Statistica Sinica Preprint No: SS-2021-0112	
Title	Shape Constrained Kernel PDF and PMF Estimation
Manuscript ID	SS-2021-0112
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0112
Complete List of Authors	Pang Du, Christopher F. Parmeter and Jeffrey S. Racine
Corresponding Authors	Jeffrey S. Racine
E-mails	racinej@mcmaster.ca
Notice: Accepted version subject to English editing.	

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Shape Constrained Kernel PDF and PMF Estimation

Pang Du* Christopher F. Parmeter† Jeffrey S. Racine‡

May 10, 2022

Abstract

We consider shape constrained kernel-based probability density function (PDF) and probability mass function (PMF) estimation. Our approach is of widespread potential applicability and includes, separately or simultaneously, constraints on the PDF (PMF) function itself, its integral (sum), and derivatives (finite-differences) of any order. We also allow for pointwise upper and lower bounds (i.e., inequality constraints) on the PDF and PMF in addition to more popular equality constraints, and the approach handles a range of transformations of the PDF and PMF including, for example, logarithmic transformations (which allows for the imposition of log-concave or log-convex constraints that are popular with practitioners). Theoretical underpinnings for the procedures are provided. A simulation-based comparison of our proposed approach with those obtained using Grenander-type methods is favourable to our approach when the DGP is itself smooth. As far as we know, ours is also the only *smooth* framework that handles PDFs and PMFs in the presence of inequality bounds, equality constraints, and other popular constraints such as those mentioned above. An implementation in R exists that incorporates constraints such as monotonicity (both increasing and decreasing), convexity and concavity, and log-convexity and log-concavity, among others, while respecting finite-support boundaries via explicit use of boundary kernel functions.

*Department of Statistics, Virginia Tech, pangdu@vt.edu
†Department of Economics, University of Miami, c.parmeter@miami.edu
‡Department of Economics and Graduate Program in Statistics, McMaster University, racinej@mcmaster.ca

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

1 Introduction

Shape constraints play a vital role in identification, estimation, and inference in econometric and statistical applications (see, e.g., Chetverikov, Santos, and Shaikh (2018) for a review of recent developments and their importance in applied work). Such constraints sometimes emerge naturally due to the nature of the data, but increasingly often are required when replacing parametric models by more versatile semi- and nonparametric models. The ability to preserve qualitative shape properties present in a parametric model is a key component of any alternative method. The adoption of nonparametric methods capable of retaining qualitative shape properties is largely driven by concerns over misspecification of the parametric model and the unforgiving consequences thereof while simultaneously allowing incorporation of such qualitative features into the robust nonparametric alternative.

There are two reasons why one might wish to integrate shape constraints into a nonparametric estimation procedure. The first is to achieve potential gains in estimator *efficiency* that can arise when one imposes *valid* shape constraints on some statistical object of interest (i.e., if one's assumption about a shape constraint on a otherwise unspecified curve is correct, then incorporating this information into the estimation procedure can improve the finite-sample performance of the corresponding estimator). The second is to assess their validity through formal quantitative inference or simply to explore the extent to which imposing constraints qualitatively impacts the resulting estimate.

The imposition of shape constraints on an otherwise unrestricted nonparametric curve is a key element of sound empirical analysis that encompasses a range of approaches, and we direct the interested reader to the book by Groeneboom and Jongbloed (2014) who detailed a wide array of shape constrained estimators and algorithms along with their theoretical properties. Perhaps the most common applications of enforcing shape constraints arise when modelling a conditional mean function (i.e., regression), which is understandable given the popularity of regression analysis among practitioners. However, the density function is also a popular object of interest and one that necessitates a separate treatment from that of regression due to its unique nature. Shape-constrained density estimation, like its regression-based counterpart, has a rich history which can be traced to

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

the seminal work of Grenander (1956), who analyzed the maximum likelihood estimator (MLE) of a decreasing density on the non-negative half-line (see also Groeneboom and Jongbloed (2018) for recent theoretical work in this direction). It is of interest to note that Prakasa Rao (1969) showed that this estimator exhibits nonstandard asymptotic behaviour (i.e., it converges at cube rate ($n^{-1/3}$) at points at which the true decreasing density is differentiable with negative derivative) which is slower than competing local kernel-based estimators that assume smoothness ($n^{-2/5}$), a common assumption among practitioners and one of two kernel-based estimators that we consider herein. While the density function is our main object of interest, we also treat the mass function, and note that kernel-based mass function estimators for categorical data have a different (and faster, i.e., $n^{-1/2}$) rate of convergence than their kernel density-based counterparts.

In density estimation settings, a variety of innovative approaches have been proposed for imposing specific constraints including monotonicity, concavity, and log-concavity, to name but a few. Though some of these approaches admit certain combinations of shape constraints, many are tailored to a *particular* setting (e.g., monotonicity *only*). And while some existing approaches incorporate bounds on the *support* of the variable under study, others do not. Furthermore, most approaches that have been proposed are predicated on *continuously* distributed random variables, though constrained probability mass functions (PMFs) may also be of value when modelling *discrete support* random variables which frequently arise in applied settings.

Grenander-based approaches (Grenander 1956) have been widely used when imposing certain shape constraints, and one of their appealing features is that they do not require any *tuning parameters*, unlike *smooth* kernel-based nonparametric methods such as those proposed below which require specification of a *bandwidth*. However, though Grenander-based approaches are nonparametric in nature, they are *non-smooth* which runs counter to the spirit of adopting *smooth* nonparametric approaches in the first place (e.g., the approach proposed by Grenander (1956) for imposing monotonicity can be characterized as the left derivative of the least concave majorant of the empirical distribution function, which is a non-smooth function). Needless to say, practitioners who routinely assume smoothness and adopt smooth nonparametric estimators will not be attracted to non-smooth nonparametric shape constrained solutions hence the appeal of *smooth* shape constrained nonparametric solutions such as those proposed herein.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

The literature on constrained nonparametric estimation has grown exponentially over the past few decades. One approach (that which we build upon, modify and extend) has proven to be a particularly popular, versatile, and extensible method for imposing constraints on a *smooth* nonparametric object (cf Hall and Presnell 1999). This approach places weights directly on the sample realizations so that the desired constraint is effectively imposed. In kernel-based *regression* settings this amounts to starting with a standard kernel estimator and then, if the constraints are violated in some region of the support, *vertically* shifting the regressand in such a way that standard kernel regression on the *shifted* regressand delivers a regression curve that satisfies the required constraints while minimizing some distance metric from the unconstrained regression function (Hall and Huang 2001; Du, Parmeter, and Racine 2013). In kernel-based *density* settings this approach can be leveraged by placing weights on the *kernel function* associated with each sample realization (as opposed to the sample realizations themselves) to produce a density that satisfies the required constraints.[1]

Building on the work of Du, Parmeter, and Racine (2013) who considered a unified framework for *smooth* shape constrained nonparametric kernel regression, in this paper we propose a unified framework for *smooth* shape constrained kernel density and probability mass function estimation. Shape constrained kernel density (and mass) function estimation differs from shape constrained kernel regression in both its practical implementation and in its theoretical properties, hence requires a completely separate treatment. Our approach is extremely flexible and allows for a range of constraints to be *simultaneously* imposed (presuming of course that the set of constraints is internally consistent). The original implementation (Hall and Huang 2001) involved optimizing a power-divergence criterion. Du, Parmeter, and Racine (2013) proposed replacing this power-divergence criterion with an L_2 norm criterion, which delivered an estimator that retained all of the desirable features of the power-divergence based method but was far more flexible and extensible and far simpler to solve from a practical perspective. The method proposed here generalizes the seminal work of Hall and Huang (2002), who studied imposition of unimodality on a univariate kernel density

[1]A similar method known as *data sharpening* exists (Hall and Kang 2005) which instead introduces weights which shift the data *horizontally* prior to smoothing, a subtle but important distinction. We adopt the approach of Hall and Presnell (1999) because vertically shifting observations can be undertaken with standard off-the-shelf quadratic programming methods, while horizontally shifting observations typically may require full-blown nonlinear programming which may be less tractable from a practical perspective.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

estimator, and modifies it in such a way as to deliver a unified approach whose implementation is particularly straightforward. As such, we believe that this unified framework would be of particular interest to practitioners who wish to simultaneously impose a range of constraints in a smooth nonparametric setting.

Additionally, we build on the insights of Z. Li, Liu, and Li (2017) who proposed a slightly modified version of the optimization criterion proposed by Du, Parmeter, and Racine (2013). While Z. Li, Liu, and Li (2017) adopted an L_2 norm criterion per Du, Parmeter, and Racine (2013), rather than optimizing distance between optimization *weights* and their unconstrained counterparts, they instead optimize distance between the constrained *estimates* and their unconstrained counterparts. Z. Li, Liu, and Li (2017) provided convincing simulation evidence that their modification can deliver constrained estimates with improved finite-sample performance, though no theoretical justification for this modification was provided by the authors. We shall demonstrate theoretically that this modified L_2 optimization criterion delivers constraint weights which ensure *identical* asymptotic behaviour to that arising when directly optimizing the weights instead. By providing the theoretical underpinnings for the slightly modified optimization criterion proposed by Z. Li, Liu, and Li (2017), we establish that the constraint weights can be based on this improved optimization criterion with no loss in information.

Finally, we demonstrate how our method can be adapted to handle constraints on the *log-density*. This is an important generalization since constraints on the log-density, when enforced using the density function directly, can result in a difficult nonlinear optimization problem. By focusing instead *directly* on the log-density, we ensure straightforward constraint enforcement with trivial conversion back to the constrained density itself all within the same unified theoretical framework as that for constraints on the density directly.

The approach developed herein stands distinct from previous results developed in Du, Parmeter, and Racine (2013), and elsewhere, in many respects. As will become apparent, in the current setting we are dealing with density estimation and weights are applied on the kernel function while in Du, Parmeter, and Racine (2013), a regression setting, weights are applied on the dependent variable which impacts the proofs in a non-trivial way. Some of the more noteworthy developments are that we prove Theorem 3.2 for the Cramér–von Mises distance function (earlier work did not consider

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

this distance metric) which required more tedious manipulations in order to handle cross-product terms involving the constraint weights that arise in the various components of our decomposition of the constrained density estimator. Additionally, Theorem 3.3 is entirely new as it is, to the authors' knowledge, the first attempt at imposing smoothness constraints on a PMF estimator. While not a theoretical contribution, we also demonstrate how straightforward it can be to impose log-concavity on a smooth kernel density estimate in a simple quadratic programming setup.

One constraint on the log-density, specifically *log-concavity*, has been a topic of intense interest throughout statistics; see Walther (2009) for an introduction and Samworth and Sen (2018) for a recent review. Briefly, log-concave densities present an appealing and natural alternative choice to the class of unimodal densities. Though the class of log-concave densities is a subset of the class of the unimodal densities, it contains most of the commonly used parametric distributions and is therefore a rich and useful nonparametric class. Recent developments include Feng et al. (2021) who studied adaptation of the nonparametric MLE density for the class of upper semi-continuous, log-concave densities on $\mathbb{R}^d$ (the logarithm of the resulting estimate is a *piecewise-linear* non-smooth function) and Rathke and Schnörr (2019) who proposed a fast implementation of the smoothed version of this estimator.[2]

Log-concavity has also played an important role in applied microeconomic analysis. By imposing log-concavity in an otherwise unrestricted nonparametric setting, economic studies that previously relied on a specific parametric model can instead rely on less restrictive nonparametric models leading to more robust results. Examples include Bagnoli and Bergstrom (2005) who outlined how the log-concavity assumption allows *just enough* special structure to yield workable theories across various subfields, Meyer-ter-Vehn, Smith, and Bognar (2017) who explored costly deliberation by two differentially informed and possibly biased jurors exploiting an assumption that jurors' information types have a log-concave density, and Tan and Zhou (2020) who relied on log-concavity in heterogeneity of agents to establish several formal results in a model of price competition entry and multi-sided markets.

Our adaptation of Hall and Huang (2002) to log-concavity also stands in contrast to a kernel-

[2]For details see the R package `fmlogcondens` (Rathke and Schnörr 2018), though it appears that this package was removed from the *Comprehensive R Archive Network* (CRAN) and "[a]rchived on 2019-07-20 as check problems were not corrected in time": installation fails with warning "package 'fmlogcondens' is not available for this version of R."

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

based linear adjustment mechanism that has been recently proposed (Wolters and Braun 2018a, 2018b). This linear adjustment mechanism tackles constrained estimation using a specified number of inflection points, and this approach could also be used to enforce log-concavity though the authors did not consider this particular constraint. However, this approach requires either that the location of these inflection points be presumed known ex ante or else they need to be approximated through some optimization routine, which has its drawbacks. Our proposed approach to imposing log-concavity, in contrast, requires no prior knowledge nor approximation of the location of inflection points. Instead, we impose the constraints through the log-density directly which leads to a direct system of linear inequality constraints. This provides a fast and efficient algorithm for imposing log-concavity in a smooth setting.[3]

In addition to the work referenced above, the related literature includes Woodroofe and Sun (1993) who considered a penalized MLE estimate of a density on the positive half of the real number line when the density is non-increasing, Meyer and Woodroofe (2004) who developed a nonparametric maximum likelihood estimator that is consistent for the mode, Hall and Kang (2005) who considered unimodal kernel density estimation via data sharpening, Dette and Pilz (2006) who conducted a comparative study of monotone constrained estimators, Birke (2009) who considered shape constrained density estimation via monotone rearrangement (Hardy, Littlewood, and Pólya 1952; Chernozhukov, Fernandez-Val, and Galichon 2009), Dümbgen and Rufibach (2009) who considered MLE of a log-concave density, and Koenker and Mizera (2010) who considered MLE of a log-concave density formulated as a convex optimization problem and shown to have an equivalent dual formulation as a constrained maximum Shannon entropy problem. Cule, Samworth, and Stewart (2010) studied a non-smooth log-concave MLE of a probability distribution function, Meyer and Habtzghi (2011) deployed regression splines, based on earlier work of Meyer (2008), to formulate a nonparametric maximum likelihood estimator of strictly decreasing probability densities in terms of convex programming and iteratively re-weighted least squares cone projection algorithms. Chen and Samworth (2013) studied the smoothed log-concave MLE of a probability distribution function,

[3]The linear adjustment mechanism of Wolters and Braun (2018b) can be shown to be equivalent to our approach *if* the number of adjustment functions is equivalent to the number of observations *and* the adjustment functions themselves are equivalent to the kernel smoothing function of the unconstrained density estimator. However, they did not consider using their linear adjustment mechanism approach to impose log-concavity which, given its popularity in applied settings, forms the basis for one of the Monte Carlo simulations we run to compare our approach with its peers; see the R package `scdensity` (Wolters 2018) for implementation of the linear adjustment mechanism approach.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Horowitz and Lee (2017) explained how to estimate and obtain an asymptotic uniform confidence band for a conditional mean function under possibly nonlinear shape restrictions, while Koenker and Mizera (2018) considered log-concave estimation for weaker forms of concavity constraints that allow heavier tail behaviour and sharper modal peaks. More recently, Lok and Tabri (2021) developed empirical tilting method for shape constrained estimation over a data-driven grid of points to enforce stochastic dominance of a pair of cumulative distribution functions.

The rest of this paper proceeds as follows: Section 2 presents a unified framework for kernel-based PDF and PMF estimators and lays out the details of our approach, Section 2.1 examines how we determine the constraint weights, Section 2.2 briefly outlines finite-support boundary kernel functions, followed by several illustrative examples of some popular constraints, Section 3 outlines theoretical properties of the proposed approach, Section 4 presents a set of Monte Carlo simulations which indicate that the approach is competitive with and often improves upon leading methods that have been tailored to two popular constraints (log-concavity and monotonicity), while Section 5 presents summary remarks and some potential extensions for future research. Detailed theoretical proofs are relegated to a technical appendix. An open implementation in R exists to assist practitioners interested in exploring the proposed methods.

2 Shape Constrained Kernel Density Estimation

Let X_i, $i = 1, \ldots, n$ be an i.i.d. random sample drawn from $f(x)$ where n denotes the sample size. To estimate $f(x)$ via smooth nonparametric methods we begin with the standard kernel density estimator,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where h is the *bandwidth*, $K(\cdot)$ is the *kernel function* usually chosen as a symmetric, mean zero probability density function itself, and where x is a support point at which the density is estimated (Rosenblatt 1956; Parzen 1962). To help discuss our (constraint) weighted density estimator, when imposing constraints on the density function we introduce a vector of constraint weights p_i,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

$i = 1, \ldots, n$, and modify (1) as follows,

$$\hat{f}(x|p) = \frac{1}{h} \sum_{i=1}^{n} p_i K\left(\frac{x - X_i}{h}\right), \tag{2}$$

and note that for $p_i = p_{unif} = 1/n$, the *uniform* weights, $\hat{f}(x|p_{unif}) = \hat{f}(x)$ which is the standard (i.e., *unconstrained*) estimator (1) . In other words, we use the notation $\hat{f}(x|p_{unif})$ in what follows to represent (2) for the special case where the constraint weights assume the value $p_i = 1/n$ for $i = 1, \ldots, n$, and these special weights are denoted p_{unif}, and for these *and only these* weights (2) is equal to(1), the standard kernel estimator (which we call the unconstrained estimator).

To impose constraints on the density function, we let $p_i = n^{-1}(1 + a_i)$ act as the constraint weights in (2) which delivers the estimator

$$\begin{aligned}
\hat{f}(x|p) &= \frac{1}{nh} \sum_{i=1}^{n} (1 + a_i) K\left(\frac{x - X_i}{h}\right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} (1 + a_i) K(Z_i) \\
&= \frac{1}{nh} \sum_{i=1}^{n} K(Z_i) + \frac{1}{nh} \sum_{i=1}^{n} a_i K(Z_i) \\
&= \hat{f}(x|p_{unif}) + \frac{1}{nh} \sum_{i=1}^{n} a_i K(Z_i),
\end{aligned} \tag{3}$$

where $Z_i = (x - X_i)/h$ and where the *unconstrained* (i.e., *uniform*) weights are $a_i = 0$ (i.e., $p_i = 1/n$, the weights that return the unconstrained estimator).

Imposing constraints on the log-density function can be accomplished with a slightly modified setup in what follows. To impose constraints on the log-density function (or its derivatives), we instead consider an estimator of the form

$$\hat{f}(x|p) = \hat{f}(x) \prod_{i=1}^{n} \exp\left\{a_i K(Z_i)/nh\right\}. \tag{4}$$

Taking logarithms, we obtain

$$\log \hat{f}(x|p) = \log \hat{f}(x) + \frac{1}{nh} \sum_{i=1}^{n} a_i K(Z_i),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

hence the constrained density estimator when imposing constraints on the log-density is given by

$$\hat{f}(x|p) = \exp\left\{\log \hat{f}(x|p_{unif}) + \frac{1}{nh}\sum_{i=1}^{n} a_i K(Z_i)\right\},$$

where, here, the *unconstrained* weights used in the object $\hat{f}(x|p_{unif})$ correspond to $a_i = 0$ in (4) which delivers the standard kernel density estimator $\hat{f}(x)$. Regardless of the constraints considered, any constraints imposed on either the density or the log-density as expressed above will be *linear* in the a_i, which combined with a quadratic objective function leads naturally to solving a quadratic program. The resulting constrained estimator in what follows is that arising from solving a quadratic program and then replacing the arbitrary weights a_i with the feasible constrained weights delivered by the quadratic program.

So far we have simply outlined two approaches that introduce weights which are amenable to delivering constrained density or log-density estimates. Now we shall explicitly introduce the constraints themselves in a general framework. Denote the jth derivative of $\hat{f}(x|p)$, $\log \hat{f}(x|p)$ and $K(Z_i)$ with respect to x as $\hat{f}^{(j)}(x|p)$, $\log^{(j)} \hat{f}(x|p)$ and $K^{(j)}(Z_i)$, respectively (the same goes for $\hat{f}(x|p_{unif})$ and $\log \hat{f}(x|p_{unif})$). Let $l(x)$ and $u(x)$ denote *pointwise* lower and upper bounds that may change with x, $l(x) \leq u(x)$. The constraints on the jth derivative of the density or log-density, $j = 0, 1, 2, \ldots$, can be expressed as

$$l(x) \leq \hat{f}^{(j)}(x|p) \leq u(x) \tag{5}$$

and

$$l(x) \leq \log^{(j)} \hat{f}(x|p) \leq u(x),$$

respectively, so for $j = 0$ we are constraining the density or log-density, for $j = 1$ the first derivative thereof, etc. Consider, by way of illustration, the constraint $\hat{f}^{(j)}(x|p) \geq l(x)$, which we express as

$$\hat{f}^{(j)}(x|p_{unif}) + \frac{1}{nh}\sum_{i=1}^{n} a_i K^{(j)}(Z_i) \geq l(x)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

or

$$\frac{1}{nh}\sum_{i=1}^{n} a_i K^{(j)}(Z_i) \geq l(x) - \hat{f}^{(j)}(x|p_{unif}).$$

Furthermore, by way of illustration, the constraint $\log^{(j)} \hat{f}(x|p) \geq l(x)$ (the lower bound $l(x)$ may well differ from that for $\hat{f}^{(j)}(x|p)$ above) can be expressed as

$$\frac{1}{nh}\sum_{i=1}^{n} a_i K^{(j)}(Z_i) \geq l(x) - \log^{(j)} \hat{f}(x|p_{unif}).$$

One appealing feature of our approach is that we can *simultaneously* impose a set of internally consistent constraints. For instance, if we wished to impose the constraint that $\hat{f}^{(0)}(x|p) = \hat{f}(x|p) \geq 0$ (non-negativity of the constrained density) and $\log^{(2)} \hat{f}(x|p) \leq 0$ (log-concavity), we would impose the constraints

$$\frac{1}{nh}\sum_{i=1}^{n} a_i K(Z_i) \geq -\hat{f}(x|p_{unif})$$

and

$$-\frac{1}{nh}\sum_{i=1}^{n} a_i K^{(2)}(Z_i) \geq \log^{(2)} \hat{f}(x|p_{unif}).$$

When solving the quadratic program outlined in the next section, the constraint $\sum_{i=1}^{n} a_i = 0$ will also typically be imposed.

We wish to be able to handle a rich array of constraints, and we may also find ourselves in settings with random variables having either unbounded or compact support. The most popular approaches to compact support kernel estimation use one kernel function when the support is bounded above and below (e.g., Beta(a,b)), one when the support is bounded below (e.g., Gamma(a)), or multiple kernel functions when the support is bounded above and below (e.g., floating boundary kernel functions). To deal with compact support random variables we indulge in some *kernel carpentry* that will deliver a flexible kernel function that is well-suited to the current setting (this is addressed in Section 2.2).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

2.1 Selection of the Constraint Weights

Having established how one can construct the constrained estimator for an *arbitrary* set of weights, we now turn to the issue of how best to select the weights to satisfy some *particular* constraint of interest.

A variety of approaches towards constrained weight selection have been proposed in the literature, each of which minimizes some measure of *divergence* between the constrained and unconstrained *weights* or the constrained and unconstrained *estimates*. There is a fair bit of latitude available to the practitioner when it comes to choice of divergence metric; some metrics may be more computationally demanding than others, and different metrics may impose binding restrictions on the weights in order to produce valid estimates. For example, Hall and Huang (2001) suggested using the Cressie-Read power-divergence metric, Hall and Huang (2002) investigated a smoothed Cramér–von Mises metric, while Du, Parmeter, and Racine (2013) suggested an L_2 norm metric. To elaborate further, in the power-divergence and L_2 norm framework, the constrained weights are selected to be as close as possible to the unconstrained weights (also called the *uniform* weights), while in the smoothed Cramér–von Mises setting, the constrained weights are chosen so that the squared integrated difference between the unconstrained and constrained densities is minimized. As documented in Hall and Huang (2002) and Du, Parmeter, and Racine (2013), one benefit of adopting an L_2 norm (i.e., squared distance) metric is that smoothing parameter selection can be based solely on the unconstrained estimator hence standard off-the-shelf methods can be used without modification, and we maintain this practice in what follows.

Following Hall and Huang (2002), Du, Parmeter, and Racine (2013) and Z. Li, Liu, and Li (2017), we will consider two closely related approaches for optimal construction of the constraint weights and emphasize their relative strengths below. For the first approach, we minimize the L_2 norm divergence between the constraint weights and the uniform weights, where the divergence metric is defined as follows:

$$D_{L_2}(p) = (p_u - p)'(p_u - p).$$

In this case, provided the desired constraints are *linear* in p, we can solve this minimization problem via a straightforward quadratic program exercise using, say, the *quadprog* package (Berwin A.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Turlach R port by Andreas Weingessel <Andreas.Weingessel@ci.tuwien.ac.at> 2019) in R. For the second approach, we minimize a smoothed Cramér–von Mises distance metric where the squared integrated difference between the unconstrained and constrained densities is defined as follows:

$$D_{CM}(p) = (n^2|h|)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} (np_i - 1)(np_j - 1) L\left(\frac{X_i - X_j}{h}\right), \tag{6}$$

where $L(\cdot)$ is the convolution kernel of $K(\cdot)$ with itself. Regardless of the metric used, $D_{L_2}(p)$ and $D_{CM}(p)$ require that the constraint weights themselves satisfy a constraint in order to guarantee that a proper probability density is produced (i.e., for constraints on the density function using (3) or constraints on the log-density function using (4) we require $\sum_{i=1}^{n} a_i = 0$). It is useful to focus on the relative merits of each distance metric used for the selection of the constraint weights. The obvious benefit of $D_{L_2}(p)$ is the relative theoretical ease with which to assess the properties of the corresponding constraint weights. As demonstrated in Du, Parmeter, and Racine (2013), the relative magnitude of the constraint weights with the L_2 norm is of size $O(n^{-1})$. Aside from this, we can also view discrepancy from the uniform weights as a measure of relative entropy with respect to the uniform distribution. The Cramér–von Mises metric has obvious practical appeal as it selects weights that lead to the constrained density deviating as little as possible from the unconstrained density. Moreover, as shown in simulations here and in Z. Li, Liu, and Li (2017), selecting the weights to minimize $D_{CM}(p)$ naturally produces density estimates which are closer to $f(x)$ than that arising from minimization of $D_{L_2}(p)$.

It is worth pointing out that the use of the power-divergence metric (Cressie and Read 1984),

$$D_{\rho}(p) = \frac{1}{\rho(1-\rho)} \left(n - \sum_{i=1}^{n} (np_i)^{\rho}\right),$$

in this setting may not be a particularly fruitful choice; this metric requires that the p_is for estimating a density via (2) are non-negative (the p_is must satisfy $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for this approach), and some constraints may in fact require the use of negative weights. Furthermore, while $D_{\rho}(p)$ has an appealing immediate interpretation as a measure of entropy, it does require user selection of an additional tuning parameter for its implementation (ρ). Lastly, as noted by Hall and Huang (2002), issues can surface as p_i approaches 0 which is due to the fact that enforcing constraints on a curve

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

leads to "data compression" (i.e., the effective sample size that is used locally is smaller than the corresponding effective sample size for the unconstrained estimator). This difference is achieved by setting some of the constraint weights to 0. This information is not lost however, but simply reassigned to observations receiving non-zero weights. Thus, there can be substantial differences between our elected metrics and $D_\rho(p)$; while both $D_{L_2}(p)$ and $D_{CM}(p)$ are well behaved when p_i approaches 0, $D_\rho(p)$ may not be applicable for certain constraints for the reasons outlined above with particular values of ρ.

2.2 Bounded Support PDF Kernel Functions

We wish to develop an approach that will suit the many and varied needs of a range of practitioners. One issue that affects the quality of kernel density estimates arises in the presence of substantial probability mass occurring at a support boundary which leads to so-called *boundary bias*. To overcome this various approaches have been proposed, with the most well-known being *data-reflection*, *data-transformation* and the use of *kernel carpentry*. Data-reflection, as its name implies, involves duplicating symmetrically (i.e., reflecting) data around its boundary, running standard bandwidth selection and kernel estimation, then adjusting the resulting estimate to ensure it is proper (i.e., integrates to one) on its support. Data-transformation involves some mathematical transform of the data that, when rescaled, has the desired effect. Kernel carpentry, on the other hand, uses kernel functions that adapt to the presence of a boundary thereby mitigating the impact of the boundary. To some degree, all methods can reduce the amount of bias that would otherwise be present near a boundary to that which holds in the interior of support where it is free from boundary effects (in effect lying h or greater distance from the boundary in the interior). However, data-reflection and transformation require extra steps be taken by the user, which is both inconvenient and unnecessary. In what follows we take a kernel carpentry approach and adopt truncated kernel functions of the type

$$K(z,a,b) = \begin{cases} \frac{K(z)}{G(z_b)-G(z_a)} & \text{if } z_a \leq z \leq z_b, \\ 0 & \text{otherwise,} \end{cases}$$

where $z = (x-X)/h$, with X the random variable representing the X_i's, $z_b = (b-x)/h$, $z_a = (a-x)/h$, and where $G(z) = \int_{-\infty}^{z} K(t)\,dt$. Since $K(z)$ is a standard univariate kernel function, $G(z)$ is the CDF counterpart to the PDF $K(z)$ that we used to estimate $F(x)$. The astute reader will recognize

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

that if $K(z)$ is, for instance, the Gaussian density function, then $K(z,a,b)$ is simply the (doubly) truncated Gaussian density function. When $a=-\infty$ and $b=\infty$ then $K(z,a,b)=K(z)$ which is a standard kernel function such as the Gaussian (or Epanechnikov), hence this kernel function allows for unbounded or compact support without modification hence its adoption. When conducting constrained estimation it may be necessary to use the integrated version of $K(z,a,b)$ or derivatives thereof. We briefly outline some helpful relationships used to obtain these objects from the doubly truncated kernel function $K(z,a,b)$.

Integral Kernel Functions (e.g., CDF kernels)

To reduce notational burden, let $H_{ba}(z)=H(z_b)-H(z_a)$ for any function $H(\cdot)$. To estimate a CDF using kernel methods in the presence of support bounds, one can obtain the counterpart to $K(z,a,b)$ by adopting the following transformation for (doubly) truncated density functions, i.e.,

$$G(z,a,b)=\begin{cases} 0 & \text{if } z<z_a \\ \frac{G(\max(\min(z,z_b),z_a))-G(z_a)}{G_{ba}(z)} & \text{if } z_a\le z\le z_b, \\ 1 & \text{otherwise,} \end{cases}$$

Derivative Kernel Functions

Some of the constraints we consider are placed on the derivative of the kernel density estimates, and hence derivatives of the kernel function may be required. To that end we consider application of the quotient rule to obtain the first derivative of the doubly truncated kernel function which yields

$$K'(z,a,b)=\begin{cases} \frac{K'(z)}{G_{ba}(z)}-\frac{K(z)G'_{ba}(z)}{G_{ba}(z)^2} & \text{if } z_a\le z\le z_b, \\ 0 & \text{otherwise.} \end{cases}$$

Note that

$$\frac{K(z)G'_{ba}(z)}{G_{ba}(z)^2}=K(z,a,b)\frac{K_{ba}(z)}{G_{ba}(z)}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

The second derivative is found by application of the quotient and product rules which yields

$$K^{''}(z,a,b) = \begin{cases} \frac{d}{dx}\frac{K'(z)}{G_{ba}(z)} - \frac{d}{dx}\frac{K(z)G'_{ba}(z)}{G_{ba}(z)^2} & \text{if } z_a \leq z \leq z_b, \\ 0 & \text{otherwise.} \end{cases}$$

We note that the first term on the right hand side can be expressed as

$$\frac{d}{dx}\frac{K'(z)}{G_{ba}(z)} = \frac{K^{''}(z)}{G_{ba}(z)} - \frac{K'(z)K_{ba}(z)}{G_{ba}(z)^2},$$

while the second term (ignoring the minus sign) can be expressed as

$$\begin{aligned} \frac{d}{dx}\frac{K(z)K_{ba}(z)}{G_{ba}(z)^2} &= \frac{K'(z)(K(z_b) - K(z_a)) + K(z)K'_{ba}(z)}{G_{ba}(z)^2} \\ &\quad - \frac{2K(z)K_{ba}(z)G_{ba}(z)G'_{ba}(z)}{G_{ba}(z)^4} \\ &= \frac{K'(z)K_{ba}(z) + K(z)K'_{ba}(z)}{G_{ba}(z)^2} - \frac{2K(z)K_{ba}(z)^2}{G_{ba}(z)^3}. \end{aligned}$$

Therefore, we obtain

$$K^{''}(z,a,b) = \begin{cases} \frac{K^{''}(z)}{G_{ba}(z)} - \frac{2K'(z)K_{ba}(z)+K(z)K'_{ba}(z)}{G_{ba}(z)^2} + \frac{2K(z)K_{ba}(z)^2}{G_{ba}(z)^3} & \text{if } z_a \leq z \leq z_b, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, for the Gaussian kernel, if $a = -\infty$ and $b = \infty$, then $K(z_a) = K(z_b) = K'(z_a) = K'(z_b) = 0$, and $G(z_b) - G(z_a) = 1$, hence $K^{'}(z,a,b) = K^{'}(z)$ and $K^{''}(z,a,b) = K^{''}(z)$ in the unbounded support case, as expected.

The utility of this doubly truncated kernel function is that it can directly admit unbounded support (i.e., on $(-\infty,\infty)$), support on $[a,\infty)$ with a finite, on $(-\infty,b]$ with b finite and on $[a,b]$ with both a and b finite without further modification. Using this kernel function allows us to deliver an approach that directly admits support bounds *and* shape constraints which we believe enhances its practical appeal by increasing its potential application.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

2.3 Hypothesis Testing

The validity of the shape constraints being imposed can be tested following the insights of Hall et al. (2001) and Du, Parmeter, and Racine (2013) via a bootstrap inferential procedure. Briefly summarizing, the test statistic is the value of the objective function taken from solving the quadratic program when imposing the constraints. The bootstrap procedure entails drawing bootstrap resamples from the null (i.e., constrained) density in order to construct the null distribution of the test statistic (i.e., the value of the objective function taken from solving the quadratic program when imposing the constraints on the bootstrap resamples). The test involves computing the P-value constructed from a comparison of the test statistic to those obtained from the empirical distribution that was constructed under the null or, alternatively, a comparison of the test statistic with the desired $1-\alpha$ quantile obtained from the empirical null distribution where α is the desired size of the test procedure (the test is one-sided with right-tailed rejection region). Further details are provided below.

More specifically, this bootstrap approach involves estimating the constrained density $\hat{f}(\mathbf{x}|p)$ based on the sample realizations $\{\mathbf{X}_i\}$ and then rejecting H_0 if the observed value of $D_\jmath(\hat{p})$ is too large, where $\jmath \in \{L_2, CM\}$. To ensure that the constraints are satisfied, we propose sampling from $\hat{f}(\mathbf{x}|p)$ rather than $\hat{f}(\mathbf{x}|p_{unif})$. A simple way to do this is via rejection sampling.

These resamples are generated under H_0, hence we recompute $\hat{f}(\mathbf{x}|p)$ for the bootstrap sample $\{\mathbf{X}_i^*\}$ which we denote $\hat{f}(\mathbf{x}|p^*)$ which then yields $D_\jmath(p^*)$. We then repeat this process B times. Finally, we compute the empirical P value, P_B, which is simply the proportion of the B bootstrap resamples $D_\jmath(p^*)$ that exceed $D_\jmath(\hat{p})$, i.e.,

$$P_B = 1 - \hat{F}(D_\jmath(\hat{p})) = \frac{1}{B}\sum_{j=1}^{B} I(D_\jmath(p^*) > D_\jmath(\hat{p})),$$

where $I(\cdot)$ is the indicator function and $\hat{F}(D_\jmath(\hat{p}))$ is the empirical distribution function of the bootstrap statistics. Then one rejects the null hypothesis if P_B is less than α, the level of the test.

We now consider a few illustrative applications of imposing shape restrictions before turning to the theoretical underpinnings of the proposed method.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

2.4 Illustrative Applications: Monotonicity and Concavity

Monotonicity and concavity constraints are but two popular shape constraint domains that our approach can cover. As in Du, Parmeter, and Racine (2013), we solve a simple quadratic program using (6) to generate the constrained estimate. Figure 1 presents results for a bounded density on $[0, 1]$ imposing monotonicity (the distribution is Beta(5,1)). For this simple illustration we generate 100 observations and select the bandwidth via Silverman's rule-of-thumb approach. We see little difference between the constrained and the unconstrained estimators for $x > 0.6$; all of the constraint enforcement occurs in the left tail of the density. Given our restriction that the weights sum to 0, this leads to only minor changes in the shape of the density beyond where the constraints need to be enforced. This becomes more clear by looking at the lower plot in Figure 1, which plots the constrained and unconstrained derivative estimates.

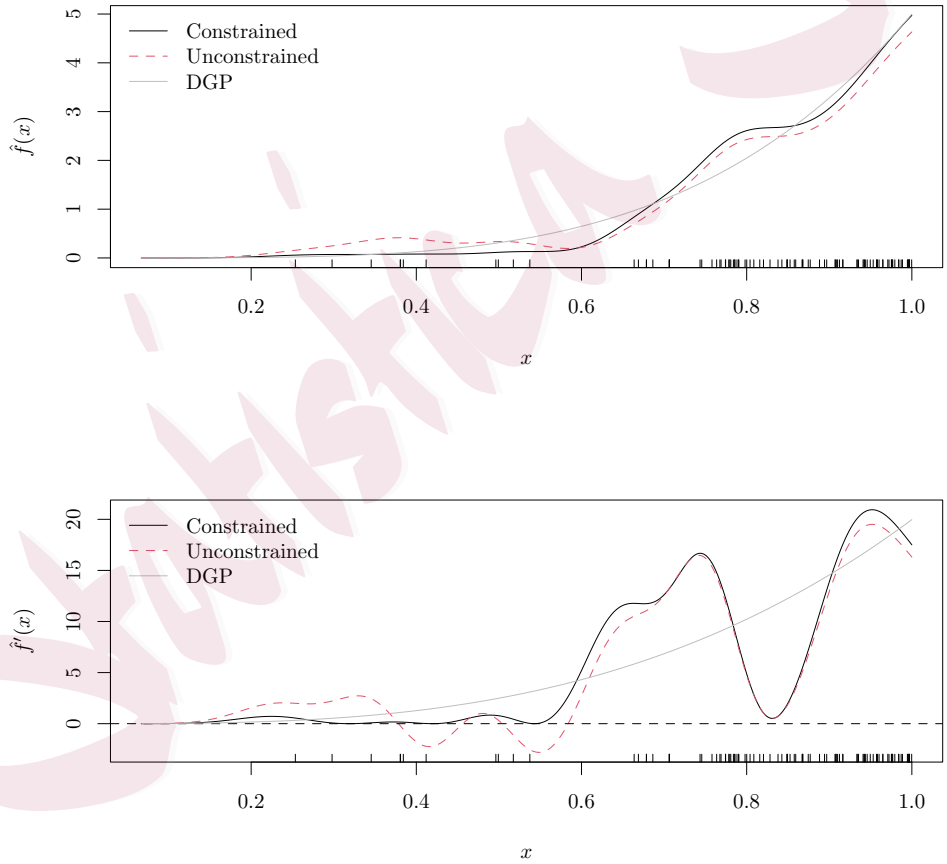

Figure 1: Monotone shape constrained density estimation ($\hat{f}'(x) \geq 0$). The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained first derivative estimates.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Figure 2 presents results for imposing concavity on an unbounded support random variable (the distribution is $N(0,1)$). Once again we generate 100 observations randomly and construct the bandwidth using Silverman's rule-of-thumb. Here we enforce concavity on the density, which is *not* a property of the Gaussian density (though it *is* log-concave). We see that enforcing *invalid* constraints produces substantial distortions in both the density and the corresponding first derivative, as expected.

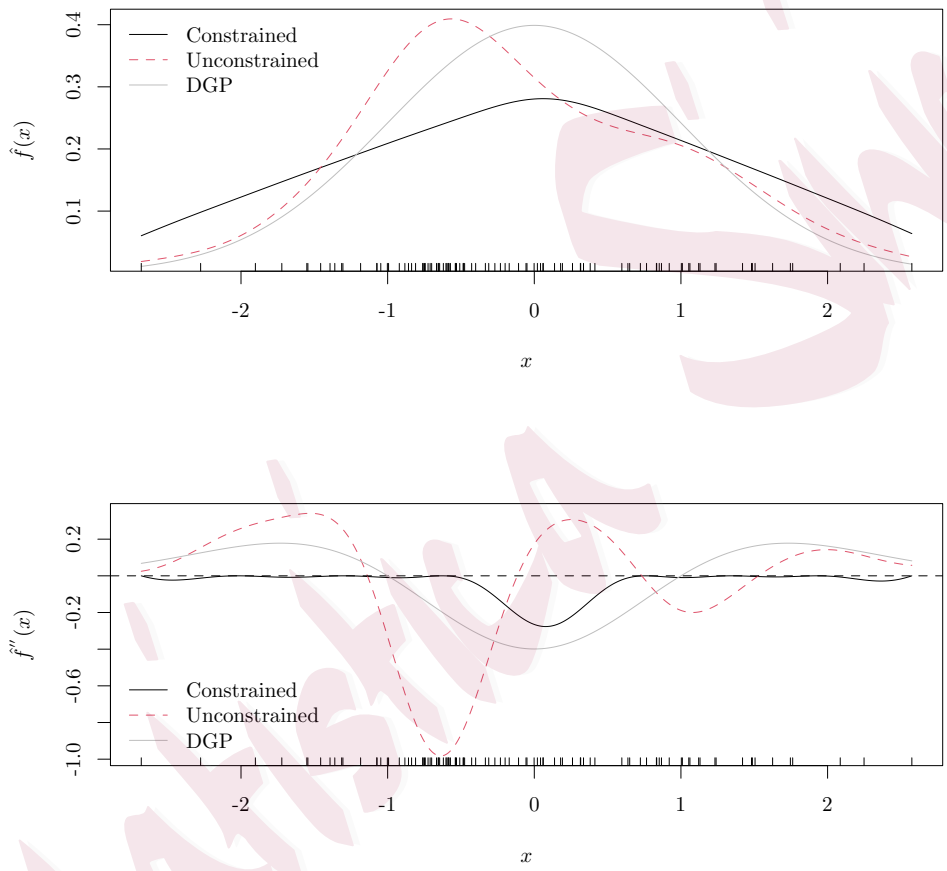

Figure 2: Concave shape constrained density estimation ($\hat{f}''(x) \leq 0$). The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained second derivative estimates.

2.5 Log-Concave Kernel Density Estimation

Log-concavity is an extremely popular constraint among practitioners (but only one of a multitude of shape constraint domains that our approach can cover). For imposing log-concavity/convexity,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

we require $d^2 \log(\hat{f}(x))/dx^2$ and $d^2K(Z_i)/dx^2$. The former is given by

$$\frac{d^2 \log(\hat{f}(x))}{dx^2} = \frac{\hat{f}''(x)\hat{f}(x) - (\hat{f}'(x))^2}{(\hat{f}(x))^2}$$

while the latter is given by

$$\frac{d^2K(Z_i)}{dx^2} = K''(Z_i).$$

Note that

$$\hat{f}'(x) = \frac{1}{nh}\sum_{i=1}^{n} K'(Z_i),$$
$$\hat{f}''(x) = \frac{1}{nh}\sum_{i=1}^{n} K''(Z_i).$$

Illustrative Application: Log-Concavity

Figure 3 presents results for a draw from the $N(0,1)$ Gaussian distribution. The Gaussian density is log-concave but the kernel estimate need not be, as the following example illustrates. We generate 250 observations from a standard normal distribution and use Silverman's rule-of-thumb bandwidth to smooth the density. As in Figure 1, there is little difference between the constrained and the unconstrained estimates. Moreover, the log-densities are also quite similar aside from the one region of non-concavity of the log-density for $-2.5 < x < -1.9$. Both the constrained and unconstrained densities integrate to 1 and are proper.

2.6 Categorical (ordered) Probability Mass Functions

The approach we consider for shape constrained PDF estimation can also be applied to shape constrained PMF estimation (Aitchison and Aitken 1976; Racine, Li, and Yan 2020). When X is an ordered categorical variable ($X \in \mathbb{D} = \{D_0, D_1, \ldots, D_{c-1}\}$ where c is the number of (ordered) outcomes) there is only the need for one value of a_i per outcome (as $a_i = a_j$ when $X_i = X_j$). When placing shape constraints on derivatives we adopt the classical convention that for discrete support variables derivatives are defined in terms of simple finite differences. For an ordered discrete random variable, we use the notation $P(x) = Pr(X = x)$ to denote the PMF. Let $\hat{P}(x)$ denote the kernel

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

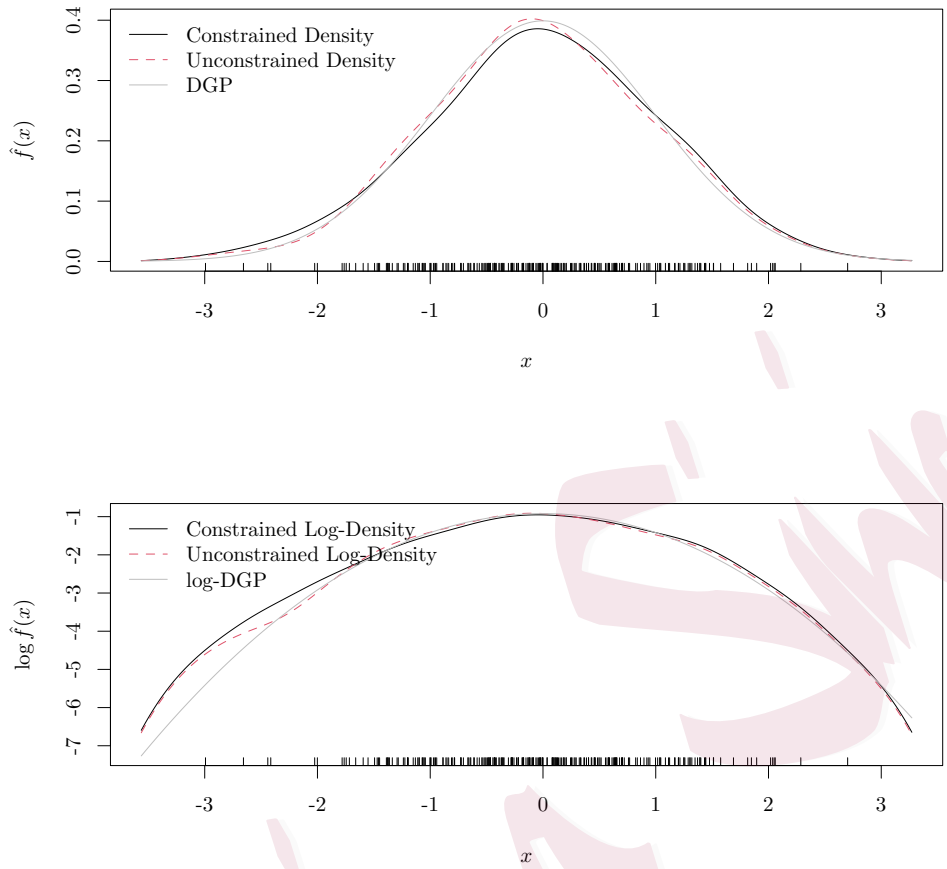

Figure 3: Log-concave shape constrained density estimation. The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained log-density estimates.

estimate of $P(x)$ given by

$$\hat{P}(x) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, x, \lambda),$$

where $l(X_i, x, \lambda)$ is an appropriate kernel function for ordered discrete support random variables. The counterpart of the first derivative in this setting is $\Delta_j(x) = (P(x_{(j)}) - P(x_{(j-1)}))/(x_{(j)} - x_{(j-1)})$ where the $x_{(j)}$ are the order statistics, and this quantity can be computed directly from an unconstrained estimate (as can higher order derivatives if needed). As was the case for shape constrained PDF estimation, the counterpart to (3) for PMF estimation can be written as

$$\hat{P}(x|p) = \hat{P}(x) + \sum_{i=1}^{n} a_i l(X_i, x, \lambda), \tag{7}$$

where λ is the smoothing parameter analogous to the bandwidth h for its continuous support counterpart. The mechanics of the shape constrained PMF estimator are the same as those for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

shape constrained PDF estimation described previously and will not be repeated here (see Racine, Li, and Yan (2020) for further details). We now consider an empirical illustration based on count data which has ordered discrete support.

Empirical Application: Shape Constrained PMFs

We consider a dataset collected by Hausman, Hall, and Griliches (1984) which records the number (count) of successful patent applications by 128 U.S. firms across a seven-year period (1968-1974). We model the kernel smoothed PMF for the number of successful patent applications with likelihood cross-validated bandwidth selection and present results in Figure 4. The non-smooth estimate is quite noisy, while the smooth estimate is much less so. Like its empirical counterpart, the smooth estimate delivers probability estimates that *sum* to 1, but the smooth estimate is expected to be more efficient from the square error perspective.

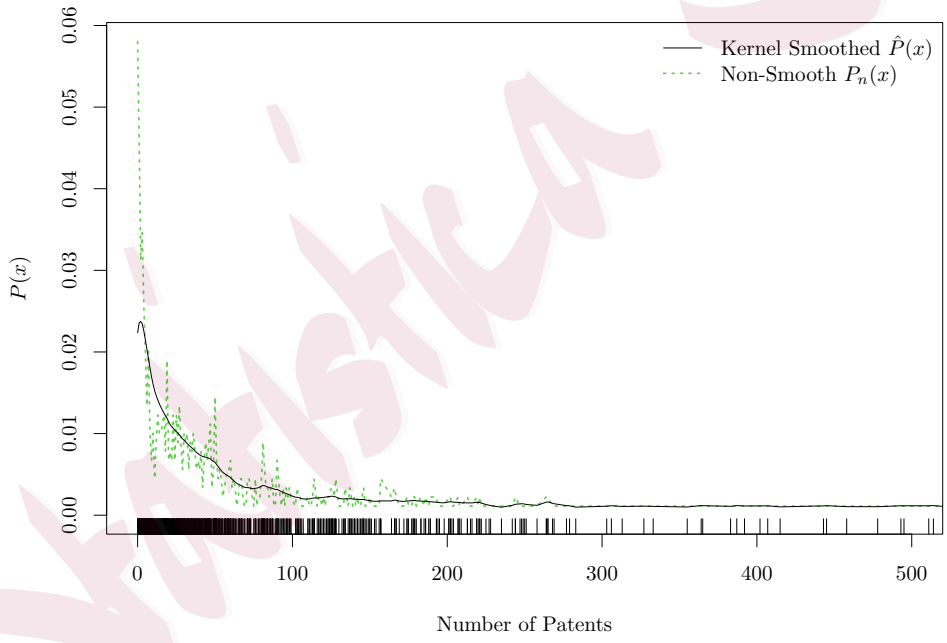

Figure 4: Unconstrained smooth and non-smooth probability mass function estimates for the patent data. The smooth estimate appears as a solid line, the non-smooth estimate as a dotted line.

Figure 4 reveals that the unconstrained kernel PMF estimator, though perhaps more plausible an estimate than the non-smooth empirical estimator, implausibly changes sign in many places. A perhaps more reasonable assumption is that the estimate is monotonically decreasing, hence we consider imposing this shape constraint on the kernel PMF estimate. Figure 5 presents the smooth

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

unconstrained and monotonically constrained estimates. As noted above, the derivatives for the PMF estimate are given by $\Delta_j(x) = (P(x_{(j)}) - P(x_{(j-1)}))/(x_{(j)} - x_{(j-1)})$ where $x_{(j)}$ are the order statistics, which can be computed directly. The weight matrix required for solving the quadratic program is then the difference between kernel functions evaluated at $x_{(j)}$ and $x_{(j-1)}$ divided by the difference $x_{(j)} - x_{(j-1)}$. For imposing the monotonically decreasing constraint we define $\Delta_1(x) \leq 0$ (we reverse this definition for monotonically increasing constraints).

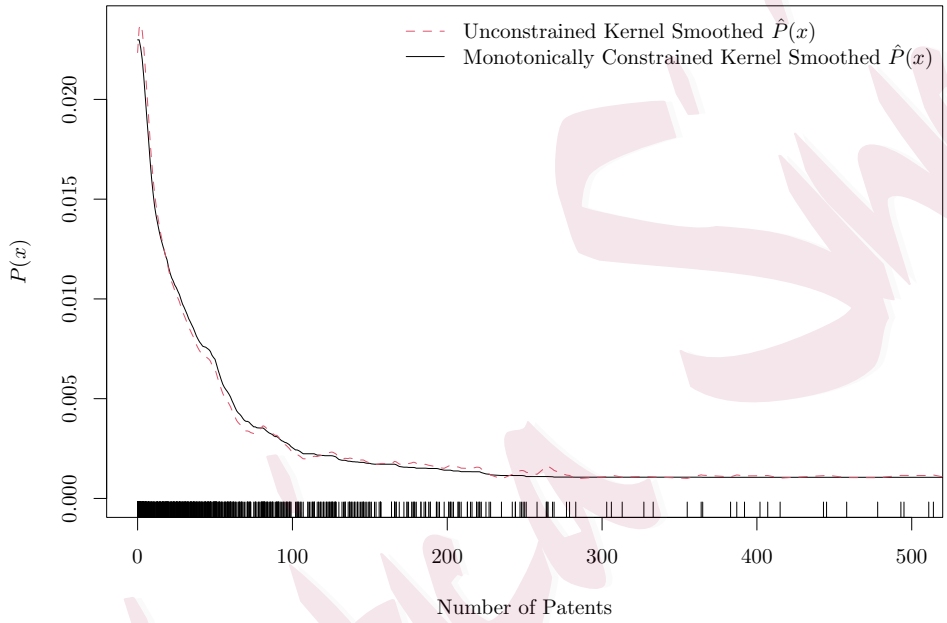

Figure 5: Unconstrained and constrained smooth probability function estimates for the patent data.

3 Theoretical Properties of the Constrained Estimator

In this section we provide four key theoretical results. First, under weak conditions, the constraint weights generated by our approach are shown to be well defined and unique. Second, we demonstrate consistency of the constrained density estimator where appropriate in terms of its closeness to the unconstrained density estimator which is well-known to be consistent. We consider three distinct settings: (i) when the constraints are indeed true on the entire support of X; (ii) when the constraints are satisfied everywhere except at points of measure zero; and (iii) when the constraints are violated on a set with positive measure. For (i) and (ii) we establish consistency of the constrained density estimator under weak conditions on the order of derivatives of the true density and on the bandwidth (naturally (iii) does not allow for consistent estimation). Third, we extend our results in the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

continuous case to those for ordered PMFs. Here we are only able to establish consistency when the constraints hold on the entire support of the discrete random variable, however these results are novel and of practical value. Fourth, we provide the asymptotic distribution of our proposed test statistic when testing the null hypothesis of the validity of the shape constraints being imposed.

Our theoretical results for the continuous data setting are similar to that in Hall and Huang (2001) and Du, Parmeter, and Racine (2013) but with four important differences. First, while Hall and Huang (2001) and Du, Parmeter, and Racine (2013) focused on imposing constraints in the regression setting, the density setting is complicated by the lack of an error term such that direct application of the existing theory is not available. Second, whereas Hall and Huang (2001) used the power-divergence measure of Cressie and Read (1984) and Du, Parmeter, and Racine (2013) used the L_2 metric, here we establish consistency of the constrained estimator using the objective function proposed by Z. Li, Liu, and Li (2017). This objective function, rather than selecting constraint weights which are as close as possible to the uniform weights (as in Du, Parmeter, and Racine 2013), selects weights which are as close as possible to the unconstrained estimator. Intuitively, this modification makes sense given that the unconstrained estimator is consistent to begin with. While Z. Li, Liu, and Li (2017) have shown impressive finite-sample properties of their objective function for selecting constraint weights for a constrained K_{nn} regression estimator, the change in the objective function also necessitates changes in existing theory. Third, while the theory that is developed works quite well for constraints on the density, when imposing constraints on, for example, the log-density, several additional modifications are required. Fourth, we develop the appropriate theory for constrained estimation of the PMF. To our knowledge this is the first application of these types of constrained methods to kernel smoothed discrete data.

To begin, $\mathbf{X}_i$ is of dimension r. Our goal is to impose constraints on the density (or log-density) of the form $f^{(\mathbf{s})}(\mathbf{x}) = [\partial^{s_1} f(\mathbf{x}) \cdots \partial^{s_r} f(\mathbf{x})]/[\partial x_1^{s_1} \cdots \partial x_r^{s_r}]$ (or $\log f^{(\mathbf{s})}(\mathbf{x})$) where $\mathbf{s}$ is an r-vector corresponding to the dimension of $\mathbf{x}$. One can observe that the general two-sided constraints in (5) can be expressed as one-sided constraints of the form

$$\sum_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k} f^{(\mathbf{s})}(\mathbf{x}) - c_k(\mathbf{x}) \geq 0, \quad k = 1, \ldots, T, \tag{8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

where T is the total number of restrictions, with the sum being taken over all density derivative vectors in $\mathbf{S}_k$ while $\alpha_{\mathbf{s},k}$ is used to generate the appropriate constraints imposed on the density derivatives ($j = 0, 1, \ldots$). This notation admits an arbitrary number of internally consistent constraints imposed simultaneously on the density and its derivatives, though for most circumstances we expect that a single constraint (i.e., $T = 1$) will suffice.[4]

Before more formally developing the theory for our general constrained density estimator, we introduce some additional simplifying notation. Denote the domain of interest by $\mathcal{J} \equiv [\mathbf{m}, \mathbf{b}] = \prod_{i=1}^{r}[m_i, b_i]$. We also define a differential operator $f \mapsto f^{\mathcal{D}}$ such that $f^{\mathcal{D}}(\mathbf{x})$ is a length-T vector with kth entry $\sum\limits_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k} f^{(\mathbf{s})}(\mathbf{x})$. We take $|\mathbf{s}| = \sum_{i=1}^{r} s_i$ as the *order* for a derivative vector $\mathbf{s} = (s_1, \ldots, s_r)$, and say a derivative $\mathbf{s}_1$ has a *higher order* than $\mathbf{s}_2$ if $|\mathbf{s}_1| > |\mathbf{s}_2|$. Let $\mathbf{S} = \cup_{k=1}^{T} \mathbf{S}_k$ and $\mathbf{d}_{\mathbf{S}}$ be the derivative of the *maximum order* among all the derivatives in $\mathbf{S}$; for simplicity, we drop the subscript $\mathbf{S}$ from $\mathbf{d}_{\mathbf{S}}$. Without loss of generality, we set $c_k(x) = 0$ in what follows. Plugging (2) into (8) yields

$$\sum_{i=1}^{n} p_i K_i^{\mathcal{D}}(\mathbf{x}) \geq 0. \tag{9}$$

Here $K_i^{\mathcal{D}}(\mathbf{x})$ represents the form of the constraints based on the appropriate kernel derivatives, i.e., it subsumes the appropriate entries of derivative vector $f^{\mathcal{D}}(\mathbf{x})$. Lastly, we define $\tilde{f}(x) = \hat{f}(x|p_{unif})$ to further simplify notation in what follows.

While the theory we present is capable of imposing constraints on either the density or the log-density, for notational simplicity only we presume that the practitioner is interested in one or the other.

3.1 Existence of the Constrained Probability Density Function Estimator

The first result that we establish is an existence result, i.e., that a set of weights exists provided that the constraints imposed are internally consistent and satisfy the constraints in (9).

Theorem 3.1 (Existence)**.** *Assume that the set* $\{1, \ldots, n\}$ *contains a sequence* $\{i_1, \ldots, i_k\}$ *with the following properties.*

i) For each $\ell = 1, \ldots, k$, $K_{i_\ell}^{\mathcal{D}}(\mathbf{x})$ *is strictly positive and continuous on an open set* $\mathbf{O}_{i_\ell} \subset \mathbb{R}^r$, *and*

[4]For $r = 1$ and the imposition of monotonicity we would have $T = 1$ with $\mathbf{s} = (1)$, $\mathbf{S}_k = \{(1)\}$, $\alpha_{\mathbf{s},k} = 1$ and $c_k(\mathbf{x}) = 0$ for all $\mathbf{x}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

vanishes on $\mathbb{R}^r \setminus \mathbf{O}_{i_\ell}$,

ii) Every $\mathbf{x} \in \mathcal{J}$ *is contained in at least one open set* $\mathbf{O}_{i_k}$,

iii) For $1 \leq \ell \leq n$, $K_{i_\ell}^{\mathcal{D}}(\mathbf{x})$ *is continuous on* $(-\infty, \infty)^r$.

Then there exists a vector $p = (p_1, \ldots, p_n)$ *such that the constraints are satisfied for all* $\mathbf{x} \in \mathcal{J}$.

Conditions *i)* and *ii)* of Theorem 3.1 are to ensure the existence of an open cover of the domain $\mathcal{J}$ by the open sets $\mathbf{O}_{i_\ell}$ on which $K_{i_\ell}^{\mathcal{D}}$ is positively supported for some i_ℓ. We note that the above conditions are sufficient but not necessary for the existence of a set of weights that satisfy the constraints for all $\mathbf{x} \in \mathcal{J}$. For example, if $\operatorname{sign} K_{j_n}^{\mathcal{D}}(\mathbf{x}) = 1 \; \forall \mathbf{x} \in \mathcal{J}$ for some sequence j_n in $\{1, \ldots, n\}$ and $\operatorname{sign} K_{l_n}^{\mathcal{D}}(\mathbf{x}) = -1 \; \forall \mathbf{x} \in \mathcal{J}$ for another sequence l_n in $\{1, \ldots, n\}$, then for those observations that switch signs, p_i may be set equal to zero, while $p_{j_n} > 0$ and $p_{l_n} < 0$ is sufficient to ensure existence of a set of p's satisfying the constraints. The proof of Theorem 3.1 can be found in the technical appendix.

3.2 Consistency of the Constrained Probability Density Function Estimator

Here we detail the consistency of our constrained estimator. To begin, define a *hyperplane subset* of $\mathcal{J}$ to be a subset of the form $\mathcal{S} = \left\{ x_{0k} \times \prod_{i \neq k} [m_i, b_i] \right\}$ for some $1 \leq k \leq r$ and some $x_{0k} \in [m_k, b_k]$. We call $\mathcal{S}$ an *interior hyperplane subset* if $x_{0k} \in (m_k, b_k)$. For what follows, $f(\cdot)$ (or $f^{\mathcal{D}}(\cdot)$) is the true density (or its derivative), $\hat{p}$ is the optimal weight vector satisfying the constraints, $\hat{f}(\cdot|\hat{p})$ (or $\hat{f}^{\mathcal{D}}(\cdot|\hat{p})$) is the constrained estimator defined in (3), and $\tilde{f}(\cdot)$ (or $\tilde{f}^{\mathcal{D}}(\cdot)$) is the unconstrained estimator defined in (3).

Assumption A1.

i) The sample $\mathbf{X}_i$ either form a regularly spaced grid on a compact set $\mathcal{I} \equiv [\mathbf{c}, \mathbf{e}] = \prod_{i=1}^{r} [c_i, e_i]$ or constitute independent random draws from a distribution whose density f is continuous and non-vanishing on $\mathcal{I}$; the kernel function $K(\cdot)$ is a symmetric, compactly supported density such that $K^{\mathcal{D}}$ is Hölder-continuous on $\mathcal{J} \subset \mathcal{I}$.

ii) $f^{\mathcal{D}}$ is continuous on $\mathcal{J}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

iii) The bandwidth associated with each variable, h_j, satisfies $h_j \propto n^{-1/(3r+2|\mathbf{d}|)}$, $1 \leq j \leq r$, where $|\mathbf{d}|$ is the maximum order of the derivative vector $\mathbf{d}$.

iv) The true density f is bounded away from 0, say, $f(\mathbf{x}) > \tau$ for some fixed constant $\tau > 0$.

Assumption A1 i) is standard in the kernel density literature.[5] Assumption A1 ii) assures requisite smoothness of $f^{\mathcal{D}}$. Note that the bandwidth rate in Assumption A1 iii) is generally higher than the standard optimal rate $n^{-1/(r+4)}$. However, this is not surprising for our restricted problem. The optimal rate only guarantees the convergence of our unrestricted function estimator $\tilde{f}$. But the restricted problem also requires the convergence of the derivative $\tilde{f}^{\mathcal{D}}$, which often needs a higher bandwidth rate. In the single-predictor monotone regression problem considered in Hall and Huang (2001), this rate happens to coincide with the optimal rate $n^{-1/5}$. Furthermore, when the bandwidths all share the same rate, one can rescale each component of $\mathbf{x}$ to ensure a uniform bandwidth $h \propto n^{-1/(3r+2|\mathbf{d}|)}$ for all components. This simplification is made without loss of generality. Thus we use h^r rather than $\prod_{j=1}^{r} h_j$ for notational simplicity. If we consider densities on a compact interval then Assumption A1 iv) is not so restrictive. However, it may not work for common densities such as normal, exponential, etc.

Theorem 3.2 (Consistency)**.** *Suppose that Assumption A1 1.-4. holds.*

i) If $f^{\mathcal{D}} > 0$ on $\mathcal{J}$ then, with probability 1, $\hat{p} = 1/n$ for all sufficiently large n and $\hat{f}^{\mathcal{D}}(\cdot|\hat{p}) = \tilde{f}^{\mathcal{D}}$ on $\mathcal{J}$ for all sufficiently large n. Hence, $\hat{f}(\cdot|\hat{p}) = \tilde{f}$ on $\mathcal{J}$ for all sufficiently large n.

ii) Suppose that $f^{\mathcal{D}} > 0$ except on an interior hyperplane subset $\mathcal{X}_0 \subset \mathcal{J}$ where we have $f^{\mathcal{D}}(\mathbf{x}_0) = 0, \forall \mathbf{x}_0 \in \mathcal{X}_0$. Also, for any $\mathbf{x}_0 \in \mathcal{X}_0$, suppose that $f^{\mathcal{D}}$ has second order continuous derivatives in the neighbourhood of $\mathbf{x}_0$ with $\frac{\partial f^{\mathcal{D}}}{\partial \mathbf{x}}(\mathbf{x}_0) = \mathbf{0}$ and $\frac{\partial^2 f^{\mathcal{D}}}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0)$ nonsingular; then $|\hat{f}(\cdot|\hat{p}) - \tilde{f}| = O_p\left(h^{|\mathbf{d}|+\frac{r+1}{2}}\right)$ uniformly on $\mathcal{J}$.

iii) Under the conditions in ii), there exist random variables $\Theta = \Theta(n)$ and $Z_1 = Z_1(n) \geq 0$ satisfying $\Theta = O_p\left(h^{|\mathbf{d}|+r+1}\right)$ and $Z_1 = O_p(1)$, such that $1 - \Theta \leq \hat{f}(x|\hat{p})/\tilde{f}(x) \leq 1 + \Theta$ uniformly for $\mathbf{x} \in \mathcal{J}$ with $\inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| > Z_1 h^{\frac{r+1}{4}}$.

In Theorem 3.2, part i) suggests that when the constraint is strictly satisfied by the true function,

[5] At the expense of a more tedious proof, the same results can be demonstrated if the density were assumed to exist on an r-dimensional ball instead of a hypercube.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

the constrained estimator $\hat{f}(\cdot|\hat{p})$ and the unconstrained estimator $\tilde{f}$ are essentially the same and thus share the same rate of convergence. Part ii) gives the order of difference between $\hat{f}(\cdot|\hat{p})$ and $\tilde{f}$ when $f^{\mathcal{D}} = 0$ on an interior hyperplane. Note that the order in ii) indicates a different convergence rate of $\hat{f}(\cdot|\hat{p})$ from that of $\tilde{f}$ in such a case. Part iii) is concerned with the asymptotic behaviour of the weights $\hat{p}$ in such a case. Also note that the results are easily extendable to the case of $f^{\mathcal{D}} \leq 0$ with a switch of sign in f.

The proof of Theorem 3.2 appears in the Appendix. Theorem 3.2 is the multivariate, multi-constraint, hyperplane subset adaptation of Du, Parmeter, and Racine (2013) to density estimation using the metric in (6).

3.3 Theoretical Properties of the Constrained Probability Mass Function Estimator

Theorems 3.1 and 3.2 can be extended to the ordered discrete support setting under similar assumptions, though with some important modifications required.

Assumption B1.

i) Assume that the set $\{1, \ldots, n\}$ contains a sequence $\{i_1, \ldots, i_k\}$ with the following properties:
   (i) For each k, $\ell_{i_k}^{\mathcal{D}}(x)$ is strictly positive on a non-empty set $\mathbf{O}_{i_k} \subset \mathbb{D}$, and vanishes on $\mathbb{D} \setminus \mathbf{O}_{i_k}$;
   (ii) Every $x \in \mathbb{D}$ is contained in at least one non-empty set $\mathbf{O}_{i_k}$.

ii) Assume that the kernel function $l(\cdot)$ in (7) is an ordered kernel function and that the smoothing parameter λ in (7) is of order $\lambda = O_p(n^{-1})$, a standard result in the literature.

Assumption B1 i) is similar to the sufficient conditions in Theorem 3.1 for the continuous case. For the smoothing parameter condition in Assumption B1 ii), Ouyang, Li, and Racine (2006) have shown that a cross-validation selected smoothing parameter λ can have order $O_p(n^{-1})$ so long as not all the marginal distributions of X are uniform.

Theorem 3.3 (PMF Estimator)**.** *Suppose that Assumption B1 holds. Then we have the following properties for the constrained PMF estimate $\hat{P}^{\mathcal{D}}(\cdot|\hat{p})$. Our use here of the differential is with respect to difference order as opposed to differentiation.*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

i) There exists a vector $p = (p_1, \ldots, p_n)$ *such that the constraints are satisfied for all* $x \in \mathbb{D}$.

ii) If $P^{\mathcal{D}} > 0$ *on* $\mathbb{D}$ *then, with probability 1,* $\hat{p} = 1/n$ *for all sufficiently large* n *and* $\hat{P}^{\mathcal{D}}(\cdot|\hat{p}) = \tilde{P}^{\mathcal{D}}$ *on* $\mathbb{D}$ *for all sufficiently large* n. *Hence,* $\hat{P}(\cdot|\hat{p}) = \tilde{P}$ *on* $\mathbb{D}$ *for all sufficiently large* n.

We remark here that the proof of existence requires only minor changes to the proof for the continuous data setting, and is thus omitted. The proof for consistency still requires taking differences across the different cells of the discrete random variable, which suggests that our imposition of constraints corresponds to an ordered discrete random variable (Q. Li and Racine 2003).

For the proof of consistency, note that parts ii) and iii) cannot be generalized. This result has a straightforward intuitive explanation. In the continuous only setting these parts focus on the case where the constraint is violated on a set of measure 0. The argument is that even if the constraint is violated, so long as it occurs on an interior subset hyperplane the constrained estimator is still a consistent estimator for the unknown density (except on a set of measure 0). In the discrete data setting this suite of results no longer holds. This is due to the fact that for a discretely supported random variable, a measure zero event is equivalent to an outcome not in the support, or, more directly, a violation of the constraint is more troubling when considering discrete data.

3.4 Asymptotic Properties of $D(\hat{p})$

Our discussion on inference of smoothness constraints will follow the same setup as Du, Parmeter, and Racine (2013). We will focus on the use of the L_2 norm rather than CM since a closed form solution for the optimal weights is mathematically more tedious due to cross-products of the weights in the objective function. We note that the asymptotic expansion of the weights will be of the same order between L_2 and CM but will obviously be of a slightly different form. Let $\psi_i(\mathbf{x}) = K_i^{\mathcal{D}}(\mathbf{x})$, $i = 1, \ldots, n$.

Recall that our minimization problem is

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} (n^{-1} - p_i)^2, \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \psi_i(\mathbf{x}) \geq 0, \forall \mathbf{x}.$$

In practice, this minimization is carried out by taking a fine grid $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, where N is large,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

and solving

$$\min_{p_i,\ldots,p_n} \sum_{i=1}^{n}(n^{-1}-p_i)^2, \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i\psi_i(\mathbf{x}_j) \geq 0, 1 \leq j \leq N. \tag{10}$$

We place the same assumption on the grid points $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ as Du, Parmeter, and Racine (2013).

Assumption B2.

i) $N \to \infty$ as $n \to \infty$ and $N = O(n)$.

ii) Let $d_N = \inf_{1\leq j_1, j_2 \leq N} |\mathbf{x}_{j_1} - \mathbf{x}_{j_2}|$ be the minimum distance between grid points. We require $d_N \to 0$ and $h^{-1}d_N \to \infty$.

Assumption B2 essentially dictates how the grid points behave. We need to ensure that the grid becomes effectively dense as n increases (Assumption B2 i)) while we also need the speed at which the smallest distance between grid points decays to be slower than the rate of decay of the smoothing parameters (Assumption B2 ii)). This last assumption is necessary to eliminate correlation across $\psi_i(\mathbf{x})$ as n grows (Chacón, Duong, and Wand 2011).

Let $\hat{p}_i, i = 1, \ldots, n$, be the solution to the quadratic programming problem in (10). Then the asymptotic distribution of $D(\hat{p})$ is given in the following theorem, whose proof is relegated to the supplementary material.

Theorem 3.4. *Suppose that assumptions A1 i)-iv) and B1 i)-iv) hold. Then, as $n \to \infty$, we have*

$$\frac{n^2\sigma^2_{K^{(\mathbf{d})}}}{h^{2|\mathbf{d}|+r}\left(\sum_{j=1}^{M} f^{\mathcal{D}}(\mathbf{x}_j^*)\right)^2} D(\hat{p}) \sim \chi^2(n), \tag{11}$$

where $\sigma^2_{K^{(\mathbf{d})}} = \int \left[K^{(\mathbf{d})}(y)\right]^2 dy$, and $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_M^*\} \subset \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are the slack points as defined in the supplementary material.

Theorem 3.4 is the density equivalent of the regression based test proposed by Du, Parmeter, and Racine (2013). Aside from several structural details, the main result follows from their initial theory. The diverging degrees of freedom is expected as both H_0 and H_1 are evaluated on infinite dimensional parameter spaces (see also Fan, Zhang, and Zhang (2001)). Also note that similar to

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

the generalized likelihood ratio test statistic in Fan, Zhang, and Zhang (2001), the distributional convergence in (11) is equivalent to $\sqrt{2n}(T_n - n) \xrightarrow{\mathcal{L}} N(0,1)$, where T_n is the statistic on the left hand side of (11).

Given the well-known issues with the speed of convergence of nonparametric tests, it is recommended to deploy a bootstrap algorithm instead.[6] Du, Parmeter, and Racine (2013) showed the consistency of the hypothesis test using $D(\hat{p})$ as the test statistic, which implies consistency of the bootstrap version. In the constrained density setting, implementation of the test consists of two steps:

i) If the true density f satisfies the shape constraints, then as $n \to \infty$,

$$P\{D(\hat{p}) \leq n\epsilon\} \to 1 \quad \text{for all } \epsilon > 0.$$

ii) If the true function f does not satisfy the shape constraints on $\mathcal{J}$, then

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} P\{D(\hat{p}) \geq n\epsilon\} = 1.$$

This result has a simple intuitive explanation. If the unconstrained estimator satisfies the constraints, then $D(\hat{p}) = 0$ and clearly there is no need to construct the constrained estimator since clearly the implication is that the constraints are most likely true. However, if the constraints are not initially satisfied, then $D(\hat{p})$ can be used to test the validity of the constraints.

One might consider generalizing the result provided above to admit different metrics such as, for example, those strictly tailored to probability weights (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$). Although our theory for consistency (Theorem 3.3) is developed for the Cramér–von Mises statistic, it could instead be developed using a power-divergence metric by following Hall and Huang (2001) or by using an L_2 norm metric following Du, Parmeter, and Racine (2013). The difference will lie mainly in specific algebraic manipulations for the different metrics. For our theory for the limiting distribution of our test statistic (Theorem 3.4), we relied on the L_2 norm in part because it delivers a tractable solution from the Karush-Kuhn-Tucker conditions. A similar result for, say, the power-divergence

[6] Another reason to prefer the bootstrap is that the normalizing constant in (11) requires the determination of slack points which can be rather difficult in practice.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

metric is possible, though some degree of approximation will still be necessary in order to obtain a suitable expression for the weights underlying the corresponding test statistic.

4 Monte Carlo Finite-Sample Performance

In this section we assess the finite-sample performance of the proposed estimator via comparison with popular peers implemented in currently supported R packages available through CRAN as of this writing. The reader will note that the proposed estimator is extremely flexible in terms of the type of constraint and number of simultaneous constraints that could be imposed. For the sake of brevity we therefore restrict our focus to a few leading test cases, and we also restrict the peer group to the most popular and promising methods of which we are currently aware. The test cases we shall consider involve two popular constraints, namely the *log-concavity* constraint and the *monotonicity* constraint (i.e., monotonically increasing). Given that our approach supports both smooth constrained PDFs and PMFs, we also shall restrict attention here to constrained PDF estimation only given the lack of peers for smooth constrained PMFs of which we are aware, however, we did provide an illustrative example involving the PMF for the interested reader and R code for constrained PDF and PMF estimation is available. The proposed approach can be found in the R package `np` (Hayfield and Racine 2008) which is available on CRAN.[7]

For comparison purposes, in the log-concave constraint setting we shall compare the proposed approach with that of Cule, Samworth, and Stewart (2010) who studied a non-smooth log-concave PDF MLE and Chen and Samworth (2013) who studied the associated smoothed log-concave estimator; these methods can be found in the R package `LogConcDEAD` (Cule, Gramacy, and Samworth 2009) (see the functions `mlelcd()` and `dslcd()` - note that similar results were obtained with the comparable functions in the R package `logcondens` (Dümbgen and Rufibach 2011) and are therefore not included in the analysis below). For an informative overview we direct the interested reader to Samworth (2017) for a recent survey on log-concave estimation and its importance in

[7]See, in particular, the functions `npuniden.sc()` and `npuniden.boundary()`, which as of this writing support the constraints monotonically increasing (`constraint="mono.incr"`), decreasing (`constraint="mono.incr"`), convex (`constraint="convex"`), concave (`constraint="concave"`), log-convex (`constraint="log-convex"`), or log-concave (`constraint="log-concave"`), in additional to general inequality constraints placed directly on the density function itself (`constraint="density"` and the upper and lower bound arguments `lb=` and `ub=`). The Cramér–von Mises (`function.distance="TRUE"`) or the L_2 metric (`function.distance="FALSE"`) can be selected for weight enforcement.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

statistical analysis. For comparison purposes, in the monotonically increasing constraint setting we compare the proposed approach with the monotone rearrangement approach of Birke (2009) which can be found in the R package `Rearrangement` (Graybill et al. 2016) (see the function `rearrangement()`).

As noted in the introduction, the constrained MLE estimates have a rather nonstandard $n^{-1/3}$ rate of convergence compared with the $n^{-2/5}$ rate for the kernel estimator. One strength of the MLE approaches is the ease with which they can handle log-concavity in higher dimensions (from the practical perspective the kernel approach is limited to perhaps $d = 3$ or $d = 4$ dimensions at the most), yet another is that they are tuning-parameter free (the kernel approach requires the selection of bandwidths). In the log-concave constraint simulations that follow, cross-validation is used for automatic bandwidth selection for the proposed kernel-based methods and we optimize the distance from the unconstrained to the constrained *function* as discussed previously. However, in order to assess the degree to which being tuning-parameter-free matters, we begin with a brief comparison of the proposed approach based on the *infeasible optimal bandwidths* (which would be essentially tuning-parameter free) versus what is used in practice which employs *data-driven* smoothing parameter selection. Naturally the optimal bandwidths will present the method in the best possible light, albeit an unrealistic one, which is why the *data-driven* bandwidth-based results are used as the reference in the tables that follow and not the *infeasible* optimal bandwidth-based ones. It can be seen, however, that the difference between using the infeasible optimal versus the feasible data-driven tuning parameter (i.e., the bandwidth) is most apparent in small sample settings (e.g., $n = 100$) though this becomes asymptotically negligible and the discrepancy shrinks as the sample size grows.

In what follows we shall consider a modest number of DGPs and, as noted above, restrict attention to log-concave and monotonically increasing constraints (the DGPs are presented in Figure 6). The DGPs and a brief description are as follows:

1. The data is drawn from the standard *smooth* unbounded support $N(0, 1)$ univariate Gaussian distribution ($X \in [-\infty, \infty]$), which is log-concave, and we report results based on the (unknown) optimal bandwidth along with the data-driven bandwidth and compare results with the non-smooth MLE estimator and smooth MLE estimator under the log-concavity

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

constraint (Section S.6.1);

2. The data is drawn from a *smooth* unbounded support $N(0, \Sigma)$ bivariate Gaussian distribution ($X \in [-\infty, \infty]^2$), which is log-concave, and we compare results with the non-smooth and smooth MLE estimators under the log-concavity constraint (Section S.6.2).

3. The data is drawn from the *smooth* left-bounded support univariate Exponential distribution ($X \in [0, \infty]$), which is log-concave, and we compare results with the non-smooth MLE estimator and smooth MLE estimator under the log-concavity constraint (Section S.6.3);

4. The data is drawn from a *smooth* bounded support univariate Beta(3,3) distribution ($X \in [0, 1]$), which is log-concave, and we compare results with the non-smooth MLE estimator and smooth MLE estimator under the log-concavity constraint (Section S.6.4);

5. The data is drawn from a *non-smooth* bounded support univariate Triangular distribution ($X \in [0, 1]$), which is log-concave but *non-smooth*, and we compare results with the non-smooth MLE estimator and smooth MLE estimator under the log-concavity constraint - note we include this DGP in order to gauge robustness as it violates assumptions invoked when using kernel smoothing methods (i.e., continuous differentiability of the density up to some particular order > 2) (Section S.6.5);

6. The data is drawn from a *smooth* bounded support univariate Uniform distribution ($X \in [0, 1]$), which is log-concave, and we compare results with the non-smooth MLE estimator and smooth MLE estimator under the log-concavity constraint (Section S.6.6);

7. The data is drawn from a *smooth* bounded support univariate Beta(3,1) distribution ($X \in [0, 1]$), which is monotonically increasing, and we compare results with the monotone-rearrangement estimator under the monotonic increasing constraint (Section S.6.7).

Note that in each of these scenarios we report mean square error (MSE, computed as $n^{-1} \sum_{i=1}^{n} (f(X_i) - \hat{f}(X_i)^2)$ where $f(\cdot)$ is the true simulation density and $\hat{f}(\cdot)$ an estimate thereof) for the smooth unconstrained version of our estimator (SU) (which is simply standard kernel density estimation), the smooth constrained version (SC), the non-smooth MLE estimator (LNS), smooth MLE estimator (LS), and monotone rearrangement estimator (MR). We report results in both

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

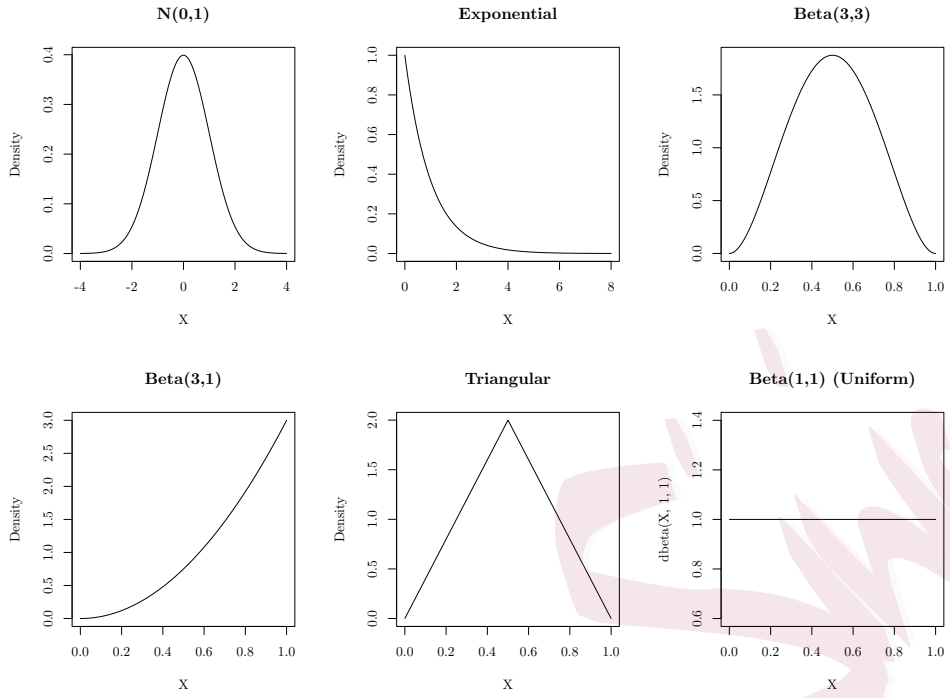

Figure 6: Monte Carlo Densities

tables (mean/median relative MSE over the M Monte Carlo replications) and in figures (boxplots of MSE for the M Monte Carlo replications). We present both mean and median *relative* MSE values so as to provide a complete impression of performance since the relative mean values may not be robust in the presence of outlying values which might occur if, say, data-driven bandwidth selection performs particularly poorly for some fraction of the resamples while the relative median would naturally be less affected by outlying values.

The proposed kernel approach admits known finite boundary points (i.e., boundary points of $\pm\infty$ have no impact on the estimate) which are used for the exponential (which uses $(a, b) = (0, \infty)$) and the beta and triangular (each of which use $(a, b) = (0, 1)$) simulations (all other cases use $(a, b) = (-\infty, \infty)$). The peer function `mlelcd()` in the `LogConcDEAD` package does not support known boundary points. Though one might consider modifying the peer function using standard correction methods, it is unclear that log-concavity would always be preserved afterwards. Regardless, any such extension of the peer method lies beyond the scope of this paper.

In order to meet page length constraints, particulars of the Monte Carlo simulations have been moved to the supplementary appendix that also contains the technical proofs (see Section S.6). In

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

brief, the proposed method is shown to be competitive with its non-smooth and smooth peers and, most importantly, provides an extensible and general approach to constraining kernel-based density estimates in a unified framework.

5 Summary

This paper presents a versatile procedure designed to impose a variety of shape constraints on a *smooth* nonparametric kernel density estimator. Via both simulations and real world data we demonstrate that the method can deliver practical and useful estimates of an unknown density that satisfies a range of constraints, and provide the theoretical underpinnings thereof. Additionally, for the constraint of log-concavity, our proposed approach performs convincingly well overall compared with popular existing approaches. Furthermore, for the constraint of monotonicity our approach is competitive with its peers, if anything performing somewhat better. But unlike many of its peers that are tailored for a *single* constraint, our approach is far more flexible and can encompass each of its peers in a unified framework. Moreover, these constraints can be applied to settings involving both continuous and ordered discrete data settings. An R implementation is available for the interested reader on CRAN (see the R package `np` (Hayfield and Racine 2008) and the functions `npuniden.sc()` and `npuniden.boundary()` contained therein).

There are many exciting and important directions in which the methods proposed herein could be extended. One would be to use the insights of Mammen (1991) (which is in the regression setting) to look at higher order asymptotic comparisons between the unconstrained and constrained estimators. As the constrained density estimator that we propose here is expected to equal the unconstrained estimator *if* the constraints imposed are valid, then in this case for sufficiently large n these two coincide (to first order) hence one would not expect large sample gains to arise. However, a more nuanced and detailed asymptotic analysis may reveal important higher order gains that could prove useful for the construction of confidence intervals in small sample settings. Another possible extension is to consider functions supported on a ball rather than a hypercube as considered here. This would require a revamping of existing tools, such as considering kernel functions supported on a ball. Both of these extensions lie beyond the scope of this paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Acknowledgements

We would like to sincerely thank the Editor, the Associate Editor, and two anonymous reviewers for their insightful comments that have significantly improved the paper. Du's research was partly supported by the U.S. National Science Foundation grant DMS-1916174.

References

Aitchison, J., and C. G. G. Aitken. 1976. "Multivariate Binary Discrimination by the Kernel Method." *Biometrika* 63: 413–20.

Bagnoli, Mark, and Ted Bergstrom. 2005. "Log-Concave Probability and Its Applications." *Economic Theory* 26 (2): 445–69.

Berwin A. Turlach R port by Andreas Weingessel <Andreas.Weingessel@ci.tuwien.ac.at>, S original by. 2019. *Quadprog: Functions to Solve Quadratic Programming Problems.* https://CRAN.R-project.org/package=quadprog.

Birke, M. 2009. "Shape Constrained Kernel Density Estimation." *Journal of Statistical Planning and Inference* 139 (8): 2851–62.

Chacón, José E., Tarn Duong, and M. P. Wand. 2011. "Asymptotics for General Multivariate Kernel Density Derivative Estimators." *Statistica Sinica* 21: 807–40.

Chen, Y., and R. Samworth. 2013. "Smoothed Log-Concave Maximum Likelihood Estimation with Applications." *Statistica Sinica*, 1373–98.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon. 2009. "Improving Point and Interval Estimators of Monotone Functions by Rearrangement." *Biometrika* 96 (3): 559–75.

Chetverikov, D., A. Santos, and A. M. Shaikh. 2018. "The Econometrics of Shape Restrictions." *Annual Review of Economics* 10: 31–63.

Cressie, N. A. C., and T. R. C. Read. 1984. "Multinomial Goodness-of-Fit Tests." *Journal of the Royal Statistical Society, Series B* 46: 440–64.

Csörgő, M., and P. Révész. 1981. *Strong Approximations in Probability and Statistics.* New York: Academic Press.

Cule, M., R. Gramacy, and R. Samworth. 2009. "LogConcDEAD: An R Package for Maximum Likelihood Estimation of a Multivariate Log-Concave Density." *Journal of Statistical Software* 29 (2). http://www.jstatsoft.org/v29/i02/.

Cule, M., R. Samworth, and M. Stewart. 2010. "Maximum Likelihood Estimation of a Multi-Dimensional Log-Concave Density." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (5): 545–607.

Dette, H., and K. F. Pilz. 2006. "A Comparative Study of Monotone Nonparametric Kernel Estimates." *Journal of Statistical Computation and Simulation* 76 (1): 41–56.

Du, P., C. F. Parmeter, and J. S. Racine. 2013. "Nonparametric Kernel Regression with Multiple Predictors and Multiple Shape Constraints." *Statistica Sinica* 23 (3): 1343–72.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Dümbgen, L., and K. Rufibach. 2009. "Maximum Likelihood Estimation of a Log-Concave Density and Its Distribution Function: Basic Properties and Uniform Consistency." *Bernoulli* 15 (1): 40–68.

———. 2011. "logcondens: Computations Related to Univariate Log-Concave Density Estimation." *Journal of Statistical Software* 39 (6): 1–28. http://www.jstatsoft.org/v39/i06/.

Fan, Jianqing, Chunming Zhang, and Jian Zhang. 2001. "Generalized Likelihood Ratio Statistics and Wilks Phenomenon." *Annals of Statistics* 29 (1): 153–93.

Feng, O. Y., A. Guntuboyina, A. K. H. Kim, and R. J. Samworth. 2021. "Adaptation in Multivariate Log-Concave Density Estimation." *Annals of Statistics* 49: 129–53.

Graybill, Wesley, Mingli Chen, Victor Chernozhukov, Ivan Fernandez-Val, and Alfred Galichon. 2016. *Rearrangement: Monotonize Point and Interval Functional Estimates by Rearrangement.* https://CRAN.R-project.org/package=Rearrangement.

Grenander, U. 1956. "On the Theory of Mortality Measurement." *Scandinavian Actuarial Journal* 1956 (2): 125–53.

Groeneboom, P., and G. Jongbloed. 2014. *Nonparametric Estimation Under Shape Constraints.* Cambridge University Press.

———. 2018. "Some Developments in the Theory of Shape Constrained Inference." *Statistical Science* 33: 473–92.

Hall, P., and H. Huang. 2001. "Nonparametric Kernel Regression Subject to Monotonicity Constraints." *Annals of Statistics* 29 (3): 624–47.

———. 2002. "Unimodal Density Estimation Using Kernel Methods." *Statistica Sinica* 12: 965–90.

Hall, P., H. Huang, J. Gifford, and I. Gijbels. 2001. "Nonparametric Estimation of Hazard Rate Under the Constraint of Monotonicity." *Journal of Computational and Graphical Statistics* 10: 592–614.

Hall, P., and K.-H. Kang. 2005. "Unimodal Kernel Density Estimation by Data Sharpening." *Statistica Sinica* 15: 73–98.

Hall, P., and B. Presnell. 1999. "Density Estimation Under Constraints." *Journal of Computational and Graphical Statistics* 8: 259–77.

Hall, P., J. Racine, and Q. Li. 2004. "Cross-Validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association* 99 (2): 1015–26.

Hardy, G. H., J. E. Littlewood, and G. Pólya. 1952. *Inequalities.* Cambridge university press.

Hausman, J., B. H. Hall, and Z. Griliches. 1984. "Econometric Models for Count Data with an Application of the Patents-R&D Relationship." *Econometrica* 52 (4): 909–38.

Hayfield, T., and J. S. Racine. 2008. "Nonparametric Econometrics: The np Package." *Journal of Statistical Software* 27 (5). http://www.jstatsoft.org/v27/i05/.

Horowitz, J. L., and S. Lee. 2017. "Nonparametric Estimation and Inference Under Shape Restrictions." *Journal of Econometrics* 201 (1): 108–26.

Koenker, R., and I. Mizera. 2010. "Quasi-Concave Density Estimation." *The Annals of Statistics*, 2998–3027.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

———. 2018. "Shape Constrained Density Estimation via Penalized Rényi Divergence." *Statistical Science* 33: 510–26.

Komlós, J., P. Major, and G. Tusnády. 1975. "An Approximation of Partial Sums of Independent Random Variables and the Sample Distribution Function, Part I." *Zeitschrift für Wahrscheinlichskeitstheorie Und Verwandte Gebiete* 32 (1-2): 111–31.

Li, Q., and J. S. Racine. 2003. "Nonparametric Estimation of Distributions with Categorical and Continuous Data." *Journal of Multivariate Analysis* 86: 266–92.

Li, Z., G. Liu, and Q. Li. 2017. "Nonparametric Knn Estimation with Monotone Constraints." *Econometric Reviews* 36: 988–1006.

Lok, T. M., and R. V. Tabri. 2021. "An Improved Bootstrap Test for Restricted Stochastic Dominance." *Journal of Econometrics* 224 (2): 371–93.

Mammen, E. 1991. "Estimating a Smooth Monotone Regression Function." *The Annals of Statistics* 19 (2): 724–40.

Meyer, M. C. 2008. "Inference using shape-restricted regression splines." *The Annals of Applied Statistics* 2 (3): 1013–33.

Meyer, M. C., and D. Habtzghi. 2011. "Nonparametric Estimation of Density and Hazard Rate Functions with Shape Restrictions." *Journal of Nonparametric Statistics* 23 (2): 455–70.

Meyer, M. C., and M. Woodroofe. 2004. "Consistent Maximum Likelihood Estimation of a Unimodal Density Using Shape Restrictions." *Canadian Journal of Statistics* 32 (1): 85–100.

Meyer-ter-Vehn, Moritz, Lones Smith, and Katalin Bognar. 2017. "A Conversational War of Attrition." *The Review of Economic Studies* 85 (3): 1897–1935.

Ouyang, D., Q. Li, and J. S. Racine. 2006. "Cross-Validation and the Estimation of Probability Distributions with Categorical Data." *Journal of Nonparametric Statistics* 18 (1): 69–100.

Parzen, E. 1962. "On Estimation of a Probability Density Function and Mode." *The Annals of Mathematical Statistics* 33: 1065–76.

Prakasa Rao, B. L. S. 1969. "Estimation of a Unimodal Density." *Sankhya Series A* 31: 23–36.

Racine, J., Q. Li, and K. X. Yan. 2020. "Kernel Smoothed Probability Mass Functions for Ordered Datatypes." *Journal of Nonparametric Statistics* 32 (3): 563–86.

Rathke, F., and C. Schnörr. 2018. *fmlogcondens: Fast Multivariate Log-Concave Density Estimation*. https://github.com/FabianRathke/fmlogcondens.

———. 2019. "Fast Multivariate Log-Concave Density Estimation." *Computational Statistics & Data Analysis* 140: 41–58.

Rosenblatt, M. 1956. "Remarks on Some Nonparametric Estimates of a Density Function." *The Annals of Mathematical Statistics* 27: 832–37.

Samworth, R. 2017. "Recent Progress in Log-Concave Density Estimation." *arXiv Preprint arXiv:1709.03154*.

Samworth, R., and B. Sen. 2018. "Editorial: Special Issue on 'Nonparametric Inference Under Shape Constraints'." *Statistical Science* 33 (4): 469–72.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Singh, R. 1987. "MISE of Kernel Estimates of a Density and Its Derivatives." *Statistics & Probability Letters* 5: 153–59.

Tan, Guofu, and Junjie Zhou. 2020. "The Effects of Competition and Entry in Multi-Sided Markets." *The Review of Economic Studies*.

Walther, Guenther. 2009. "Inference and Modeling with Log-Concave Distributions." *Statistical Science* 24 (3): 319–27. https://doi.org/10.1214/09-STS303.

Wolters, M. A. 2018. *scdensity: Shape-Constrained Kernel Density Estimation*. https://CRAN.R-project.org/package=scdensity.

Wolters, M. A., and W. J. Braun. 2018a. "A Practical Implementation of Weighted Kernel Density Estimation for Handling Shape Constraints." *Stat* 7 (1): e202.

———. 2018b. "Enforcing Shape Constraints on a Probability Density Estimate Using an Additive Adjustment Curve." *Communications in Statistics - Simulation and Computation* 47 (3): 672–91.

Woodroofe, M., and J. Sun. 1993. "A Penalized Maximum Likelihood Estimate of $f(0+)$ When f Is Non-Increasing." *Statistica Sinica*, 501–15.