Semiparametric Estimation of the Shape of the Limiting Bivariate Point Cloud

Reetam Majumder[*,†], Benjamin A. Shaby[‡], Brian J. Reich[§], and Daniel S. Cooley[‡]

Abstract. We propose a model to flexibly estimate joint tail properties by exploiting the convergence of an appropriately scaled point cloud onto a compact limit set. Characteristics of the shape of the limit set correspond to key tail dependence properties. We directly model the shape of the limit set using Bézier splines, which allow flexible and parsimonious specification of shapes in two dimensions. We fit the Bézier splines to data in pseudo-polar coordinates using Markov chain Monte Carlo sampling, utilizing a limiting approximation to the conditional likelihood of the radii given angles. We propose a novel prior on the shape of the limit set via constraints on the paramters of the Bézier splines. A direct advantage of our Bayesian approach is that the support of this prior guarantees that each posterior sample is a valid limit set boundary, allowing direct posterior analysis of any quantity derived from the shape of the curve. Furthermore, we obtain interpretable inference on the asymptotic dependence class by using mixture priors with point masses on the corner of the unit box. Finally, we apply our model to bivariate datasets of extremes of variables related to fire risk and air pollution.

Keywords: Bayesian inference, Bézier curves, extreme values, Gauge functions, limit set.

1 Introduction

Multivariate tail risk calculations require knowledge of the strength of dependence in the joint tail of the relevant distribution. Here, we propose a model to flexibly estimate joint tail characteristics in a way that coherently links several existing measures of tail dependence. To do this, we describe tail dependence of a multivariate distribution through its associated *gauge function* (Balkema et al., 2010; Balkema and Nolde, 2010; Nolde, 2014). The homogeneity property of the gauge function allows us to recover the entire gauge function from its unit level set, which bounds the support of the appropriately scaled data points in the limit (Nolde and Wadsworth, 2022; Wadsworth and Campbell, 2024). We represent the unit level set of the gauge function using a semiparametric model specified within a Bayesian framework wherein required constraints on such functions are automatically satisfied. We obtain a posterior sample of gauge functions, and our prior specification ensures that each member of the sample is a valid gauge function not requiring any *post hoc* adjustments such as re-scaling or truncation.

Efforts to exploit the limit set representation of multivariate extreme values (Davis et al., 1988; Kinoshita and Resnick, 1991; Balkema et al., 2010; Balkema and Nolde,

[*]Southeast Climate Adaptation Science Center, North Carolina State University

[†]Department of Mathematical Sciences, University of Arkansas

[‡]Department of Statistics, Colorado State University

[§]Department of Statistics, North Carolina State University

2010) have appeared only recently. Wadsworth and Campbell (2024) decompose the data into pseudo-polar coordinates and use a limiting argument to approximate the distribution of the radii with a truncated gamma distribution whose parameters depend on the gauge function. They first transform the data to unit exponential margins, and assuming a parametric form for the gauge function, perform maximum likelihood estimation with the truncated gamma likelihood. They extend this approach using mixtures of parametric forms, but need to perform post-hoc re-scaling of the mixtures to satisfy the required properties of valid gauge functions.

In contrast to Wadsworth and Campbell (2024), whose primary focus is estimating probabilities of sets in the joint tail region, Simpson and Tawn (2022) focus on inference for the limit set boundary itself, taking a flexible semiparametric approach. They estimate the sample limit set by approximating the limiting upper endpoint of the distribution of radii with an estimated high quantile, as a function of the angle. To do this, they fit a generalized Pareto distribution, whose scale parameter varies by angle, to the large radii. The radii are calculated by decomposing the bivariate data points transformed to unit exponential margins with a rank transformation. As the result is not a valid limit set, they perform a subsequent scaling and truncation procedure based on a Hill estimator (Hill, 1975) to force their estimate to satisfy the required conditions.

Like Simpson and Tawn (2022), our focus is flexible estimation of the limit set boundary, though our methodology is quite different. Here, we directly model the boundary of the limiting scaled point cloud, which is prescribed by the unit level set of the gauge function, as a Bézier spline. Bézier splines are constituted of Bézier curves, with points on the curve that can be represented as Bernstein polynomials (for reviews, see Hazewinkel, 2012; Farouki, 2012). Similar semiparametric approaches have been used previously in multivariate extremes to characterize the Pickands dependence function for extremal dependence (Marcon et al., 2014, 2017; Vettori et al., 2018), the angular density in the context of multivariate regular variation (Hanson et al., 2017), and the angular dependence function (Murphy-Barltrop et al., 2024), which we will explore below as a direct byproduct of the gauge function. Bézier splines are convenient here because they allow parsimonious specification of shapes in $\mathbb{R}^2$ which are defined by a small number of control points. Placing appropriate constraints on the control points can ensure that the resultant shapes satisfy the conditions required of limit set boundaries. To estimate the parameters of the Bézier spline, we use the result from Wadsworth and Campbell (2024) which says that, given a gauge function evaluated at the data points, the distribution of the large radial components decays like a gamma distribution whose rate parameter depends on the gauge function. We then use standard Markov chain Monte Carlo (MCMC) machinery to sample from the posterior distribution.

In this work, we place prior constraints on parameters of the Bézier splines that define the limit set boundary, which induces a prior on the shape of the limit set. This work is thus far unique in that the support of this prior guarantees that every element of the posterior satisfies the challenging restrictions on the shape of the limit set (Nolde, 2014). Our approach has several advantages. First, since we model the shape of the limiting point cloud in a way that automatically results in a valid limit set, we can use standard MCMC methods to obtain valid posterior inference. Second, our model

allows the boundary of the limit set to exactly touch the corners of the unit box; this in particular gives a clean interpretation of the distinction between asymptotic independence (AI) and asymptotic dependence (AD) classes, since this distinction essentially corresponds to whether or not the boundary touches the upper right corner. Third, our approach produces a posterior sample of valid limit set curves which yields a realistic picture of the state of knowledge about the joint tail region given the data. Bayesian hypothesis testing can then be used to study the properties of the limit set. In addition, we also note that our work builds on a growing literature on Bayesian approaches to multivariate extreme value analysis (e.g. Boldi and Davison, 2007; Sabourin et al., 2013; Sabourin and Naveau, 2014; de Carvalho et al., 2022; Padoan and Rizzelli, 2022, to name a few).

The rest of the paper is arranged as follows. Section 2 introduces the limiting scaled point cloud and how it can be used to model the tail behavior of bivariate data. Section 3 develops the modeling of the limit set boundary using Bézier splines. Section 4 contains a simulation study demonstrating our approach. Section 5 contains two applications—the Santa Ana Winds dataset, and ozone concentration data for the contiguous US—where we use Bézier splines to model the tail dependence in the data. Section 6 concludes.

2   The Limiting Scaled Point Cloud

Consider a collection of n independent random vectors in $\mathbb{R}^2_+$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, each having joint density $f_{\boldsymbol{X}}$, with standard exponential margins. At times, it will be convenient to transform the components of $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}}$ into pseudo-polar coordinates $(R, \boldsymbol{W})$, as $R = X_1 + X_2$, and $\boldsymbol{W} = \boldsymbol{X}/R$. Note that for $\boldsymbol{W} = (W_1, W_2), W_2 = 1 - W_1$.

Now define the scaled point cloud as the collection of points divided by $\log n$, $\{\boldsymbol{X}_1/\log n, \ldots, \boldsymbol{X}_n/\log n\}$. If we assume that $\lim_{t\to\infty} -\log f_{\boldsymbol{X}}(t\boldsymbol{x})/t = g(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^2_+$, for some continuous function g, then the scaled point cloud converges onto a compact limit set

$$G = \{\boldsymbol{x} \in \mathbb{R}^2 : g(\boldsymbol{x}) \leq 1\}$$

as $n \to \infty$ (Davis et al., 1988; Kinoshita and Resnick, 1991; Balkema and Nolde, 2010; Nolde, 2014; Nolde and Wadsworth, 2022). The function g is called the *gauge function* associated with the density $f_{\boldsymbol{X}}$. Denote the boundary of G as ∂G.

Every gauge function g is homogeneous of order one, with $g(c\boldsymbol{x}) = cg(\boldsymbol{x})$ for any $c > 0$ (Nolde, 2014). We will exploit this property by modeling the limit set boundary ∂G directly and using its associated gauge function, induced by homogeneity, for estimation (see Section 3.2). Any valid limit set G must satisfy the following constraints on its shape:

1. G is *star-shaped*, meaning that for any $t \in (0, 1)$, if $\boldsymbol{x}$ is in G, then $t\boldsymbol{x}$ is also in G.

2. The supremum of the boundary ∂G is 1 in each component direction. That is, ∂G touches, but does not cross, the upper and right-hand sides of the unit box.

4

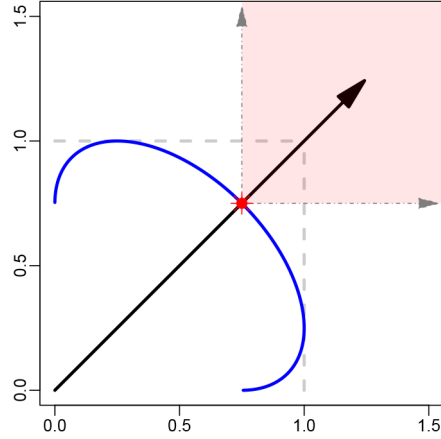
Figure 1: Schematic of η. The blue curve is the unit level set of gauge function $g(\boldsymbol{x})$, which forms the limit set boundary ∂G, and the proportional distance to the red point from the origin is the tail dependence coefficient η. While the red point is always on the diagonal, the intersection of the shaded red region and the blue curve does not necessarily occur on the diagonal.

We seek a flexible way of representing the boundary ∂G of the limit set G that satisfies conditions 1 and 2 and can be estimated from *iid* samples of the random vector $\boldsymbol{X}$. The shape of the limit set contains useful information about the extremal dependence of the distribution of the data. Nolde (2014) linked particular features of the shape of G with various indices of joint tail dependence in the literature. The residual tail dependence coefficient (Ledford and Tawn, 1996), the angular dependence function (Wadsworth and Tawn, 2013), components of the conditional extremes model (Heffernan and Tawn, 2004), and the index $\tau_1(\delta)$ (Simpson et al., 2020) all have direct connections to the shape of G. Our primary focus is on the residual tail dependence coefficient, $\eta \in (0, 1]$, which is defined by assuming that, for $\boldsymbol{X}$ in exponential margins, its survivor function satisfies

$$P(X_1 > x, X_2 > x) = \mathcal{L}(e^x)e^{-x/\eta}$$

as $x \to \infty$, for some function $\mathcal{L}$ that is slowly varying at infinity (Ledford and Tawn, 1996). Then the coefficient η describes the strength of dependence in the joint tail, with $\eta \in (1/2, 1)$ indicating positive tail dependence but AI, and $\eta = 1$ indicating AD, assuming $\mathcal{L}(x) \not\to 0$. The dependence class (AI vs. AD) is defined by the limiting conditional probability $\chi \in [0, 1]$, where

$$\chi := \lim_{x \to \infty} \frac{P(X_1 > x, X_2 > x)}{P(X_1 > x)},$$

with $\chi = 0$ characterizing AI, and $\chi > 0$ characterizing AD.
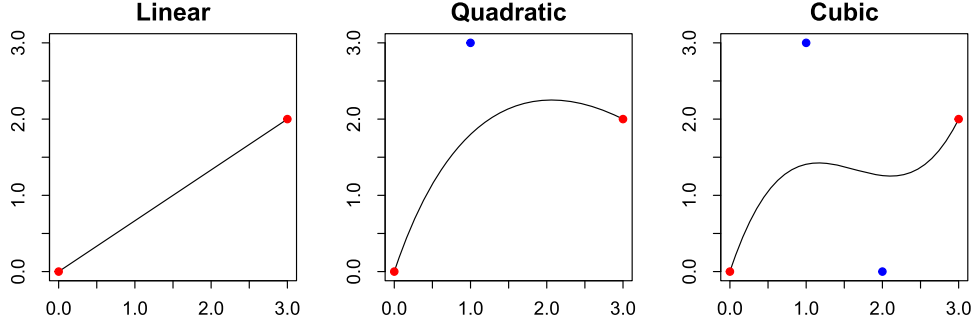
Figure 2: Examples of Bézier curves of orders 1, 2, and 3. The red control points (end points) $\mathbf{p}_0$ and $\mathbf{p}_m$ always lie on the curve, while the blue control points usually do not.

The residual tail dependence coefficient, η, can be calculated (Nolde, 2014; Nolde and Wadsworth, 2022) from shape of the limit set as

$$\eta = \min\{r : r \times [1, \infty]^2 \cap G = \emptyset\}.$$

This is illustrated schematically in Figure 1, where one can think of sliding the shaded box down the ray with slope 1 until it first touches the boundary ∂G. The radius corresponding to this first point of intersection is η. Assuming as above that $\mathcal{L}(x) \nrightarrow 0$, when $\boldsymbol{X}$ is asymptotically dependent, $\eta = 1$, so ∂G necessarily touches the upper right-hand corner of the unit box. Conversely, when $\boldsymbol{X}$ is asymptotically independent, $\eta < 1$, so ∂G does not touch the upper right-hand corner, and is referred to as *blunt*.

3 Modeling the Shape Using Bézier Splines

3.1 A Bézier Spline Representation of the Limit Set Boundary

Bézier curves (e.g. Hazewinkel, 2012; Farouki, 2012) are a class of parametric functions that can be used as building blocks to represent complex shapes. Bézier curves are defined by a set of *control points* $\mathbf{p}_0$ to $\mathbf{p}_m$, where m is the order of the curve. Figure 2 plots examples of Bézier curves of orders 1–3. The end points (red) define the beginning and end of the curve; intermediate control points (blue) of each curve control its shape but generally do not lie on the curve. A quadratic Bézier curve, for example, traces the path:

$$B(t) = (1 - t)[(1 - t)\mathbf{p}_0 + t\mathbf{p}_1] + t[(1 - t)\mathbf{p}_1 + t\mathbf{p}_2],$$

for $0 \le t \le 1$. Rearranging this equation simplifies it to:

$$B(t) = (1 - t)^2 \mathbf{p}_0 + 2t(1 - t)\mathbf{p}_1 + t^2 \mathbf{p}_2.$$

A useful property is that if the three points are co-linear, then a quadratic Bézier curve simplifies to a linear Bézier curve. Several Bézier curves can in turn be linked together at the end points to form a Bézier spline. The end points of each Bézier curve

within the spline now function as *knots* for the spline. Splines comprised of quadratic Bézier curves are particularly useful since analytical solutions for quadratic equations are straightforward to obtain, and therefore represent a flexible yet parsimonious way to approximate the shapes of limit set boundaries. In particular, increasing the order to cubic splines would make it difficult to constrain the shapes to the unit box, and would prevent the shapes from having the sharp corners required to represent asymptotically dependent limit set boundaries.

Because they are parsimoniously parameterized and straightforward to constrain, quadratic Bézier splines are convenient for modeling the boundary ∂G of the limit set G. We specify ∂G as a Bézier spline comprised of three quadratic Bézier curves $g_B = \{B_1(t), B_2(t), B_3(t)\}$, where $B_1(t) := B(t; \mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2)$, $B_2(t) := B(t; \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4)$, and $B_3(t) := B(t; \mathbf{p}_4, \mathbf{p}_5, \mathbf{p}_6)$, for $\mathbf{p}_i \in \mathbb{R}^2$, $i = 0, 1, \ldots, 6$. The three curves trace the paths:

$$B_1(t) = (1-t)^2 \mathbf{p}_0 + 2t(1-t)\mathbf{p}_1 + t^2 \mathbf{p}_2,$$
$$B_2(t) = (1-t)^2 \mathbf{p}_2 + 2t(1-t)\mathbf{p}_3 + t^2 \mathbf{p}_4,$$
$$B_3(t) = (1-t)^2 \mathbf{p}_4 + 2t(1-t)\mathbf{p}_5 + t^2 \mathbf{p}_6,$$

for $0 \le t \le 1$. We denote the point $\mathbf{p}_i := (p_{i,1}, p_{i,2})$, $0 \le p_{i,1}, p_{i,2} \le 1$, and place two sets of constraints on the curves in order to elicit valid gauge functions which satisfy conditions 1 and 2. The first set of constraints ensure that the Bézier spline touches all four edges of the unit square:

$$p_{0,1} = p_{6,2} = 0,$$
$$p_{2,2} = p_{4,1} = 1.$$

The second set of constraints are sufficient conditions to ensure that the star-shaped property holds for the spline:

$$p_{1,1} \le p_{2,1},$$
$$m(\mathbf{0}, \mathbf{p}_1) \ge m(\mathbf{0}, \mathbf{p}_2),$$
$$m(\mathbf{0}, \mathbf{p}_4) \ge m(\mathbf{0}, \mathbf{p}_5),$$
$$p_{4,2} \ge p_{5,2},$$
$$p_{3,1} = p_{3,2},$$
$$p_{3,1} \ge \max(p_{2,1}, p_{4,2})$$

where $m(\mathbf{p}, \mathbf{p}')$ denotes the slope of the line connecting the points $\mathbf{p}$ and $\mathbf{p}'$, and $\mathbf{0} = (0,0)$ is the origin. The final condition ensures that if $p_{2,1}$ or $p_{4,2}$ are 1, $p_{3,1}$ is also 1. Thus, we arrive at a model for the limit set boundary ∂G, indexed by the 9 univariate parameters $\boldsymbol{\theta}_g = (p_{0,2}, p_{1,1}, p_{1,2}, p_{2,1}, p_{3,1}, p_{4,2}, p_{5,1}, p_{5,2}, p_{6,1})^{\mathrm{T}}$. In practice, these constraints are enforced within a Metropolis sampler. If a candidate MCMC sample for one of the parameters does not satisfy the constraints, the sampler rejects it automatically.
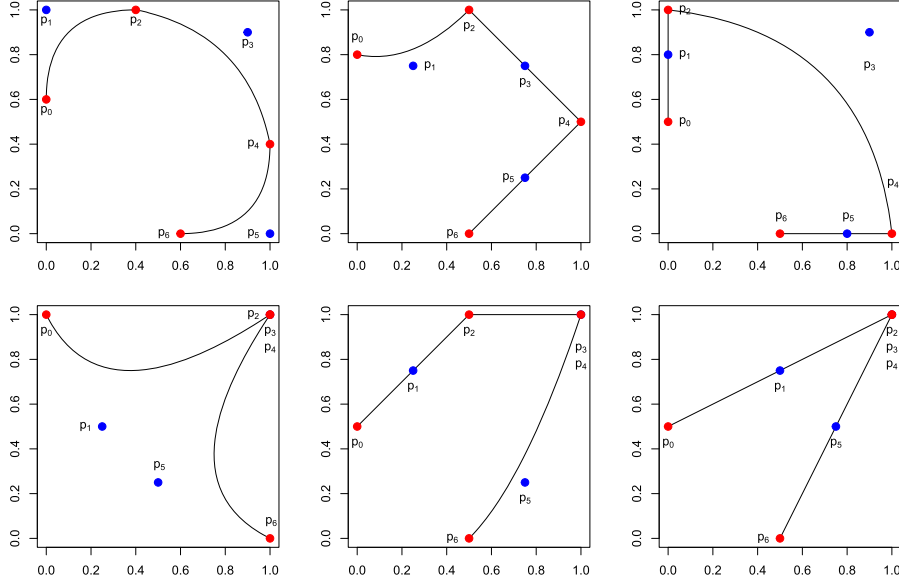
Figure 3: Examples of unit level sets of gauge functions that can be expressed using Bézier splines comprised of 3 quadratic Bézier curves. The red points are the end points of each curve, while the blue points are intermediate points controlling the shapes of the curves.

Figure 3 plots Bézier splines under these constraints, each representing a gauge function with different dependence properties. Top row plots correspond to AI scenarios, whereas plots in the bottom row correspond to AD scenarios. The four red control points are the knots of the spline. The three blue control points affect the shape, and the spline passes through them only if they are co-linear with the preceding and proceeding control points. In the general case, there are 9 coordinates, each admitting a uniform support, which need to be estimated to fully specify a valid gauge function. Richer models can be achieved using more control points; however, it would come with increased computational cost and additional constraints to ensure that conditions 1 and 2 hold. The quadratic Bézier spline with 4 knots therefore constitutes a parsimonious representation for ∂G which is still flexible enough to capture multiple dependence regimes and mimic most of the common parametric models.

3.2   Statistical Inference For the Limit Set Boundary

With a model defined for the limit set boundary ∂G, we turn to the question of how to estimate the shape from *iid* copies of the random vector $\boldsymbol{X}$ in standard exponential margins. After transforming $\boldsymbol{X}$ to pseudo-polar coordinates $(R, \boldsymbol{W})$, a convenient form (Wadsworth and Campbell, 2024) for the conditional density of a large radius R, given

the angle $\boldsymbol{W}$, is

$$f_{R\,|\,W}(r\,|\,\boldsymbol{w}) \propto r^{d-1}\exp\{-rg(\boldsymbol{w})[1+o(1)]\}, \quad r \to \infty,$$

where d is the dimension of $\boldsymbol{X}$ (we have only considered $d = 2$ here). For likelihood-based inference, Wadsworth and Campbell (2024) show that the $o(1)$ term can be moved outside the exponent in most cases and therefore ignored; they consequently consider the approximation adequate for radii larger than a threshold $r_0(\boldsymbol{w})$. This yields a truncated gamma likelihood:

$$R\,|\,\boldsymbol{W} = \boldsymbol{w}, R > r_0(\boldsymbol{w}), \boldsymbol{\theta}_g \sim \text{truncGamma}\left(\alpha, g_{\boldsymbol{\theta}_g}(\boldsymbol{w})\right). \tag{1}$$

For most common bivariate copulas, $\alpha = 2$. However, the quality of the approximation tends to vary (for further details, see Wadsworth and Campbell, 2024), and α is usually treated as a parameter to be estimated. Thus, given a gauge function $g_{\boldsymbol{\theta}_g}$ an approximate likelihood for the large radii given the angles is

$$L(\boldsymbol{\theta}_g, \alpha; (r_1, \boldsymbol{w}_1), \ldots, (r_{n_0}, \boldsymbol{w}_{n_0})) = \prod_{i=1}^{n_0} \frac{g_{\boldsymbol{\theta}_g}(\boldsymbol{w})^\alpha}{\Gamma(\alpha)} \frac{r_i^{\alpha-1}\exp\{-r_i g_{\boldsymbol{\theta}_g}(\boldsymbol{w}_i)\}}{1 - F(r_0(\boldsymbol{w}_i); \alpha, g_{\boldsymbol{\theta}_g}(\boldsymbol{w}_i))}, \tag{2}$$

where n_0 is the number of points exceeding the threshold $r_0(\boldsymbol{w})$, and $F(\,\cdot\,; \alpha, g_{\boldsymbol{\theta}_g}(\boldsymbol{w}_i))$ is the CDF of a gamma distribution with shape parameter α and rate parameter $g_{\boldsymbol{\theta}_g}(\boldsymbol{w}_i)$.

To calculate the gauge function at each data point, as required in the likelihood (2), we exploit the homogeneity property of g. This gives us that the value of the gauge function evaluated at a point $\boldsymbol{x}$ is the distance from the origin to $\boldsymbol{x}$, relative to the distance from the origin of the intersection of the ray connecting $\boldsymbol{x}$ with the origin and the boundary ∂G. In the schematic in Figure 4, the intersection with ∂G is denoted as $\boldsymbol{x}_{\partial G}$, so that

$$g_{\boldsymbol{\theta}_G}(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}_{\partial G}\|}. \tag{3}$$

We also need to select a threshold $r_0(\boldsymbol{w})$, as a function of angle. Wadsworth and Campbell (2024) and Simpson and Tawn (2022) both chose thresholds as functions of the angle, first as empirical quantiles of moving windows of angle and then using smooth semi parametric quantile regression. We employ a much simpler approach, and choose a high quantile in each threshold marginal component. We have found that this very basic strategy results in estimation performance at least comparably as good as more complicated alternatives. In addition, choosing marginal thresholds has two key advantages. First, it is simple to implement and requires no intricate tuning. Second, it permits, in principle, transformation to standard exponential margins within a hierarchical model, whereas thresholds that depend jointly on both components do not. With this in mind, we choose a value $\tau \in (0,1)$, and then set each marginal threshold at the τth marginal empirical quantile q_{τ,X_1} for X_1 and q_{τ,X_2} for X_2. In pseudo-polar coordinates, this gives a radial threshold of

$$r_0(\boldsymbol{w}) = \begin{cases} \frac{q_{\tau,X_2}}{1-w}, & w \in \left[0, \frac{q_{\tau,X_1}}{q_{\tau,X_1}+q_{\tau,X_2}}\right] \\[2ex] \frac{q_{\tau,X_1}}{w}, & w \in \left(\frac{q_{\tau,X_1}}{q_{\tau,X_1}+q_{\tau,X_2}}, 1\right]. \end{cases}$$
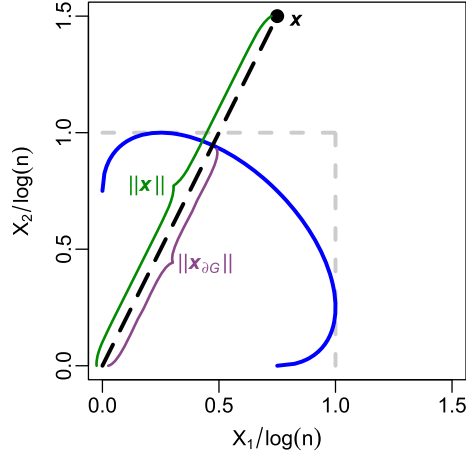
Figure 4: Schematic of how to calculate $g_{\boldsymbol{\theta}_G}$ at a data point ($\boldsymbol{x}$), given a boundary curve ∂G. The value of the gauge function is the distance from the origin to $\boldsymbol{x}$, relative to the distance from the origin of the intersection of the ray connecting $\boldsymbol{x}$ with the origin and the boundary ∂G.

While our simple strategy for threshold construction could potentially fail in pathological cases where it spuriously ignores important observations, we have not encountered such cases. A sensitivity study comparing the effect of the choice of threshold on estimation of η is provided in the Supplementary Material (Majumder et al., 2024a, Section A).

3.3 Prior Distributions For Control Points

Our model for the limit set boundary ∂G is indexed by 9 univariate parameters, viz. $\boldsymbol{\theta}_g = (p_{0,2}, p_{1,1}, p_{1,2}, p_{2,1}, p_{3,1}, p_{4,2}, p_{5,1}, p_{5,2}, p_{6,1})^{\mathrm{T}}$. To inform prior selection for these control points, we examine the boundary ∂G of four parametric copula models, and learn the conditions on the control points which allow the Bézier curve to mimic their shapes. The four copulas that we consider are the Gaussian, inverted logistic, logistic, and asymmetric logistic. The first two are asymptotically independent, while the final two are asymptotically dependent; analytical expressions of dependence measures for these models are provided in Table 1, replicated from Simpson and Tawn (2022). Since the limit set boundary ∂G can take on a variety of shapes, including the AD case where it touches the upper right-hand corner of the unit box (see e.g., Nolde and Wadsworth, 2022, Figure 2), we let the coordinates of the control points (i.e., the members of $\boldsymbol{\theta}_g$) vary in $[0, 1]$ (subject to the constraints presented in Section 3.1) for flexibility. This permits the possibility of them being exactly equal to 0 or 1 to accommodate AD as seen in the logistic and asymmetric logistic copulas and the very weak dependence as seen in the inverted logistic copula. We now outline the support for the distributions of the control points which will allow the Bézier splines to mimic our four copulas of

interest, and then specify priors for all parameters to be used in the remainder of this study.

The limit set boundary for an asymptotically dependent copula is obtained whenever $p_{2,1} = 1$ or $p_{4,1} = 1$; additionally, the logistic copula is implied by $\mathbf{p}_2 = \mathbf{p}_3 = \mathbf{p}_4 = (1,1)$ (Figure 3, bottom-right). This is equivalent to collapsing the second curve of the Bézier spline to a single point, and can be incorporated into our model by having a semi-continuous prior distribution for $p_{2,1}, p_{3,1},$ and $p_{4,2}$ with support over (0,1] which includes a point mass at 1. Similarly, collapsing the second curve and having a semi-continuous prior on $p_{0,2}$ and $p_{6,1}$ can incorporate an approximation of the limit set boundary for an asymmetric logistic copula. The theoretical shape of an asymmetric logistic copula is represented using a red line in Figure 5b (bottom right panel), while Bézier spline approximations can be seen in the form of the blue line in the same figure, as well as in Figure 3 (bottom left panel). In both cases, the Bézier splines approximate the sharp angles in the interior using smooth quadratic curves. Finally, approximating the limit set boundary for an inverted logistic copula using a Bézier spline would require the first and third curves to collapse onto the $x_1 = 0$ and $x_2 = 0$ lines respectively (Figure 3, top-right). To accommodate this case, we set semi-continuous priors with point masses at 0 for $p_{1,1}, p_{2,1}, p_{4,2},$ and $p_{5,2}$. To flexibly accommodate the wide range of limit set boundary shapes possible within our framework, we set priors as follows:

$$\alpha \sim \text{LogNormal}(1,1),$$

$$p_{0,2}, p_{1,2}, p_{5,1}, p_{6,1} \overset{\text{iid}}{\sim} \text{Uniform}(0,1).$$

The LogNormal prior on α has its density concentrated near $\alpha = 2$. The remaining control points have priors that are the mixture of a standard Uniform distribution and at least one point mass (to allow important geometric features of ∂G with positive probability). They have the following forms:

$$p_{1,1} \overset{\text{iid}}{\sim} 0.1 \cdot \mathbb{I}(p_{1,1} = 0) + 0.9 \cdot \text{Uniform}(0,1),$$

$$p_{2,1} \overset{\text{iid}}{\sim} 0.1 \cdot \mathbb{I}(p_{2,1} = 0) + 0.8 \cdot \text{Uniform}(0,1) + 0.1 \cdot \mathbb{I}(p_{2,1} = 1),$$

$$p_{3,1} \,|\, p_{2,1}, \, p_{4,2} \sim 0.6 \cdot \text{Uniform}(0,1) + 0.4 \cdot \mathbb{I}(\max(p_{2,1}, p_{4,2}) = 1).$$

The points $p_{5,2}$ and $p_{4,2}$ are distributed identically to $p_{1,1}$ and $p_{2,1}$, respectively. The point mass probabilities were chosen on the basis of a sensitivity study whose aim was to have good discrimination between the asymptotically dependent and asymptotically independent cases based on posterior probabilities of $\eta = 1$, while also simultaneously being able to provide unbiased, consistent estimates of η when $\eta < 1$. In particular, the prior on $p_{3,1}$ ensures that it is 1 only if AD is implied by either $p_{2,1}$ or $p_{4,2}$. These prior assumptions can accommodate the logistic, inverted logistic, and asymmetric logistic copulas, as well as intermediate forms such as the Gaussian copula which do not require point masses.

3.4 Additional Bivariate Extremal Dependence Measures

Alongside the tail dependence coefficient η (Ledford and Tawn, 1996) discussed in Section 2, we consider two additional indices of tail dependence which can be derived from

the gauge function. The first of these is the angular dependence function $\lambda(\omega)$ which considers different scalings for the two components of $\boldsymbol{X}$. In particular, Wadsworth and Tawn (2013) considered asymptotic probabilities of the following form:

$$P(X_1 > \omega x, X_2 > (1 - \omega)x) = \mathcal{L}_\omega(e^x)e^{-x\lambda(\omega)},$$

for some function $\mathcal{L}_\omega$ that is slowly varying at infinity, $\omega \in [0, 1]$, and $\lambda(\omega) \in (0, 1]$ when X_1 and X_2 have positive dependence. The function $\lambda(\omega)$ therefore captures both extremal dependence regimes, with AD implying the pointwise lower bound of $\lambda(\omega)$. Evaluation of $\lambda(\omega)$ for rays ω near 0 and 1 corresponds to regions where one variable is larger than the other. In particular, $\lambda(\omega)$ is a generalization of η, with $\eta = 1/\{2\lambda(1/2)\}$. Murphy-Barltrop et al. (2024) found that global estimators (such as the Simpson-Tawn estimator) which simultaneously estimate $\lambda(\omega)$ for all values of ω tend to provide better estimates compared to pointwise estimators (such as the Hill estimator). The Bézier spline estimator is also a global estimator of $\lambda(\omega)$, and examining the estimation of $\lambda(\omega)$ is therefore a useful measure to compare it with the Simpson-Tawn estimator.

We also investigate the dependence measure $\tau_1(\delta)$ (Simpson et al., 2020), given by:

$$P(X_1 > x, X_2 \le \delta x) = \mathcal{L}_\delta(e^x)e^{-x/\tau_1(\delta)},$$

for some function $\mathcal{L}_\delta$ that is slowly varying at infinity, with $\delta \in [0, 1]$. $\tau_1(\delta)$ is monotonically increasing in δ, with $\tau_1(1) = 1$. This dependence measure characterizes the probability of X_1 being large while X_2 is of a smaller order. Specifically, if there exists a $\delta^* < 1$ such that $\tau_1(\delta^*) = 1$, it implies that X_1 can be large while X_2 is small (with δ determining just *how* small). If no such δ^* exists, then X_1 can be large only if X_2 is also large. We can define $\tau_2(\delta)$ analogously, and both $\tau_1(\delta), \tau_2(\delta) \in (0, 1]$.

Table 1 provides the analytical expressions for these measures for the four copulas that we consider in our study. Like with η, these dependence measures can be exactly deduced from limit set boundaries, and hence from gauge functions. Since the Bézier splines are quadratic polynomials, we can easily calculate $\tau_1(\delta)$ and $\lambda(\omega)$ for any estimated limit set boundary simply by finding the intersections of polynomials and lines, which have closed-form solutions. While we do not present results for $\tau_2(\delta)$ in our study, it can be calculated in the same manner as $\tau_1(\delta)$. We refer the reader to Simpson and Tawn (2022) for a detailed discussion on how each of these measures can be obtained from gauge functions in more general settings.

4 Simulation Study

4.1 Study Setup

We demonstrate the appropriateness of using Bézier splines to model the limit set boundary corresponding to the gauge function by means of a simulation study comparing its performance with the Simpson-Tawn estimator (Simpson and Tawn, 2022).

We consider four bivariate copulas to generate data with exponential marginal distributions: the Gaussian, the logistic, the inverted logistic, and the asymmetric logistic.

Copula	$g(\boldsymbol{x}) = g(x_1, x_2)$
Gaussian	$\{x_1 + x_2 - 2\rho(x_1 x_2)^{1/2}\}/(1 - \rho^2)$
Logistic	$\gamma^{-1} \max(x_1, x_2) + (1 - \gamma^{-1}) \min(x_1, x_2)$
Inv-Logistic	$(x_1^{1/\gamma} + x_2^{1/\gamma})^{\gamma}$
Asy-Logistic	$\min\big(x_1 + x_2, \gamma^{-1} \max(x_1, x_2) + (1 - \gamma^{-1}) \min(x_1, x_2)\big)$

(a) Gauge function g for the four bivariate copulas.

Copula	η	$\lambda(\omega)$	$\tau_1(\delta) = \tau_2(\delta)$
Gaussian	$(1 + \rho)/2$	$\begin{cases} \max(\omega, 1 - \omega), & \text{if } t_\omega \leq \rho^2 \\ \frac{1 - 2\rho\sqrt{\omega(1-\omega)}}{1 - \rho^2}, & \text{if } t_\omega \geq \rho^2 \end{cases}$	$\begin{cases} 1, & \text{if } \delta \geq \rho^2 \\ \frac{1 - \rho^2}{1 + \delta - 2\rho\sqrt{\delta}}, & \text{if } \delta \leq \rho^2 \end{cases}$
Logistic	1	$\max(\omega, 1 - \omega)$	$(\gamma^{-1} + 1 - \gamma^{-1}\delta)^{-1}$
Inv-Logistic	$2^{-\gamma}$	$\{\omega^{1/\gamma} + (1 - \omega)^{1/\gamma}\}^{\gamma}$	1
Asy-Logistic	1	$\max(\omega, 1 - \omega)$	1

(b) Dependence measures for the four bivariate copulas. Here, $t_\omega = \min(\omega, 1-\omega)/\max(\omega, 1-\omega)$.

Table 1: Gauge function g, and a summary of dependence measures for the four bivariate copulas used in our study. Table has been reproduced from Simpson and Tawn (2022).

The Gaussian copula is parameterized by its correlation $\rho \in [0, 1)$, while the dependence parameter for the remaining three copulas is $\gamma \in (0, 1)$. Table 1 lists the gauge functions associated with each copula, as well as the set of corresponding extremal dependence coefficients $\{\eta, \lambda(\omega), \tau_1(\delta)\}$. Further details for these four copulas can be found in Simpson and Tawn (2022). For each copula, we consider five parameter settings: $\rho, \gamma = \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Smaller values of γ correspond to stronger tail dependence, whereas larger values of ρ lead to stronger tail dependence.

For each copula and parameter combination, we generate 100 datasets of $n = 5,000$ data points each. The data are converted to pseudo-polar coordinates $(R, \boldsymbol{W})$, and the n_0 points that are above the $\tau = 0.75$ quantile marginal threshold for at least one variable are used to model the gauge function for each dataset. The n_0 radii are assumed to approximately distributed according to a truncated gamma distribution with a common shape parameter α and a rate parameter equal to an appropriate gauge function evaluated at the data point. We use Metropolis updates for all parameters and run 11,000 MCMC iterations for each dataset, discarding the first 1,000 as burn-in. All Metropolis updates are tuned to give an acceptance probability of 0.4, and posterior convergence is diagnosed based on the visual inspection of trace plots. We compare our estimated limit set boundaries with the Simpson-Tawn estimator in terms of how well they estimate the set of dependence coefficients outlined in Table 1. The Simpson-Tawn

estimator for the datasets was evaluated using the default settings recommended by the authors in Simpson and Tawn (2022). We use the root mean square error (RMSE) to compare estimates of the scalar η, and the root mean integrated square error (RMISE) to compare estimates of the functions $\lambda(\omega)$ and $\tau_1(\delta)$. The methodology is implemented in R (R Core Team, 2023); the code is available on GitHub through the `BezELS` package (Majumder et al., 2024b).

4.2   Parameter Estimates

Each panel in Figure 5 plots the limit set boundaries elicited by the estimated Bézier splines based on the posterior distribution from a single dataset with the dependence parameter set to 0.5. Plots in Figure 5a display the dependence modeling in pseudo-polar coordinates. The dashed grey line corresponds to the threshold $r_0(\boldsymbol{w})$ for angles $\boldsymbol{w}$. Each estimated limit set boundary is represented as a functional boxplot (Hyndman and Shang, 2010; Sun and Genton, 2011), a visual representation of functional data analogous to a classical boxplot. Each functional quantile depicted in the figure is a function contained in the sample; in this case, the sample consists of limit set boundaries based on Bézier splines, in pseudo-polar coordinates, drawn from the posterior. The curves are ordered according to a notion of modified band depth (López-Pintado and Romo, 2009). The median is shown in dark blue, and the limit set boundary corresponding to the data-generating model is shown in red. The envelope represents the 50% central region, analogous to the box of a classical boxplot. The outer light blue lines of the functional boxplot correspond to the whiskers of a classical boxplot. Finally, the vertical lines indicate the maximum envelope of the functions except outliers. Plots in Figure 5b display the dependence models in Euclidean coordinates, with the median Bézier spline in dark blue and the limit set boundary corresponding to the data-generating model in red. They are overlaid on Bézier splines evaluated from 500 draws from the posterior distribution, plotted in gray. The boxplots on the top and right margins correspond to the posterior distributions of $p_{2,1}$ and $p_{4,2}$ respectively, which serve as a visual indicator of asymptotic dependence in the data. Specifically, if the median of either boxplot is 1, the posterior median of η is 1. The Bézier splines are able to adequately represent the geometric form of all four copulas.

Figure 6 shows boxplots of the posterior median of η for the four copulas based on the Bézier spline (blue) and Simpson-Tawn estimators (green). Analytical values of η obtained using expressions in Table 1 are shown as red dots in each plot, and the coverage of equi-tailed 95% intervals are noted in plain text below each boxplot. Coverage for the Simpson-Tawn estimator is based on 100 bootstrapped samples from each of the 100 datasets. We plot the median instead of the mean for the Bézier spline estimates since the posterior distributions are often highly asymmetric due to point-mass prior distributions. The Bézier spline estimator has low bias and nominal coverage for estimates of η. The Simpson-Tawn estimator has nominal or near-nominal coverage in most cases except for the asymmetric logistic copula when the dependence parameter γ is 0.5 or higher. It also has noticeably higher bias than the Bézier spline estimator for both asymptotically dependent copulas, and shows a sharp decline in coverage as the strength of dependence drops for the asymmetric logistic copula. We evaluate $\lambda(\omega)$ and $\tau_1(\delta)$
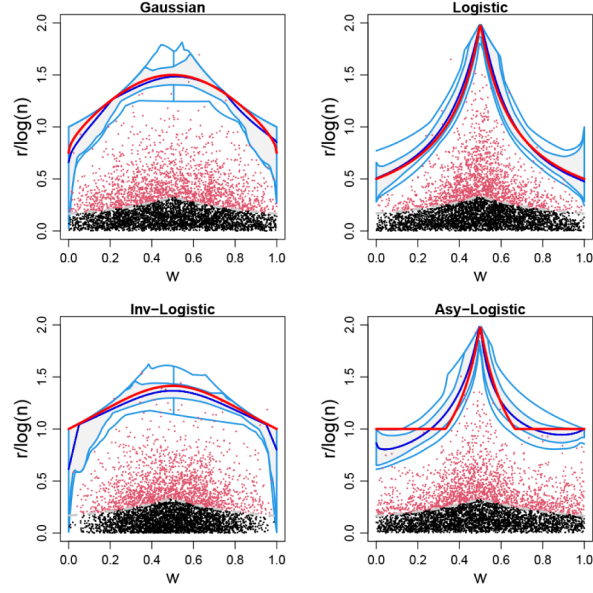
(a) Estimated gauge functions in pseudo-polar space.
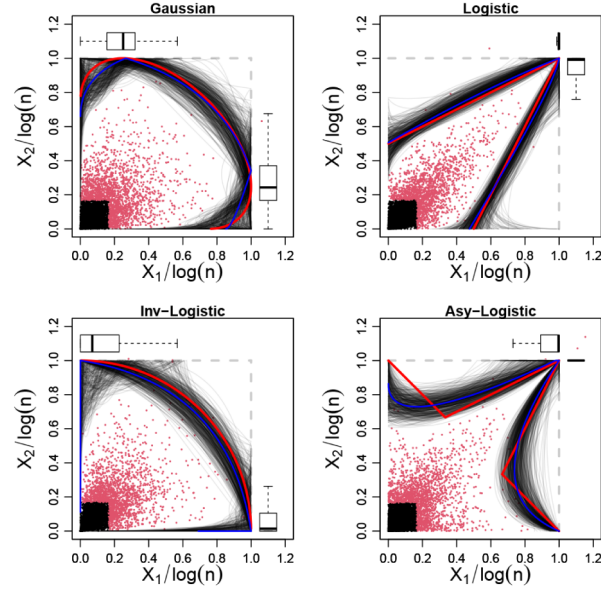
(b) Estimated gauge functions in Euclidean space.

Figure 5: Limit set boundaries based on Bézier splines (blue) and corresponding functions for the data-generating model (red) in pseudo-polar and Euclidean space for the Gaussian, logistic, inverted logistic, and asymmetric logistic copulas with dependence parameters set to 0.5.
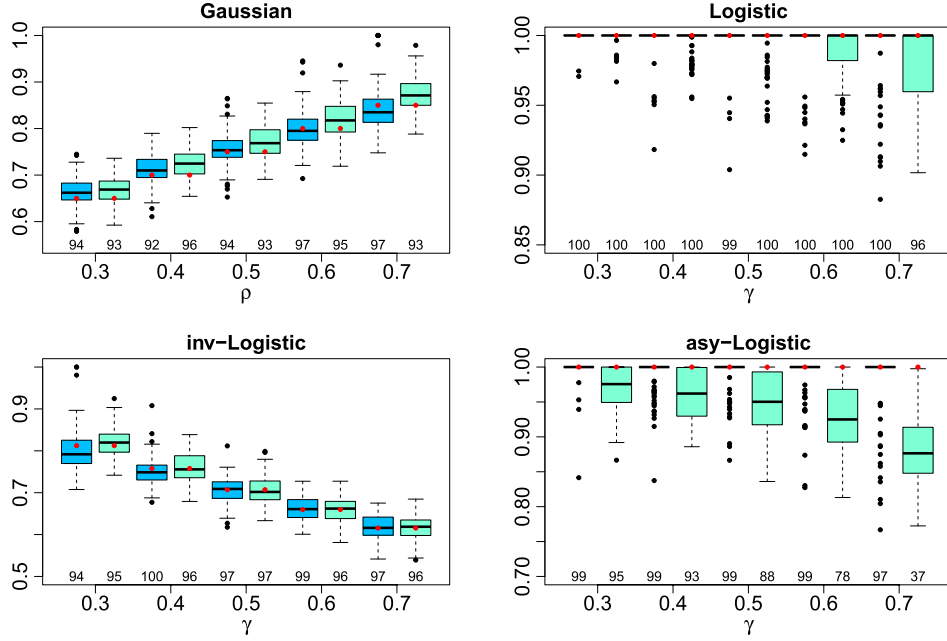
Figure 6: Sampling distribution of the posterior medians of η based on the Bézier spline estimate (blue) for the four copulas, alongside estimates using the Simpson-Tawn estimator (green). The red dots indicate the true values, and coverage of equi-tailed 95% intervals are noted below each boxplot.

based on the Bézier spline and Simpson-Tawn estimators for $\omega, \delta = 0.01, 0.02, \ldots, 0.99$. Figure 7a shows boxplots for the posterior medians of $\lambda(0.40)$ based on the two estimators, with the corresponding dependence parameter set to 0.5. Figure 7b similarly plots the distribution of $\tau_1(0.25)$ for the two estimators. Both estimators are better at estimating $\lambda(\omega)$ than $\tau_1(\delta)$, and the Bézier spline estimator tends to have better coverage than the Simpson-Tawn estimator.

Table 2 summarizes the RMSE ratio for estimates of η and RMISE ratios for estimates of $\lambda(\omega)$ and $\tau_1(\delta)$ based on the Bézier spline and Simpson-Tawn estimators. Most of the values are greater than 1, indicating that dependence measures based on Bézier spline estimates of the gauge function have comparable or better RMSE/RMISE than those based on the Simpson-Tawn estimator. However, for the inverted logistic copula, the Simpson-Tawn estimator outperforms the Bézier spline estimator when estimating of $\tau_1(\delta)$. Our experiments indicate that the error arises when the posterior $p_{4,2} > \delta$; this leads to $\tau_1(\delta)$ estimates to be less than 1, whereas the theoretical value for the inverted logistic copula is 1 for all δ. Our approach, however, is still able to correctly estimate $\tau(\delta) = 1$ in all the inverted logistic scenarios for all but extremely small values of δ.

Table 3 provides the number of datasets (out of 100) in each scenario where the posterior median of η is estimated to be 1. The values in the parentheses are corre-
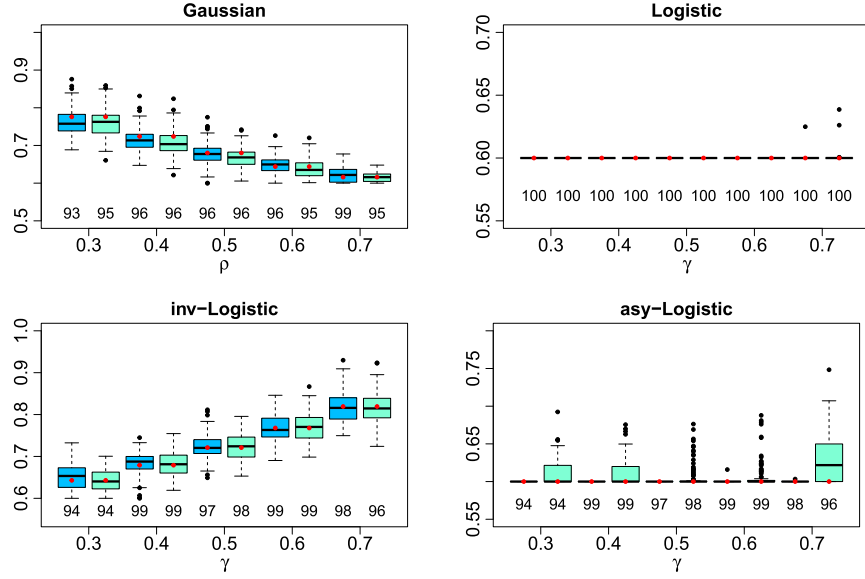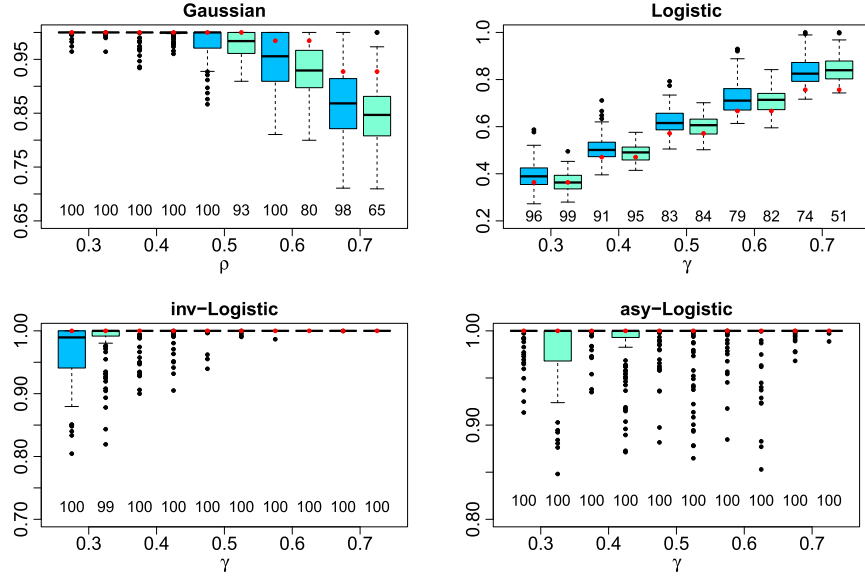
(a) Sampling distribution of the posterior medians of $\lambda(0.40)$.

(b) Sampling distribution of the posterior medians of $\tau_1(0.25)$.

Figure 7: Sampling distribution of the posterior medians of $\lambda(0.40)$ and $\tau_1(0.25)$ based on the Bézier spline (blue) and Simpson-Tawn (green) estimators for four copulas, with dependence parameters set to 0.5. The red lines indicate the true values, and coverage of equi-tailed 95% intervals are noted below each boxplot.

Measure	Copula	Dependence parameter value				
		0.3	0.4	0.5	0.6	0.7
$\frac{RMSE(\tilde{\eta})}{RMSE(\hat{\eta})}$	Gaussian	1.02	1.11	1.04	1.06	0.81
	Logistic	1.38	0.81	1.36	1.17	1.07
	Inv-Logistic	0.70	0.99	1.11	1.03	1.01
	Asy-Logistic	2.39	2.01	2.23	2.37	2.39
$\frac{RMISE(\tilde{\lambda})}{RMISE(\hat{\lambda})}$	Gaussian	1.19	1.18	1.05	0.94	0.75
	Logistic	1.35	0.61	1.10	1.04	1.28
	Inv-Logistic	0.82	1.10	1.20	1.10	1.04
	Asy-Logistic	6.17	6.65	10.91	7.43	5.67
$\frac{RMISE(\tilde{\tau})}{RMISE(\hat{\tau})}$	Gaussian	0.85	1.00	1.08	1.17	1.18
	Logistic	1.00	0.95	0.89	0.83	1.00
	Inv-Logistic	0.23	0.27	0.17	0.28	0.17
	Asy-Logistic	2.38	2.96	1.77	1.72	1.23

Table 2: RMSE ratios for estimates of η and RMISE ratios for estimates of $\lambda(\omega)$ and $\tau_1(\delta)$ based on the Bézier spline ($\hat{\eta}, \hat{\lambda},$ and $\hat{\tau}$) and Simpson-Tawn ($\tilde{\eta}, \tilde{\lambda},$ and $\tilde{\tau}$) estimators over simulated datasets for four copulas and five dependence levels.

	Dependence parameter value				
	0.3	0.4	0.5	0.6	0.7
Gaussian	00 (00)	00 (00)	00 (00)	00 (00)	06 (00)
Logistic	98 (91)	93 (79)	96 (77)	91 (74)	82 (63)
Inv-Logistic	02 (00)	00 (00)	00 (00)	00 (00)	00 (00)
Asy-Logistic	96 (30)	80 (25)	82 (22)	83 (14)	84 (00)

Table 3: Number of datasets (out of 100) where the posterior median of η is 1 for each scenario. Values in parenthesis correspond to the Simpson-Tawn estimator.

sponding point estimates of η based on the Simpson-Tawn estimator. For the Bézier spline estimators, the values were always estimated to be near 0 for the asymptotically independent copulas, and always high for the asymptotically dependent copulas. While there were some cases where the Bézier spline estimates the posterior of η as 1 when the dependence is high in an asymptotically independent copula, both methods are good at estimating AI correctly. On the other hand, the Simpson-Tawn estimates show a decline in their ability to estimate the correct value of $\eta = 1$ for the asymptotically dependent copulas when dependence is low. This is especially noticeable for the asymmetric logistic copula, and has been documented by Simpson and Tawn (2022) as well. The Bézier spline estimator is much better at predicting AD correctly across all scenarios. We conclude that the Bézier splines are adept at representing limit set boundaries associated with common parametric copula models, and are also flexible enough to represent a wider variety of edge cases. In all cases, the true value of η was well-estimated from the posterior distribution, and the model is particularly adept at identifying AD.

4.3   Additional Simulation Studies

Two additional simulation studies are presented in the Supplementary Material (Majumder et al., 2024a). In both cases, the results are compared based on RMSE/RMISE values of bivariate extremal dependence coefficients, as well as how often it estimates $\eta = 1$. The first study considers how our relatively straightforward method of selecting the quantile threshold affects the estimation process. This is carried out by comparing it against an 'oracle' threshold which is an asymptotic approximation to the true conditional quantile $q_\tau(w)$ and requires knowledge of the true gauge function to compute. We are unable to find any meaningful improvement in the estimates when we consider an oracle threshold, which indicates that our choice of threshold is adequate for the scenarios we have considered. The second study repeats the simulation study presented in this section, but for a sample size $n = 600$. This is carried out to ensure that our approach is still valid for small data sizes like the one that arises in our second application, presented in the following section. Our results indicate that despite slightly higher RMISE values and slightly lower coverage, the Bézier splines can still capture the shape of the true limit set boundary with low bias. Comparisons with the Simpson-Tawn estimator provide results that are quite similar to the ones presented in this section.

5   Applications

5.1   Analysis of the Santa Ana Winds Data

We apply our method to the Santa Ana winds and dryness data (Cooley et al., 2019). The Santa Ana winds are a multivariate meteorological regime that has been implicated as a major driver of large wildfires in southern California (Billmire et al., 2014). Wildfires are related to several conditions like temperature, humidity, wind speed, and fuel supply (Littell et al., 2018). Historically, the autumn months of September, October, and November have had a higher number of wildfires compared to the winter months, and are associated with warm temperatures, low humidity, and high winds. Cooley et al. (2019) surmised that the data exhibits AD and used the framework of regular variation to estimate probabilities associated with two different risk regions. The regular variation structure employed by them, however, cannot capture the nuance of AI. Our analysis could produce more accurate estimates of the probabilities of joint tail events that require extrapolation beyond the range of the data. In particular, if the data were asymptotically independent, an assumption of AD would overestimate the probability of such jointly extreme events under the approach used by Cooley et al. (2019).

We consider daily dryness (%) and wind speed (m/s) data collected at the March Air Reserve Base station in Riverside County from the HadISD dataset (Dunn et al., 2012). The dryness is defined in this case as the $100-$ RH, where RH is the relative humidity measured as a percentage. The bivariate time series represents a measure of the daily risk of fire. The station's variables have appeared to be associated with known Santa Ana events. The data consists of 3,902 days for the months of September–November from 1972–2015; we assume temporal stationarity and independence as in Cooley et al. (2019). Figure 8 plots the data both in its original scale as well as the rank transformed
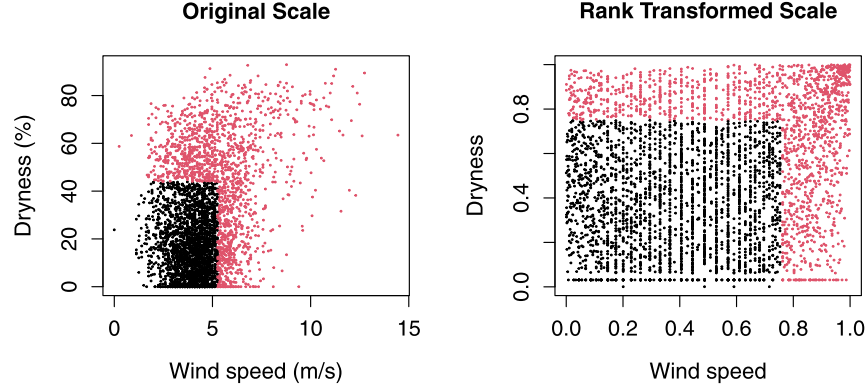
Figure 8: Santa Ana wind speeds and dryness measured at the March Air Force Base station. Data above the 0.75 marginal quantile threshold are in red.

scale. The data shows tail dependence, noticeable in the rank transformed data with a cluster of values in the upper right corner. Our goal is to study the tail dependence between the two variables by estimating a gauge function for the data after a further transformation to unit exponential margins.

We analyze this data at a threshold of $\tau = 0.75$, providing us with 1,529 points that are above the threshold in at least one margin. The conditional distribution of the radii are assumed to be a truncated gamma distribution for this analysis. We run 2 MCMC chains of 11,000 iterations each, discarding the first 1,000 from each chain as burn-in. The priors and the remainder of the MCMC settings are identical to the simulation study.

Figure 9 plots the estimated limit set boundaries, with the median curve plotted in blue. The plot on the left is in pseudo-polar coordinates and depicts a functional boxplot of the estimated limit set boundary, while the plot on the right is in Euclidean coordinates. The posterior median of η is estimated to be 1, and $P(\eta = 1 \mid \boldsymbol{X}) = 0.60$, suggesting AD between wind speed and dryness. Figure 10 evaluates the goodness-of-fit for the truncated gamma model in terms of PP plots as well as QQ plots, based on post burn-in MCMC samples α^* and $\boldsymbol{\theta}_g^*$ from the posterior distribution of the parameters. The PP plot for n_0 observations with $R_i > r_0(\boldsymbol{w}_i)$ is the set of points $[i/(n_0 + 1), u_{(n_0-i+1)}]$, where $u_{(1)} \geq u_{(2)} \geq \ldots \geq u_{(n_0)}$ is the ordered sample of $u_i = \hat{F}_t\big(r_i; \boldsymbol{w}_i, r_0(\boldsymbol{w}_i)\big)$, and $\hat{F}_t$ is the fitted CDF of the truncated gamma model with the likelihood function as in (2):

$$\hat{F}_t\big(r|\boldsymbol{w}, r_0(\boldsymbol{w})\big) := \Pr\big(R \leq r|W = \boldsymbol{w}, R > r_0(\boldsymbol{w})\big) = 1 - \frac{1 - F(r; \alpha^*, g_{\boldsymbol{\theta}_g^*}(\boldsymbol{w}))}{1 - F(r_0(\boldsymbol{w}); \alpha^*, g_{\boldsymbol{\theta}_g^*}(\boldsymbol{w}))}.$$

For the corresponding QQ plot, each point on the PP plot is transformed to unit exponential margins through an inverse CDF transformation. In both cases, the black points are obtained from gamma parameters implied by the median curve of the limit set boundary, while the gray lines correspond to $n_0 = 100$ random draws from the posterior. Both plots suggest that the model fits the data adequately.
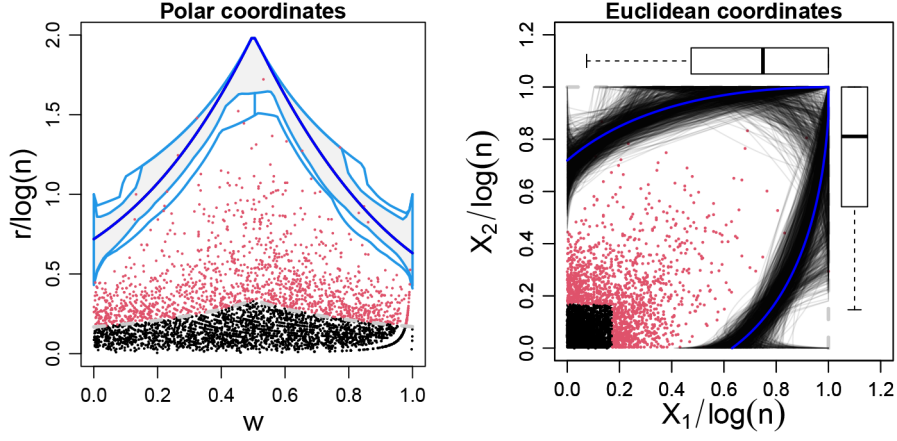
Figure 9: Limit set boundaries based on Bézier splines in pseudo-polar space (left) and Euclidean space (right) for assessing tail dependence between Santa Ana wind speed (X_1) and dryness (X_2). Median curves are plotted in dark blue.
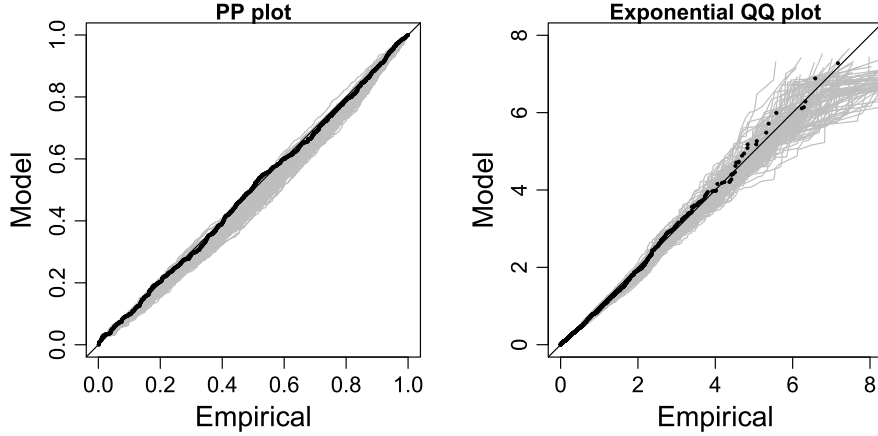

Figure 10: PP plot (left) and exponential QQ plot (right) for the truncated gamma model fitted to the Santa Ana winds and dryness data. The black points correspond to the fit for the median Bézier spline. Gray lines are based on 100 random draws from the posterior.

To test sensitivity to the threshold level, we repeated the experiment at additional threshold levels of $\tau = 0.70, 0.80,$ and 0.90. In all 3 additional cases, $P(\eta = 1 \mid \boldsymbol{X}) > 0.50$, with the smallest value (of 0.56) occurring for $\tau = 0.90$ and the largest value (of 0.64) for $\tau = 0.70$. Taken together, we conclude that dryness and wind speed for Santa Ana are asymptotically dependent with high probability, consistent with the results of Cooley et al. (2019).
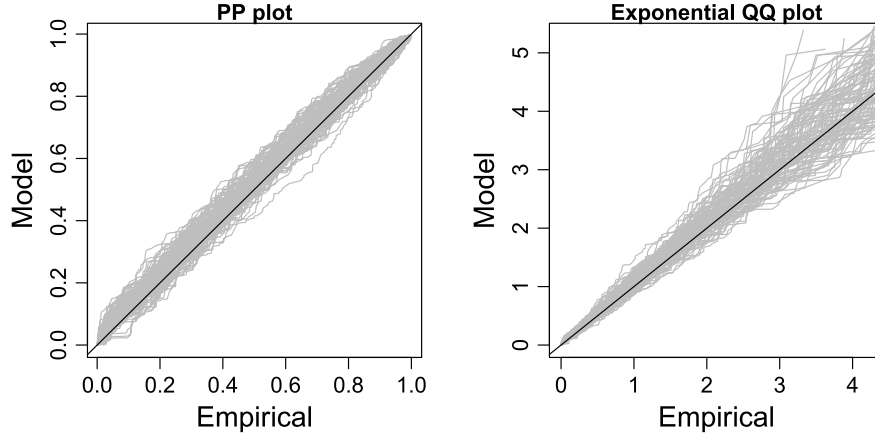
Figure 11: PP plot (left) and exponential QQ plot (right) for the truncated gamma model fitted to the Ozone concentration data. The gray lines correspond to the fit for the median Bézier splines at 100 random locations.

5.2 Analysis of Ozone Concentration Data

For our second study, we consider air pollution measurement data across the US from the Community Multiscale Air Quality (CMAQ) model (Binkowski and Roselle, 2003; Wyat Appel et al., 2007, 2008) as well as EPA Air Quality System (AQS) (US EPA, 2017) data for the contiguous US. While CMAQ is a numerical model available across the entire country at a 12km resolution, the AQS dataset consists of observations monitored at 1,376 stations across the US. Among them, only 519 stations had over 600 observations, which is what we use for this analysis. The full dataset has previously been used by Gong et al. (2021) to develop a combined data product for 12 air pollutants. When fusing data products, it is important to calibrate the model data to ground truth. In our application, we will verify how strong the dependence is between the AQS and CMAQ datasets for ozone, one of the 12 pollutants made available by both datasets.

Our data consists of daily ozone readings for the months of July–September from 2010–2014, resulting in a bivariate time series of CMAQ and AQS data for up to 610 days at each station. The sample correlations between the AQS and CMAQ data for the 519 stations range from 0.29–0.86 with a median of 0.69, suggesting a high level of agreement in the bulk of the distribution. To assess tail dependence, we fit a gauge function with truncated gamma likelihood for data beyond the $\tau = 0.75$ threshold, independently at every station. We run 2 MCMC chains for 11,000 iterations each for each station's data, discarding the first 1,000 as burn-in.

The posterior median of η has an average value of 0.81 across the 519 locations, and is 1 (asymptotically dependent) for 79 of those stations. This suggests that the CMAQ data product can adequately represent the tail behavior of observational ambient ozone. Figure 12 plots the posterior probability of AD based on the truncated gamma model. While we are unable to discern any spatial pattern for high or low posterior values of
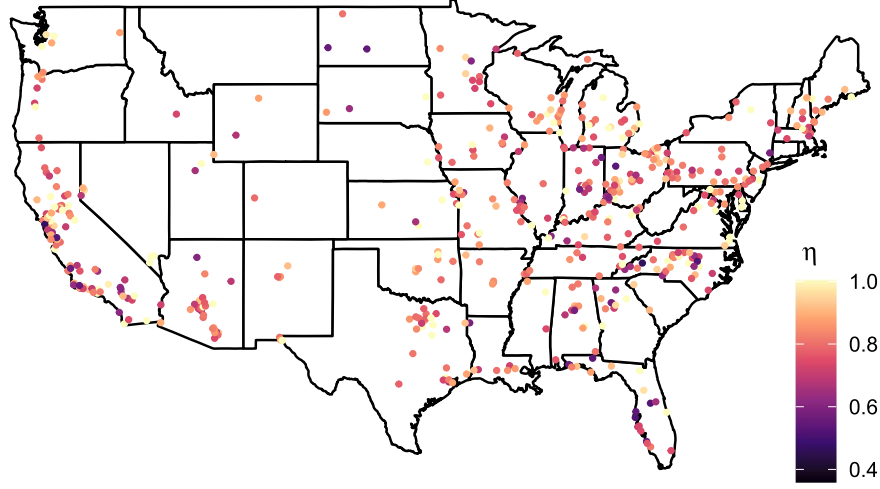
Figure 12: Posterior median of η at 519 AQS monitoring stations in the US.

η from the map, we do note that several of the low values are in urban areas with high population densities.

Finally, to study the sensitivity of our posterior to the threshold level τ, we repeated our analysis at two additional values of $\tau = 0.70$ and 0.80. Estimates from both these cases were quite similar to the baseline case of $\tau = 0.75$, with correlations > 0.90 for both the posterior median of η and $\mathrm{P}(\eta = 1)$. There were 76 and 80 locations respectively where the posterior median of η was 1, and 62 of those locations were shared with the baseline case. Thus, our results are not very sensitive to the choice of threshold for small data sizes.

6   Discussion

Key aspects of tail dependence in multivariate distributions can be described through their corresponding gauge functions. In this study, we propose a semiparametric method for estimating gauge functions by modeling their unit level sets as Bézier splines comprised of three quadratic Bézier curves. The splines can represent the gauge function unit level sets, and hence limit set boundaries, of varying shapes, and are parsimoniously parameterized by a small number of control points. The quadratic specification makes it straightforward to obtain analytical solutions for the shape of the limit set, and constraints on the control points ensure that the resultant shapes are valid limit set boundaries. Bayesian estimation of the Bézier splines requires only standard MCMC techniques and allows important cases on the edge of the parameter space to be represented by employing mixture priors with point masses. We demonstrate the efficacy of our model using numerical studies as well as two real data applications involving fire weather in California, and ambient air pollution from ozone across the US.

We have only considered bivariate random vectors here, but the modeling strategy can scale to three dimensions by using Bézier surfaces, with the control points of constituent Bézier curves set in $\mathbb{R}^3$ instead of $\mathbb{R}^2$. It will, however, require more complicated constraints to ensure that the star shaped property holds, and dimensions greater than three appear to be infeasible. In addition, it appears possible to extend our modeling framework to include negative dependence by transforming to Laplace margins rather than exponential margins. This has been previously suggested (Simpson and Tawn, 2022; Wadsworth and Campbell, 2024) and recently implemented in the context of radially stable Pareto distributions (Papastathopoulos et al., 2023), and within a semiparametric angular-radial model (Mackay and Jonathan, 2023). Implementing our Bézier model in Laplace margins (in two dimensions) would require somewhere between two to four times the number of control points as we have now, specifying appropriate constraints on their support, and more sophisticated sampling algorithms to ensure convergence. Finally, while three Bézier curves are sufficient to ensure that the boundary of our estimated limit set in the two dimensional case exactly touches the corners of the unit box, the curves themselves don't necessarily need to be quadratic. It is possible to use a nonparametric Bayesian framework to construct curves that have an arbitrary number of intermediate control points (i.e., control points excluding the start and end points). Though this would be computationally more expensive, the resulting Bézier spline is likely to have broader support over the space of all limit set boundaries.

Funding

Supplementary Material

Semiparametric Estimation of The Shape of The Limiting Bivariate Point Cloud (DOI: 10.1214/25-BA1514SUPP; .pdf). The Supplementary Material (Majumder et al., 2024a) contains additional results from the simulation study presented in Section 4, as well as an additional simulation study which compares the efficacy of our approach under different approaches of choosing the truncation threshold.

References

Balkema, A. A., Embrechts, P., and Nolde, N. (2010). "Meta densities and the shape of their sample clouds." *J. Multivariate Anal.*, 101(7): 1738–1754. URL https://doi.org/10.1016/j.jmva.2010.02.010 MR2610743. doi: https://doi.org/10.1016/j.jmva.2010.02.010. 1

Balkema, G. and Nolde, N. (2010). "Asymptotic independence for unimodal densi-

24

ties." *Adv. in Appl. Probab.*, 42(2): 411–432. URL https://doi.org/10.1239/aap/1275055236 MR2675110. doi: https://doi.org/10.1239/aap/1275055236. 1, 3

Billmire, M., French, N. H. F., Loboda, T., Owen, R. C., and Tyner, M. (2014). "Santa Ana winds and predictors of wildfire progression in southern California." *Int. J. Wildland Fire*, 23(8): 1119–1129. 18

Binkowski, F. S. and Roselle, S. J. (2003). "Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description." *J. Geophys. Res. – Atmos.*, 108(D6). 21

Boldi, M.-O. and Davison, A. C. (2007). "A mixture model for multivariate extremes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 217–229. MR2325273. doi: https://doi.org/10.1111/j.1467-9868.2007.00585.x. 3

Cooley, D., Hunter, B. D., and Smith, R. L. (2019). "Univariate and multivariate extremes for the environmental sciences." In *Handbook of environmental and ecological statistics*, 153–180. Chapman and Hall/CRC. MR3889895. 18, 20

Davis, R. A., Mulrow, E., and Resnick, S. I. (1988). "Almost sure limit sets of random samples in $\mathbf{R}^d$." *Adv. in Appl. Probab.*, 20(3): 573–599. URL https://doi.org/10.2307/1427036 MR0955505. doi: https://doi.org/10.2307/1427036. 1, 3

de Carvalho, M., Kumukova, A., and dos Reis, G. (2022). "Regression-type analysis for multivariate extreme values." *Extremes*, 25(4): 595–622. MR4505181. doi: https://doi.org/10.1007/s10687-022-00446-6. 3

Dunn, R. J. H., Willett, K. M., Thorne, P. W., Woolley, E. V., Durre, I., Dai, A., Parker, D. E., and Vose, R. S. (2012). "HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011." *Clim. Past*, 8(5): 1649–1679. URL https://cp.copernicus.org/articles/8/1649/2012/ 18

Farouki, R. T. (2012). "The Bernstein polynomial basis: a centennial retrospective." *Comput. Aided Geom. D.*, 29(6): 379–419. URL https://doi.org/10.1016/j.cagd.2012.03.001 MR2921860. doi: https://doi.org/10.1016/j.cagd.2012.03.001. 2, 5

Gong, W., Reich, B. J., and Chang, H. H. (2021). "Multivariate spatial prediction of air pollutant concentrations with INLA." *Environ. Res. Commun.*, 3(10): 101002. 21

Hanson, T. E., de Carvalho, M., and Chen, Y. (2017). "Bernstein polynomial angular densities of multivariate extreme value distributions." *Statistics & Probability Letters*, 128: 60–66. MR3656376. doi: https://doi.org/10.1016/j.spl.2017.03.030. 2

Hazewinkel, M. (2012). *Encyclopaedia of Mathematics.* Number Vol. 1 in Encyclopaedia of Mathematics. Springer Netherlands. URL https://books.google.com/books?id=IEnqCAAAQBAJ MR1492593. doi: https://doi.org/10.1007/978-94-015-1288-6. 2, 5

Heffernan, J. E. and Tawn, J. A. (2004). "A conditional approach for multivariate extreme values." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3): 497–546. With dis-

cussions and reply by the authors. URL https://doi.org/10.1111/j.1467-9868. 2004.02050.x MR2088289. doi: https://doi.org/10.1111/j.1467-9868.2004. 02050.x. 4

Hill, B. M. (1975). "A simple general approach to inference about the tail of a distribution." *Ann. Stat.*, 3(5): 1163–1174. MR0378204. 2

Hyndman, R. J. and Shang, H. L. (2010). "Rainbow plots, bagplots, and boxplots for functional data." *J. Computat. Graph. Stat.*, 19(1): 29–45. MR2752026. doi: https:// doi.org/10.1198/jcgs.2009.08158. 13

Kinoshita, K. and Resnick, S. I. (1991). "Convergence of scaled random samples in $\mathbf{R}^d$." *Ann. Probab.*, 19(4): 1640–1663. MR1127719. 1, 3

Ledford, A. W. and Tawn, J. A. (1996). "Statistics for near independence in multivariate extreme values." *Biometrika*, 83(1): 169–187. URL https://doi.org/10.1093/ biomet/83.1.169 MR1399163. doi: https://doi.org/10.1093/biomet/83.1.169. 4, 10

Littell, J. S., McKenzie, D., Wan, H. Y., and Cushman, S. A. (2018). "Climate change and future wildfire in the western United States: An ecological approach to nonstationarity." *Earth's Future*, 6(8): 1097–1111. URL https://doi.org/10.1029/ 2018EF000878 18

López-Pintado, S. and Romo, J. (2009). "On the concept of depth for functional data." *J. Am. Stat. Assoc.*, 104(486): 718–734. MR2541590. doi: https://doi.org/ 10.1198/jasa.2009.0108. 13

Mackay, E. and Jonathan, P. (2023). "Modelling multivariate extremes through angular-radial decomposition of the density function." *arXiv preprint*, arXiv:2310.12711. 23

Majumder, R., Reich, B. J., Shaby, B. A., and Cooley, D. S. (2024a). "Supplement to "Semiparametric estimation of the shape of the limiting Bivariate point cloud"." *Bayesian Analysis.* 9, 18, 23

Majumder, R., Shaby, B. A., Reich, B. J., and Cooley, D. S. (2024b). *BezELS: Bezier splines for Estimating Limit Sets*. R package version 0.1.0. URL https://github. com/reetamm/BezELS 13

Marcon, G., Padoan, S., Naveau, P., Muliere, P., and Segers, J. (2017). "Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials." *J. Stat. Plan. Infer.*, 183: 1–17. MR3589492. doi: https://doi.org/10. 1016/j.jspi.2016.10.004. 2

Marcon, G., Padoan, S. A., Naveau, P., and Muliere, P. (2014). "Nonparametric estimation of the dependence among multivariate rainfall maxima." In *METMA VII-GRASPA 14.* 2

Murphy-Barltrop, C. J. R., Wadsworth, J. L., and Eastoe, E. F. (2024). "Improving estimation for asymptotically independent bivariate extremes via global estimators for the angular dependence function." *Extremes*, 27(4): 643–671. MR4816049. doi: https:// doi.org/10.1007/s10687-024-00490-4. 2, 11

Nolde, N. (2014). "Geometric interpretation of the residual dependence coefficient." *J. Multivariate Anal.*, 123: 85–95. URL https://doi.org/10.1016/j.jmva.2013.08.018 MR3130422. doi: https://doi.org/10.1016/j.jmva.2013.08.018.   1, 2, 3, 4, 5

Nolde, N. and Wadsworth, J. L. (2022). "Linking representations for multivariate extremes via a limit set." *Adv. Appl. Probab.*, 54: 688–717. MR4497395. doi: https://doi.org/10.1017/apr.2021.51.   1, 3, 5, 9

Padoan, S. A. and Rizzelli, S. (2022). "Consistency of Bayesian inference for multivariate max-stable distributions." *The Annals of Statistics*, 50(3): 1490–1518. MR4441129. doi: https://doi.org/10.1214/21-aos2160.   3

Papastathopoulos, I., De Monte, L., Campbell, R., and Rue, H. (2023). "Statistical inference for radially-stable generalized Pareto distributions and return level-sets in geometric extremes." *arXiv preprint*, arXiv:2310.06130.   23

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/   13

Sabourin, A. and Naveau, P. (2014). "Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization." *Computational Statistics & Data Analysis*, 71: 542–567. MR3131989. doi: https://doi.org/10.1016/j.csda.2013.04.021.   3

Sabourin, A., Naveau, P., and Fougères, A.-L. (2013). "Bayesian model averaging for multivariate extremes." *Extremes*, 16: 325–350. MR3101876. doi: https://doi.org/10.1007/s10687-012-0163-0.   3

Simpson, E. S. and Tawn, J. A. (2022). "Estimating the limiting shape of bivariate scaled sample clouds for self-consistent inference of extremal dependence properties." *arXiv preprint*, arXiv:2207.02626. MR4828446. doi: https://doi.org/10.1214/24-ejs2300.   2, 8, 9, 11, 12, 13, 17, 23

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). "Determining the dependence structure of multivariate extremes." *Biometrika*, 107(3): 513–532. URL https://doi.org/10.1093/biomet/asaa018 MR4138974. doi: https://doi.org/10.1093/biomet/asaa018.   4, 11

Sun, Y. and Genton, M. G. (2011). "Functional Boxplots." *J. Computat. Graph. Stat.*, 20(2): 316–334. MR2847798. doi: https://doi.org/10.1198/jcgs.2011.09224.   13

US EPA (2017). "Air Quality System Database." Accessed on 3 August, 2017. URL http://epa.gov/ttn/airs/aqsdatamaty   21

Vettori, S., Huser, R., and Genton, M. G. (2018). "A comparison of dependence function estimators in multivariate extremes." *Stat. Comput.*, 28: 525–538. MR3761339. doi: https://doi.org/10.1007/s11222-017-9745-7.   2

Wadsworth, J. L. and Campbell, R. (2024). "Statistical inference for multivariate extremes via a geometric approach." *Journal of the Royal Statistical Society Series*

B: Statistical Methodology, qkae030. MR4825002. doi: https://doi.org/10.1093/jrsssb/qkae030.   1, 2, 7, 8, 23

Wadsworth, J. L. and Tawn, J. A. (2013). "A new representation for multivariate tail probabilities." *Bernoulli*, 19(5B): 2689–2714. URL https://doi.org/10.3150/12-BEJ471 MR3160568. doi: https://doi.org/10.3150/12-BEJ471.   4, 11

Wyat Appel, K., Bhave, P. V., Gilliland, A. B., Sarwar, G., and Roselle, S. J. (2008). "Evaluation of the community multiscale air quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II—particulate matter." *Atmos. Environ.*, 42(24): 6057–6066.   21

Wyat Appel, K., Gilliland, A. B., Sarwar, G., and Gilliam, R. C. (2007). "Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance: Part I—Ozone." *Atmos. Environ.*, 41(40): 9603–9615.   21