Taylor & Francis
Taylor & Francis Group

Check for updates

Tail density estimation for exploratory data analysis using kernel methods

B. Béranger [a,b], T. Duong[a,*], S. E. Perkins-Kirkpatrick [c] and S. A. Sisson [b]

[a]Theoretical and Applied Statistics Laboratory (LSTA), University Pierre and Marie Curie - Paris 6, Paris, France; [b]School of Mathematics and Statistics, University of New South Wales, Sydney, Australia; [c]Climate Change Research Centre, University of New South Wales, Sydney, Australia

ABSTRACT

It is often critical to accurately model the upper tail behaviour of a random process. Nonparametric density estimation methods are commonly implemented as exploratory data analysis techniques for this purpose and can avoid model specification biases implied by using parametric estimators. In particular, kernel-based estimators place minimal assumptions on the data, and provide improved visualisation over scatterplots and histograms. However kernel density estimators can perform poorly when estimating tail behaviour above a threshold, and can over-emphasise bumps in the density for heavy tailed data. We develop a transformation kernel density estimator which is able to handle heavy tailed and bounded data, and is robust to threshold choice. We derive closed form expressions for its asymptotic bias and variance, which demonstrate its good performance in the tail region. Finite sample performance is illustrated in numerical studies, and in an expanded analysis of the performance of global climate models.

1. Introduction

The extreme values (very large or very small values) of a dataset are frequently of interest as they are closely related to uncommon events with important consequences. For climate data, these tail events include heat waves (prolonged extreme high temperatures), cold snaps (extreme low temperatures), floods (extreme high levels of waterways or tides or waves), storms (extreme high wind speeds or amounts of precipitation) and droughts (prolonged extreme low amounts of rainfall) (e.g. Kotz and Nadarajah 2000; Coles 2001).

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is a d-dimensional random vector with cumulative distribution function (c.d.f.) $F_{\boldsymbol{X}}$ and probability density function (p.d.f.) $f_{\boldsymbol{X}}$. A common representation of the upper tail values arising from this distribution focuses on examining those values which exceed some high threshold $\boldsymbol{u} = (u_1, \ldots, u_d)^\top$, which determines the

---

CONTACT B. Béranger ✉ B.Beranger@unsw.edu.au 📍 Theoretical and Applied Statistics Laboratory (LSTA), University Pierre and Marie Curie - Paris 6, F-75005 Paris, France
*Present address: Computer Science Laboratory (LIPN) University Paris-Nord - Paris 13, F-93430, Villetaneuse

support on which values of X are considered to be large. We can denote these upper tail values by $X^{[u]} \equiv X \mid X > u$, under which each marginal inequality must hold i.e. $X_j > u_j$ for $j = 1, \ldots, d$. Estimating the asymptotic tail behaviour of X is one of the standard goals of extreme value theory.

In the simplest case of univariate extremes ($d = 1$), common approaches to distributional tail estimation rely on parametric models, typically based on generalised extreme value (GEV) or generalised Pareto distributions (GPD), for which numerous estimation procedures are available. These methods include maximum likelihood (Prescott and Walden 1980; Hosking 1985; Smith 1985; Macleod 1989), probability weighted moments (Hosking, Wallis, and Wood 1985), maximum product spacing (Cheng and Amin 1983), least squares estimation (Maritz and Munro 1967), estimation based on order statistics and records (Hill 1975; Pickands 1975), the method of moments (Christopeit 1994) and Bayesian estimation (Lye, Hapuarachchi, and Ryan 1993). As with all parametric estimators, these approaches suffer from the possibility of misspecification, particularly when the asymptotic GEV and GPD models cannot be assumed to hold. This potential can be avoided by nonparametric estimation which does not make assumptions on particular parametric forms. See Markovich (2007, Chapter 3) for a summary of nonparametric estimation of univariate heavy tailed densities.

For multivariate extremes ($d > 1$), nonparametric estimation of indicators of extremal dependence is an intensively studied field, and includes estimation of the Pickands or extremal dependence function (Pickands 1975; Hall and Tajvidi 2000; Marcon, Padoan, Naveau, and Muliere 2014), the tail dependence function (Huang 1992; Drees and Huang 1998; Einmahl, Krajina, and Segers 2008, 2012), and the spectral measure (Einmahl, de Haan, and Piterbarg 2001; Einmahl and Segers 2009; de Carvalho, Oumow, Segers, and Warchoł 2013). The motivation for nonparametric estimators is stronger for multivariate extremes than in the univariate case as there is no general parametric form to describe the range of extremal behaviour of max-stable processes. This means that a choice of any particular parametric family has a possibility of resulting in model misspecification.

In this article we focus on density estimation for data that exceed some upper quantile of the data for which asymptotic, extreme value theory-derived models may not hold. Hence our motivation differs from that in the traditional extreme value theory literature in the sense that the our main primary interest here is inference at sub-asymptotic levels i.e. around and slightly beyond the range of the observed data.

This manuscript proposes a new class of multivariate kernel-based nonparametric density estimators to model the upper tail behaviour of X, without having to pre-specify a parametric family. Kernel density estimators are among the most widely used nonparametric estimators (e.g. Silverman 1986; Wand and Jones 1995) and they possess excellent visualisation properties which can naturally form part of an exploratory data analysis. However, standard kernel estimators can produce spurious bumps in the estimated tails of f_X if it has heavy tails, as is particularly the case when analysing extremes and moderate extremes. On the other hand, if we focus on the tail sample $X^{[u]}$, which is truncated at the threshold u, a standard estimator of the tail density $f_{X^{[u]}}$ is strongly influenced by the boundary effects due to the truncated support, as well as the choice of this threshold.

Here we focus on modifications to standard kernel density estimation which attenuate these spurious bumps, and accommodate a truncated support when estimating $f_{X^{[u]}}$. Our approach is based on the standard identity $f_{X^{[u]}}(x) = f_X(x)/\bar{F}_X(u)$ where $\bar{F}_X(u)$ is

the corresponding survival function. We estimate the complete density f_X using transformation kernel density estimation techniques (see e.g. Silverman 1986; Charpentier and Flachaire 2015): this approach has the double advantage of being able to handle bounded supported data as well as reducing the spurious bumps in the tail. Estimating the normalisation constant $\bar{F}_X(u)$ is straightforward once an estimator of f_X is established.

While multivariate transformation estimators are a well-known method for complete densities, our contribution consists of modifying them for the estimation of tail densities. We also supply new results for the pointwise bias and variance which describe the behaviour of the estimator at the boundary and in the tails. In the context of moderate extremes, this permits the construction and theoretical justification of more efficient kernel based nonparametric density estimators for the tails of observed processes. It additionally allows these estimators to be used within existing goodness-of-fit measures (e.g. Perkins, Pitman, and Sisson 2013) in place of more poorly performing histogram estimates of tail behaviour.

One may argue that in the univariate extreme value theory setup, the Pickands-Balkema-de Haan theorem (Balkema and de Haan 1974; Pickands 1975) can be used. This states that, under some mild conditions on the underlying c.d.f., all observations above some large threshold u are well approximated by the GPD. We will show that our proposed kernel estimator produces a comparable fit to the one produced by the GPD on the exceedances and can even occasionally outperform it.

Nonparametric estimation of heavy-tail distribution via transformation kernel density methods has previously been considered (e.g. Wand, Marron, and Ruppert 1991; Bolancé, Guillen, and Nielsen 2003; Clements, Hurn, and Lindsay 2003). Buch-Larsen, Nielsen, Guillén, and Bolancé (2005) introduced a class of transformation functions based on the Champernowne distribution function as a generalisation of the shifted-power family functions (of which the logarithm is the simplest case) to define a semiparametric estimation procedure for heavy-tailed distributions. Density estimation based on this approach has mainly been undertaken in the univariate setting, see e.g. Bolancé, Guillén, Gustafsson, and Nielson (2012) and the references therein. In the bivariate setting Bolancé, Guillen, Pelican, and Vernic (2008) proposed a transformation kernel estimation using product kernels to estimate conditional tail expectations on an insurance claims datasets, while Pitt and Guillen (2010) outline a method for bivariate density estimation via a weighted likelihood (without performing any data analysis). Buch-Kromann, Guillén, Linton, and Nielsen (2011) apply a multiplicative bias reduction method to a multivariate version of the Champernowne transformation estimation procedure of Buch-Larsen et al. (2005), with a normal scale vector method for bandwidth selection. This strategy can be compared to taking a copula as an auxiliary function for density estimation, although this introduces methodological complexity. Most of this work aims to improve the performance of heavy-tailed distribution density estimators via optimisation of the transformation function (Buch-Kromann et al. 2011, Section 3.5). In contrast, in this article we focus on a general method for multivariate bandwidth selection, by defining a consistent bandwidth matrix selector rather than simply a vector selector. Furthermore, we treat the goodness-of-fit problem of measuring a density estimate via reference to a hypothesised density for model selection purposes.

The layout of this article is as follows. Our primary contribution is presented in Section 2, which develops the transformation kernel estimator for tail density estimation,

establishes its pointwise bias and variance (with proofs deferred to the Appendix) and examines optimal bandwidth estimation. We also assess histogram based tail density estimation, and develop the role of tail density estimators in goodness-of-fit (model selection) procedures. In Section 3 we verify our results on finite samples for simulated data in both univariate and multivariate settings, and in Section 4 we expand the work of Perkins et al. (2013) by performing an analysis of 22 global climate models (GCMs) and assess how well they are able to reproduce observed temperature extremes. Section 5 concludes with a discussion.

2. Tail densities for upper tail values

2.1. Transformation tail density estimation

Let $X_1, \ldots, X_n$ be a random sample drawn from a common univariate distribution F_X with density f_X. If f_X has heavy tails, standard kernel estimators are susceptible to producing spurious bumps in the tails of the density estimates, as they apply a fixed amount of smoothing over the entire sample space. A common approach is to apply a transformation on the data sample to reduce the inter-point distances in these upper tail values so that a global constant smoothing is more appropriate. We focus on transformation kernel estimators, where a known monotonic transformation $t(\cdot)$ maps the data support to the real line where standard kernel estimators are well-established, before back-transforming to the original data support. See e.g. Silverman (1986) and Charpentier and Flachaire (2015).

Let $Y = t(X)$ be a transformed random variable, with distribution F_Y and density f_Y. The relationship between the transformed random variable Y and the original X at a non-random point x is given by

$$f_X(x) = |t'(x)| f_Y(t(x))$$

where t' is the first derivative of t. Consider the transformed sample $Y_1, \ldots, Y_n$ where $Y_i = t(X_i)$, $i = 1, \ldots, n$, and $y = t(x)$. Since many upper tail data samples are also bounded, e.g. $X_1, \ldots X_n$ are supported on (u_0, ∞), a suitable transformation would be $t(x) = \log(x - u_0)$. In the case for unbounded data, the logarithm transformation can still be used if we set $u_0 < \min\{X_1, \ldots, X_n\}$. As $Y_1, \ldots, Y_n$ are supported on the real line, f_Y can then be estimated by the standard kernel density estimator

$$\hat{f}_Y(y; h) = n^{-1} \sum_{i=1}^{n} K_h(y - Y_i)$$

where $K_h(y) = h^{-1} K(y/h)$ is a scaled kernel, $h > 0$ is the bandwidth or smoothing parameter and K is a symmetric kernel density function. The estimator for f_X can then be defined by replacing the true density by its kernel estimator

$$\hat{f}_X(x; h) = |t'(t^{-1}(y))| \hat{f}_Y(y; h).$$

Using the standard identity $f_{X[u]}(x) = f_X(x)/\bar{F}_X(u)$, our proposed estimator of the tail density is

$$\hat{f}_{X[u]}(x; h) = \hat{f}_X(x; h)/\hat{\bar{F}}_X(u; h)$$

where $\hat{\bar{F}}_X(u; h) = \int_{-\infty}^{u} \hat{f}_X(x; h) \, dx$ can be numerically approximated.

A generalisation of this transformation kernel estimator to multivariate data is established through a d-dimensional random vector $X = (X_1, \ldots, X_d)^\top$ with distribution function F_X and density function f_X. The random variable of values greater than a vector threshold $u = (u_1, \ldots, u_d)^\top$ is denoted as $X^{[u]} \equiv X|X > u$ under which each marginal inequality must hold, i.e. $X_j > u_j$ for $j = 1, \ldots, d$. The support of $X^{[u]}$ is the Cartesian product $(u, \infty) = (u_1, \infty) \times \cdots \times (u_d, \infty)$. For $x \in (u, \infty)$, the corresponding tail density is $f_{X^{[u]}}(x) = f_X(x)/\bar{F}_X(u)$ and tail distribution is $F_{X^{[u]}}(x) = F_X(x)/\bar{F}_X(u)$, where $\bar{F}_X(u) = \int_{(u,\infty)} f_X(w)dw$ is the survival function of X evaluated at u.

Let $X_1, \ldots, X_n$ form a random sample drawn from the common d-variate distribution F_X. Consider the transformed random variable $Y = t(X)$ where $t : (u_0, \infty) \to \mathbb{R}^d$ is defined by $t(x) = (t_1(x_1), \ldots, t_d(x_d))^\top$ where the t_j are monotonic functions on (u_{0j}, ∞) e.g. $t_j(x_j) = \log(x_j - u_{0j}), j = 1, \ldots, d$. The density of X is then related to the density of Y by

$$f_{X^{[u]}}(x) = f_Y(t(x))|\mathbf{J}_t(x)|$$

where $|\mathbf{J}_t|$ is the Jacobian of t. Denoting the transformed data sample as $Y_1, \ldots, Y_n$, with $Y_i = t(X_i), i = 1, \ldots, n$, the kernel estimator of f_Y at a non-random point $y = (y_1, \ldots, y_d)^\top = t(x)$ is then given by

$$\hat{f}_Y(y; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(y - Y_i)$$

where K is a symmetric d-variate density function, the bandwidth matrix $\mathbf{H}$ is a $d \times d$ positive definite symmetric matrix of smoothing parameters, and the scaled kernel $K_{\mathbf{H}}(y) = |\mathbf{H}^{-1/2}|K(\mathbf{H}^{-1/2}y)$. The tail density can then be defined by replacing the true density function by its kernel estimator

$$\hat{f}_X(x; \mathbf{H}) = |\mathbf{J}_t(t^{-1}(y))|\hat{f}_Y(y; \mathbf{H})$$

where $t^{-1}(y) = (t_1^{-1}(y_1), \ldots, t_d^{-1}(y_d))^\top$ is the element-wise inverse of $t(y)$. Therefore

$$\hat{f}_{X^{[u]}}(x; \mathbf{H}) = \hat{f}_X(x; \mathbf{H})/\hat{\bar{F}}_X(u; \mathbf{H}) \tag{1}$$

where $\hat{\bar{F}}_X(u; \mathbf{H}) = \int_{(-\infty, u)} \hat{f}_X(x; \mathbf{H}) \, dx$ can be numerically approximated, for example by a Riemann sum.

In this approach, the threshold u is only required to be specified in Equation (1). The statistical properties of $\hat{f}_{X^{[u]}}$ are almost completely determined by those of $\hat{f}_X$ which do not rely on the choice of the threshold u. This is in contrast to an estimator of $f_{X^{[u]}}$ based on only the truncated sample $\{X_i : X_i > u\}$, as this is highly dependent on the choice (and the estimation) of the threshold. Conveniently, for our proposed estimator, it is possible to efficiently explore the tail behaviour for several thresholds, as the most onerous calculations are carried out to compute $\hat{f}_X$, and need not be repeated for each threshold choice. Furthermore, with this decoupling of the density estimation from the threshold estimation, this leaves the potential for the incorporation of more sophisticated estimators of u, although this is beyond the scope of this paper.

2.2. Tail density estimator performance

Under standard regularity conditions and using standard analysis techniques, Lemma A.1 in the Appendix demonstrates that the pointwise bias and variance of the kernel density with unbounded data support $\hat{f}_Y$ are given by

$$\text{Bias}\{\hat{f}_Y(\boldsymbol{y}; \mathbf{H})\} = \tfrac{1}{2} m_2(K) \text{tr}(\mathbf{H} \mathsf{D}^2 f_Y(\boldsymbol{y}))\{1 + o(1)\}$$

$$\text{Var}\{\hat{f}_Y(\boldsymbol{y}; \mathbf{H})\} = n^{-1} |\mathbf{H}|^{-1/2} f_Y(\boldsymbol{y}) R(K)\{1 + o(1)\},$$

where $m_2(K) = \int_{\mathbb{R}^d} y_1^2 K(\boldsymbol{y}) \, d\boldsymbol{y}$, $R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{y})^2 \, d\boldsymbol{y}$ and $\mathsf{D}^2 f_Y$ is the Hessian matrix of second order partial derivatives of f_Y with respect to $\boldsymbol{y}$. The equivalent result for the transformation kernel estimator $\hat{f}_X$ is more difficult to establish, especially for a general transformation $\boldsymbol{t}$, so we focus on the logarithm transformation, $\boldsymbol{t}(\boldsymbol{x}) = (\log(x_{1d}), \ldots, \log(x_d))^\top$.

Theorem 2.1: *Suppose that $\mathbf{X}$ is supported on $(\mathbf{0}, \infty)$. Under the regularity conditions (A1)–(A3) in the Appendix, the bias and variance of the logarithm transformation kernel estimator $\hat{f}_X$ at an estimation point $\boldsymbol{x} \in (\mathbf{0}, \infty)$ are*

$$\text{Bias}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} = \tfrac{1}{2} m_2(K) \big[\pi(\boldsymbol{x})^{-1} f_X(\boldsymbol{x}) \text{tr}(\mathbf{H} \, \text{Diag}(\boldsymbol{x})) + 2\pi(\boldsymbol{x})^{-1} \text{tr}(\mathbf{H}\boldsymbol{x}\mathsf{D}f_X(\boldsymbol{x})^\top \text{Diag}(\boldsymbol{x}))$$

$$+ \text{tr}(\mathbf{H} \, \text{Diag}\,(\boldsymbol{x}) \text{Diag}(\mathsf{D}f_X(\boldsymbol{x}))) + \text{tr}(\mathbf{H} \, \text{Diag}(\boldsymbol{x}) \mathsf{D}^2 f_X(\boldsymbol{x}) \text{Diag}(\boldsymbol{x})) \big] \{1 + o(1)\}$$

$$\text{Var}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} = n^{-1} |\mathbf{H}|^{-1/2} R(K) \pi(\boldsymbol{x})^{-1} f_X(\boldsymbol{x})\{1 + o(1)\},$$

where $\pi(\boldsymbol{x}) = \prod_{j=1}^d x_j$, $\text{Diag}(\boldsymbol{x})$ is the $d \times d$ diagonal matrix with main diagonal given by $\boldsymbol{x}$, and $\mathsf{D}f_X$ and $\mathsf{D}^2 f_X$ are the gradient vector and Hessian matrix of f_X with respect to $\boldsymbol{x}$.

Proof: See Appendix.   ∎

Without loss of generality, the above results for $\mathbf{X}$ supported on $(\mathbf{0}, \infty)$ may be extended to the general case for $\mathbf{X}$ supported on $(\boldsymbol{u}_0, \infty)$ following a suitable translation.

For $d = 1$, the results under Theorem 2.1 reduce to

$$\text{Bias}\{\hat{f}_X(x; h)\} = \frac{1}{2} m_2(K) h^2 [f_X(x) + 3x f_X'(x) + x^2 f_X''(x)]\{1 + o(1)\}$$

$$\text{Var}\{\hat{f}_X(x; h)\} = \frac{R(K)}{nhx} f_X(x)\{1 + o(1)\},$$

which agree with those in Charpentier and Flachaire (2015, Equations (14) and (17)). These authors note that if $f_X(0), f_X'(0), f_X''(0)$ are all finite, then the bias tends to $\tfrac{1}{2} m_2(K) h^2 f_X(0)$ as $x \to 0$. So if $f_X(0) \neq 0$, then bias and variance problems may persist when approaching the boundary. On the other hand, away from the boundary the bias and variance tend to 0 as $x \to \infty$.

The multivariate expressions are not as straightforward to interpret in general, however computing the $d = 2$ case explicitly is instructive. Writing $\mathbf{H} = [h_1^2, h_{12}; h_{12}, h_2^2]$ as a 2×2

matrix, then

$$\text{Bias}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} = \frac{1}{2}m_2(K)\left[\left(\frac{h_1^2}{x_1} + \frac{h_2^2}{x_2}\right)f_X(\boldsymbol{x}) + 2\left(\frac{h_1^2 x_1}{x_2}\frac{\partial f_X(\boldsymbol{x})}{\partial x_1} + \frac{h_2^2 x_2}{x_1}\frac{\partial f_X(\boldsymbol{x})}{\partial x_2}\right)\right.$$

$$+ \left(h_1^2 x_1 \frac{\partial f_X(\boldsymbol{x})}{\partial x_1} + h_2^2 x_2 \frac{\partial f_X(\boldsymbol{x})}{\partial x_2}\right)$$

$$\left. + \left(h_1^2(x_1^2 + x_1 x_2)\frac{\partial^2 f_X(\boldsymbol{x})}{\partial x_1^2} + h_2^2(x_1 x_2 + x_2^2)\frac{\partial^2 f_X(\boldsymbol{x})}{\partial x_2^2}\right)\right]\{1 + o(1)\}$$

$$\text{Var}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} = \frac{R(K)f_X(\boldsymbol{x})}{n(h_1^2 h_2^2 - h_{12}^2)^{1/2}x_1 x_2}\{1 + o(1)\}.$$

The variance is a straightforward extension of the univariate expression. However this is not the case for the bias: the coefficient for f_X now involves $(h_1^2/x_1 + h_2^2/x_2)$ and $\mathsf{D}f_X$ involves $[h_1^2 x_1/x_2, h_2^2 x_2/x_1]$ in addition to $[h_1^2 x_1, h_2^2 x_2]$, due to the action of the Jacobian $|\mathbf{J}_t(\boldsymbol{x})| = \pi(\boldsymbol{x})^{-1}$. If $f_X(\mathbf{0}), \mathsf{D}f_X(\mathbf{0}), \mathsf{D}^2 f_X(\mathbf{0})$ are all finite then the bias tends to

$$\frac{1}{2}m_2(K)\left[\left(\frac{h_1^2}{x_1} + \frac{h_2^2}{x_2}\right)f_X(\mathbf{0}) + \left(\frac{h_1^2 x_1}{x_2}\frac{\partial f_X(\mathbf{0})}{\partial x_1} + \frac{h_2^2 x_2}{x_1}\frac{\partial f_X(\mathbf{0})}{\partial x_2}\right)\right]$$

as $x_1, x_2 \to 0$. Hence if $f_X(\mathbf{0}) \neq 0$, then the bias grows without bound; and likewise for the variance. Away from the boundary, the MSE tends to 0 as $x_1, x_2 \to \infty$. Furthermore, for general d, for a fixed $\boldsymbol{x}$ in the tail region, then we have $\text{MSE}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} = O(n^{-1}|\mathbf{H}|^{-1/2} + \text{tr}^2(\mathbf{H}))$ as $n \to \infty$.

Returning to our proposed tail density estimator, we have

$$\hat{f}_{X^{[u]}}(\boldsymbol{x}; \mathbf{H}) = \hat{f}_X(\boldsymbol{x}; \mathbf{H})/\hat{\bar{F}}_X(\boldsymbol{u}; \mathbf{H}) = \hat{f}_X(\boldsymbol{x}; \mathbf{H})/\bar{F}_X(\boldsymbol{u})\{1 + o_p(1)\}$$

as $\hat{\bar{F}}_X(\boldsymbol{u})$ is pointwise MSE convergent to $\bar{F}_X(\boldsymbol{u})$ – see Jin and Shao (1999). Under the regularity conditions in Theorem 2.1, this implies that $\text{MSE}\{\hat{f}_{X^{[u]}}(\boldsymbol{x}; \mathbf{H})\} = \text{MSE}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\}/\bar{F}_X(\boldsymbol{u})^2\{1 + o(1)\}$, so the properties of the tail density estimator $\hat{f}_{X^{[u]}}$ largely carry over from the transformation kernel density estimator $\hat{f}_X$, with the important difference that $\hat{f}_{X^{[u]}}$ suffers much less from boundary problems than $\hat{f}_X$. This is because $\hat{f}_X$ has potentially undesirable behaviour near its boundary $\boldsymbol{u}$, whereas we only require $\hat{f}_X$ to be calculated on $(\boldsymbol{u}_0, \infty)$, with $\boldsymbol{u} \gg \boldsymbol{u}_0$. Note that $\boldsymbol{u}_0$ is fixed, and there is no true value to estimate; see the simulation studies of Section 3 for an example.

An alternative for density estimation in heavy tails is to vary the amount of smoothing, rather than to apply a stabilising transformation e.g. Loftsgaarden and Quesenberry (1965) and Abramson (1982), though these estimators do not account for data boundedness. To handle the boundedness of the data sample, another approach is based on modifying the kernel function itself to avoid assigning probability mass outside the data support, e.g. Gasser and Müller (1979) and Chen (1999). These techniques are focussed on boundary behaviour and do not address the issue of spurious bumps in the tails which are far away from this boundary. Our proposed logarithm transformation kernel estimator is able to handle both of these issues simultaneously.

2.3. *Optimal bandwidth computation*

The complicated form of the pointwise $\mathrm{MSE}\{\hat{f}_X(x; \mathbf{H})\}$ does not facilitate the computation of a closed form mean integrated squared error, so it is not feasible to define an oracle bandwidth for the transformation density estimator $\hat{f}_X$. Since the estimation is carried out in the unbounded space of $Y_1, \ldots, Y_n$, then our strategy is to carry out the bandwidth selection on these transformed data, as there is large body of data-based bandwidth selectors which lead to consistent density estimates. The back-transformation to the original data scale does not require any adjustment to this bandwidth to compute the transformation density estimator, and subsequently to the tail density estimator.

From Lemma A.1 in the Appendix, the mean integrated squared error (MISE) of the density estimator $\hat{f}_Y$ is

$$\mathrm{MISE}\{\hat{f}_Y(\cdot; \mathbf{H})\} = \left[\tfrac{1}{4} m_2^2(K)(\mathrm{vec}^\top \mathbf{H} \otimes \mathrm{vec}^\top \mathbf{H})\boldsymbol{\psi}_{Y,4} + n^{-1}|\mathbf{H}|^{-1/2}R(K)\right]\{1 + o(1)\},$$

where $\boldsymbol{\psi}_{Y,4} = \int_{\mathbb{R}^d} \mathsf{D}^{\otimes 4} f_Y(y) f_Y(y)\, \mathrm{d}y$, as defined in Chacón and Duong (2010), and vec is the vectorisation operator which stacks the columns of matrix into a single column. Using this MISE expression, we can then define an oracle optimal bandwidth choice as the minimiser of the MISE

$$\mathbf{H}^* = \operatorname*{argmin}_{\mathbf{H} \in \mathcal{F}} \mathrm{MISE}\{\hat{f}_Y(\cdot; \mathbf{H})\} = O(n^{-2/(d+4)}) \tag{2}$$

where $\mathcal{F}$ is the space of $d \times d$ symmetric positive definite matrices. Furthermore, utilising this optimal bandwidth in $\hat{f}_Y$, the minimal MISE is $\inf_{\mathbf{H} \in \mathcal{F}} \mathrm{MISE}\{\hat{f}_Y(\cdot; \mathbf{H})\} = O(n^{-4/(d+4)})$. With this bandwidth matrix order, for a non-random point x in the tail region, the minimal MSE for the tail density estimator is $\inf_{\mathbf{H} \in \mathcal{F}} \mathrm{MSE}\{\hat{f}_{X^{[u]}}(x; \mathbf{H})\} = O(n^{-4/(d+4)})$ also, as $n \to \infty$.

The optimal bandwidth selector defined in Equation (2) is mathematically intractable as it depends on unknown quantities. Accordingly a vast body of research in the density estimation literature has focussed on providing data-based bandwidth selectors which estimate or approximate the optimal bandwidth. There are three main classes: (i) normal scale (or rule of thumb), (ii) plug-in and (iii) cross validation.

The class of normal scale selectors is an extension to the multivariate case of the quick and simple bandwidth selectors where the unknown density f is replaced by a normal density, leading to

$$\hat{\mathbf{H}}_{\mathrm{NS}} = \left[\frac{4}{(d+2)n}\right]^{2/(d+4)} \mathbf{S} n^{-2/(d+4)}$$

where $\mathbf{S}$ is the sample covariance matrix of $Y_1, \ldots, Y_n$ (see e.g. Wand and Jones 1995, p. 111).

The class of plug-in selectors consists of a generalisation of the work of Sheather and Jones (1991) for univariate data by Wand and Jones (1994) and Duong and Hazelton (2003) for multivariate data. Plug-in selectors use as a starting point the AMISE formula (Asymptotic MISE) where the only unknown quantity is the $\boldsymbol{\psi}_{Y,4}$ functional. The fourth order differential $\mathsf{D}^{\otimes 4}$ is expressed as a vector of length d^4, resulting from a four-fold Kronecker

product of the first order differential D. Replacing this by an estimator $\hat{\boldsymbol{\psi}}_{Y,4}$ yields the plug-in criterion

$$\text{PI}(\mathbf{H}) = \tfrac{1}{4}m_2^2(K)(\text{vec}^\top \mathbf{H} \otimes \text{vec}^\top \mathbf{H})\hat{\boldsymbol{\psi}}_{Y,4}(\mathbf{G}) + n^{-1}R(K)|\mathbf{H}|^{-1/2}$$

where $m_2(K)$ is defined in Section 2.2, vec is the operator that stacks the element of a matrix column-wise into a vector, $\hat{\boldsymbol{\psi}}_{Y,4}(\mathbf{G}) = n^{-2}\sum_{i,j=1}^n \mathsf{D}^{\otimes 4}L_\mathbf{G}(\boldsymbol{Y}_i - \boldsymbol{Y}_j)$, $L_\mathbf{G}$ is an initial pilot kernel with pilot bandwidth matrix $\mathbf{G}$ and $R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{x})^2 \, d\boldsymbol{x}$. The plug-in selector $\hat{\mathbf{H}}_{\text{PI}}$ is the minimiser over $\mathcal{F}$ of $\text{PI}(\mathbf{H})$.

For the class of cross validation selectors we focus on unbiased (or least squares) cross validation and smoothed cross validation. Unbiased cross validation (UCV) was introduced by Bowman, Hall, and Titterington (1984) and Rudemo (1982) for the univariate case. The unbiased cross validation selector, $\hat{\mathbf{H}}_{\text{UCV}}$ for the multivariate case (Sain, Baggerly, and Scott 1994), is defined as the minimiser over $\mathcal{F}$ of

$$\text{UCV}(\mathbf{H}) = \int_{\mathbb{R}^d} \hat{f}_\boldsymbol{Y}(\boldsymbol{y};\mathbf{H})^2 \, d\boldsymbol{x} - 2n^{-1}\sum_{i=1}^n \hat{f}_{\boldsymbol{Y},-i}(\boldsymbol{Y}_i;\mathbf{H}),$$

where $\hat{f}_{\boldsymbol{Y},-i}(\boldsymbol{Y}_i;\mathbf{H}) = [n(n-1)]^{-1}\sum_{j=1}^n K_\mathbf{H}(\boldsymbol{Y}_i - \boldsymbol{Y}_j)$. The smoothed cross validation (SCV) selector $\hat{\mathbf{H}}_{\text{SCV}}$, is defined as the minimiser over $\mathcal{F}$ of

$$\text{SCV}(\mathbf{H}) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n (K_\mathbf{H} * K_\mathbf{H} * L_\mathbf{G} * L_\mathbf{G} - 2K_\mathbf{H} * L_\mathbf{G} * L_\mathbf{G} + L_\mathbf{G} * L_\mathbf{G})(\boldsymbol{Y}_i - \boldsymbol{Y}_j)$$
$$+ n^{-1}R(K)|\mathbf{H}|^{-1/2},$$

where $*$ is the convolution operator, as introduced by Hall, Marron, and Park (1992) for univariate data, and by Sain et al. (1994) for multivariate data. If there are no replications in the data, then SCV with $\mathbf{G} = 0$ is identical to UCV as the pilot kernel L_0 can then be thought of as the Dirac delta function.

The UCV selector can be directly computed as it contains no unknown quantities, however specification of the bandwidth $\mathbf{G}$ of the pilot kernel is required for the plug-in and SCV selectors. Computational data-based algorithms which address this are found in Wand and Jones (1995) and Duong and Hazelton (2003) for plug-in selectors and Hall et al. (1992) and Duong and Hazelton (2005) for SCV selectors.

2.4. Tail density estimation via histograms

Histograms, especially for univariate data, are widely used as alternatives to kernel estimators for visualising data samples, even when focussing on distributional tails (see e.g. Perkins, Pitman, Holbrook, and McAneney 2007; Perkins et al. 2013). Their advantages include computational and mathematical simplicity, and that they do not suffer from the boundary bias problems of standard kernel estimators. In the context of tail density estimation, we divide the data range of the sample $X_1, \ldots, X_n$ into a regular partition of hypercubes A_i of size $b_1 \times \cdots \times b_d$, and define the binwidth as $\boldsymbol{b} = (b_1, \ldots, b_d)^\top \in \mathbb{R}^d$.

The histogram estimator $\tilde{f}_X$ at a point x in a bin A_i is

$$\tilde{f}_X(x; b) = \frac{\gamma_i}{nb_1 \cdots b_d}$$

where γ_i represents the number of observations in the hypercube A_i. The histogram estimator of the tail density $\tilde{f}_{X^{[u]}}$ is

$$\tilde{f}_{X^{[u]}}(x; b) = \tilde{f}_X(x; b)/\tilde{\bar{F}}_X(u; b)$$

where $\tilde{\bar{F}}_X(u; b)$ counts the number of observations in the hypercubes covered by (u, ∞), divided by $nb_1 \cdots b_d$. If conditions similar to (A1) and (A3) in the Appendix are fulfilled then, by Scott (2015, Theorem 3.5), the MISE of the histogram estimator is $\text{MISE}\{\tilde{f}_{X^{[u]}}(\cdot; b)\} = O((nb_1 \cdots b_d)^{-1} + b^\top b)$ with minimal MISE $\inf_{b>0} \text{MISE}\{\tilde{f}_{X^{[u]}}(\cdot; b)\} = O(n^{-2/(d+2)})$. This is asymptotically slower than the $O(n^{-4/(d+4)})$ minimal MSE rate for the kernel estimator $\hat{f}_{X^{[u]}}(x)$ for x not in the boundary region. Hence, from the mean squared error perspective, the kernel density estimator is preferable to a histogram for density estimation in the tail region, especially as the dimension d increases.

Analogous with the data-based optimal bandwidth selectors, the normal scale optimal binwidth (Scott 2015, Theorem 3.5) is

$$\hat{b}_j = 2 \times 3^{1/(d+2)} \pi^{d/(d+4)} s_j n^{-1/(d+2)} \tag{3}$$

where $s_j, j = 1, \ldots, d$ are the marginal sample standard deviations of $X_1^{[u]}, \ldots, X_n^{[u]}$. There is no equivalent variety of binwidth selectors which generalise Equation (3) compared to bandwidth selectors (Section 2.3) due the slower asymptotic performance of histograms as compared to kernel estimators.

2.5. Model assessment via tail density estimation

Tail density estimation can provide one way to assess the fidelity of the observed dataset to one or more candidate models. For example, in climate science different climate models commonly produce competing predictions of environmental variables. The performance of these models is often validated by comparing the model predicted output, with that of the observed data (Flato et al. 2013). These comparisons may be based on the full body of the predicted variables, or focus primarily on the extremes (e.g. Perkins et al. 2007, 2013). Similarly, in the context of extreme value theory, the analyst is regularly required to determine which of multiple competing parametric families, such as max-stable distributions, provides the best fit to an observed extremal dataset (e.g. Coles and Tawn 1994).

Suppose that we have a suite of parametric models indexed by $\mathcal{M} = \{1, \ldots, M\}$, and we wish to determine which of them most appropriately describe the tails of the underlying distribution of the observed dataset, f_X. Perkins et al. (2013) utilised the histogram estimator $\tilde{f}_{X^{[u]}}$ of the observed data sample as a surrogate for the unknown target $f_{X^{[u]}}$, and so the fit of the parametric models was assessed according to the discrepancy of the parametric (tail) density functions $g_1, \ldots, g_M$ defined over (u, ∞) and the histogram $\tilde{f}_{X^{[u]}}$. Their tail

index (generalised here to d dimensions) is given by

$$\tilde{T}_1(g_j) = \int_{(\boldsymbol{u},\infty)} |g_j(\boldsymbol{x}) - \tilde{f}_{\boldsymbol{X}^{[u]}}(\boldsymbol{x};\boldsymbol{b})| \, d\boldsymbol{x},$$

with the preferred models being those which give the smaller or smallest discrepancy

$$\underset{j \in \mathcal{M}}{\operatorname{argmin}} \ \tilde{T}_1(g_j).$$

Note that the subscript of $\tilde{T}_1$ indicates the L_1 error measure used in its definition. We prefer to use the L_2 error to assess a model fit:

$$\tilde{T}_2(g_j) = \int_{(\boldsymbol{u},\infty)} [g_j(\boldsymbol{x}) - \tilde{f}_{\boldsymbol{X}^{[u]}}(\boldsymbol{x};\boldsymbol{b})]^2 \, d\boldsymbol{x}. \tag{4}$$

An improvement to this procedure is to replace the histogram in Equation (4) with the transformation kernel estimator $\hat{f}_{\boldsymbol{X}^{[u]}}$:

$$\hat{T}_2(g_j) = \int_{(\boldsymbol{u},\infty)} [g_j(\boldsymbol{x}) - \hat{f}_{\boldsymbol{X}^{[u]}}(\boldsymbol{x};\mathbf{H})]^2 \, d\boldsymbol{x}. \tag{5}$$

This will accordingly allow the usual artefacts of histogram estimators to be avoided or at least reduced. These include the anchor point problem (i.e. how to specify the locations of the histogram bins) and the empty bin problem (where it is unclear whether histogram bins with empty counts should be interpreted as a true zero probability or are due to insufficient observed data). This latter case is important for upper tail values as they are sparsely distributed in the tail regions.

In the following section, we highlight the purpose of working with transformed density estimators, by directly contrasting $\tilde{T}_2(g_j)$ and $\hat{T}_2(g_j)$ with the index based on the standard kernel density estimator

$$\hat{T}_2^*(g_j) = \int_{(\boldsymbol{u},\infty)} [g_j(\boldsymbol{x}) - \hat{f}_{\boldsymbol{X}^{[u]}}^*(\boldsymbol{x};\mathbf{H})]^2 \, d\boldsymbol{x}, \tag{6}$$

where $\hat{f}_{\boldsymbol{X}^{[u]}}^*(\boldsymbol{x})$ represents the standard kernel density estimator constructed without applying the transformation $\boldsymbol{t}$. As presented in the Introduction, in the univariate case, under the Pickands-Balkema-de Haan theorem, the observations above some high threshold u can be approximated by the GPD distribution. We thus define by $\check{f}_{X^{[u]}}$ the GPD tail density estimator constructed from the observations above the threshold u and respective tail index using the L_2 error by

$$\check{T}_2(g_j) = \int_{(u,\infty)} [g_j(x) - \check{f}_{X^{[u]}}(x)]^2 \, dx. \tag{7}$$

The integrals in Equations (4)–(7) can be approximated by (weighted) Reimann sums.

3. Numerical studies

3.1. Simulated data - univariate

We now numerically examine the performance of the kernel density estimator introduced in Section 2.1 for estimating upper tails, and demonstrate that it is a good surrogate for the true tail distribution. We additionally evaluate the estimator's performance through the model assessment procedure of Section 2.5. Gaussian kernels are adopted throughout given the usual secondary level of importance given to kernel choice in standard kernel methods (see e.g. Table 2.1 of Wand and Jones 1995 where the difference between the most efficient (Epanechnikov) kernel and the least efficient (uniform) kernel is less than 7%).

To examine a range of tail behaviours we generate a dataset of size $n = 2,000$ from each of the Gumbel, Fréchet and generalised Pareto (GPD) target distributions, and set the threshold u at the 95% upper sample quantile. For each sample, we compute:

(1) The appropriate maximum-likelihood based, parametric estimator: Fréchet (FRE), Gumbel (GUM) or generalised Pareto (GPD);
(2) The maximum likelihood estimator of the generalised Pareto distribution, $\check{f}_{X[u]}$, using only the observations above the threshold u (GPD+);
(3) The histogram estimator $\tilde{f}_{X[u]}$ with normal scale optimal binwidth (HIS);
(4) The transformation kernel based estimator $\hat{f}_{X[u]}$ with transformation $t(x) = \log(x - u_0)$, where $u_0 = \min(X_1, \ldots, X_n) - 0.05 \operatorname{range}(X_1, \ldots, X_n)$ using the normal scale (KNS), plug-in (KPI), unbiased cross validation (KUC) and smoothed cross validation (KSC) optimal bandwidth selectors;
(5) The standard kernel based estimators $\hat{f}^*_{X[u]}$, using the normal scale (KNS*), plug-in (KPI*), unbiased cross validation (KUC*) and smoothed cross validation (KSC*) optimal bandwidth selectors.

The top row of Figure 1 illustrates the various tail density estimates for the three target distributions, with the true density shown as a solid black line. Displayed are the generalised Pareto estimator $\check{f}_{X[u]}$ (GPD+; grey long-dashed), the histogram estimator $\tilde{f}_{X[u]}$ (HIS; green dotted line), and the transformed $\hat{f}_{X[u]}$ (KPI; red dashed) and standard $\hat{f}^*_{X[u]}$ (KPI*; blue dot-dashed) kernel density estimates, both with plug in estimators only for clarity. Visually, the transformed kernel estimators appear to be the more accurate nonparametric estimators of tail behaviour in each case, being noticeably smoother and less noisy. That the kernel density estimators are naturally continuous functions also leads to better visualisations than histogram based estimators, and they are more helpful when comparing to a continuous target density.

Both the transformed kernel and the GPD (using the largest observations only) estimators appear to provide comparable fit in each case.

The bottom row of Figure 1 examines the extremal performance of the same estimators through qq-plots of the target quantiles versus the GPD+ and nonparametric estimated quantiles, for target quantiles ranging from 95% to 99.9%. Of all nonparametric estimators, the histogram estimator most consistently approximates the true quantiles. This performance compared to the kernel-based estimators is not unexpected, however, as the latter
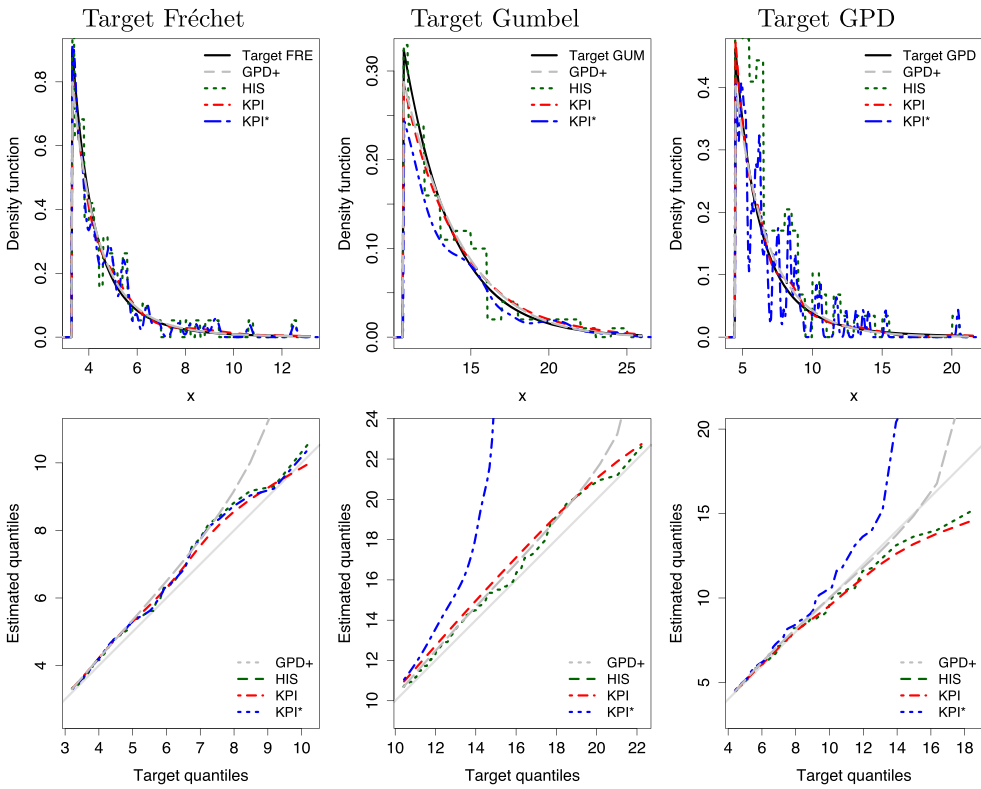
Figure 1. Generalised Pareto estimator $\check{f}_{\chi[u]}$ (GPD+; grey long-dashed) and nonparametric estimators of the univariate tail density (top) and of the tail quantiles (bottom) when the target density is Fréchet (left), Gumbel (centre) and generalised Pareto (right). Sample size is $n = 2000$. Fréchet ($\mu = 1, \sigma = 0.5, \xi = 0.25$), Gumbel ($\mu = 1.5, \sigma = 3$) and Pareto ($\mu = 0, \sigma = 1, \xi = 0.25$) target densities are represented by a solid black line. The histogram estimator $\tilde{f}_{\chi[u]}$ with normal scale binwidth (HIS) is represented by a dotted green line, the transformed kernel plug-in estimator $\hat{f}_{\chi[u]}$ (KPI) by a short dashed red line and the standard kernel estimator $\hat{f}^*_{\chi[u]}$ (KPI*) by a dot-dash blue line.

aim to optimally estimate the density function rather than the quantile function. Comparing the two kernel-based estimators, the transformed kernel estimator tends to either outperform (centre, right panels) or perform as well as (left panel) the standard estimator, which can be attributed to the standard estimator's natural boundary bias. The tail quantiles obtained from the transformed kernel estimators appear to perform better than those of the generalised Pareto estimator (GPD+) when the target density is Fréchet, and they are comparable for the Gumbel target. Unsurprisingly, the GPD+ estimator performs the strongest when the data are in fact GPD distributed.

Note that the transformed kernel estimator has produced estimates with lighter tails than the true density. For large n this is possibly due to the choice of a Gaussian kernel K_h to construct the density estimates, so that the upper tail of this estimate (mapped through the inverse transform t^{-1}) is light compared to the true Fréchet and Pareto tails. For smaller n, finite sample variation can produce a density estimate with either lighter or heavier tails in the body of the data (see Supplementary Information).
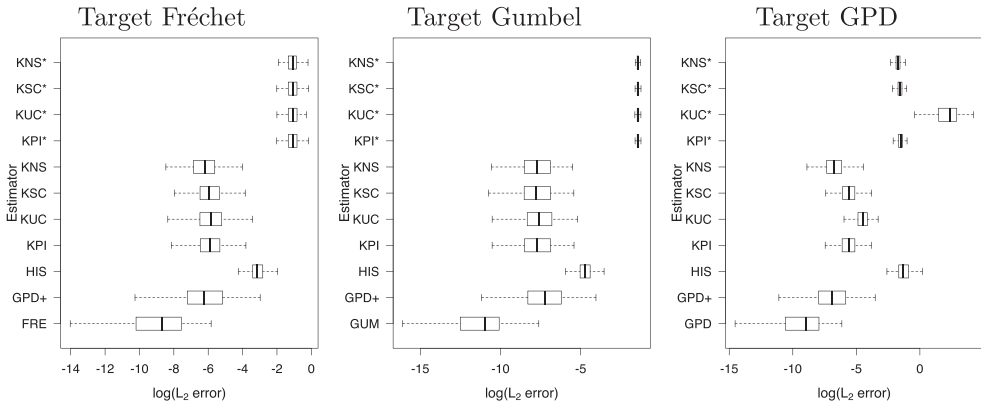
Figure 2. Box-plots of the log L_2 errors for the parametric Fréchet (FRE), Gumbel (GUM), generalised Pareto (GPD), and histogram (HIS) tail density estimators as well as the generalised Pareto $\check{f}_{\chi[u]}$ (GPD+). Transformed kernel density estimators $\hat{f}_{\chi[u]}$ use the plug-in (KPI), unbiased cross validation (KUC), smoothed cross validation (KSC) and normal scale kernel (KNS) optimal bandwidth selectors. Standard kernel density estimators $\hat{f}^*_{\chi[u]}$ are indicated by an asterisk (*). True target densities are (left panel) Fréchet, (centre) Gumbel and (right) GPD. Box plots are based on 400 replicates of $n = 2,000$ observations.

For more quantitative results we repeat this process over 400 replicates for different dataset sizes $n = 500, 1000$ and 2000, producing tail samples of size $m = 25, 50$ and 100, with the threshold u set at the 95% upper quantile. As these three sample sizes gave similar results, we only present those for $n = 2000$ here for brevity. See the Supplementary Information for results with $n = 500, 1000$. We take a numerical approximation (Reimann sum) of the L_2 loss ($\tilde{T}_2$, $\hat{T}_2$, $\hat{T}^*_2$ and $\check{T}_2(g_j)$).

Figure 2 presents box-plots of the accuracy of each tail density estimator for each true tail distribution. As expected, for each target distribution the most accurate density estimator is the correctly specified parametric model. The transformed kernel density estimators systematically perform better than their standard kernel counterparts, although they can be more variable. The standard kernel estimators and the histogram estimator compete for the worst estimate of the tail density, depending on the true target distribution. The differences in the accuracy between kernel estimators with different bandwidth selectors is small, in contrast to studies where the bandwidth selection class is a crucial factor (see e.g. Sheather and Jones 1991; Wand and Jones 1995, Chapter 3), indicating that it is the difference between estimators that is dominating performance. Using the normal scale bandwidth selector (KNS) provides a greater accuracy compared to other bandwidth selectors when the target distribution is GPD, and this is also slightly evident for Fréchet distributed data. The best transformed kernel and the GPD tail density estimators (KNS and GPD+) appear to perform equally well when the target distribution is Fréchet or Gumbel, with a slight advantage to the KNS estimator in the case of Gumbel distributed data. Clearly the GPD+ estimator is over-performing when the target is GPD.

Finally, we examine the density estimator performance in terms of its ability to correctly select the true, data-generating model (Section 2.5). For each of the 3×400 datasets generated previously, we compute the tail indices of the L_2 loss ($\tilde{T}_2$, $\hat{T}_2$, $\hat{T}^*_2$ and $\check{T}_2(g_j)$) with respect to each parametric model.

Table 1. Proportion of 400 simulated datasets from each known target distribution (Fréchet, Gumbel and GPD) that are correctly identified as coming from each of these distributions by having the smallest tail index value. Bold text indicates the highest proportion for each target model. Nonparametric density estimators are the histogram ($\tilde{T}_2$), the transformed kernel ($\hat{T}_2$) and the standard kernel ($\hat{T}_2^*$). The parametric GPD estimator on tail data is $\check{T}_2$. Tail indices are calculated according to the L_2 loss.

Target	$\tilde{T}_2$	$\hat{T}_2$	$\hat{T}_2^*$	$\check{T}_2$
FRE	0.74	0.89	0.80	**0.92**
GUM	**1.00**	**1.00**	0.00	**1.00**
GPD	0.19	**0.98**	0.00	0.95

Table 1 displays the proportion of times that samples from a given true distribution are identified as coming from either Fréchet, Gumbel or GPD distributions (i.e. by having the smallest tail index value), as a function of tail density estimator. In each case, the highest proportion of replicates selecting the correct model is given in bold. As the Gumbel distribution ($\xi = 0$) is on the limiting border of the parameter space of the Fréchet distribution ($\xi > 0$), to avoid possible model misidentification, we additionally perform a deviance test. If the Gumbel provides a significantly better fit than the Fréchet distribution, meaning that the shape parameter is not significantly different from zero, then we only consider Gumbel and GPD distributions as candidate models.

As might be expected, for any target distribution in Table 1, using the transformation based estimator ($\hat{T}_2$) as a surrogate for the target density generally selects the correct target in the vast majority of cases, with proportions substantially higher than those achieved through the standard kernel and histogram tail indices, $\hat{T}_2^*$ and $\tilde{T}_2$. Both the transformation and GPD based estimators ($\hat{T}_2$ and $\check{T}_2$) have comparable abilities in correctly identifying the underlying distribution with proportions around 0.90 and higher. When the true density is Fréchet, the best performing estimator determined by the L_2 error measure favours $\check{T}_2$ while it favours $\hat{T}_2$ when the target is GPD. When this study is repeated with smaller sample sizes ($n = 1{,}000$ and 500), the superiority of the transformation-based tail index compared to the GPD-based tail index is clearer (see Tables A & B in the Supplementary Information). Overall, the transformation kernel-based index $\hat{T}_2$ performs as strongly as, and in some cases better than the GPD-based index $\check{T}_2$ and consistently better than the nonparametric-based indices $\hat{T}_2^*$ and $\tilde{T}_2$.

3.2. Simulated data - multivariate

The analysis of multivariate upper tail values is considerably more challenging than its univariate counterpart. Within extreme value theory, a powerful motivation for exploratory data analysis using kernel-based estimation is that no single parametric family exists for max-stable distributions. See e.g. Kotz and Nadarajah (2000), Coles (2001), Beirlant, Goegebeur, Teugels, and Segers (2004), de Haan and Ferreira (2006), Falk, Hüsler, and Reiss (2011) and Beranger and Padoan (2015) for theoretical details and applications. Although a multivariate extension of the GPD distribution is available (see for example, Rootzén and Tajvidi 2006; Rootzén, Segers, and Wadsworth 2017) we do not consider it here, as its principles are based on at least one marginal component exceeding some high threshold rather than considering all components to be above a threshold, which is our focus here.
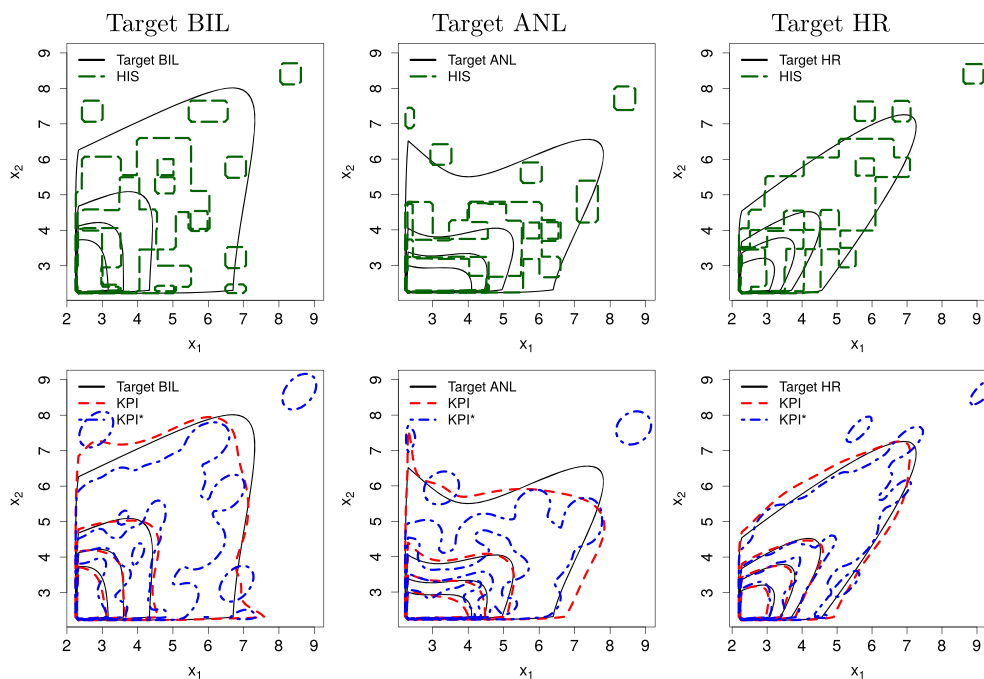
Figure 3. Nonparametric estimators of the bivariate tail density when the target density is bilogistic (BIL), asymmetric negative logisitic (ANL) and Hüsler-Reiss (HR). Sample size is $n = 4000$. Bilogistic ($\alpha = 0.8$, $\beta = 0.52$), asymmetric negative logistic (dependence parameter $= 1.3$, asymmetry parameter $= (0.2, 0.7)$) and Hüsler-Reiss (dependence parameter $= 2.4$) target quantiles are represented by a solid black line. [Top panels] The histogram estimator $\tilde{f}_{X^{[u]}}$ with a normal scale bin width (HIS) is represented by a long dashed green line, [bottom panels] the transformation kernel estimator $\hat{f}_{X^{[u]}}$ with plug-in bandwidth estimator (KPI) by the short dashed red line and the standard kernel estimator $\hat{f}^*_{X^{[u]}}$ with plug-in bandwidth estimator (KPI*) by the dot-dashed blue line.

We now numerically examine the performance of the bivariate transformation-based density estimator for a range of upper tail behaviours. We generate datasets of size $n = 4,000$ from the asymmetric negative logistic (ANL; Joe 1990), the bilogistic (BIL; Smith, Tawn, and Yuen 1990) and the Hüsler-Reiss (HR; Hüsler and Reiss 1989) distributions. The threshold $\boldsymbol{u}$ is determined as each dataset's marginal 90% upper quantiles.

For each dataset we compute the appropriate maximum likelihood based parametric estimator (assuming simultaneously estimated generalised extreme value distribution margins), a 2-dimensional histogram with normal scale optimal bandwidth (HIS) and the transformation and standard kernel estimators with plug-in optimal bandwidth selectors (respectively KPI and KPI*) and transformation $\boldsymbol{t}(\boldsymbol{x}) = (\log(x_1 - u_{01}), \log(x_2 - u_{02}))^\top$, where $u_{0j} = \min\{X_{1j}, \ldots, X_{nj}\} - 0.05\,\mathrm{range}\{X_{1j}, \ldots, X_{nj}\}$, $j = 1, 2$. The results from the other bandwidth selectors are not displayed both for clarity, and due to the limited impact of the bandwidth selector method on the performance of the density estimator.

Figure 3 illustrates the 25%, 50%, 75% and 99% highest density level sets of the histogram (long dashed green line, top panels), transformed kernel density estimates (dashed red line, bottom panels) and standard kernel density estimates (dot-dashed blue line,
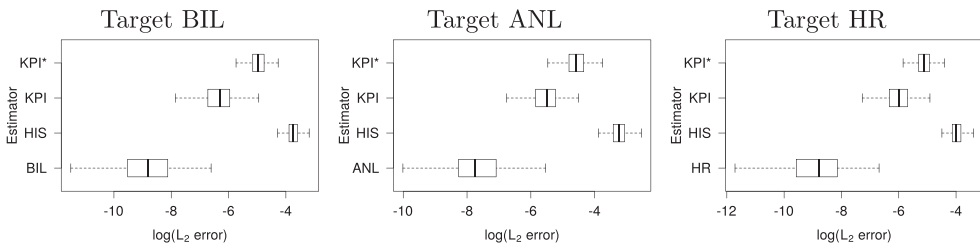
Figure 4. Box-plots of the $\log L_2$ errors for the bivariate bilogistic (BIL), asymmetric negative logistic (ANL) and Hüsler-Reiss (HR) parametric estimators, the 2-dimensional histogram (HIS), and the bivariate transformation (KPI) and standard (KPI*) kernel estimator with plug-in optimal bandwidth selector. True target densities are (left) the bivariate biologistic, (centre) the asymmetric negative logistic and (right) the Hüsler-Reiss models. Box plots are based on 400 replicates of $n = 4,000$ observations.

bottom panels) in comparison with the target distribution (solid black line). Visually, the transformed kernel estimator performs extremely well – it is able to identify and describe most of the features of the target densities as its contours follow the target contours very closely. In contrast, the blocky, discrete nature of the histogram estimator makes it difficult to discern the nature of the underlying target, and the standard kernel estimator is clearly unable to capture the features of the tail density as accurately as the transformation kernel estimator, displaying many spurious bumps in the tail.

Similarly to Figure 2, Figure 4 report the $\log L_2$ performance of the parametric (BIL, ANL, HR), histogram (HIS), and transformation and standard kernel (KPI and KPI*) tail density estimators in approximating the known target distribution, based on 400 replicate simulations, for each of the target distributions considered in Figure 3. As for the univariate case, the correctly specified parametric estimator of each distribution generates the smallest error. The transformation kernel density estimator produces the next most efficient estimator, with the bivariate histogram and the bivariate standard kernel performing the most poorly in each case.

Finally, we examine the ability of the density estimator to correctly select the true data generating model. Similarly to Table 1, Table 2 show the proportion of times that each model was selected based on the bivariate histogram ($\tilde{T}_2$), transformation kernel ($\hat{T}_2$) and standard kernel $\hat{T}_2^*$ based tail indices using data generated from a known model, where the parametric fitted models $g_j(x)$ are each of BIL, ANL and HR. The bold figures indicate the estimator most often correctly selecting each target model.

In contrast to the univariate analyses, the bivariate results are mixed. The standard kernel based index appears to be able to choose the correct model slightly more consistently than the histogram based index. However, while the transformation density estimator outperforms the standard kernel based index for Hüsler-Reiss distributed data, it underperforms in other circumstances, particularly for ANL data. In general it seems that the best performing estimator for model selection is dataset dependent.

A more detailed examination of these results reveals that the transformation-based estimator is clearly the best performer in terms of its ability to estimate the true density precisely. Table 3 presents the mean L_2 error when fitting each model to data generated under each of the BIL, ANL and HR models, taken over 400 replicate datasets. For any true model and fitted model (i.e. any row in Table 3), the transformation kernel density

Table 2. Proportion of 400 simulated datasets from each known target distribution (BIL bilogistic, ANL asymmetric negative logistic, and HR Hüsler-Reiss) that are identified as coming from each of these distributions by having the smallest tail index value, as a function of nonparametric density estimator. Bold text indicates the highest proportion for each target model. Nonparametric density estimators are the bivariate histogram ($\tilde{T}_2$), the transformed kernel ($\hat{T}_2$) and the standard kernel ($\hat{T}_2^*$).

Target	$\tilde{T}_2$	$\hat{T}_2$	$\hat{T}_2^*$
BIL	0.69	0.62	**0.74**
ANL	0.83	0.15	**0.85**
HR	0.89	**0.99**	0.81

Table 3. Mean L_2 errors of the nonparametric estimators for 400 simulated datasets from each known true target distribution (BIL bilogistic, ANL asymmetric negative logistic, and HR Hüsler-Reiss), compared to each parametric fitted model. Nonparametric density estimators are the bivariate histogram ($\tilde{T}_2$), the transformed kernel ($\hat{T}_2$) and the standard kernel ($\hat{T}_2^*$). Bold text highlights the minimum L_2 error for each estimator, indicating the fitted model most often selected.

True Model	Fitted Model	$\tilde{T}_2$	$\hat{T}_2$	$\hat{T}_2^*$
	BIL	**0.024**	**0.002**	**0.007**
BIL	ANL	0.026	0.003	0.009
	HR	0.030	0.004	0.012
	BIL	0.045	**0.002**	0.015
ANL	ANL	**0.040**	0.004	**0.011**
	HR	0.060	0.012	0.030
	BIL	0.021	0.007	0.008
HR	ANL	0.022	0.007	0.009
	HR	**0.019**	**0.003**	**0.006**

estimate provides the most accurate density estimate (on average). This echoes the high performance findings for our density estimator in Figure 4.

For a given true model, and for a specified density estimator, the bold figure indicates the fitted model that is chosen most often (on average) in terms of minimising the L_2 error. Thus for e.g. BIL data, the BIL model is likely to be selected most often, regardless of the choice of density estimator. (Note that as these are mean values, there is some overlap of the distribution of L_2 errors within each density estimator, which ultimately produces the proportions observed in Table 2.) This is also the case, on average for HR data tending to choose the HR model most often for each estimator. However, for ANL data, the transformation density estimator $\hat{T}_2$ will select the BIL model most often (resulting in the low 0.15 correct classification rate in Table 2), even though it is by far the better estimator of the ANL density (with a mean L_2 score of 0.004, compared to 0.040 and 0.011), simply because this estimator is also a slightly closer match to the fitted BIL model in this case. In general this suggests that while the transformation based kernel density estimator clearly outperforms both the standard kernel and histogram based density estimators in terms of the quality of the tail density estimation, care should be taken when using these estimators in a model selection scenario, particularly for models in more than one dimension.

4. Exploratory data analysis of climate models

Perkins et al. (2007, 2013) previously used univariate histogram density estimators for both visualisation and model selection to evaluate the ability of global climate models (GCMs) to simulate extreme temperatures (minima and maxima) over Australia. The models which they considered are the climate models assessed by the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) to investigate changes in temperature extremes. A well-known challenge for these models is to be able to accurately project extreme temperatures (Perkins et al. 2007, 2013; Fischer, Beyerle, and Knutti 2013; Sillmann, Kharin, Zhang, Zwiers, and Bronaugh 2013; Sillmann, Kharin, Zwiers, Zhang, and Bronaugh 2013; Cowan, Purich, Perkins, Pezza, Boschat, and Sadler 2014).

Following earlier work, Perkins et al. (2013) developed a univariate tail index (see Section 2.5, Equation (4)) which evaluates the amount of overlap between a model-predicted distributional tail, g_i, and the distribution of the observed extreme data. This index reflects the discrepancy between two distributional tails, whereby a model perfectly fitting the observed data has zero score, and increasing scores imply an increasing lack-of-fit of the model to the observed data. Unlike for the simulated parametric models in Section 3.1 and 3.2, there is no closed form for the density function g_i, $i = 1, \ldots, M$, to characterise the data values generated by the climate models. Perkins et al. (2013) replaced the unknown target density g_i with a histogram $\tilde{g}_i$, based on model generated data, when comparing to the histogram of the observed data $\tilde{f}_{X[u]}$ in Equation (4), i.e. they used the index $\tilde{T}_1(\tilde{g}_i)$ to determine the most appropriate model. Because of this difference with the model selection analysis in Sections 3.1 and 3.2, there is reason to believe that this procedure is more reliable in model selection terms, as the comparison is between two data-based tail density estimators, and it is accordingly likely that the better the density estimator, the more credible the comparison between the two datasets will be.

We extend this previous histogram estimator-based analysis by considering a wider and more modern ensemble of global climate models than those in Perkins et al. (2013), as well as exploring alternatives to $\tilde{T}_1(\tilde{g}_i)$ as the model selection criterion. Here we use $M = 22$ climate models participating in the World Climate Research Programme's 5th phase Coupled Model Intercomparison Project (CMIP5; see Flato et al. 2013), which currently underpin global and regional climate projections of extremes (e.g. Sillmann et al. 2013). The choice of models was based on the availability of daily maximum and minimum temperature data for the historical experiment (∼ 1860–2005; see Taylor, Stouffer, and Meehl 2012). Other targeted temperature extreme evaluation studies on the CMIP5 ensemble have found generally well-simulated changes in observed trends of specific indices (e.g. Flato et al. 2013; Sillmann et al. 2013). However unlike this study, no consideration has been given to the full underlying distribution of tail values.

The observed data sample are the daily observed maximum temperatures for Sydney, Australia, from 01/01/1911 to 31/12/2005 yielding a sample of $n = 34,699$ observations. Observations were obtained from the Australian Water Availability Project dataset (Jones, Wang, and Fawcett 2009), a gridded product covering all of Australia. All AR4 climate models were run to generate data in this same time frame, and the GCM grid box in which Sydney is located was extracted. The threshold determining the extreme maximum temperatures is the 95% upper quantile $u = 30.98°$C. Additionally, note that the climate models are physical, not statistical, and run their own climate. Hence when run for long

Table 4. Univariate histogram- $\tilde{T}_2(\tilde{g}_i)$, kernel- $\hat{T}_2(\hat{g}_i)$ and GPD-based $\check{T}_2(\check{g}_i)$ tail index scores, based on histogram $\tilde{g}_i$, kernel $\hat{g}_i$ and GPD $\check{g}_i$ density estimators, for the moderately extreme maximum temperatures produced by the twenty-two AR4 climate models. The models displayed are the ten best performing models in one dimension. Bold figures indicate the four best performing models under each model selection index.

Model	Model selection index		
	$\tilde{T}_2(\tilde{g})$	$\hat{T}_2(\hat{g})$	$\check{T}_2(\check{g})$
CanESM2	0.0042	**0.0006**	0.0015
CMCC-CESM	0.0055	0.0039	0.0009
CMCC-CM	0.0053	0.0031	0.0011
CNRM-CMS	**0.0033**	**0.0005**	**0.0003**
HadGEM2-CC	0.0060	0.0036	**0.0004**
HadGEM2-ES	0.0039	0.0018	**0.0002**
MIROC5	**0.0037**	0.0009	0.0020
MPI-ESM-LR	**0.0029**	**0.0003**	0.001
MPI-ESM-MR	**0.0018**	**0.0002**	**0.0005**
MPI-ESM-P	0.0063	0.0030	0.0063

enough their properties of non stationarity are very clear. Furthermore, they are forced via anthropogenic climate emissions, which induce a highly non stationary climate.

Table 4 displays the modified Perkins et al. (2013) histogram-based tail indices, $\tilde{T}_2(\tilde{g}_i)$, the transformation kernel density estimator based index, $\hat{T}_2(\hat{g}_i)$ and the GPD based tail index $\check{T}_2(\check{g}_i)$ for ten out of the 22 models. Note that $\hat{T}_2(\hat{g}_i)$ implements the transformation kernel estimator for both the observed data ($\hat{f}_{X[u]}$) and the GCM generated data (within $\hat{T}_2$). The bold figures indicate the four best performing models (out of 22) for each tail index. In this one-dimensional analysis, both histogram- and transformation kernel density-based estimators strongly identify the same two models (i.e. with lowest tail index): MPI-ESM-LR and MPI-ESM-MR, as best describing the observed univariate extremes. All three tail indices share models MPI-ESM-MR and CNRM-CMS in their top four best models to simulate moderate extremes. The three tail indices also have eight models in common out of their top ten.

Figure 5 illustrates both the tail density estimators and qq-plots for the common top performing models across all tail indices. The histogram-, kernel- and GPD-based estimates are represented by the solid, dashed and dotted lines respectively, whereas the observed and GCM data are denoted by black and grey lines. For both models (CNRM-CMS and MPI-ESM-LR), each of the three density estimates of the simulated data closely follow their respective density estimator of the observed data. The quality of the density estimates is also evident in the qq-plots. Here the smoother transformation kernel density estimator (dashed lines) is able to find a better match between observed and GCM model data than the histogram (solid lines) for both models, with MPI-ESM-MR providing a better overall fit (in particular for large quantiles). The quality of the fit provided by the transformation kernel and GPD density estimators appears to be very similar.

For a bivariate analysis, we consider the upper tail of pairs of maximum and minimum temperatures over the same time period, in order to investigate which of the climate models can best predict joint upper tail behaviour. (The largest minimum temperatures are important in understanding the duration and severity of heatwaves.) The threshold for the maximum temperatures are the 90% marginal upper quantiles $\boldsymbol{u} = (28.77^\circ\text{C}, 18.07^\circ\text{C})^\top$.
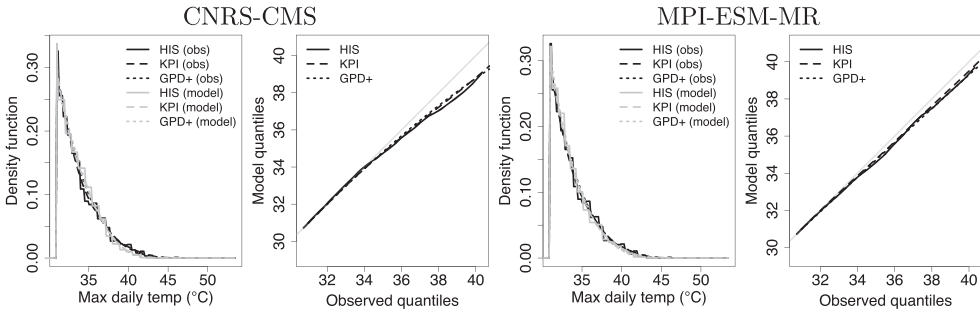
Figure 5. Histogram ($\tilde{f}_{\chi^{[u]}}$), transformation kernel ($\hat{f}_{\chi^{[u]}}$) and GPD ($\check{f}_{\chi^{[u]}}$) based estimators of the tail densities for two of the best AR4 models: (left to right) CNRM-CMS and MPI-ESM-MR. Histogram estimators (HIS) are denoted by solid lines, kernel plug-in estimators (KPI) by dashed lines and GPD estimators (GPD+) by dotted lines. Observed data (obs) is illustrated in black and GCM data (model) in grey.

Table 5. Bivariate histogram- $\tilde{T}_2(\tilde{g}_i)$ and transformation kernel-based $\hat{T}_2(\hat{g}_i)$ tail index scores, based on histogram $\tilde{g}_i$ and kernel $\hat{g}_i$ density estimators, for the extreme (minimum, maximum) temperatures produced by the twenty-two AR4 climate models. The models displayed are the ten best performing models. Bold figures indicate the four best performing models under each model selection index.

	Model selection index	
Model	$\tilde{T}_2(\tilde{g})$	$\hat{T}_2(\hat{g})$
CMCC.CM	0.0149	0.0092
CNRM-CMS	**0.0076**	**0.0039**
HadCM3	0.0123	0.0066
HadGEM2.ES	0.0133	0.0081
IPSL-CM5A-LR	0.0123	0.0060
IPSL-CM5B-LR	**0.0079**	**0.0041**
MIROC5	0.0124	0.0070
MPI-ESM-LR	**0.0096**	**0.0048**
MPI-ESM-MR	0.0106	0.0059
MPI-ESM-P	**0.0083**	**0.0040**

Table 5 presents the same information as Table 4 but for the bivariate data (without the GPD-based tail index).

Here, both model selection indices selecting the same best four models CNRMS-CMS, IPSL-CM5B-LR, MPI-ESM-LR and MPI-ESM-P. Two of these were already identified in Table 4 for their ability to simulate univariate temperature upper tail behaviour in comparison with the observed data. In particular, the CNRM-CMS model is clearly identified by both indices, achieving the lowest tail index scores, and has, along with the MPI-ESM-P model, the best ability to simulate bivariate upper tail temperatures.

Figure 6 illustrates the bivariate tail density estimators for the CNRM-CMS and MPI-ESM-P models, two of the top-performing bivariate models. Similarly to Figure 5, histograms are shown by solid lines, and transformation kernel density estimators by dashed lines. Observed and GCM data are represented by black and grey lines respectively. It is immediately apparent that the kernel-based density estimates are visually much cleaner,
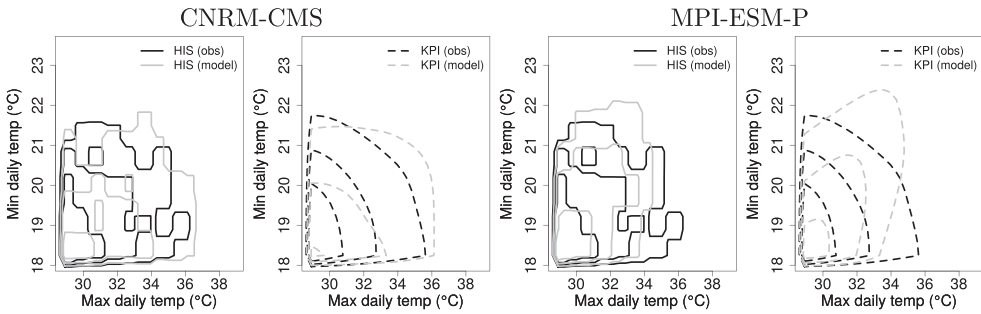
Figure 6. Histogram (left panel) and kernel estimators (right panel) of the tail densities for the CNRM-CMS and MPI-ESM-P climate models. Histograms estimators (HIS) are denoted by solid lines and kernel plug-in estimators (KPI) by dashed lines. Observed data (obs) is illustrated with black lines and climate model data (model) by grey lines. Kernel estimator contours indicate the 25%, 50% and 75% highest density level sets.

and easier to evaluate than their histogram counterparts. In particular, it is immediate that the CNRM-CMS model is a visually better match to the observed data than MPI-ESM-P.

The top performing CNRM-CMS model appears to simulate the bivariate temperature upper joint tail quite well, although the second best performing MPI-ESM-P model visually performs much more poorly. This indicates two possibilities. That the kernel estimators need to be further refined at the boundary or, perhaps more likely, that physical parameters in the GCMs need to be revised for a more realistic simulation of minimum temperature extremes. While minimum temperatures are physically simpler for a climate model to simulate than maximum temperature (Perkins et al. 2007), the mis-representation of the observed temperature distribution is a well-known issue for GCMs, which is at least in part explained by their coarse resolution (Seneviratne et al. 2012). For example, dynamically downscaled regional climate models that are run at finer resolutions for a limited spatial domain can offer some improvement in the simulation of extreme temperatures (e.g. Seneviratne et al. 2012; Vautard et al. 2013; Perkins, Moise, Whetton, and Katzfey 2014).

The goal of this exploratory data analysis is to propose feasible geophysical models which adequately describe the observed temperature data maxima and minima. As these geophysical models are expressed as a set of differential equations, their overall statistical properties are not well-known. Our estimates and visualisations of the tail densities of these geophysical models are a first step in elucidating their statistical properties, upon which more sophisticated data analysis can be subsequently applied.

5. Discussion

Nonparametric density estimation is a useful exploratory data analysis tool for the analysis of the tail behaviour of an observed process. In this article we have introduced a nonparametric kernel estimator for the analysis of the tail density of univariate and multivariate data by applying a logarithm transformation to account for the heavy tails and boundedness of distributional tail samples. Our proposed tail density estimator does not suffer from the usual boundary problems associated with kernel estimators. It is also robust in terms of the choice of the high threshold value u.

Our theoretical results (centred on Theorem 2.1) indicate the good performance of this transformation kernel density estimator in the upper tails. Numerical illustrations of its performance were given in Sections 3 and 4. This tail density estimator provides visually useful representations of upper tail sample behaviour compared to, say, histogram estimators – consider the contrast in visual clarity between the histogram and kernel estimators illustrated in Figure 6. Furthermore, it can reliably be incorporated into existing diagnostic and performance measures, such as the tail index of Perkins et al. (2013).

There is, of course, scope for further development and analysis of these ideas. As our proposed tail density estimator is decoupled from the threshold estimation, a promising avenue for amelioration would be the inclusion of more sophisticated threshold estimators than the simple quantile thresholds we have utilised.

Throughout we have constructed our kernel density estimates based on Gaussian kernels. In general, this means that the tail behaviour of the kernel-based density estimators is necessarily the tail behaviour of the kernel K mapped through the inverse of the transformation t. This implies that if the tail behaviour of the sample does not correspond to that of the transformed kernel, then the kernel density estimate will poorly represent the true behaviour of the observed data distribution far beyond the range of the observed data. A natural approach to resolving this problem could be to adapt the form of K and t to directly correspond to the (estimated) tail behaviour of the observed sample.

Additional improvements could be obtained by incorporating a local polynomial adjustment (e.g. Geenens 2014) to the boundary to improve over the transformation kernel approach, although here our primary interest is in the behaviour of the upper tail. Similarly, while the logarithm transformation is widely used due it conveniently mapping a semi-infinite interval to the real line, alternative transformations could be considered. Possibilities include the shifted power family of Wand et al. (1991)

$$t(x) = \begin{cases} (x + \lambda_1)^{\lambda_2} \, \mathrm{sign}(\lambda_2) & \lambda_2 \neq 0 \\ \log(x + \lambda_2) & \lambda_2 = 0 \end{cases}$$

where the log-transformation fixes $\lambda_1 = 0, \lambda_2 = -u_0$, and so other values may lead to better estimation, and the richer family of transformations proposed by Wand et al. (1991) and those posited by Geenens (2014).

Finally, the performance of this transformation-based kernel density estimator is limited by the performance of standard kernel density estimator methods. In particular, its performance will decline as the dimension of the random vector X increases. While this is unavoidable, if one wishes to perform kernel density estimation in this setting, it is important that it is implemented as efficiently as possible. The results presented in this article provide one step towards achieving this.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

ORCID

B. Béranger http://orcid.org/0000-0002-7944-3925

S. E. Perkins-Kirkpatrick http://orcid.org/0000-0001-9443-4915

S. A. Sisson http://orcid.org/0000-0001-8943-067X

References

Abramson, I.S. (1982), 'Bandwidth Variation in Kernel Estimates-A Square Root Law', *The Annals of Statistics*, 10, 1217–1223.

Balkema, A.A., and de Haan, L. (1974), 'Residual Life Time at Great Age', *The Annals of Probability*, 2, 792–804.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004), *Statistics of Extremes: Theory and Applications*, Chichester: John Wiley & Sons.

Beranger, B., and Padoan, S. (2015), 'Extreme Dependence Models', in *Extreme Value Modeling and Risk Analysis*, eds. Dipak K. Dey, and Jun Yan, New York: Chapman and Hall/CRC, pp. 325–352.

Bolancé, C., Guillén, M., Gustafsson, J., and Nielson, J.P., *Quantitative Operational Risk Models*, Chapman & Hall/CRC Finance Series, Boca Raton, FL: CRC Press, 2012.

Bolancé, C., Guillen, M., and Nielsen, J.P. (2003), 'Kernel Density Estimation of Actuarial Loss Functions', *Insurance: Mathematics and Economics*, 32, 19–36.

Bolancé, C., Guillen, M., Pelican, E., and Vernic, R. (2008), 'Skewed Bivariate Models and Nonparametric Estimation for the CTE Risk Measure', *Insurance: Mathematics and Economics*, 43, 386–393.

Bowman, A.W., Hall, P., and Titterington, D.M. (1984), 'Cross-validation in Nonparametric Estimation of Probabilities and Probability Densities', *Biometrika*, 71, 341–351.

Buch-Kromann, T., Guillén, M., Linton, O., and Nielsen, J.P. (2011), 'Multivariate Density Estimation using Dimension Reducing Information and Tail Flattening Transformations', *Insurance: Mathematics and Economics*, 48, 99–110.

Buch-Larsen, T., Nielsen, J.P., Guillén, M., and Bolancé, C. (2005), 'Kernel Density Estimation for Heavy-tailed Distributions using the Champernowne Transformation', *Statistics*, 39, 503–516.

Chacón, J.E., and Duong, T. (2010), 'Multivariate Plug-in Bandwidth Selection with Unconstrained Bandwidth Matrices', *Test*, 19, 375–398.

Charpentier, A., and Flachaire, E. (2015), 'Log-Transform Kernel Density Estimation of Income Distribution', *L'Actualité Économique*, 91, 141–159.

Chen, S.X. (1999), 'Beta Kernel Estimators for Density Functions', *Computational Statistics & Data Analysis*, 31, 131–145.

Cheng, R.C.H., and Amin, N.A.K. (1983), 'Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin', *Journal of the Royal Statistical Society. Series B* , 45, 394–403.

Christopeit, N. (1994), 'Estimating Parameters of an Extreme Value Distribution by the Method of Moments', *Journal of Statistical Planning and Inference*, 41, 173–186.

Clements, A., Hurn, S., and Lindsay, K. (2003), 'Mobius-like Mappings and Their Use in Kernel Density Estimation', *Journal of the American Statistical Association*, 98, 993–1000.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, London: Springer-Verlag.

Coles, S.G., and Tawn, J.A. (1994), 'Statistical Methods for Multivariate Extremes: An Application to Structural Design', *Journal of the Royal Statistical Society. Series C* , 43, 1–48.

Cowan, T., Purich, A., Perkins, S., Pezza, A., Boschat, G., and Sadler, K. (2014), 'More Frequent, Longer, and Hotter Heat Waves for Australia in the Twenty-First Century', *Journal of Climate*, 27, 5851–5871.

de Carvalho, M., Oumow, B., Segers, J., and Warchoł, M. (2013), 'A Euclidean Likelihood Estimator for Bivariate Tail Dependence', *Communications in Statistics - Theory and Methods*, 42, 1176–1192.

de Haan, L., and Ferreira, A. (2006), *Extreme Value Theory: An Introduction*, New York: Springer.

Drees, H., and Huang, X. (1998), 'Best Attainable Rates of Convergence for Estimators of the Stable Tail Dependence Function', *Journal of Multivariate Analysis*, 64, 25–46.

Duong, T., and Hazelton, M.L. (2003), 'Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation', *Journal of Nonparametric Statistics*, 15, 17–30.

Duong, T., and Hazelton, M.L. (2005), 'Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation', *Scandinavian Journal of Statistics*, 32, 485–506.

Einmahl, J.H.J., de Haan, L., and Piterbarg, V.I. (2001), 'Nonparametric Estimation of the Spectral Measure of an Extreme Value Distribution', *The Annals of Statistics*, 29, 1401–1423.

Einmahl, J.H.J., Krajina, A., and Segers, J. (2008), 'A Method of Moments Estimator of Tail Dependence', *Bernoulli*, 14, 1003–1026.

Einmahl, J.H.J., Krajina, A., and Segers, J. (2012), 'An M-Estimator for Tail Dependence in Arbitrary Dimensions', *The Annals of Statistics*, 40, 1764–1793.

Einmahl, J.H.J., and Segers, J. (2009), 'Maximum Empirical Likelihood Estimation of the Spectral Measure of an Extreme-value Distribution', *The Annals of Statistics*, 37, 2953–2989.

Falk, M., Hüsler, J., and Reiss, R.-D. (2011), *Laws of Small Numbers: Extremes and Rare Events* (extended ed.), Basel: Birkhäuser/Springer Basel AG.

Fischer, E.M., Beyerle, U., and Knutti, R. (2013), 'Robust Spatially Aggregated Projections of Climate Extremes', *Nature Climate Change*, 3, 1033–1038.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M. (2013), 'Evaluation of Climate Models', in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds. T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 741–866.

Gasser, T., and Müller, H.-G. (1979), 'Kernel Estimation of Regression Functions', in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Berlin: Springer, pp. 23–68.

Geenens, G. (2014), 'Probit Transformation for Kernel Density Estimation on the Unit Interval', *Journal of the American Statistical Association*, 109, 346–358.

Hall, P., Marron, J.S., and Park, B.U. (1992), 'Smoothed Cross-validation', *Probability Theory and Related Fields*, 92, 1–20.

Hall, P., and Tajvidi, N. (2000), 'Distribution and Dependence-function Estimation for Bivariate Extreme-value Distributions', *Bernoulli*, 6, 835–844.

Hill, B.M. (1975), 'A Simple General Approach to Inference About the Tail of a Distribution', *The Annals of Statistics*, 3, 1163–1174.

Hosking, J.R.M. (1985), 'Algorithm AS 215: Maximum-Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution', *Journal of the Royal Statistical Society. Series C*, 34, 301–310.

Hosking, J.R.M., Wallis, J.R., and Wood, E.F. (1985), 'Estimation of the Generalized Extreme-value Distribution by the Method of Probability-weighted Moments', *Technometrics*, 27, 251–261.

Huang, X. (1992), 'Statistics of Bivariate Extreme Values', Ph.D. thesis, Erasmus University.

Hüsler, J., and Reiss, R.-D. (1989), 'Maxima of Normal Random Vectors: Between Independence and Complete Dependence', *Statistics & Probability Letters*, 7, 283–286.

Jin, Z., and Shao, Y. (1999), 'On Kernel Estimation of a Multivariate Distribution Function', *Statistics & Probability Letters*, 41, 163–168.

Joe, H. (1990), 'Families of Min-stable Multivariate Exponential and Multivariate Extreme Value Distributions', *Statistics & Probability Letters*, 9, 75–81.

Jones, D.A., Wang, W., and Fawcett, R. (2009), 'High-quality Spatial Climate Data-sets for Australia', *Australian Meteorological and Oceanographic Journal*, 58, 233–248.

Kotz, S., and Nadarajah, S. (2000), *Extreme Value Distributions: Theory and Applications*, London: Imperial College Press.

Loftsgaarden, D.O., and Quesenberry, C.P. (1965), 'A Nonparametric Estimate of a Multivariate Density Function', *The Annals of Mathematical Statistics*, 36, 1049–1051.

Lye, L., Hapuarachchi, K., and Ryan, S. (1993), 'Bayes Estimation of the Extreme-value Reliability Function', *IEEE Transactions on Reliability*, 42, 641–644.

Macleod, A.J. (1989), 'Remark AS R76: A Remark on Algorithm AS 215: Maximum-Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution', *Journal of the Royal Statistical Society. Series C*, 38, 198–199.

Magnus, J.R., and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics* (rev. ed.), Chichester: John Wiley and Sons.

Marcon, G., Padoan, S.A., Naveau, P., and Muliere, P. (2014), 'Multivariate Nonparametric Estimation of the Pickands Dependence Function using Bernstein Polynomials', *arXiv:1405.5228*.

Maritz, J.S., and Munro, A.H. (1967), 'On the Use of the Generalised Extreme-value Distribution in Estimating Extreme Percentiles', *Biometrics*, 23, 79–103.

Markovich, N. (2007), *Nonparametric Analysis of Univariate Heavy-tailed Data: Research and Practice*, Chichester: John Wiley & Sons.

Perkins, S.E., Moise, A., Whetton, P., and Katzfey, J. (2014), 'Regional Changes of Climate Extremes Over Australia - A Comparison of Regional Dynamical Downscaling and Global Climate Model Simulations', *International Journal of Climatology*, 34, 3456–3478.

Perkins, S.E., Pitman, A.J., Holbrook, N.J., and McAneney, J. (2007), 'Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions', *Journal of Climate*, 20, 4356–4376.

Perkins, S.E., Pitman, A.J., and Sisson, S.A. (2013), 'Systematic Differences in Future 20 Year Temperature Extremes in AR4 Model Projections over Australia as a Function of Model Skill', *International Journal of Climatology*, 33, 1153–1167.

Pickands III, J. (1975), 'Statistical Inference using Extreme Order Statistics', *The Annals of Statistics*, 3, 119–131.

Pitt, D., and Guillen, M. (2010), 'An Introduction to Parametric and Nonparametric Models for Bivariate Positive Insurance Claim Severity Distributions', Tech. rep.

Prescott, P., and Walden, A.T. (1980), 'Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-value Distribution', *Biometrika*, 67, 723–724.

Rootzén, H., Segers, J., and Wadsworth, J.L. (2017), 'Multivariate Peaks over Thresholds Models', *Extremes*, 21, 115–145.

Rootzén, H., and Tajvidi, N. (2006), 'Multivariate Generalized Pareto Distributions', *Bernoulli*, 12, 917–930.

Rudemo, M. (1982), 'Empirical Choice of Histograms and Kernel Density Estimators', *Scandinavian Journal of Statistics*, 9, 65–78.

Sain, S.R., Baggerly, K.A., and Scott, D.W. (1994), 'Cross-validation of Multivariate Densities', *Journal of the American Statistical Association*, 89, 807–817.

Scott, D.W. (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization* (2nd ed.), Hoboken, NJ: John Wiley & Sons.

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X. (2012), 'Changes in Climate Extremes and their Impacts on the Natural Physical Environment', in *IPCC WGI/WGII Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX)*, eds. C. Field, V. Barros, T. Stocker, D. Qin, D. Dokken, K. Ebi, M. Mastrandrea, K. Mach, G.-K. Plattner, S. Allen, M. Tignor, and P. Midgley, Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 190–230.

Sheather, S.J., and Jones, M.C. (1991), 'A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation', *Journal of the Royal Statistical Society. Series B* , 53, 683–690.

Sillmann, J., Kharin, V.V., Zhang, X., Zwiers, F.W., and Bronaugh, D. (2013a), 'Climate Extremes Indices in the CMIP5 Multimodel Ensemble: Part 1. Model Evaluation in the Present Climate', *Journal of Geophysical Research: Atmospheres*, 118, 1716–1733.

Sillmann, J., Kharin, V.V., Zwiers, F.W., Zhang, X., and Bronaugh, D. (2013b), 'Climate Extremes Indices in the CMIP5 Multimodel Ensemble: Part 2. Future Climate Projections', *Journal of Geophysical Research: Atmospheres*, 118, 2473–2493.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Smith, R.L. (1985), 'Maximum Likelihood Estimation in a Class of Nonregular Cases', *Biometrika*, 72, 67–90.

Smith, R.L., Tawn, J.A., and Yuen, H.K. (1990), 'Statistics of Multivariate Extremes', *International Statistical Review*, 58, 47–58.

Taylor, K.E., Stouffer, R.J., and Meehl, G.A. (2012), 'An Overview of CMIP5 and the Experiment Design', *Bulletin of the American Meteorological Society*, 93, 485–498.

Vautard, R., Gobiet, A., Jacob, D., Belda, M., Colette, A., Déqué, M., Fernández, J., García-Díez, M., Goergen, K., Güttler, I., Halenka, T., Karacostas, T., Katragkou, E., Keuler, K., Kotlarski, S., Mayer, S., van Meijgaard, E., Nikulin, G., Patarčić, M., Scinocca, J., Sobolowski, S., Suklitsch, M., Teichmann, C., Warrach-Sagi, K., Wulfmeyer, V., and Yiou, P. (2013), 'The Simulation of European Heat Waves from an Ensemble of Regional Climate Models within the EURO-CORDEX Project', *Climate Dynamics*, 41, 2555–2575.

Wand, M.P. (1992), 'Error Analysis for General Multivariate Kernel Estimators', *Journal of Nonparametric Statistics*, 2, 1–15.

Wand, M.P., and Jones, M.C. (1994), 'Multivariate Plug-in Bandwidth Selection', *Computational Statistics*, 9, 97–116.

Wand, M.P., and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Wand, M., Marron, J., and Ruppert, D. (1991), 'Transformations in Density Estimation', *J. Amer. Statist. Assoc.*, 86, 343–353.

Appendix. Proofs

The below assumptions will be used to establish the optimality properties of our transformation kernel density estimators. These assumptions are usually expressed for random variables with unbounded support, which in our case is the transformed variable $Y = t(X)$. This set of conditions do not form a minimal set, but they serve as a convenient starting point to state our results.

(A1) The d-variate density f_Y is continuous, square integrable and ultimately monotone for all element-wise partial second derivatives.

(A2) The d-variate kernel K is a positive, symmetric, square integrable p.d.f. such that $\int_{\mathbb{R}^d} \boldsymbol{y}\boldsymbol{y}^\top K(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = m_2(K)\mathbf{I}_d$ where $m_2(K)$ is finite and $\mathbf{I}_d$ is the $d \times d$ identity matrix.

(A3) The bandwidth matrix $\mathbf{H} = \mathbf{H}(n)$ forms a sequence of symmetric and positive definite matrices such that $n^{-1}|\mathbf{H}|^{-1/2}$ and every element of $\mathbf{H}$ approaches zero as $n \to \infty$.

The proof of Theorem A.1 requires Lemma A.1 (below) which establishes the minimal rate of MISE convergence of $\hat{f}_Y$. This result has already been established (e.g. Wand 1992), however we include details of a proof using an alternative notation for fourth order derivatives of a multivariate function via four-fold Kronecker product, which is simpler to code than tensors.

Lemma A.1: *Suppose that the conditions* (A1–A3) *hold. The MISE of the the kernel density estimator with unbounded data support $\hat{f}_Y$ is*

$$\mathrm{MISE}\{\hat{f}_Y(\cdot; \mathbf{H})\} = \left[\tfrac{1}{4} m_2^2(K)(\mathrm{vec}^\top \mathbf{H} \otimes \mathrm{vec}^\top \mathbf{H})\boldsymbol{\psi}_{Y,4} + n^{-1}|\mathbf{H}|^{-1/2}R(K) \right]\{1 + o(1)\}.$$

where $\boldsymbol{\psi}_{Y,4} = \int_{\mathbb{R}^d} \mathsf{D}^{\otimes 4} f_Y(\boldsymbol{y})f_Y(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$.

Proof of Lemma A.1.: The expected value of $\hat{f}_Y$ is

$$\mathbb{E}\hat{f}_Y(y; \mathbf{H}) = \mathbb{E}K_{\mathbf{H}}(y - Y) = \int_{\mathbb{R}^d} K_{\mathbf{H}}(y - w)f(y)\,dw = K_{\mathbf{H}} * f_Y(x)$$

where $*$ denotes the convolution operator between two functions. Asymptotically, using a Taylor series expansion, we have

$$\mathbb{E}\hat{f}_Y(y; \mathbf{H}) = \int_{\mathbb{R}^d} |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}(y - w))f_Y(y)\,dw$$

$$= \int_{\mathbb{R}^d} K(w)f_Y(y - \mathbf{H}^{1/2}w)\,dw$$

$$= \int_{\mathbb{R}^d} K(w)\left[f_Y(y) - w^{\top}\mathbf{H}^{1/2}\mathsf{D}f_Y(x)\right.$$

$$\left. + \frac{1}{2}w^{\top}\mathbf{H}^{1/2}\mathsf{D}^2 f_Y(y)\mathbf{H}^{1/2}w\right]\{1 + o(1)\}\,dw$$

$$= \left[f_Y(y) + \frac{1}{2}\int_{\mathbb{R}^d} K(w)\mathrm{tr}(ww^{\top}\mathbf{H}\mathsf{D}^2 f_Y(y))\,dw\right]\{1 + o(1)\}$$

$$= \left[f_Y(y) + \frac{1}{2}m_2(K)\mathrm{tr}(\mathbf{H}\mathsf{D}^2 f_Y(y))\right]\{1 + o(1)\}.$$

This allows us to write the bias of $\hat{f}_Y(y; \mathbf{H})$ as

$$\mathbb{E}\hat{f}_Y(y; \mathbf{H}) - f_Y(y) = \tfrac{1}{2}m_2(K)\mathrm{tr}(\mathbf{H}\mathsf{D}^2 f_Y(y))\{1 + o(1)\}.$$

For the variance, we have $\mathrm{Var}\hat{f}(y; \mathbf{H}) = n^{-1}\mathbb{E}[K_{\mathbf{H}}(y - Y)^2] - n^{-1}[\mathbb{E}K_{\mathbf{H}}(y - Y)]^2$. The second term is given by the above, so we are required to evaluate

$$E\hat{f}(y; \mathbf{H})^2 = \mathbb{E}[K_{\mathbf{H}}(y - Y)^2] = \int_{\mathbb{R}^d} K_{\mathbf{H}}(y - w)^2 f_Y(w)\,dw$$

$$= \int_{\mathbb{R}^d} |\mathbf{H}|^{-1}K(\mathbf{H}^{-1/2}(y - w))^2 f_Y(w)\,dw$$

$$= \int_{\mathbb{R}^d} |\mathbf{H}|^{-1/2}K(w)^2 f_Y(y - \mathbf{H}^{-1/2}w)\,dw$$

$$= |\mathbf{H}|^{-1/2}f_Y(y)\int_{\mathbb{R}^d} K(w)^2\,dw\{1 + o(1)\}$$

$$= |\mathbf{H}|^{-1/2}f_Y(y)R(K)\{1 + o(1)\}.$$

Thus the variance term is

$$\mathrm{Var}\{\hat{f}_Y(y; \mathbf{H})\} = n^{-1}\{|\mathbf{H}|^{-1/2}f_Y(y)R(K) - [f_Y(y) + \tfrac{1}{2}m_2(K)\mathrm{tr}(\mathbf{H}\mathsf{D}^2 f_Y(y))]^2\}\{1 + o(1)\}.$$

Since $\mathbf{H} \to 0$ then $|\mathbf{H}|^{-1/2}$ dominates both the constant term $f_Y(y)$ and the $\mathrm{tr}(\mathbf{H})$ term so we can write

$$\mathrm{Var}\{\hat{f}_Y(y; \mathbf{H})\} = n^{-1}|\mathbf{H}|^{-1/2}f_Y(y)R(K)\{1 + o(1)\}.$$

The integrated square bias (ISB) is then

$$\mathrm{ISB}\{\hat{f}_Y(\cdot;\mathbf{H})\} = \int_{\mathbb{R}^d} \mathrm{Bias}^2 \hat{f}_Y(\boldsymbol{y};\mathbf{H})\, d\boldsymbol{y} = \int_{\mathbb{R}^d} \frac{1}{4} m_2^2(K) \mathrm{tr}^2(\mathbf{H} D^2 f_Y(\boldsymbol{y}))\, d\boldsymbol{y}\{1 + o(1)\}$$

$$= \frac{1}{4} m_2^2(K) \int_{\mathbb{R}^d} \mathrm{tr}^2(\mathbf{H} D^2 f_Y(\boldsymbol{y}))\, d\boldsymbol{y}\{1 + o(1)\}$$

$$= \frac{1}{4} m_2^2(K) (\mathrm{vec}^\top \mathbf{H} \otimes \mathrm{vec}^\top \mathbf{H}) \boldsymbol{\psi}_{Y,4}\{1 + o(1)\},$$

and similarly the integrated variance (IV) is

$$\mathrm{IV}\{\hat{f}_Y(\cdot;\mathbf{H})\} = \int_{\mathbb{R}^d} n^{-1} |\mathbf{H}|^{-1/2} f_Y(\boldsymbol{y}) R(K)\, d\boldsymbol{y}\{1 + o(1)\}$$

$$= n^{-1} |\mathbf{H}|^{-1/2} R(K) \int_{\mathbb{R}^d} f_Y(\boldsymbol{y})\, d\boldsymbol{y}\{1 + o(1)\}$$

$$= n^{-1} |\mathbf{H}|^{-1/2} R(K)\{1 + o(1)\},$$

using the integrability assumptions in conditions (A1) and (A2). Hence we obtain the result as $\mathrm{MISE}\{\hat{f}_Y(\cdot;\mathbf{H})\} = \mathrm{ISB}\{\hat{f}_Y(\cdot;\mathbf{H})\} + \mathrm{IV}\{\hat{f}_Y(\cdot;\mathbf{H})\}$. ∎

Proof of Theorem 2.1.: Let $\boldsymbol{y} = \boldsymbol{t}(\boldsymbol{x}) = (\log(x_{1d}),\ldots,\log(x_d))^\top$, and inversely $\boldsymbol{x} = \exp(\boldsymbol{y}) = (\exp(y_1),\ldots,\exp(y_d))^\top$. The Jacobian is $|\mathbf{J}_t(\boldsymbol{x})| = 1/(x_1 \cdots x_d) = \exp(-|\boldsymbol{y}|)$ where $|\boldsymbol{y}| = y_1 + \cdots + y_d$. Thus $f_Y(\boldsymbol{y}) = 1/|\mathbf{J}_t(\boldsymbol{x})| f_X(\boldsymbol{x}) = \exp(|\boldsymbol{y}|) f_X(\exp(\boldsymbol{y}))$. This representation will allow us to determine the Hessian matrix of $D^2 f_Y$ since it the previous lemma shows that it is a crucial element in $\mathrm{MISE}\{\hat{f}_Y(\cdot;\mathbf{H})\}$.

To evaluate derivatives of $f_Y(\boldsymbol{y})$ with respect to $\boldsymbol{y}$, we require the following preliminary differentials:

$$d\exp(|\boldsymbol{y}|) = D[\exp(y_1 + \cdots + y_d)]^\top d\boldsymbol{y} = [\exp(y_1),\ldots,\exp(y_d)]^\top d\boldsymbol{y} = \exp(\boldsymbol{y})^\top d\boldsymbol{y},$$

$$d\exp(\boldsymbol{y}) = [d\exp(y_1),\ldots,d\exp(y_d)] = [\exp(y_1)dy_1,\ldots,\exp(y_d)dy_d] = \mathrm{Diag}(\exp(\boldsymbol{y}))d\boldsymbol{y},$$

and

$$d\mathrm{Diag}(\exp(\boldsymbol{y})) = \mathrm{Diag}(d\exp(y_1),\ldots,d\exp(y_d) = \mathrm{Diag}(\exp(y_1)dy_1,\ldots,\exp(y_d)dy_d)$$

$$= \mathrm{Diag}(\exp(\boldsymbol{y}))\mathrm{Diag}(d\boldsymbol{y}).$$

where $\mathrm{Diag}(\boldsymbol{a})$ is the diagonal matrix whose elements are $\boldsymbol{a}$. It can be decomposed as $\mathrm{Diag}(\boldsymbol{a}) = \sum_{j=1}^d \boldsymbol{e}_j^\top \boldsymbol{a} \boldsymbol{e}_j \boldsymbol{e}_j^\top$ in terms of $\boldsymbol{e}_j$, the jth elementary d-vector which is all zero except for 1 at the jth element. So then

$$d\,\mathrm{vec}(\mathrm{Diag}(\exp(\mathrm{y}))) = \sum_{j=1}^d \mathrm{vec}(\mathrm{Diag}(\exp(\boldsymbol{y}))\boldsymbol{e}_j\boldsymbol{e}_j^\top)\boldsymbol{e}_j^\top d\boldsymbol{y}.$$

The differential of f_Y is

$$df_Y(\boldsymbol{y}) = (d\exp(|\boldsymbol{y}|))f_X(\exp(\boldsymbol{y})) + \exp(|\boldsymbol{y}|)df_X(\exp(\boldsymbol{y}))$$

$$= f_X(\exp(\boldsymbol{y}))\exp(\boldsymbol{y})^\top d\boldsymbol{y} + \exp(|\boldsymbol{y}|)Df_X(\exp(\boldsymbol{y}))^\top d\exp(\boldsymbol{y})$$

$$= f_X(\exp(\boldsymbol{y}))\exp(\boldsymbol{y})^\top d\boldsymbol{y} + \exp(|\boldsymbol{y}|)Df_X(\exp(\boldsymbol{y}))^\top \mathrm{Diag}(\exp(\boldsymbol{y}))d\boldsymbol{y}$$

which implies that the first derivative is

$$Df_Y(\boldsymbol{y}) = f_X(\exp(\boldsymbol{y}))\exp(\boldsymbol{y}) + \exp(|\boldsymbol{y}|)\mathrm{Diag}(\exp(\boldsymbol{y}))D_X(\exp(\boldsymbol{y}))$$

$$= f_X(\exp(\boldsymbol{y}))\exp(\boldsymbol{y}) + \exp(|\boldsymbol{y}|)[Df_X(\exp(\boldsymbol{y}))^\top \otimes \mathbf{I}_d]\mathrm{vec}(\mathrm{Diag}(\exp(\mathrm{y}))),$$

using the first identification table in Magnus and Neudecker (1999, p. 176) to convert these differentials to derivatives. The second form of $\mathsf{D}f_Y(\mathbf{y})$ derives from the identity $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{I}_d)\mathrm{vec}\mathbf{B}$ for conformable matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$.

The differential of $\mathsf{D}f_Y(\mathbf{y})$ is

$$
\begin{aligned}
d\mathsf{D}f_Y(\mathbf{y}) = {} & f_X(\exp(\mathbf{y}))d\exp(\mathbf{y}) + (df_X(\exp(\mathbf{y})))\exp(\mathbf{y}) \\
& + (d\exp(|\mathbf{y}|))\mathrm{Diag}(\exp(\mathbf{y}))\mathsf{D}f_X(\exp(\mathbf{y})) \\
& + \exp(|\mathbf{y}|)[\mathsf{D}f_X(\exp(\mathbf{y}))^\top \otimes \mathbf{I}_d]d\mathrm{vecDiag}(\exp(\mathbf{y})) \\
& + \exp(|\mathbf{y}|)[(d\exp(\mathbf{y}))^\top \mathsf{D}^2 f_X(\exp(\mathbf{y})) \otimes \mathbf{I}_d]\mathrm{vecDiag}(\exp(\mathbf{y})) \\
= {} & f_X(\exp(\mathbf{y}))\mathrm{Diag}(\exp(\mathbf{y}))d\mathbf{y} + \exp(\mathbf{y})\mathsf{D}f_X(\exp(\mathbf{y}))^\top \mathrm{Diag}(\exp(\mathbf{y}))d\mathbf{y} \\
& + \mathrm{Diag}(\exp(\mathbf{y}))\mathsf{D}f_X(\exp(\mathbf{y}))\exp(\mathbf{y})^\top d\mathbf{y} \\
& + \exp(|\mathbf{y}|)[\mathsf{D}f_X(\exp(\mathbf{y}))^\top \otimes \mathbf{I}_d]\left\{ \sum_{j=1}^d \mathrm{vec}[\mathrm{Diag}(\exp(\mathbf{y}))\mathbf{e}_j\mathbf{e}_j^\top]\mathbf{e}_j^\top d\mathbf{y} \right\} \\
& + \exp(|\mathbf{y}|)[d\mathbf{y}^\top \mathrm{Diag}(\exp(\mathbf{y}))\mathsf{D}^2 f_X(\exp(\mathbf{y})) \otimes \mathbf{I}_d]\mathrm{vecDiag}(\exp(\mathbf{y})).
\end{aligned}
$$

This can be simplified by noting that for d-vectors $\mathbf{a}, \mathbf{b}$,

$$
\sum_{j=1}^d (\mathbf{a}^\top \otimes \mathbf{I}_d)\mathrm{vec}[\mathrm{Diag}(\mathbf{b})\mathbf{e}_j\mathbf{e}_j^\top]\mathbf{e}_j^\top = \sum_{j=1}^d \mathrm{Diag}(\mathbf{b})\mathbf{e}_j\mathbf{e}_j^\top \mathbf{a}\mathbf{e}_j^\top = \mathrm{Diag}(\mathbf{b})\mathrm{Diag}(\mathbf{a})
$$

that is,

$$
\begin{aligned}
d\mathsf{D}f_Y(\mathbf{y}) = {} & \{ f_X(\exp(\mathbf{y}))\mathrm{Diag}(\exp(\mathbf{y})) + \exp(\mathbf{y})\mathsf{D}f_X(\exp(\mathbf{y}))^\top \mathrm{Diag}(\exp(\mathbf{y})) \\
& + \mathrm{Diag}(\exp(\mathbf{y}))\mathsf{D}f_X(\exp(\mathbf{y}))\exp(\mathbf{y})^\top d\mathbf{y} \\
& + \exp(|\mathbf{y}|)\mathrm{Diag}(\exp(\mathbf{y}))\mathrm{Diag}(\mathsf{D}f_X(\exp(\mathbf{y}))) \\
& + \exp(|\mathbf{y}|)\mathrm{Diag}(\exp(\mathbf{y}))\mathsf{D}^2 f_X(\exp(\mathbf{y}))\mathrm{Diag}(\exp(\mathbf{y})) \} d\mathbf{y}.
\end{aligned}
$$

This implies that the Hessian matrix of $\mathsf{D}f_Y(\mathbf{y})$ is

$$
\begin{aligned}
\mathsf{D}^2 f_Y(\mathbf{y}) = {} & f_X(\mathbf{x})\mathrm{Diag}(\mathbf{x}) + \mathbf{x}\mathsf{D}f_X(\mathbf{x})^\top \mathrm{Diag}(\mathbf{x}) + \mathrm{Diag}(\mathbf{x})\mathsf{D}f_X(\mathbf{x})\mathbf{x}^\top \\
& + \pi(\mathbf{x})\mathrm{Diag}(\mathbf{x})\mathrm{Diag}(\mathsf{D}f_X(\mathbf{x})) + \pi(\mathbf{x})\mathrm{Diag}(\mathbf{x})\mathsf{D}^2 f_X(\mathbf{x})\mathrm{Diag}(\mathbf{x}) \quad\quad \text{(A1)}
\end{aligned}
$$

as $\exp(|\mathbf{y}|) = |\mathbf{J}_t(\mathbf{x})|^{-1} = x_1 x_2 \cdots x_d = \pi(\mathbf{x})$.

Firstly using the definition of $\hat{f}_X$, its expected value is $\mathbb{E}\hat{f}_X(\mathbf{x}; \mathbf{H}) = |\mathbf{J}_t(\mathbf{x})|\mathbb{E}\hat{f}_Y(\mathbf{y}; \mathbf{H})$ and its associated bias, from combining Lemma A.1 and Equation (A1), is

$$
\begin{aligned}
& \mathrm{Bias}\{\hat{f}_X(\mathbf{x}; \mathbf{H})\} \\
& = \mathbb{E}\hat{f}_X(\mathbf{x}; \mathbf{H}) - f_X(\mathbf{x}) = |\mathbf{J}_t(\mathbf{x})|\mathrm{Bias}\{\hat{f}_Y(t(\mathbf{x}); \mathbf{H})\} \\
& = \tfrac{1}{2}m_2(K)\pi(\mathbf{x})^{-1}\mathrm{tr}(\mathbf{H}\mathsf{D}^2 f_Y(\mathbf{y}))\{1 + o(1)\} \\
& = \tfrac{1}{2}m_2(K)\pi(\mathbf{x})^{-1}\mathrm{tr}\{\mathbf{H}[f_X(\mathbf{x})\mathrm{Diag}(\mathbf{x}) + \mathbf{x}\mathsf{D}f_X(\mathbf{x})^\top \mathrm{Diag}(\mathbf{x}) + \mathrm{Diag}(\mathbf{x})\mathsf{D}f_X(\mathbf{x})\mathbf{x}^\top \\
& \quad + \pi(\mathbf{x})\mathrm{Diag}(\mathbf{x})\mathrm{Diag}(\mathsf{D}f_X(\mathbf{x})) + \pi(\mathbf{x})\mathrm{Diag}(\mathbf{x})\mathsf{D}^2 f_X(\mathbf{x})\mathrm{Diag}(\mathbf{x})]\}\{1 + o(1)\} \\
& = \tfrac{1}{2}m_2(K)[\pi(\mathbf{x})^{-1}f_X(\mathbf{x})\mathrm{tr}(\mathbf{H}\mathrm{Diag}(\mathbf{x})) + 2\pi(\mathbf{x})^{-1}\mathrm{tr}(\mathbf{H}\mathbf{x}\mathsf{D}f_X(\mathbf{x})^\top \mathrm{Diag}(\mathbf{x})) \\
& \quad + \mathrm{tr}(\mathbf{H}\mathrm{Diag}(\mathbf{x})\mathrm{Diag}(\mathsf{D}f_X(\mathbf{x}))) + \mathrm{tr}(\mathbf{H}\mathrm{Diag}(\mathbf{x})\mathsf{D}^2 f_X(\mathbf{x})\mathrm{Diag}(\mathbf{x}))]\{1 + o(1)\}.
\end{aligned}
$$

Similarly we have $\mathbb{E}[\hat{f}_X(\boldsymbol{x}; \mathbf{H})^2] = |\mathbf{J}_t(\boldsymbol{x})|^2 \mathbb{E}[\hat{f}_Y(\boldsymbol{t}(\boldsymbol{x}); \mathbf{H})^2]$, leading to

$$\begin{aligned}
\mathrm{Var}\{\hat{f}_X(\boldsymbol{x}; \mathbf{H})\} &= \mathbb{E}[\hat{f}_X(\boldsymbol{x}; \mathbf{H})^2] - \{\mathbb{E}\hat{f}_X(\boldsymbol{x}; \mathbf{H})\}^2 = |\mathbf{J}_t(\boldsymbol{x})|^2 \mathrm{Var}\{\hat{f}_Y(\boldsymbol{t}(\boldsymbol{x}); \mathbf{H})\} \\
&= n^{-1} |\mathbf{H}|^{-1/2} R(K) |\mathbf{J}_t(\boldsymbol{x})|^2 f_Y(\boldsymbol{y})\{1 + o(1)\} \\
&= n^{-1} |\mathbf{H}|^{-1/2} R(K) \pi(\boldsymbol{x})^{-1} f_X(\boldsymbol{x})\{1 + o(1)\}.
\end{aligned}$$

∎