other quantiles for m^* (results not reported here) and found that BQR, AME and ZQR typically produce highly inaccurate scale estimates. In contrast, SPEUS is much more successful at accurately recovering the true location and scales, and is not sensitive to different values of m^* even in the heavy-tailed cases.

5.2 Abalone Age

Determining the age of abalone is a tedious and challenging task that often involves counting growth rings under a microscope. We use 7 physical measures of blacklip abalone, including length, diameter, height, weight and others, to predict the age of 1,526 infant samples originally collected by Nash et al. (1994). We split the data into a training set (70%) and a test set (30%). The abalone age dataset is usually analyzed using a regression model based on ordinary least squares (OLS) (Gitman et al., 2018; Chang and Joe, 2019), but the histogram in the left panel of Figure 5 suggests the possibility of skewness. To address this, we employed SPEUS by applying skewed pivot-blend to the normal density function. Moreover, we included the skewed methods AME (Yang et al., 2019) and ESN (Mudholkar and Hutson, 2000) for comparison.

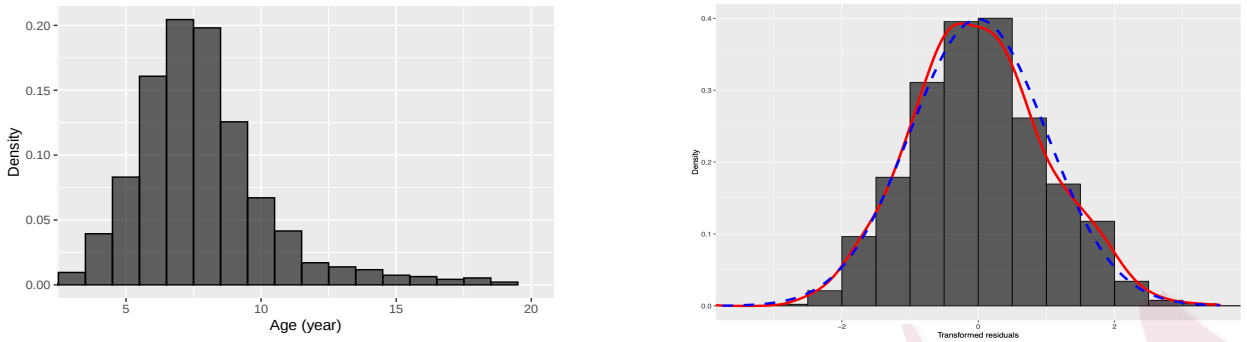
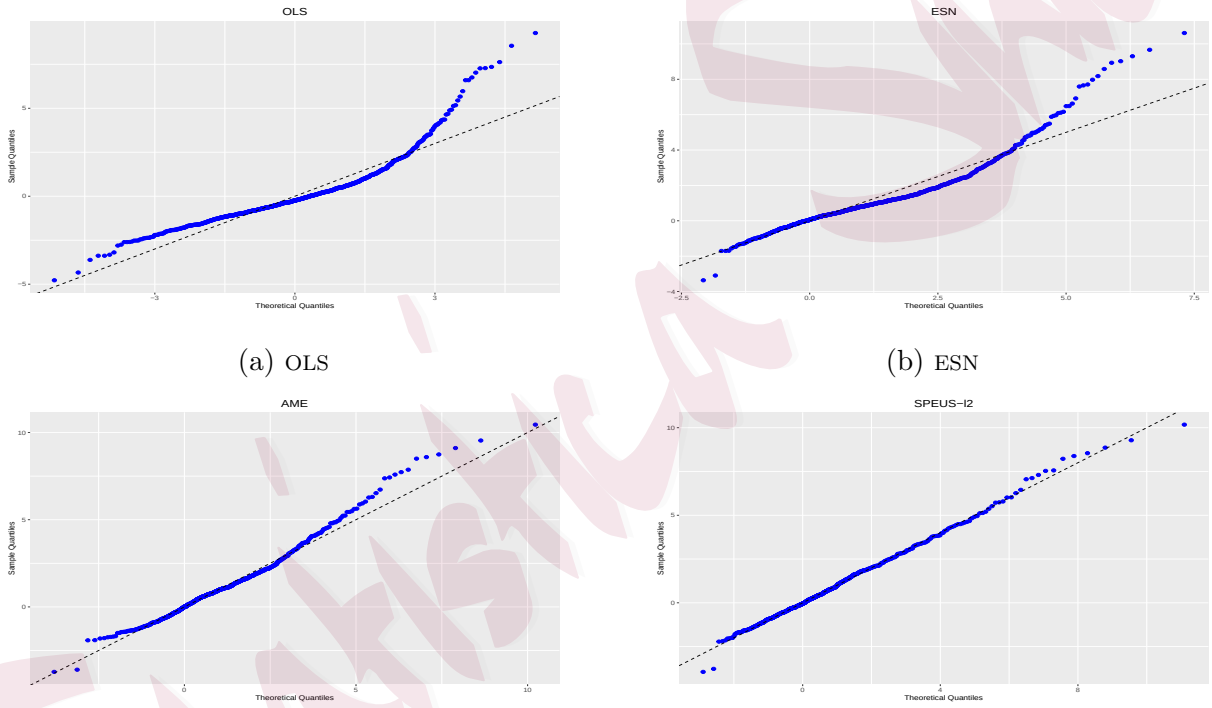


Figure 5: Left: histogram of abalone age. Right: "back-transformed" SPEUS residuals, the associated density estimate (red solid curve), and the standard normal density (blue dashed curve).



(a) OLS (b) ESN

(c) AME (d) SPEUS

Figure 6: Q-Q plots of model residuals on the abalone data.

To assess the goodness of fit of several different models, we present Q-Q plots of the residuals in Figure 6. Although OLS is widely adopted for the data, it clearly demonstrates a lack of fit. The ESN and AME models offer significant improvement through scale and tail adjustments, but their Q-Q plots still exhibit substantial right-skewness. In contrast, the sample quantiles in the SPEUS model nearly match the theoretical quantiles. The symmetry

after the back-transform, as shown in the right panel of Figure 5, corroborates this point.

To compare the fit of the methods, which optimize different criteria based on various distributional assumptions, we conducted Kolmogorov-Smirnov tests and calculated the associated p-values: 0.008 for OLS, 0.003 for ESN, 0.08 for AME, and 0.3 for SPEUS. These findings validate our model's superior fit. It effectively addresses pivotal point and skew effects in the data, surpassing alternative approaches that rely solely on adjustments to intercept and scale.

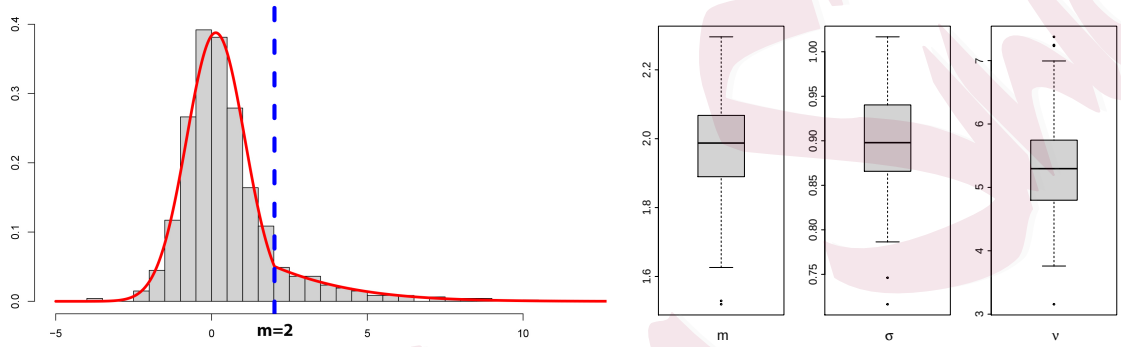


Figure 7: Abalone data. Left: Histogram of SPEUS residuals with a labeled pivotal point. Right: Bootstrap estimates of m, σ, ν (with 100 replications), where $m \neq 0$ and $\sigma \neq \nu$ are significant.

Figure 7 shows the histogram of the SPEUS residuals along with the bootstrap results for m, σ, ν. The scale estimates $\hat{\sigma} = 0.9$, $\hat{\nu} = 5.4$, with 90% confidence intervals $[0.8, 1.0]$ and $[4.0, 6.5]$, respectively, suggest significant skewness in the data compared to a standard Gaussian distribution. The estimated pivotal point, $\hat{m} = 2$ (with a 90% confidence interval $[1.7, 2.2]$) is likely associated with the legal minimum size limits in Tasmania during the 1980s (where and when the data were collected). The determination of size limits included adding an estimated two years' growth to the size at which abalone reached sexual maturity in different areas, aiming to ensure abalone could reproduce before being harvested (Tarbath, 1999). However, blacklip abalone do not mature in size, and the significant growth variability among various abalone stocks led to frequent changes in size limits, impacting abalone of various ages. Our estimated pivotal point appears to correspond with the 2-year protection regulation.

5.3 Medical Expenditure

Modeling medical cost data and identifying relevant predictors are valuable yet demanding tasks. MEPS conducts large-scale surveys across the United States and provides nationally representative information about medical expenditures. We model medical expenditures on 17 features on a subset (stratum ID 2109, third PSU) of the 2019 MEPS data, which includes 150 participants. Of the 17 features, 15 are from Linero et al. (2020), including, for example, the amount of total utilization of prescribed medications (`RXTOT19`), the number of dental care visits in 2019 (`DVTOT19`), age (`AGE19X`), each participant's rating about their own health status (`RTHLTH31`), and categorized family income (`POVCAT19`). We also include the variables `SEX` and `ACTLIM31`, with the latter being a binary variable indicating whether a participant has any physical restrictions that impede his/her ability to engage in physical labor.

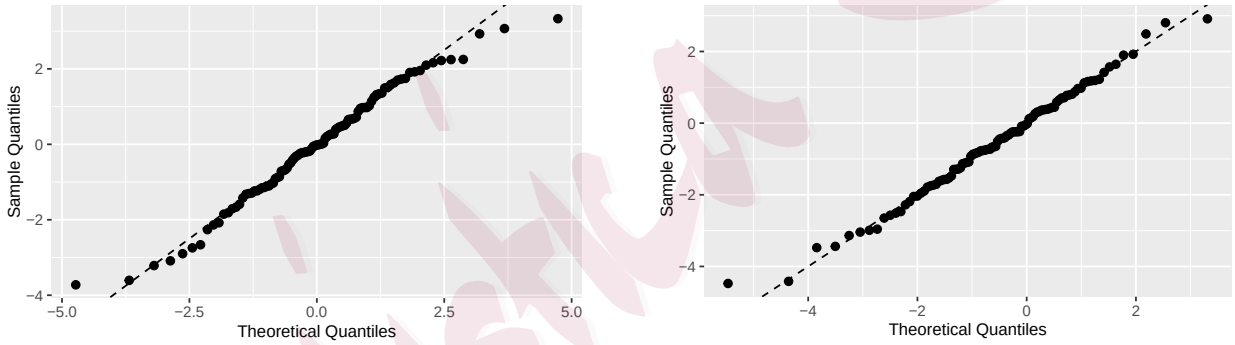


Figure 8: Q-Q plots of the continuous-part residuals on MEPS. The left panel corresponds to the standard log-normal two-part model and the right panel corresponds to its skewness-enhanced counterpart.

Traditional medical cost data analysis often employs a two-part model with logistic and log-normal components. We compared this to the sparse skewed two-part model (cf. Section 3) using a logarithmic function for T and Huber's loss for ρ. The regularization parameter is tuned by 5-fold selective cross-validation (She and Tran, 2019). Figure 8 demonstrates the superior fit of the latter in terms of the continuous component. In binary component analysis, 100 repeated classification tests on 75/25 training/test splits show that our method improved accuracy from 78% to 84%.

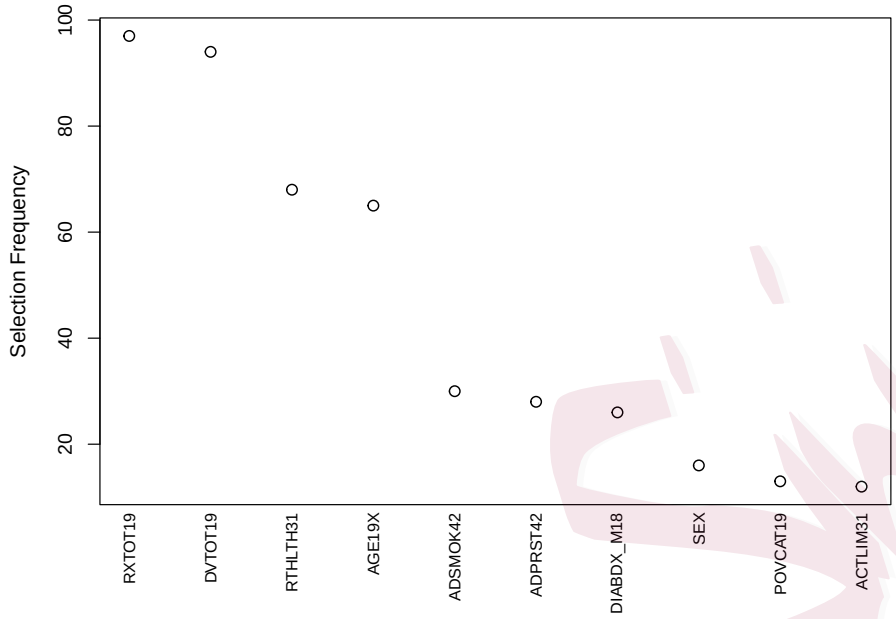


Figure 9: Selection frequencies of the top 10 MEPS features over the 100 bootstrap replications.

Next, we analyze the variable selection outcomes using the proposed model. By bootstrapping the data 100 times, we plot the selection frequencies of the top 10 variables in Figure 9. It is worth noting that setting $\sigma = \nu$ resulted in all variables being selected at low frequencies, less than 23%. This emphasizes the profound influence of skewness on variable selection.

According to Figure 9, the first 4 variables, `RXTOT19`, `DVTOT19`, `AGE19X`, and `RTHLTH31`, exhibit high selection frequencies ($> 60\%$). In contrast, other variables exhibited significantly lower selection frequencies. Below, we provide some practical explanation and guidance regarding these four variables.

First, `RXTOT19`, representing the count of a person's total prescribed medications, is identified as highly influential in predicting medical expenditures. This predictor has the highest correlation (0.43) with the response among all the predictors. Our conclusion is consistent with Holle et al. (2021), highlighting that prescribed medication expenses constitutes a substantial portion of medical costs in the USA. Moreover, the patient's age (`AGE19X`) and

self-perceived health status (RTHLTH31) emerge as significant predictors influencing medical costs. This discovery aligns with Axon and Kamel (2021).

Perhaps interestingly, our analysis also reveals that the number of dental care visits (DVTOT19), selected over 90% of the time, plays a significant role in determining the total medical costs. Its contribution appears to be unique, as it has low correlations (< 0.09) with the other three major predictors (the number of prescriptions, self-perception, and age). Beyond the direct costs of dental care, a plausible explanation could be that individuals with regular dental visits may be more health-conscious and have higher incomes, making them more willing to spend on healthcare.

6 Summary

Skewness poses a significant challenge in data science, and many approaches attempt to model skewness by introducing different scales based on the median of a symmetric, unimodal density. This paper introduced a novel two-piece density family constructed through skewed pivot-blend. "Pivot" refers to the central reference point around which different affine transformations are applied to two conditional densities, and it can be positioned anywhere. "Blend" signifies the merging of these asymmetrically scaled densities using appropriate mixings to create a new continuous density.

We proposed a joint modeling framework that simultaneously estimates scales, the pivotal point, and other location parameters. In particular, we argued that the pivotal point does not correspond to the intercept when skewness is present, a key aspect previously overlooked in the literature. In practice, the inclusion of a single pivotal point parameter significantly enhances a model's capacity in real-world applications.

As an important application, the paper also investigated sparse skewed two-part models, a problem that has recently gained much attention in biomedical and econometric studies. Our non-asymptotic analysis showcases how skewness in random samples, especially those

with potentially heavy tails, can affect statistical accuracy. The quantification of the impact of asymmetrical scales on the choice of regularization parameters and the rates of statistical error provides an insightful examination of skewness within a finite-sample context.

We aim to raise data analysts' awareness of data skew, as well as potential distortions that can arise when applying common transformations and conventional log-likelihoods. The technique of skewed pivot-blend offers an effective strategy for mitigating these challenges.

References

Agency for Healthcare Research and Quality (2015). MEPS HC-181: 2015 Full Year Consolidated Data File.

Agency for Healthcare Research and Quality (2019). MEPS HC-216: 2019 Full Year Consolidated Data File.

Arellano-Valle, R. B., Gómez, H. W., and Quintana, F. A. (2005). Statistical inference for a general class of asymmetric distributions. *Journal of Statistical Planning and Inference*, 128(2):427–443.

Axon, D. R. and Kamel, A. (2021). Patterns of healthcare expenditures among older United States adults with pain and different perceived health status. *Healthcare*, 9(10):1327.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 171–178.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188.

Benoit, D. F. and Van den Poel, D. (2017). bayesQR: A Bayesian approach to quantile regression. *Journal of Statistical Software*, 76(1):1–32.

Bera, A. K., Galvao, A. F., Montes-Rojas, G. V., and Park, S. Y. (2016). Asymmetric Laplace regression: Maximum likelihood, maximum entropy and quantile regression. *Journal of Econometric Methods*, 5(1):79–101.

Bernardi, A. and Bernardi, M. (2018). Two-sided skew and shape dynamic conditional score models. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 121–124.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Boos, D. D. (1987). Detecting skewed errors from regression residuals. *Technometrics*, 29(1):83–90.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

Chai, H. S. and Bailey, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Statistics in Medicine*, 27(18):3643–3655.

Chang, B. and Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics & Data Analysis*, 139:45–63.

Cooray, K. and Ananda, M. M. (2005). Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal*, 2005(5):321–334.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 829–844.

Dominicy, Y. and Sinner, C. (2017). Distributions and composite models for size-type data. In Hokimoto, T., editor, *Advances in Statistical Methodologies and Their Application to Real Problems*, chapter 8, pages 159–183. IntechOpen, Rijeka, Croatia.

Fernández, C. and Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Gitman, I., Chen, J., Lei, E., and Dubrawski, A. (2018). Novel prediction techniques based on clusterwise linear regression. *arXiv preprint:1804.10742*.

Götze, F., Sambale, H., and Sinulis, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26:1–22.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*. New Jersey: John Wiley & Sons.

Holle, M., Wolff, T., and Herant, M. (2021). Trends in the concentration and distribution of health care expenditures in the US, 2001-2018. *JAMA Network Open*, 4(9):e2125179–e2125179.

Hyndman, R. J. and Grunwald, G. K. (2000). Applications: Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Australian & New Zealand Journal of Statistics*, 42(2):145–158.

Jones, M. (2014). Generating distributions by transformation of scale. *Statistica Sinica*, 749–771.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. John Wiley & Sons.

Koenker, R. (2009). Package 'quantreg'. http://CRAN.R-project.org/package=quantreg.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456):1458–1468.

Kuchibhotla, A. K. and Chakrabortty, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11:1389–1456.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lee, M. J. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2020). Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76(1):131–144.

Liu, L., Ma, J. Z., and Johnson, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Statistics in Medicine*, 27(18):3528–3539.

Ma, Y. and Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, 31(3):459–468.

Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New Jersey: John Wiley & Sons.

Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, 83(2):291–309.

Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17(3):247–281.

Nadarajah, S. and Kotz, S. (2003). Skewed distributions generated by the normal kernel. *Statistics & Probability Letters*, 65(3):269–277.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. (1994). The population biology of abalone (Haliotis species) in Tasmania. i. blacklip abalone (H. rubra) from the North Coast and islands of Bass Strait. *Sea Fisheries Division, Technical Report*, 48:411.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Rubio, F. J. and Steel, M. F. (2020). The family of two-piece distributions. *Significance*, 17(1):12–13.

Rubio, F. J. and Steel, M. F. J. (2015). Bayesian modelling of skewness and kurtosis with two-piece scale and shape distributions. *Electronic Journal of Statistics*, 9(2):1884–1912.

Sarul, L. S. and Sahin, S. (2015). An application of claim frequency data using zero inflated and hurdle models in general insurance. *Journal of Business Economics and Finance*, 4(4).

Scollnik, D. (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, 2007(1):20–33.

She, Y. and Tran, H. (2019). On cross-validation for sparse reduced rank regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):145–161.

She, Y., Wang, Z., and Jin, J. (2021). Analysis of generalized bregman surrogate algorithms for nonsmooth nonconvex statistical learning. *The Annals of Statistics*, 49(6):3434–3459.

Stacy, E. W. (1962). A generalization of the Gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Tarbath, D. (1999). *Estimates of Growth and Natural Mortality of the Blacklip Abalone (Haliotis Rubra) in Tasmania*. Technical report series. Tasmanian Aquaculture and Fisheries Institute.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 24–36.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van der Vaart, A. W. and Wellner, J. A. (2013). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Berlin: Springer Science & Business Media.

Wang, J., Boyer, J., and Genton, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica*, 1259–1270.

Yang, T., Gallagher, C. M., and McMahan, C. S. (2019). A robust regression methodology via M-estimation. *Communications in Statistics-Theory and Methods*, 48(5):1092–1107.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zhu, D. and Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, 148(1):86–99.