| 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,<br>Christopher F. Parmeter and<br>Jeffrey S. Racine |
| **Corresponding Authors** | Jeffrey S. Racine |
| **E-mails** | racinej@mcmaster.ca |

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

# Shape-Constrained Kernel PDF and PMF Estimation

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

August 26, 2022

**Abstract**

We present an approach for estimating shape-constrained kernel-based probability density functions (PDFs) and probability mass functions (PMFs) that includes 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. Furthermore the approach handles a range of transformations of the PDFs and PMFs including, for example, logarithmic transformations, which allow us to impose log-concave or log-convex constraints. We also provide the theoretical underpinnings for the procedures. The results of a simulation-based comparison between our proposed approach and those Grenander-type methods favor our approach when the data-generating process is smooth. To the best of our knowledge, ours is also the only *smooth* framework that handles PDFs and PMFs in the presence of inequality bounds, equality constraints, and other popular constraints. An implementation in R 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 by using boundary kernel functions.

## 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

*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

owing to the nature of the data, but increasingly often are required when replacing parametric models with more versatile semi- and nonparametric models. The ability to preserve the qualitative shape properties present in a parametric model is a key component of any alternative method. However, the consequences of misspecifying the parametric model can be severe, and influence the choice of the 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* by imposing *valid* shape constraints on some statistical object of interest. That is, if one's assumption about a shape constraint on an 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 reason is to assess the validity of the shape constraints using formal quantitative inference, or to determine the qualitative effect of the constraints on the resulting estimate.

Imposing shape constraints on an otherwise unrestricted nonparametric curve is a key element of a sound empirical analysis that encompasses a range of approaches; see Groeneboom and Jongbloed (2014) for examples of shape-constrained estimators and algorithms, along with their theoretical properties. Perhaps the most common applications of enforcing shape constraints arise when modeling a conditional mean function (i.e., a regression), which is understandable, given the popularity of regression analysis. However, the density function is also a popular object of interest that necessitates a separate treatment from that of regression, owing to its unique nature. Shape-constrained density estimation, like its regression-based counterpart, has a rich history that can be traced to the seminal work of Grenander (1956), who analyzes the maximum likelihood estimator (MLE) of a decreasing density on the nonnegative half-line (see also Groeneboom and Jongbloed (2018) for recent theoretical work in this direction). Note that Prakasa Rao (1969) shows that this estimator exhibits nonstandard asymptotic behavior, because it converges at a cube rate ($n^{-1/3}$) at points at which the true decreasing density is differentiable with a negative derivative. This is slower than competing local kernel-based estimators that assume smoothness ($n^{-2/5}$), a common assumption among practitioners that we adopt in one of the two kernel-based estimators we consider here. Although 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

(and faster, i.e., $n^{-1/2}$) rate of convergence than that of their kernel density-based counterparts.

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

Grenander-based approaches (Grenander 1956) have been widely used to impose 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 the specification of a *bandwidth*. However, although Grenander-based approaches are nonparametric in nature, they are *nonsmooth* which runs counter to the spirit of adopting a *smooth* nonparametric approach in the first place. For example, the approach Grenander (1956) proposes for imposing monotonicity can be characterized as the left derivative of the least concave majorant of the empirical distribution function, which is a nonsmooth function. Practitioners who routinely assume smoothness and adopt smooth nonparametric estimators are not likely to be attracted to nonsmooth nonparametric shape-constrained solutions, hence the appeal of *smooth* shape-constrained nonparametric solutions, such as those proposed herein.

The literature on constrained nonparametric estimation has grown significantly over the past few decades. The approach that we extend here has proven to be a particularly popular, versatile, and extensible method for imposing constraints on a *smooth* nonparametric object (see Hall and Presnell 1999). This approach places weights directly on the sample realizations so that the desired constraint is imposed effectively. In kernel-based *regression* settings, this amounts to starting with a standard kernel estimator. Then, if the constraints are violated in some region of the support we shift the regressand *vertically* in such a way that a 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 (P. Hall and Huang 2001; Du, Parmeter,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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. A similar method, known as *data sharpening* (Hall and Kang 2005), instead introduces weights that 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 using standard off-the-shelf quadratic programming methods, whereas horizontally shifting observations may require full-blown nonlinear programming, which may be less tractable from a practical perspective.

Building on the work of Du, Parmeter, and Racine (2013), who consider a unified framework for *smooth* shape-constrained nonparametric kernel regression, we propose a unified framework for *smooth* shape-constrained kernel density and PMF estimation. Shape-constrained kernel density (and mass) function estimation differs from shape-constrained kernel regression in terms of both its practical implementation and in its theoretical properties, and hence requires a separate treatment. Our approach is extremely flexible, and allows for a range of constraints to be imposed *simultaneously* (presuming, of course, that the set of constraints is internally consistent). The original implementation (P. Hall and Huang 2001) involves optimizing a power-divergence criterion. Du, Parmeter, and Racine (2013) propose replacing this criterion with an $L_2$-norm criterion, which delivers an estimator that retains all of the desirable features of the power-divergence-based method, but is 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 impose unimodality on a univariate kernel density estimator, and modify it in such a way as to deliver a unified approach with a straightfoward implementation. We believe that this unified framework will 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 Li, Liu, and Li (2017), who propose a slightly modified version of the optimization criterion proposed by Du, Parmeter, and Racine (2013). While Li, Liu, and Li (2017) adopt an $L_2$-norm criterion, as per Du, Parmeter, and Racine (2013), rather than optimizing the distance between the optimization *weights* and their unconstrained counterparts, they instead optimize the distance between the constrained *estimates* and their unconstrained

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

counterparts. Although Li, Liu, and Li (2017) provide convincing simulation evidence that their modification can deliver constrained estimates with improved finite-sample performance, they offer no theoretical justification for this modification. We demonstrate theoretically that this modified $L_2$-optimization criterion delivers constraint weights that ensure *identical* asymptotic behavior to that from optimizing the weights directly. By providing the theoretical underpinnings for the slightly modified optimization criterion proposed by Li, Liu, and Li (2017), we establish that the constraint weights can be based on this criterion with no loss of information.

Finally, we demonstrate how our method can be adapted to handle constraints on the *log-density*. This is an important generalization, because 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 directly on the density.

The proposed approach differs from that of Du, Parmeter, and Racine (2013), among others, in several ways. In our setting, we are dealing with density estimation and weights are applied on the kernel function. In contrast, in the regression setting of Du, Parmeter, and Racine (2013), weights are applied on the dependent variable, which affects the proofs in a nontrivial way. Here we prove Theorem 2 for the Cramér–von Mises distance function (earlier works have not considered this distance metric), which requires handling cross-product terms involving the constraint weights in the various components of our decomposition of the constrained density estimator. Additionally, Theorem 3 is entirely new. To the best of our knowledge, it represents the first attempt to impose smoothness constraints on a PMF estimator. While not a theoretical contribution, we also demonstrate how to impose log-concavity on a smooth kernel density estimate in a simple quadratic programming setup.

One of the constraints on the log-density, specifically *log-concavity*, has long been a topic of interest in 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 to the class of unimodal densities. Though the class of log-concave densities is a subset of the class of unimodal densities, it contains most of the commonly used parametric distributions, and is therefore a rich

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

and useful nonparametric class. Recent developments include the works of Feng et al. (2021) who study an 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* nonsmooth function), and Rathke and Schnörr (2019), who propose a fast implementation of the smoothed version of this estimator.

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 the works of Bagnoli and Bergstrom (2005), who describe 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 explore costly deliberations 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 rely on log-concavity in agent heterogeneity to establish several formal results in a model of price competition entry and multi-sided markets.

Our adaptation of the work of Hall and Huang (2002) to log-concavity also stands in contrast to a recently proposed kernel-based linear adjustment mechanism (Wolters and Braun 2018a, 2018b) that tackles constrained estimation using a specified number of inflection points. This approach can also be used to enforce log-concavity, though the authors do not consider this particular constraint. However, it would require that we know the locations of these inflection points ex ante, otherwise they need to be approximated using some optimization routine, which has its drawbacks. In contrast, our proposed approach to imposing log-concavity requires no prior knowledge or approximations of the locations of the inflection points. Instead, we impose the constraints on the log-density directly, leading to a direct system of linear inequality constraints, and thus a fast and efficient algorithm for imposing log-concavity in a smooth setting is provided. The linear adjustment mechanism of Wolters and Braun (2018b) is 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 do not consider using their linear adjustment mechanism approach to impose log-concavity which, given its popularity

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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 an implementation of the linear adjustment mechanism approach.]

In addition to the works referenced above, the related literature includes the studies of Woodroofe and Sun (1993), who consider a penalized MLE estimate of a density on the positive half of the real number line when the density is nonincreasing, Meyer and Woodroofe (2004), who develop a nonparametric MLE that is consistent for the mode, Hall and Kang (2005), who consider unimodal kernel density estimation using data sharpening, Dette and Pilz (2006), who conduct a comparative study of monotone constrained estimators, Birke (2009), who considers shape-constrained density estimation using monotone rearrangement (Hardy, Littlewood, and Pólya 1952; Chernozhukov, Fernandez-Val, and Galichon 2009), Dümbgen and Rufibach (2009), who studied the MLE of a log-concave density, and Koenker and Mizera (2010), who formulate the MLE of a log-concave density as a convex optimization problem, showing that it has an equivalent dual formulation to that of a constrained maximum Shannon entropy problem. Cule, Samworth, and Stewart (2010) study a nonsmooth log-concave MLE of a probability distribution function and Meyer and Habtzghi (2011) use regression splines, based on the work of Meyer (2008), to formulate a nonparametric MLE of strictly decreasing probability densities in terms of convex programming and iteratively re-weighted least squares cone projection algorithms. Chen and Samworth (2013) study the smoothed log-concave MLE of a probability distribution function, Horowitz and Lee (2017) explain how to estimate and obtain an asymptotic uniform confidence band for a conditional mean function under possibly nonlinear shape restrictions, and Koenker and Mizera (2018) consider a log-concave estimation for weaker forms of concavity constraints that allow for heavier tail behavior and sharper modal peaks. More recently, Lok and Tabri (2021) develop an empirical tilting method for shape-constrained estimation over a data-driven grid of points to enforce the 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 probability density function (PDF) and PMF estimators and describes our approach. Here, Section 2.1 examines how we determine the constraint weights, and Section 2.2 briefly discusses finite-support boundary kernel functions and presents several examples of popular constraints. Section 3 outlines

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

the theoretical properties of the proposed approach and Section 4 presents a set of Monte Carlo simulations that show that the proposed approach is competitive with, and often improves upon leading methods that have been tailored to two popular constraints (log-concavity and monotonicity). Section 5 concludes the paper. Detailed theoretical proofs are relegated to the online Supplementary Material, and an open implementation in R exists to assist practitioners interested in exploring the proposed methods.

## 2 Shape-Constrained Kernel Density Estimation

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

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where $h$ is the *bandwidth*, $K(\cdot)$ is the *kernel function*, usually chosen as a symmetric mean zero PDF itself, and $x$ 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 $p_i$, for $i = 1, \ldots, n$, and modify (2.1) as follows:

$$\hat{f}(x|p) = \frac{1}{h} \sum_{i=1}^{n} p_i K\left(\frac{x - X_i}{h}\right). \tag{2.2}$$

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 (2.1). In other words, we use the notation $\hat{f}(x|p_{unif})$ in what follows to represent (2.2) for the special case in which the constraint weights assume the value $p_i = 1/n$, for $i = 1, \ldots, n$. Furthermore, these special weights are denoted as $p_{unif}$, and for these *and only these*, weights (2.2) is equal to(2.1), 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

in (2.2), yielding the estimator

$$\begin{aligned}\hat{f}(x|p) &= \frac{1}{nh}\sum_{i=1}^{n}(1+a_i)K\left(\frac{x-X_i}{h}\right)\\ &= \frac{1}{nh}\sum_{i=1}^{n}(1+a_i)K(Z_i)\\ &= \frac{1}{nh}\sum_{i=1}^{n}K(Z_i)+\frac{1}{nh}\sum_{i=1}^{n}a_iK(Z_i)\\ &= \hat{f}(x|p_{unif})+\frac{1}{nh}\sum_{i=1}^{n}a_iK(Z_i),\end{aligned} \tag{2.3}$$

where $Z_i = (x - X_i)/h$ and 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. To impose constraints on the log-density function (or its derivatives), we instead consider an estimator of the form

$$\hat{f}(x|p) = \hat{f}(x)\prod_{i=1}^{n}\exp\left\{a_iK(Z_i)/nh\right\}. \tag{2.4}$$

Taking the logarithm of both sides, we obtain

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

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

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

where, the *unconstrained* weights used in the object $\hat{f}(x|p_{unif})$, correspond to $a_i = 0$ in (2.4) 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 $a_i$, which, combined with a quadratic objective function, leads naturally to solving a quadratic program. The resulting constrained estimator arises from solving a quadratic program, and then replacing the arbitrary weights $a_i$ with the feasible constrained weights determined by the quadratic program.

Thus far, we have outlined two approaches that introduce weights that deliver constrained density

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

or log-density estimates. Now, we explicitly introduce the constraints themselves in a general framework. Denote the $j$th 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, respectively, that may change with $x$, where $l(x) \leq u(x)$. The constraints on the $j$th derivative of the density and log-density, for $j = 0, 1, 2, \ldots$, can be expressed as

$$l(x) \leq \hat{f}^{(j)}(x|p) \leq u(x) \tag{2.5}$$

and

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

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

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

or

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

Furthermore, 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 wish to impose the constraints that $\hat{f}^{(0)}(x|p) = \hat{f}(x|p) \geq 0$ (nonnegativity of the constrained density) and $\log^{(2)} \hat{f}(x|p) \leq 0$ (log-concavity), we can impose the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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, we typically impose the constraint $\sum_{i=1}^{n} a_i = 0$.

We wish 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 for 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, in Section 2.2, we use *kernel carpentry* to provide a flexible kernel function that is well-suited to the current setting.

### 2.1 Selection of the Constraint Weights

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

A variety of approaches for constrained weight selection have been proposed in the literature, each of which minimizes some measure of *divergence* between the constrained and the unconstrained *weights* or the constrained and the unconstrained *estimates*. Some divergence metrics are more computationally demanding than others, and different metrics may impose binding restrictions on the weights in order to produce valid estimates. For example, P. Hall and Huang (2001) suggest using the Cressie–Read power-divergence metric, Hall and Huang (2002) investigate a smoothed Cramér–von Mises metric, and Du, Parmeter, and Racine (2013) suggest an $L_2$-norm metric. Specifically, in the power-divergence and $L_2$-norm frameworks, the constrained weights are selected to be as close as possible to the unconstrained weights (also called the *uniform* weights), whereas in the smoothed

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Cramér–von Mises setting, the constrained weights are chosen to minimize the squared integrated difference between the unconstrained and the constrained densities. As Hall and Huang (2002) and Du, Parmeter, and Racine (2013) document, one benefit of adopting an $L_2$-norm (i.e., the squared distance) metric is that we can select smoothing parameters 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 Li, Liu, and Li (2017), we consider two closely related approaches for the optimal construction of the constraint weights, and emphasize their relative strengths. 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 by means of a straightforward quadratic program exercise using, say, the *quadprog* package (Berwin A. 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:

$$D_{CM}(p) = (n^2|h|)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} (np_i - 1)(np_j - 1) L\left(\frac{X_i - X_j}{h}\right), \tag{2.6}$$

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 (2.3) or constraints on the log-density function using (2.4), we require $\sum_{i=1}^{n} a_i = 0$). It is useful to focus on the relative merits of each distance metric used to select 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 Du, Parmeter, and Racine (2013) demonstrate, the relative magnitude of the constraint weights with the $L_2$-norm is $O(n^{-1})$. We can also view a difference from the uniform weights as a measure of relative entropy with respect to the uniform distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

The Cramér–von Mises metric has obvious practical appeal, because 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 Li, Liu, and Li (2017), selecting the weights to minimize $D_{CM}(p)$ naturally produces density estimates that are closer to $f(x)$ than are estimates from minimizing $D_{L_2}(p)$.

Note that using 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 useful, because it requires that the $p_i$ used to estimate a density from (2.2) be nonnegative ($p_i$ must satisfy $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for this approach), and some constraints may require negative weights. Furthermore, although $D_\rho(p)$ has an appealing immediate interpretation as a measure of entropy, it does require that the user select an additional tuning parameter for its implementation ($\rho$). Lastly, as Hall and Huang (2002) note, problems can arise as $p_i$ approaches zero, because enforcing constraints on a curve leads to "data compression" (i.e., the effective sample size 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 zero. This information is not lost however, but simply reassigned to observations that receive nonzero weights. Thus, there can be substantial differences between our elected metrics and $D_\rho(p)$; while both $D_{L_2}(p)$ and $D_{CM}(p)$ behave well when $p_i$ approaches zero, $D_\rho(p)$ may not be applicable for certain constraints with particular values of $\rho$.

## 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. *Boundary bias* affects the quality of kernel density estimates when substantial probability mass occurs at a support boundary. The most well-known solutions to this problem are *data-reflection*, *data-transformation*, and *kernel carpentry*. Data-reflection involves duplicating data symmetrically (i.e., reflecting) around its boundary, running standard bandwidth selection and kernel estimation, and 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

desired effect. Kernel carpentry uses kernel functions that adapt to the presence of a boundary, thus mitigating the effect of the boundary. To some degree, these methods all reduce the amount of bias that would otherwise be present near a boundary to that which holds in the interior of the 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 of 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 $X_i$, $z_b = (b - x)/h$, $z_a = (a - x)/h$, and $G(z) = \int_{-\infty}^{z} K(t)\,dt$. Given that $K(z)$ is a standard univariate kernel function, $G(z)$ is the CDF counterpart to the PDF $K(z)$ that we used to estimate $F(x)$. Note 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. When conducting a 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 $K(z,a,b)$.

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

To reduce the 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, we can obtain the counterpart to $K(z,a,b)$ by adopting the following transformation for (doubly) truncated density functions:

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

### 2.2.2 Derivative Kernel Functions

Some of the constraints we consider are placed on the derivative of the kernel density estimates, and hence we may require derivatives of the kernel function. To that end, we apply the quotient rule to obtain the first derivative of the doubly truncated kernel function, yielding

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

Note that

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

The second derivative is found by applying the quotient and the product rules, yielding

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

Note that the first term on the right-hand side can be expressed as

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

and 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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, support on $(-\infty, b]$ with $b$ finite, and support 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.

### 2.3 Hypothesis Testing

We can test the validity of the shape constraints being imposed by following the insights of P. Hall et al. (2001) and Du, Parmeter, and Racine (2013), and using a bootstrap inferential procedure. Briefly, the test statistic is the value of the objective function from solving the quadratic program when imposing the constraints. The bootstrap procedure draws 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 from solving the quadratic program when imposing the constraints on the bootstrap resamples). The test involves computing a $P$-value constructed by comparing the test statistic with that obtained from the empirical distribution constructed under the null or, alternatively, by comparing the test statistic with the desired $1 - \alpha$ quantile obtained from the empirical null distribution where $\alpha$ is the desired size of the test procedure (the test is one-sided with a right-tailed rejection region).

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 from $\hat{f}(\mathbf{x}|p_{unif})$. A simple way to do this is to use 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 as $\hat{f}(\mathbf{x}|p^*)$, yielding $D_\jmath(p^*)$. We repeat this process $B$ times. Finally, we compute the empirical $P$ value, $P_B$, which is simply the proportion of the $B$ bootstrap resamples

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

$D_\jmath(p^*)$ that exceed $D_\jmath(\hat{p})$, that is,

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

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

### 2.4 Illustrative Applications: Monotonicity and Concavity

Monotonicity and concavity constraints are two popular shape constraint domains that our approach can cover. As in Du, Parmeter, and Racine (2013), we solve a simple quadratic program using (2.6) to generate the constrained estimate. Figure 1 presents the 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 using 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 zero, this leads to only minor changes in the shape of the density beyond where the constraints need to be enforced. This becomes clearer by looking at the lower plot in Figure 1, which plots the constrained and unconstrained derivative estimates.

Figure 2 presents the results when 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

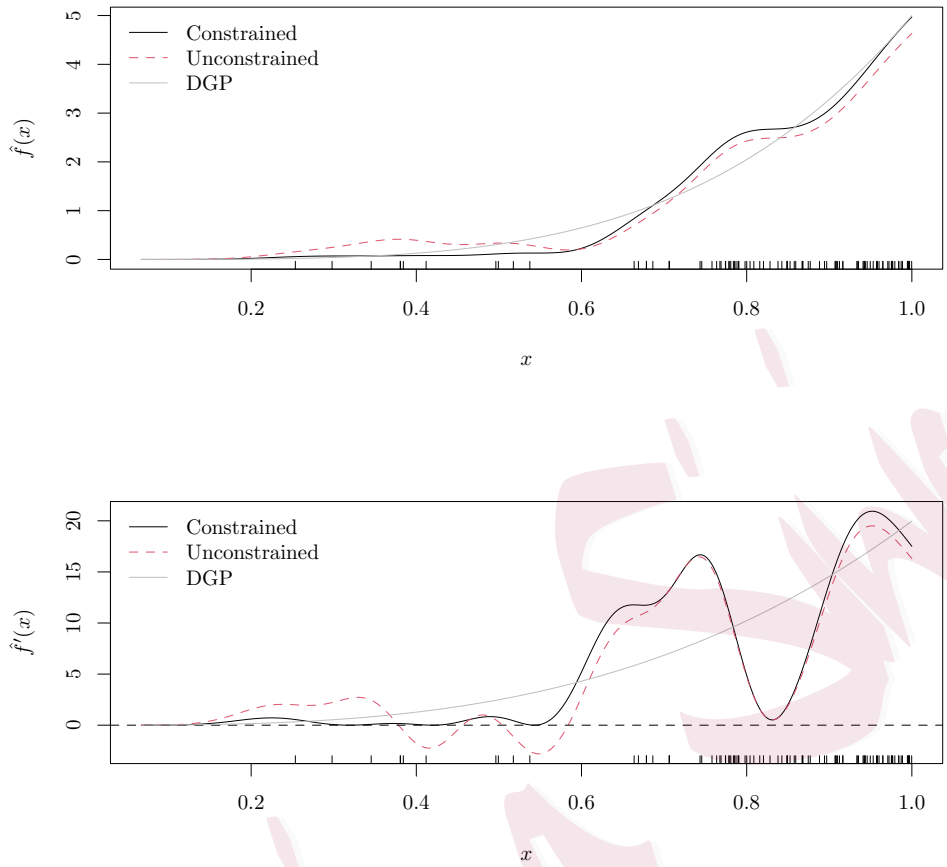


Figure 1: Monotone shape-constrained density estimation ($\hat{f}'(x) \geq 0$). The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained first derivative estimates.

### 2.5 Log-Concave Kernel Density Estimation

Log-concavity is a popular constraint, although it is only one of many shape constraint domains that our approach can cover. To impose log-concavity/convexity, we require $d^2 \log(\hat{f}(x))/dx^2$ and $d^2K(Z_i)/dx^2$. The former is given by

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

and the latter is given by

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

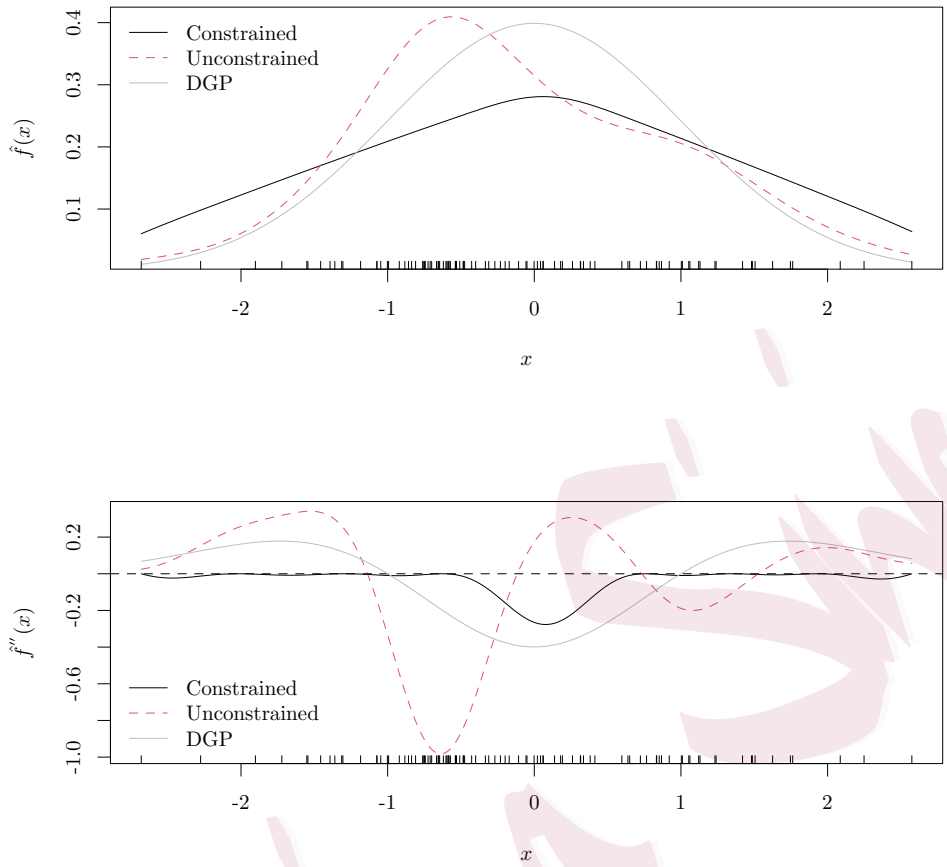


Figure 2: Concave shape-constrained density estimation ($\hat{f}''(x) \leq 0$). The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained second derivative estimates.

Note that

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

### 2.6 Illustrative Application: Log-Concavity

Figure 3 presents the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

one region of nonconcavity of the log-density for $-2.5 < x < -1.9$. Both the constrained and the unconstrained densities integrate to one and are proper.

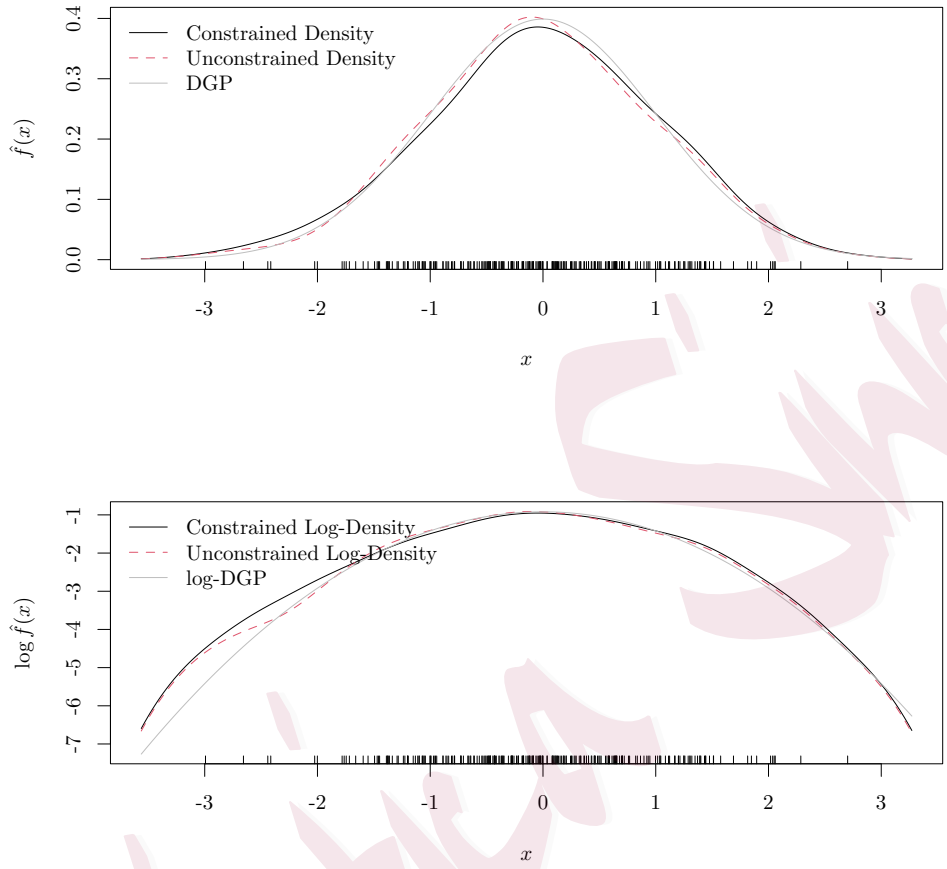


Figure 3: Log-concave shape-constrained density estimation. The upper figure plots the constrained and unconstrained density estimates, the lower figure plots the constrained and unconstrained log-density estimates.

## 2.7 Categorical (ordered) PMFs

The approach we consider for a shape-constrained PDF estimation can also be applied to a 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), we need only the one value of $a_i$ per outcome (because $a_i = a_j$ 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

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 $x_{(j)}$ are the order statistics, that can be computed directly from an unconstrained estimate (as can higher-order derivatives, if needed). As was the case for the shape-constrained PDF estimation, the counterpart to (2.3) for the PMF estimation can be written as

$$\hat{P}(x|p) = \hat{P}(x) + \sum_{i=1}^{n} a_i l(X_i, x, \lambda), \tag{2.7}$$

where $\lambda$ 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 the shape-constrained PDF estimation described previously, and so are not repeated here (see Racine, Li, and Yan (2020) for further details). We now consider an empirical illustration based on count data that have ordered discrete support.

## 2.8 Empirical Application: Shape-Constrained PMFs

We consider a data set collected by Hausman, Hall, and Griliches (1984) that 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 the results in Figure 4. The nonsmooth estimate is quite noisy, whereas the smooth estimate is much less so. Like its empirical counterpart, the smooth estimate delivers probability estimates that *sum* to one, but the smooth estimate is expected to be more efficient from a squared error perspective.

Figure 4 reveals that the unconstrained kernel PMF estimator, though perhaps more plausible an estimate than the nonsmooth 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 unconstrained and monotonically constrained estimates. As noted above, the derivatives for the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

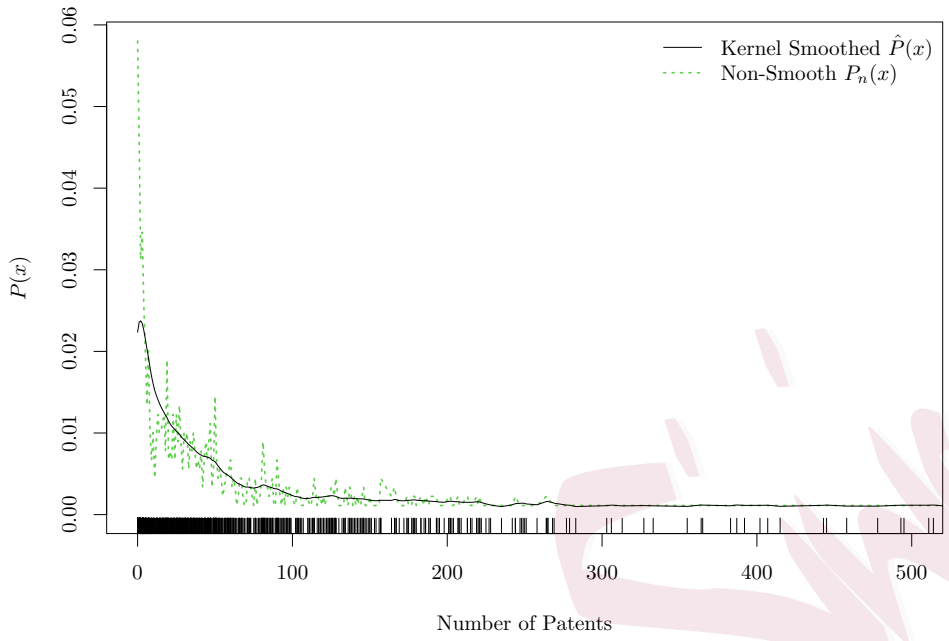


Figure 4: Unconstrained smooth and nonsmooth PMF estimates for patent data. The smooth estimate appears as a solid line, the nonsmooth estimate as a dotted line.

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 to solve the quadratic program is then the difference between kernel functions evaluated at $x_{(j)}$ and $x_{(j-1)}$ divided by the difference $x_{(j)} - x_{(j-1)}$. To impose the monotonically decreasing constraint, we define $\Delta_1(x) \leq 0$ (we reverse this definition for monotonically increasing constraints).

## 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 the 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 the consistency of the constrained density estimator under weak conditions on the order of the derivatives of the true density and on the bandwidth (naturally, (iii) does not allow for consistent estimation). Third, we

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

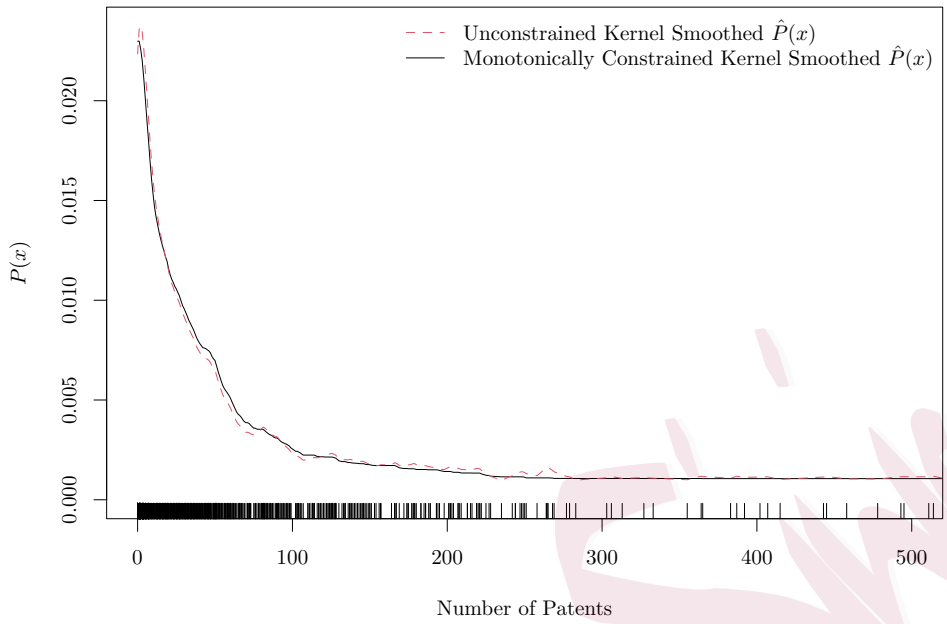


Figure 5: Unconstrained and constrained smooth probability function estimates for the patent data.

extend our results in the 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; nevertheless, 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 continuous data are similar to those of P. Hall and Huang (2001) and Du, Parmeter, and Racine (2013), but with four important differences. First, P. Hall and Huang (2001) and Du, Parmeter, and Racine (2013) impose constraints in a regression setting. The density setting is complicated by the lack of an error term, such that we cannot apply existing theory directly. Second, P. Hall and Huang (2001) use the power-divergence measure of Cressie and Read (1984) and Du, Parmeter, and Racine (2013) use the $L_2$-metric. Here, we establish the consistency of the constrained estimator using the objective function proposed by Li, Liu, and Li (2017), which, rather than selecting constraint weights as close as possible to the uniform weights (as in Du, Parmeter, and Racine 2013), selects weights as close as possible to the unconstrained estimator. Intuitively, this modification makes sense, given that the unconstrained estimator is consistent to begin with. Although Li, Liu, and Li (2017) show the impressive finite-sample properties of their objective function when selecting constraint weights for a constrained $K_{nn}$ regression estimator, the change

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

in the objective function also necessitates changes to existing theory. Third, existing theory works relatively well for constraints on the density. However, several additional modifications are required when imposing constraints on, for example, the log-density. Fourth, we develop the appropriate theory for the constrained estimation of the PMF. To the best of 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 $\mathbf{x}$. Note that the general two-sided constraints in (2.5) can be expressed as one-sided constraints of the form

$$\sum_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k} f^{(\mathbf{s})}(\mathbf{x}) - c_k(\mathbf{x}) \geq 0, \quad k = 1, \ldots, T, \tag{3.8}$$

where $T$ is the total number of restrictions, with the sum taken over all density derivative vectors in $\mathbf{S}_k$, and $\alpha_{\mathbf{s},k}$ 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 in most cases, we expect that a single constraint (i.e., $T = 1$) will suffice. As an example, for $r = 1$ and the imposition of monotonicity, we have $T = 1$ with $\mathbf{s} = (1)$, $\mathbf{S}_k = \{(1)\}$, $\alpha_{\mathbf{s},k} = 1$, and $c_k(\mathbf{x}) = 0$, for all $\mathbf{x}$.

Before 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 $k$th entry $\sum_{\mathbf{s}\in\mathbf{S}_k} \alpha_{\mathbf{s},k} f^{(\mathbf{s})}(\mathbf{x})$. 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 that of $\mathbf{s}_2$ 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 (2.2) into (3.8) yields

$$\sum_{i=1}^{n} p_i K_i^{\mathcal{D}}(\mathbf{x}) \geq 0. \tag{3.9}$$

Here, $K_i^{\mathcal{D}}(\mathbf{x})$ represents the form of the constraints based on the appropriate kernel derivatives, that is, it subsumes the appropriate entries of the derivative vector $f^{\mathcal{D}}(\mathbf{x})$. Lastly, we define

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

$\tilde{f}(x) = \hat{f}(x|p_{unif})$ to further simplify our notation.

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

### 3.1 Existence of the Constrained PDF Estimator

The first result that we establish is an existence result, that is, that a set of weights exists, provided that the constraints imposed are internally consistent and satisfy the constraints in (3.9).

**Theorem 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 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 1 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 $i_\ell$. 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, and $p_{j_n} > 0$ and $p_{l_n} < 0$ are sufficient to ensure the existence of a set of $p$ satisfying the constraints. The proof of Theorem 1 is provided in the Supplemental Material.

### 3.2 Consistency of the Constrained PDF Estimator

Here, we discuss the consistency of our constrained estimator. To begin, define a *hyperplane subset* of $\mathcal{J}$ as 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 $x_{0k} \in [m_k, b_k]$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

We call $\mathcal{S}$ an *interior hyperplane subset* if $x_{0k} \in (m_k, b_k)$. In the following, $f(\cdot)$ (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 (2.3), and $\tilde{f}(\cdot)$ (or $\tilde{f}^{\mathcal{D}}(\cdot)$) is the unconstrained estimator defined in (2.3).

**Assumption A1.**

i) The sample $\mathbf{X}_i$ either forms a regularly spaced grid on a compact set $\mathcal{I} \equiv [\mathbf{c}, \mathbf{e}] = \prod_{i=1}^{r}[c_i, e_i]$, or constitutes independent random draws from a distribution with a density $f$ that is continuous and nonvanishing 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 $\mathcal{J}$.

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

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

Assumption A1 i) is standard in the kernel density literature; at the expense of a more tedious proof, the same results can be demonstrated if the density is assumed to exist on an $r$-dimensional ball instead of on a hypercube. Assumption A1 ii) ensures the requisite smoothness of $f^{\mathcal{D}}$. Note that the bandwidth rate in Assumption A1 iii) is, in general, 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 $\tilde{f}$. However, 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 P. 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 the normal and exponential densities.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

**Theorem 2** (Consistency)**.** *Suppose that Assumption A1 1.–4. holds.*

*i) If* $f^{\mathcal{D}} > 0$ *on* $\mathcal{J}$*, then, with probability one,* $\hat{p} = 1/n$ *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* $f^{\mathcal{D}}(\mathbf{x}_0) = 0, \forall \mathbf{x}_0 \in \mathcal{X}_0$*. In addition, for any* $\mathbf{x}_0 \in \mathcal{X}_0$*, suppose that* $f^{\mathcal{D}}$ *has second-order continuous derivatives in the neighborhood 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 2, part i) suggests that when the constraint is strictly satisfied by the true function, 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 behavior of the weights $\hat{p}$ in such a case. Note that these results are easily extendable to the case of $f^{\mathcal{D}} \leq 0$ with a switch of sign in $f$.

The proof of Theorem 2 appears in the online Supplementary Material. Theorem 2 is a multivariate, multi-constraint, hyperplane subset adaptation of Du, Parmeter, and Racine (2013) to density estimation using the metric in (2.6).

### 3.3 Theoretical Properties of the Constrained PMF Estimator

Theorems 1 and 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:

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

(i) for each $k$, $\ell_{i_k}^{\mathcal{D}}(x)$ is strictly positive on a nonempty 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 nonempty set $\mathbf{O}_{i_k}$.

i) Assume that the kernel function $l(\cdot)$ in (2.7) is an ordered kernel function, and that the smoothing parameter $\lambda$ in (2.7) is of order $\lambda = O_p(n^{-1})$, which is a standard result in the literature.

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

**Theorem 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 the difference order, as opposed to differentiation.*

*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 one, $\hat{p} = 1/n$ 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 $n$.*

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 cells of the discrete random variable, which suggests that our constraints correspond to an ordered discrete random variable (Li and Racine 2003).

For the proof of the 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 zero. The argument is that, even if the constraint is violated, as 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 zero). In the discrete data setting, these results no longer hold, because for a discretely supported random variable, a measure-zero event is equivalent to an outcome not in the support; thus, a violation of the constraint is more troubling when considering discrete data.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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

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

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, and solving

$$\min_{p_i,\ldots,p_n} \sum_{i=1}^{n} (n^{-1} - p_i)^2, \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \psi_i(\mathbf{x}_j) \geq 0, 1 \leq j \leq N. \tag{3.10}$$

We place the same assumption on the grid points $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ as in 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 also needing the speed at which the smallest distance between the grid points decays to be slower than the rate of decay of the smoothing parameters (Assumption B2 ii)). The latter assumption is necessary to eliminate correlation across $\psi_i(\mathbf{x})$ as $n$ grows (Chacón, Duong, and Wand 2011).

Let $\hat{p}_i$, for $i = 1, \ldots, n$, be the solution to the quadratic programming problem in (3.10). Then,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

the asymptotic distribution of $D(\hat{p})$ is given in the following theorem, with the proof given in the Supplementary Material.

**Theorem 4.** *Suppose that assumptions A1 i)–iv) and B1 i)–iv) hold. Then, as $n \to \infty$, we have*

$$\frac{n^2 \sigma^2_{K^{(\mathbf{d})}}}{h^{2|\mathbf{d}|+r} \left( \sum_{j=1}^{M} f^{\mathcal{D}}(\mathbf{x}^*_j) \right)^2} D(\hat{p}) \sim \chi^2(n), \tag{3.11}$$

*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 defined in the Supplementary Material.*

Theorem 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, because $H_0$ and $H_1$ are both evaluated on infinite-dimensional parameter spaces (see also Fan, Zhang, and Zhang (2001)). Note too that, similarly to the generalized likelihood ratio test statistic of Fan, Zhang, and Zhang (2001), the distributional convergence in (3.11) 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 (3.11).

Given the well-known issues with the speed of convergence of nonparametric tests, we recommend using a bootstrap algorithm instead. Another reason to prefer the bootstrap is that the normalizing constant in (3.11) requires that we determine slack points, which may be difficult in practice. Du, Parmeter, and Racine (2013) show the consistency of the hypothesis test using $D(\hat{p})$ as the test statistic, which implies that the bootstrap version is consistent. In the constrained density setting, 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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, because the constraints are most likely true. However, if the constraints are not initially satisfied, then $D(\hat{p})$ can be used to test their validity.

One might consider generalizing the above result 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) is developed for the Cramér–von Mises statistic, it can instead be developed using a power-divergence metric by following P. Hall and Huang (2001) or by using the $L_2$-norm, following Du, Parmeter, and Racine (2013). The main difference lies in the algebraic manipulations required for the different metrics. For our theory on the limiting distribution of our test statistic (Theorem 4), we rely 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 metric is possible, though some degree of approximation is still necessary in order to obtain suitable expressions 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, and compare it with that of its competitors implemented in currently supported R packages available through CRAN. Note that the proposed estimator is extremely flexible in terms of the type of constraint and the number of simultaneous constraints that can be imposed. For the sake of brevity, we focus on a few test cases, and restrict the group of competitors to the most popular and promising methods of which we are currently aware. The test cases we consider involve two popular constraints, namely the *log-concavity* constraint and the *monotonicity* constraint (i.e., monotonically increasing). Although our approach supports both smooth constrained PDFs and PMFs, we focus on constrained PDF estimation, because of the lack of competing options for smooth constrained PMFs. However, we do provide an illustrative example involving the PMF; the R code for the constrained PDF and PMF estimations is available upon request. The proposed approach can be found in the R package `np` (Hayfield and Racine 2008), which is available on CRAN. See, in particular, the functions `npuniden.sc()` and `npuniden.boundary()`,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

which 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 $L_2$-norm (`function.distance="FALSE"`) can be used to enforce the weights.

For comparison purposes, in the log-concave constraint setting, we compare the proposed approach with those of Cule, Samworth, and Stewart (2010), who study a nonsmooth log-concave PDF MLE, and Chen and Samworth (2013), who study the associated smoothed log-concave estimator; these methods can be found in the R package `LogConcDEAD` (Cule, Gramacy, and Samworth 2009) in the functions `mlelcd()` and `dslcd()`. Note that we obtained similar results with the comparable functions in the R package `logcondens` (Dümbgen and Rufibach 2011), and so do not include these in the analysis below. For an informative overview, see Samworth (2017) for a recent survey of log-concave estimation and its importance in 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 a practical perspective, the kernel approach is limited to perhaps $d = 3$ or $d = 4$ dimensions. These approaches are also fre of tuning parameters, whereas the kernel approach requires the selection of bandwidths. In the log-concave constraint simulations that follow, we use cross-validation to select the bandwidths 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 free of tuning parameters matters, we begin by comparing the proposed approach based on *infeasible optimal bandwidths* (which are essentially free of tuning parameters) with *data-driven* smoothing parameter selection. Naturally, the optimal bandwidths present the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

method in the best possible light, albeit an unrealistic one, which is why we use the *data-driven* bandwidth-based results as a reference in the tables that follow, and not the *infeasible* optimal bandwidth-based results. 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 as the sample size increases.

In what follows, we consider a modest number of DGPs and, as noted above, restrict our 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 are drawn from the standard *smooth* unbounded support $N(0, 1)$ univariate Gaussian distribution ($X \in [-\infty, \infty]$), which is log-concave. We report the results based on the (unknown) optimal bandwidth and the data-driven bandwidth, and compare them those of the nonsmooth MLE estimator and the smooth MLE estimator under the log-concavity constraint (Section 6.1).

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

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

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

5. The data are drawn from a *nonsmooth* bounded support univariate triangular distribution ($X \in [0, 1]$), which is log-concave but *nonsmooth*, and we compare the results with those of the nonsmooth MLE estimator and the smooth MLE estimator under the log-concavity constraint. Note that we include this DGP in order to gauge its robustness, becaues it violates assumptions invoked when using kernel smoothing methods (i.e., the continuous differentiability of the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

density up to some particular order $> 2$) (Section 6.5).

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

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

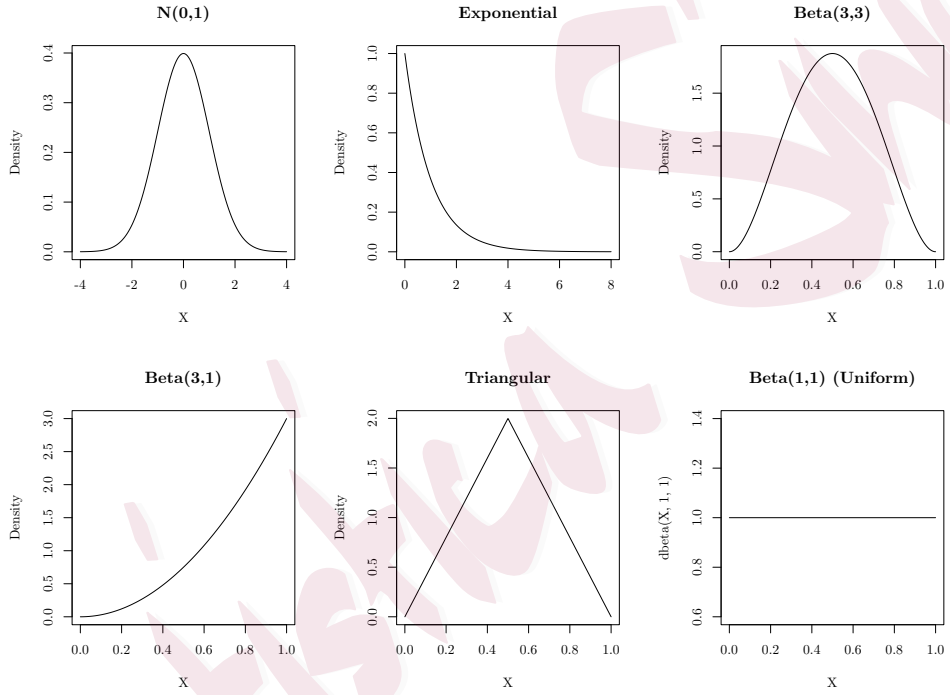


Figure 6: Monte Carlo Densities

Note that in each of these scenarios, we report the mean squared 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 $\hat{f}(\cdot)$ is an estimate thereof) for the smooth unconstrained version of our estimator (SU) (which is simply the standard kernel density estimation), smooth constrained version of our estimator (SC), nonsmooth MLE estimator (LNS), smooth MLE estimator (LS), and monotone rearrangement estimator (MR). We report the results in both tables (mean/median relative MSE over $M$ Monte Carlo replications) and figures (box plots of the MSE for the $M$ Monte Carlo replications). We present the *relative* MSE values for the mean and median to provide a complete impression of the performance, because the relative mean

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

values may not be robust in the presence of outlying values. Such values occur if, say, data-driven bandwidth selection performs poorly for some fraction of the resamples, and the relative median is naturally less affected by outlying values.

The proposed kernel approach admits known finite boundary points (i.e., the boundary points of $\pm\infty$ have no effect on the estimate), which are used for the exponential (which uses $(a, b) = (0, \infty)$), 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. Although one might consider modifying the peer function using standard correction methods, it is unclear whether log-concavity is always preserved. Regardless, any such extension of the peer method lies beyond the scope of this study.

In order to meet the page length constraints, the particulars of the Monte Carlo simulations have been moved to the Supplementary Material, which also contains the technical proofs (see Section 6). Briefly, the proposed method is shown to be competitive with its nonsmooth and smooth peers and, most importantly, provides an extensible and general approach to constraining kernel-based density estimates in a unified framework.

## 5 Conclusion

We have presented a versatile procedure designed to impose a variety of shape constraints on a *smooth* nonparametric kernel density estimator. We use simulations and real-world data to show that the method can deliver practical and useful estimates of an unknown density, satisfying a range of constraints, and provide the theoretical underpinnings thereof. Additionally, for the constraint of log-concavity, our proposed approach convincingly outperforms popular existing approaches. Furthermore, for the constraint of monotonicity, our approach is competitive with its peers, perhaps even performing somewhat better. However, 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 within a unified framework. Moreover, these constraints can be applied to settings involving both continuous and ordered discrete data settings. An R implementation is available on CRAN (see the R package `np` (Hayfield and Racine 2008), and the functions `npuniden.sc()` and `npuniden.boundary()` contained

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

therein).

There are many exciting and important directions in which the proposed methods can be extended. For example, we can use the insights of Mammen (1991) (in the regression setting) to consider higher-order asymptotic comparisons between the unconstrained and constrained estimators. Given that the constrained density estimator that we propose here is expected to equal the unconstrained estimator *if* the constraints imposed are valid, then for sufficiently large $n$, these two coincide (to the first order). Hence, one would not expect large sample gains. However, a more nuanced and detailed asymptotic analysis may reveal important higher-order gains that could prove useful for constructing of confidence intervals in small sample settings. Another possible extension is to consider functions supported on a ball, rather than on a hypercube, as considered here. This would require changing existing tools, such as considering kernel functions supported on a ball. Both of these extensions are left to future work.

## Supplementary Materials

This material contains all of the proofs for the theorems introduced in the paper and the full set of tables and figures for the Monte Carlo simulations.

## Acknowledgements

We sincerely thank the editor, the associate editor, and two anonymous reviewers for their helpful comments and suggestions. Du's research was partly supported by the U.S. National Science Foundation grant DMS-1916174.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

# 6 Supplementary Finite-Sample Performance Appendix (NOT FOR PUBLICATION)

In order to meet page length constraints, particulars of the Monte Carlo simulations have been moved from the main body of the text into this section.

## 6.1 Smooth Unbounded Gaussian DGP, Log-Concave Constraint

Table 1: Gaussian $N(0,1)$ DGP, log-concave density estimation, comparison of data-driven bandwidth smooth constrained estimator (SC), data-driven bandwidth smooth unconstrained estimator (SU), optimal bandwidth smooth constrained estimator (SCo), optimal bandwidth smooth unconstrained estimator (SUo), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SCo | SU | SUo | LNS | LS | n | SC | SCo | SU | SUo | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 0.87 | 1.10 | 0.88 | 1.58 | 1.03 | 100 | 1 | 0.88 | 1.09 | 0.90 | 1.47 | 0.95 |
| 200 | 1 | 0.90 | 1.12 | 0.91 | 1.49 | 1.10 | 200 | 1 | 0.88 | 1.06 | 0.89 | 1.44 | 1.10 |
| 400 | 1 | 0.93 | 1.09 | 0.94 | 1.39 | 1.13 | 400 | 1 | 0.92 | 1.07 | 0.94 | 1.36 | 1.11 |
| 800 | 1 | 0.93 | 1.08 | 0.95 | 1.32 | 1.14 | 800 | 1 | 0.93 | 1.06 | 0.94 | 1.34 | 1.13 |

Tables 1 and Figure 7 summarize MSE results for data drawn from the $N(0,1)$ distribution. These results reveal that, as one would expect, imposing constraints consistent with the DGP (the Gaussian distribution is log-concave) improves the efficiency of the estimate (compare columns SC with SU - since this result holds in general we do not remark on it further in what follows). Next, we observe that the nonsmooth MLE estimator (LNS), as one might expect, does not perform as well as its smooth counterparts for this smooth DGP, though it is interesting to note that the smooth kernel estimator (SC) is capable of outperforming its smooth MLE counterpart (LS) for $n > 100$. The slightly diminished performance of the constrained kernel estimator (SC) for $n = 100$ relative to the smooth MLE estimator (LS) appears to be a reflection of small-sample variability in the data-driven bandwidth procedure, nothing more (using the infeasible optimal bandwidth for the Gaussian $N(0,1)$ distribution, $h = 1.059\hat{\sigma}n^{-1/5}$, delivers an infeasible estimator (SCo) that outperforms LNS and LS for all sample sizes). We are encouraged by the performance of the proposed approach in this setting particularly since the MLE methods are *tailor made for this particular constraint* while our

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

approach is an omnibus one that can handle this and many other constraints as demonstrated in the illustrations provided in Section 2 and those that follow.

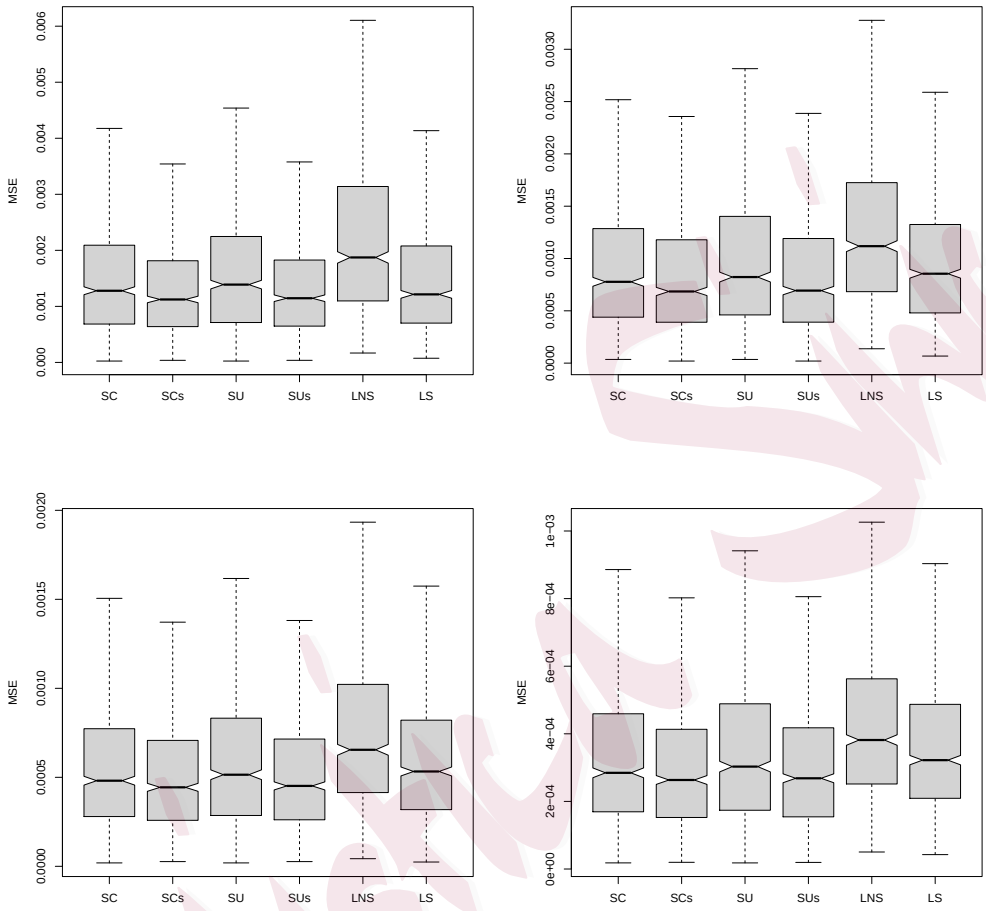


Figure 7: MSE boxplots for the $N(0,1)$ DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the data-driven bandwidth smooth constrained estimator (SC), data-driven bandwidth smooth unconstrained estimator (SU), optimal bandwidth smooth constrained estimator (SCo), optimal bandwidth smooth unconstrained estimator (SUo), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Smaller numbers are preferred.

### 6.2 Smooth Unbounded Bivariate Gaussian DGP, Log-Concave Constraint

We next consider imposing constraints in a multivariate (bivariate) setting. We mimic the univariate description in Section 6.1 so will not repeat it here, other than to say that we use a standard product kernel function, impose the log-concavity constraint in both dimensions, and use data-driven bandwidth selection identical to that used in the univariate case while allowing the bandwidths

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

to differ across dimensions. We simulate data from a correlated bivariate Gaussian with $\rho = 0.5$, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$.

Table 2 is similar to Table 1, and imposing log-concavity (which is consistent with the underlying DGP) leads to an improvement in estimator performance, though relatively less so than in the univariate case (more on this shortly). It is also interesting to note that the LS estimator (smooth constrained MLE) outperforms both kernel estimators while the LNS estimator (nonsmooth MLE) approaches that of the kernel estimators as $n$ increases. We would be remiss if we did not point out that this does not diminish the value of the proposed approach for two reasons, namely 1) the LNS and LS estimators are exclusively tailored to the log-concavity constraint (our approach is omnibus in the sense that we are able to handle a very broad array of constraints), and 2) our approach is intended for those using kernel methods who wish to further constrain their unrestricted estimator (telling them to use a non-kernel method would run counter to their needs).

Following up on the relative performance of the constrained versus unconstrained kernel estimators in this setting, it turns out that the relatively small apparent improvement appears to simply be an artifact of bivariate bandwidth selection, i.e., the appropriate selection of increasingly larger bandwidths as the dimension increases (a consequence of the so-called *curse of dimensionality*). That is, as the bandwidth increases other things equal, the unconstrained kernel estimator grows ever more smooth which, in this setting, translates to an estimator that is more likely to be log-concave than one which is less smooth. To assess the extent to which this is the case, we include Table 3 which presents the counterparts to column "SU" in Table 2 (i.e., the relative MSE of the constrained versus the unconstrained kernel estimators) where we use bandwidths that are 0.50, 0.75, 1.00, and 2.00 times the data-driven bandwidths which were used in Table 2. Note that using bandwidths that are twice as large (column "2.00" in Table 3) as the data-driven bandwidths results in a constrained and unconstrained estimator that are indistinguishable from an MSE standpoint simply because the oversmoothing leads to an unconstrained estimator that is in fact log-concave. However, as we undersmooth (i.e., moving from column "0.75" then "0.50" in Table 3) we indeed see the relative MSE improvement that one might otherwise expect. In essence, for a given sample size, the univariate kernel estimator is more likely to violate log-concavity than the bivariate estimator as the former will select smaller bandwidth(s) which is reflected in the relative performance of the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

constrained and unconstrained estimators in Table 1 (univariate) versus Table 2 (bivariate).

Table 2: Bivariate Gaussian $N(0, \Sigma)$ DGP, log-concave density estimation, comparison of data-driven bandwidth smooth constrained estimator (SC), data-driven bandwidth smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | LNS | LS | n | SC | SU | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.00 | 1.58 | 0.65 | 100 | 1 | 1.00 | 1.41 | 0.63 |
| 200 | 1 | 1.01 | 1.25 | 0.62 | 200 | 1 | 1.01 | 1.14 | 0.58 |
| 400 | 1 | 1.01 | 1.06 | 0.60 | 400 | 1 | 1.01 | 0.98 | 0.57 |
| 800 | 1 | 1.01 | 0.92 | 0.58 | 800 | 1 | 1.01 | 0.88 | 0.55 |

Table 3: Bivariate Gaussian $N(0, \Sigma)$ DGP, log-concave density estimation, comparison of constrained (SC) versus unconstrained (SU) MSE performance with respect to the optimal data-driven bandwidths. Column headers are the fraction of the data-driven bandwidth used in the construction of both estimators (e.g., 1.00 uses the raw data-driven bandwidths, 0.50 uses 1/2 of that, 2.00 twice that, etc.). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | 0.50 | 0.75 | 1.00 | 2.00 | n | 0.50 | 0.75 | 1.00 | 2.00 |
|---|---|---|---|---|---|---|---|---|---|
| 100.00 | 1.01 | 1.01 | 1.00 | 1.00 | 100.00 | 1.02 | 1.00 | 1.00 | 1.00 |
| 200.00 | 1.02 | 1.01 | 1.01 | 1.00 | 200.00 | 1.02 | 1.02 | 1.01 | 1.00 |
| 400.00 | 1.03 | 1.04 | 1.01 | 1.00 | 400.00 | 1.03 | 1.04 | 1.01 | 1.00 |
| 800.00 | 1.10 | 1.09 | 1.01 | 1.00 | 800.00 | 1.11 | 1.08 | 1.01 | 1.00 |

### 6.3 Smooth Left-Bounded Exponential DGP, Log-Concave Constraint

Table 4: Exponential DGP, log-concave density estimation, comparison of data-driven bandwidth smooth constrained estimator (SC), data-driven bandwidth smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | LNS | LS | n | SC | SU | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.13 | 0.56 | 6.96 | 100 | 1 | 1.15 | 0.46 | 8.77 |
| 200 | 1 | 1.26 | 0.58 | 10.88 | 200 | 1 | 1.37 | 0.47 | 14.08 |
| 400 | 1 | 1.34 | 0.58 | 15.57 | 400 | 1 | 1.45 | 0.48 | 20.08 |
| 800 | 1 | 1.41 | 0.63 | 23.32 | 800 | 1 | 1.57 | 0.57 | 32.01 |

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Table 4 and Figure 8 summarize MSE results for data drawn from the Exponential distribution. Note that here (and in the remaining simulation that follow) we use data-driven bandwidths only and do not report tables for the optimal (unknown) bandwidth as was done in Section 6.1. Unlike the Gaussian case summarized in Table 1 where the nonsmooth MLE estimator (LNS) and smooth MLE estimators (LS) were outperformed by the proposed method *and* the nonsmooth MLE estimator (LNS) was outperformed by the smooth MLE estimator (LS), here the nonsmooth MLE estimator (LNS) outperforms the proposed estimator yet the smooth MLE estimator (LS) performs markedly worse and is itself outperformed by the proposed estimator. This suggests that among the nonsmooth MLE (LNS) and smooth MLE (LS) estimators, relative performance cannot be taken for granted and one may substantially outperform the other depending on the underlying DGP.

### 6.4 Smooth Bounded Beta(3,3) DGP, Log-Concave Constraint

Table 5 and Figure 9 consider the Beta(3,3) distribution and repeats the above exercise. As was the case above, for $n > 100$ the fully automatic bandwidth selector delivers bandwidths that offer superior performance and can be recommended for all but the smallest of sample sizes.

Table 5: Beta(3,3) DGP, log-concave density estimation, comparison of smooth constrained estimator (SC), smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | LNS | LS | n | SC | SU | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.14 | 1.31 | 0.89 | 100 | 1 | 1.08 | 1.43 | 0.95 |
| 200 | 1 | 1.26 | 1.40 | 1.09 | 200 | 1 | 1.08 | 1.48 | 1.14 |
| 400 | 1 | 1.32 | 1.39 | 1.18 | 400 | 1 | 1.10 | 1.35 | 1.12 |
| 800 | 1 | 1.29 | 1.39 | 1.25 | 800 | 1 | 1.13 | 1.39 | 1.22 |

### 6.5 Non-Smooth Bounded Triangular DGP, Log-Concave Constraint

Table 6 and Figure 10 consider the nonsmooth Triangular distribution and repeats the above exercise. Even though the kernel method presumes that the underlying distribution is smooth, the proposed method outperforms the nonsmooth MLE estimator (LNS) but in this instance is dominated by smooth MLE estimator (LS), which is perhaps unexpected given the nonsmooth nature of the DGP.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

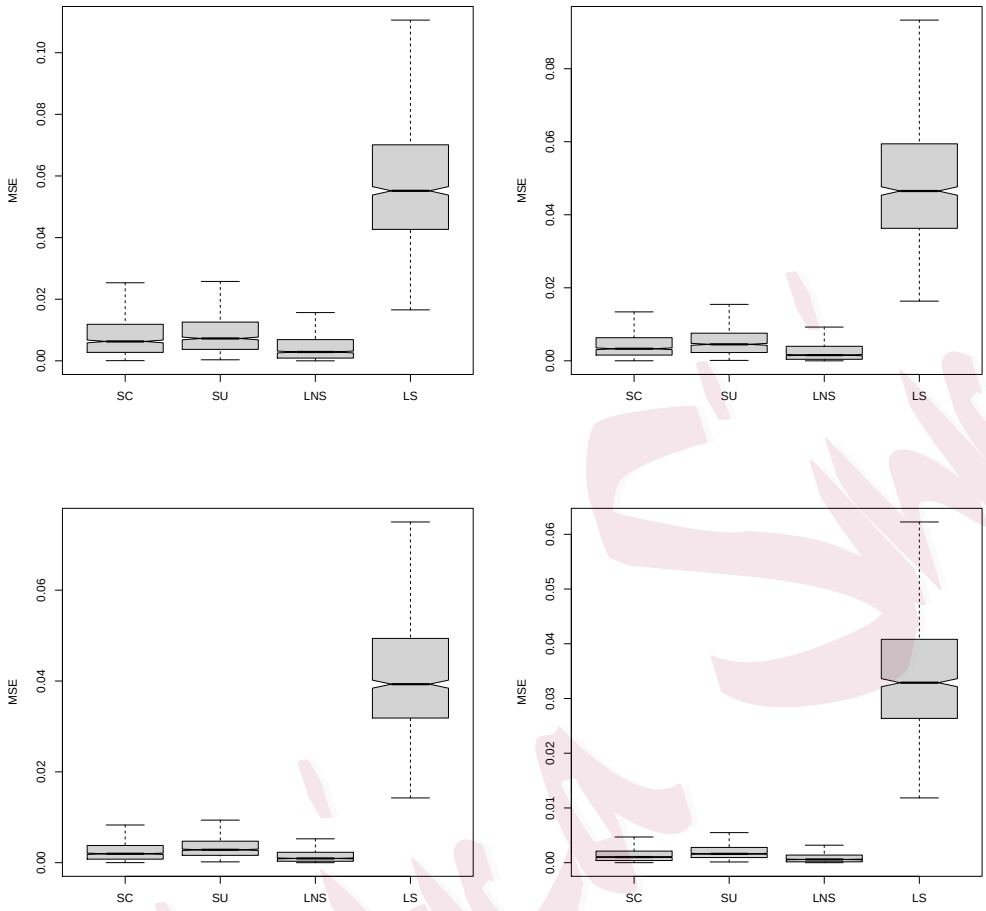


Figure 8: MSE boxplots for the Exponential DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the data-driven bandwidth smooth constrained estimator (SC), data-driven bandwidth smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Smaller numbers are preferred.

Table 6: Triangular DGP, log-concave density estimation, comparison of smooth constrained estimator (SC), smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | LNS | LS | n | SC | SU | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.11 | 1.06 | 0.77 | 100 | 1 | 1.10 | 1.06 | 0.76 |
| 200 | 1 | 1.19 | 1.11 | 0.89 | 200 | 1 | 1.14 | 1.10 | 0.89 |
| 400 | 1 | 1.21 | 1.12 | 0.96 | 400 | 1 | 1.18 | 1.10 | 0.93 |
| 800 | 1 | 1.22 | 1.09 | 0.98 | 800 | 1 | 1.19 | 1.06 | 0.94 |

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

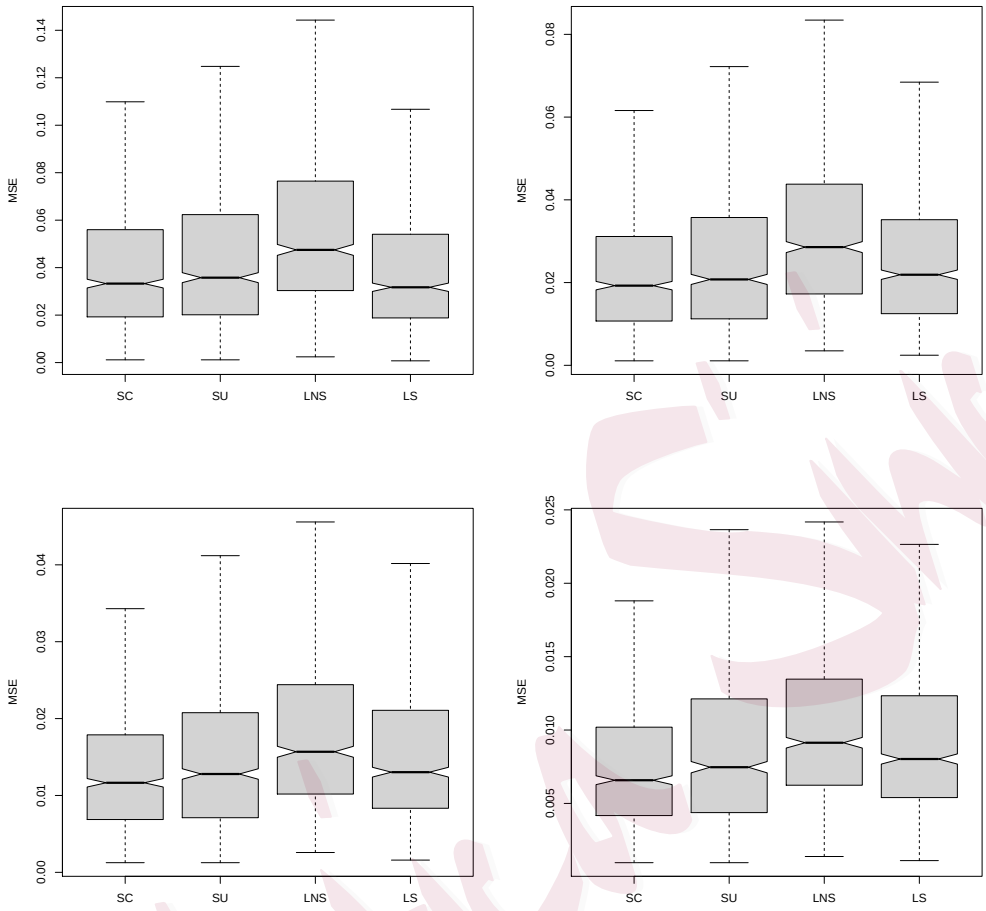


Figure 9: MSE boxplots for the Beta(3,3) DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the smooth constrained kernel estimator (SC), smooth unconstrained kernel estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Smaller numbers are preferred.

### 6.6 Non-Smooth Bounded Uniform DGP, Log-Concave Constraint

Table 7 and Figure 11 consider the Uniform distribution and repeats the above exercise. It is interesting to note that here the smooth MLE estimator (LS) performs particularly poorly relative to its peers while the proposed approach (SC) dominates its peers. However, one perhaps lesser-appreciated reason for this dominance is due to the fact that this case is particularly favourable to the kernel estimator as the doubly truncated kernel used in this setting mimics the uniform distribution when the data-driven bandwidth grows, which is in fact appropriate for this DGP hence the particularly good performance of even the unconstrained kernel estimator (SU). In fact,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

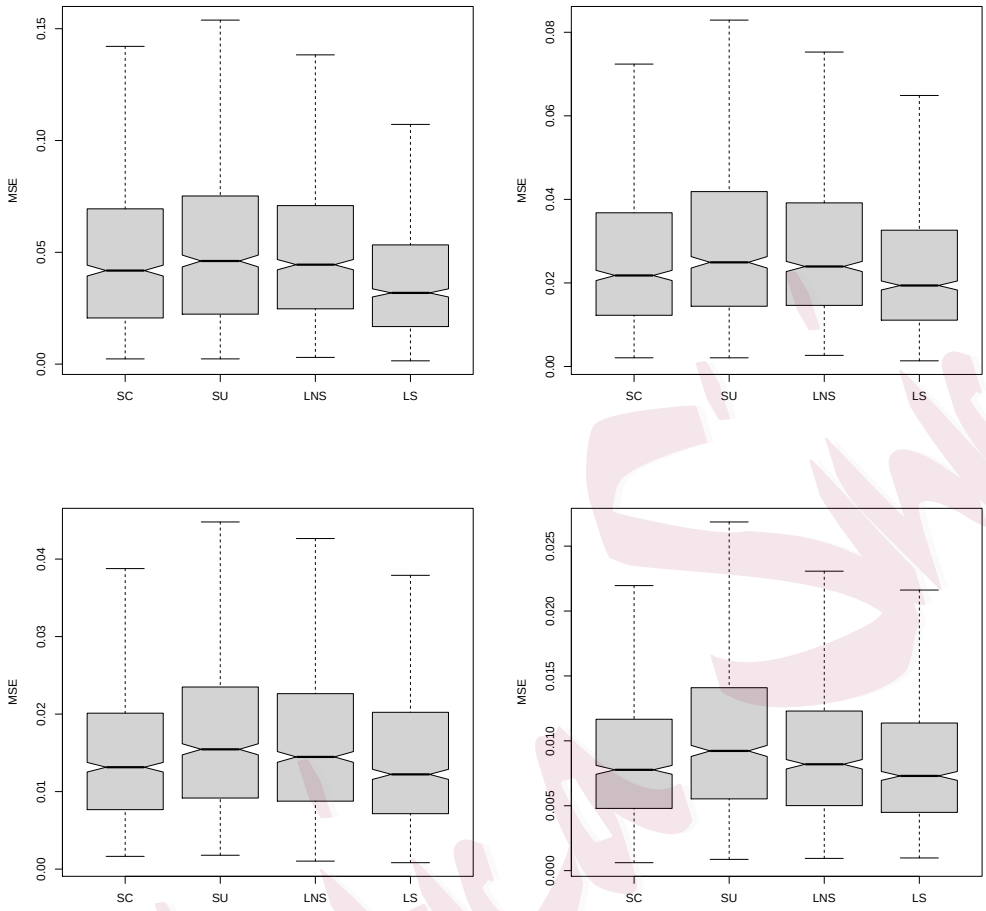


Figure 10: MSE boxplots for the Triangular DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the smooth constrained kernel estimator (SC), smooth unconstrained kernel estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Smaller numbers are preferred.

the median MSE is so low relative to its peers (almost zero) that relative efficiency is orders of magnitude better as can be seen in Table 7.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Table 7: Uniform DGP, log-concave density estimation, comparison of smooth constrained estimator (SC), smooth unconstrained estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | LNS | LS | n | SC | SU | LNS | LS |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.42 | 2.37 | 4.59 | 100 | 1 | 1 | 1.10e+17 | 2.90e+17 |
| 200 | 1 | 1.84 | 2.72 | 7.35 | 200 | 1 | 1 | 7.18e+16 | 2.73e+17 |
| 400 | 1 | 2.17 | 3.36 | 12.58 | 400 | 1 | 1 | 5.02e+16 | 2.59e+17 |
| 800 | 1 | 2.36 | 3.92 | 20.26 | 800 | 1 | 1 | 4.06e+16 | 2.57e+17 |

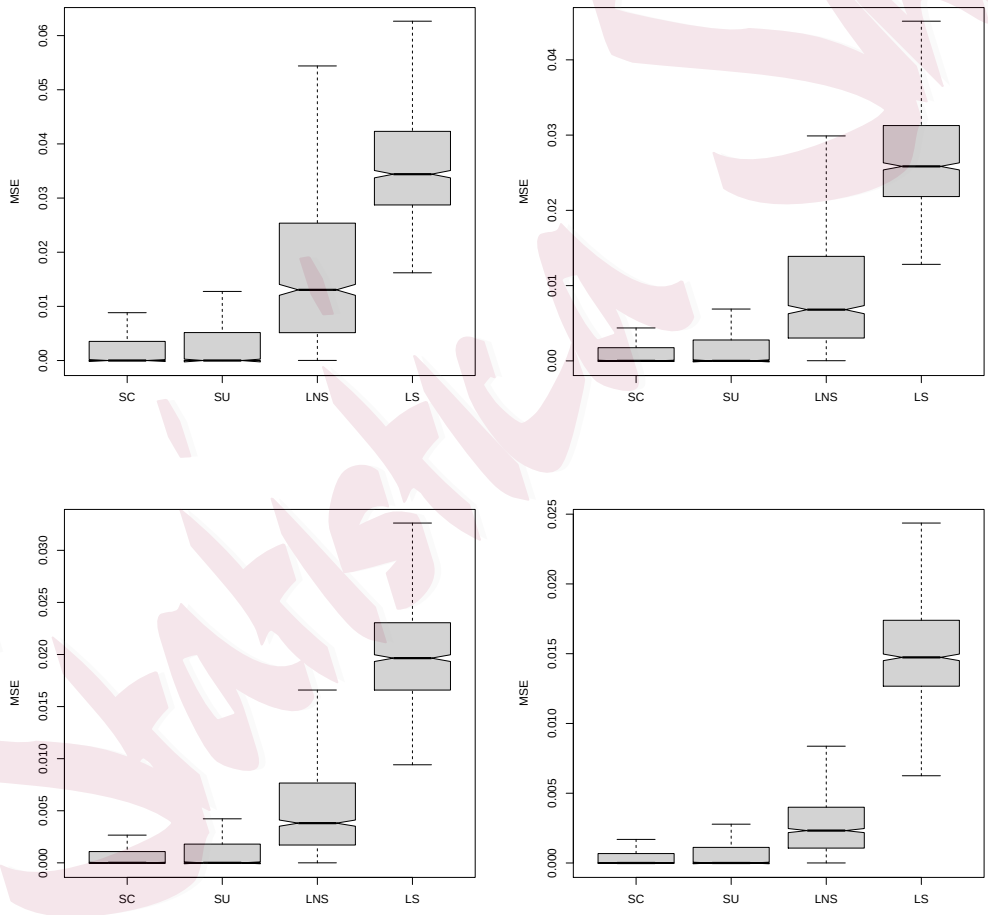


Figure 11: MSE boxplots for the Uniform DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the smooth constrained kernel estimator (SC), smooth unconstrained kernel estimator (SU), nonsmooth MLE estimator (LNS) and smooth MLE estimator (LS). Smaller numbers are preferred.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

The reader will have noted that relative MSE values that are quite high are reported in Table 7 for the LNS and LS methods. The reason for this stems from the fact that one feature of the proposed estimator is that when the true DGP is uniform $[a, b]$ with known boundary points, the kernel estimator is unbiased for any bandwidth $h$. In this case, cross-validation correctly selects a large value of $h$ which minimizes variance (the bias is proportional to $h$, the variance inversely proportional to $h$). And for this DGP a large bandwidth delivers essentially the uniform distribution with a resulting MSE that is extremely small resulting in relative performance that indeed is as high as $10^{17}$ since the numerator of the MSE ratio is close to zero. This oversmoothing phenomenon is a known feature of various kernel methods: see Hall, Racine, and Li (2004) by way of illustration.

### 6.7 Smooth Bounded Beta(3,1) DGP, Monotonicity Constraint

Table 8 and Figure 12 consider the monotonically increasing Beta(3,1) distribution, and repeats the above exercise except that here we compare the proposed approach with the monotone rearrangement approach of Birke (2009). One benefit of monotone rearrangement is that it does not disturb the probability mass since it simply rearranges the point estimates. We observe that the fully automatic bandwidth selector delivers bandwidths that offer improved performance overall when considering median MSE and mean MSE though the latter holds for $n > 100$ and likely reflects the poor data-driven bandwidth for a small fraction of the samples which might be expected given the nature of the DGP.

Table 8: Beta(3,1) DGP, monotone increasing density estimation, comparison of smooth constrained estimator (SC), smooth unconstrained estimator (SU) and monotone rearrangement (MR). Results are relative MSE (relative to the proposed method SC), the table on the left is mean relative MSE, right median relative MSE, columns with numbers $> 1$ indicate the proposed method (SC) is preferred.

| n | SC | SU | MR | n | SC | SU | MR |
|---|---|---|---|---|---|---|---|
| 100 | 1 | 1.09 | 0.95 | 100 | 1 | 1.23 | 1.05 |
| 200 | 1 | 1.15 | 1.01 | 200 | 1 | 1.10 | 1.05 |
| 400 | 1 | 1.18 | 0.98 | 400 | 1 | 1.17 | 1.09 |
| 800 | 1 | 1.16 | 1.03 | 800 | 1 | 1.09 | 1.07 |

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

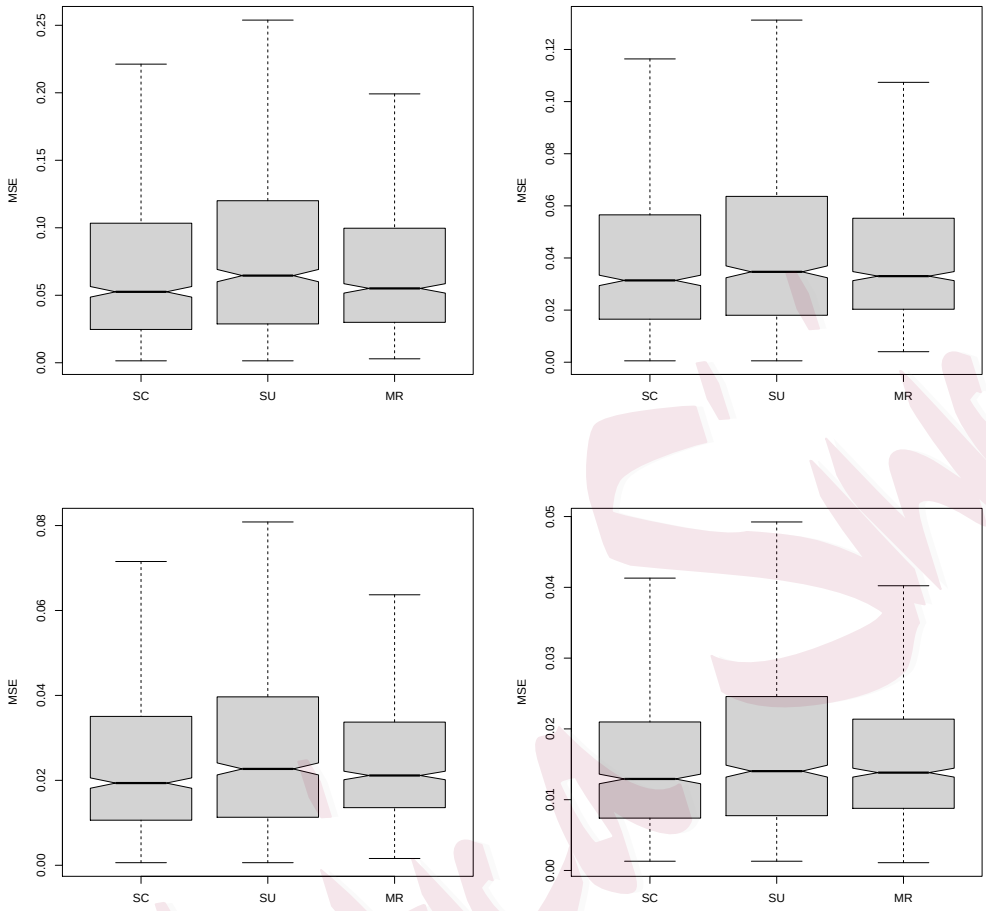


Figure 12: MSE boxplots for the Beta(3,1) DGP, clockwise from top left $n = 100$, $n = 200$, $n = 400$, $n = 800$. The legends correspond to the smooth constrained kernel estimator (SC), smooth unconstrained kernel estimator (SU) and monotone rearrangement (MR). Smaller numbers are preferred.

# 7 Supplementary Technical Appendix (NOT FOR PUBLICATION)

## 7.1 Proof of Theorem 1

Conditions (i) and (ii) in the theorem give an open cover of the domain $\mathcal{J}$ by the open sets $\mathbf{O}_{i_l}$, $l = 1, \ldots, k$. Without loss of generality, we can assume this cover is minimum, that is, there are no $l_1 \neq l_2$ such that $\mathbf{O}_{i_{l_1}} \subset \mathbf{O}_{i_{l_2}}$ or vice versa. Also, we can assume that the sets are properly ordered such that each pair of consecutive sets overlap with each other. Denote the boundary of a set $S$ by

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

$\partial S$ and let $\partial O \equiv \partial \mathcal{J} \cup (\cup_{l=1}^{k} \partial \mathbf{O}_{i_l} \cap \mathcal{J})$. Let $d(\mathbf{x}, A) = \inf\{|\mathbf{x} - \mathbf{y}| : \mathbf{y} \in A\}$ be the distance from a point $\mathbf{x}$ to a set $A$. For each $\mathbf{x} \in \partial O$, define $d_{\mathbf{x}} = \min\{d(\mathbf{x}, \partial \mathbf{O}_{i_l}) : \mathbf{x} \in \mathbf{O}_{i_l}\}$. Then $\min_{\mathbf{x} \in \partial O} d_{\mathbf{x}} > 0$ since $\{\mathbf{O}_{i_l} : l = 1, \ldots, k\}$ form a finite open cover of $\mathcal{J}$. Choose an $\epsilon \in (0, \min_{\mathbf{x} \in \partial O} d_{\mathbf{x}})$. For an open set $A$, define its $\epsilon$-closure $\overline{A}_\epsilon$ within $\mathcal{J}$ as the largest closed set in $\mathcal{J}$ whose distance to $A$ is $\epsilon$. Note that from our choice of $\epsilon$, $\mathcal{J} \subset (\overline{\cup_{l=1}^{k} \mathbf{O}_{i_l}})_\epsilon$ and for each $1 \leq l \leq k$ every point in $(\mathbf{O}_{i_l} \cap \mathcal{J}) \setminus \overline{\mathbf{O}}_{i_l, \epsilon}$ is covered by at least one $\mathbf{O}_{i_s}$ with $s \neq l$.

Similar to both the proof of Theorem 4.1 in P. Hall and Huang (2001) and Theorem 1 in Du, Parmeter, and Racine (2013), we use induction to introduce the un-normalized weights $w_{i_l}$, $l = 1, \ldots, k$. First, $w_{i_1}$ is set to be 1 or $-1$, depending on the sign of $K_{i_1}^{\mathcal{D}}$, such that $w_{i_1} K_{i_1}^{\mathcal{D}} > 0$ on $\overline{\mathbf{O}}_{i_1, \epsilon}$. Suppose we have constructed a sequence $w_{i_1}, \ldots, w_{i_j}$ such that $\sum_{l \leq j} w_{i_l} K_{i_l}^{\mathcal{D}} > 0$ on $(\overline{\cup_{l=1}^{j} \mathbf{O}_{i_l}})_\epsilon$. Since $\mathbf{O}_{i_{j+1}} \cap (\overline{\cup_{l=1}^{j} \mathbf{O}_{i_l}})_\epsilon \neq \emptyset$ (by the choice of $\epsilon$ and the fact that $\mathbf{O}_{i_j} \cap \mathbf{O}_{i_{j+1}} \neq \emptyset$), $K^{\mathcal{D}} > 0$ on $\mathbf{O}_{i_{j+1}}$ (by condition (i)), and $K^{\mathcal{D}}$ is bounded away from $-\infty$, uniformly in $l \leq j$ and in $\mathbf{x} \in \mathbb{R}^r$ (by condition (iii)), then if $w_{i_{j+1}}$ is chosen with the same sign as $K_{i_{j+1}}^{\mathcal{D}}$ and with $|w_{i_{j+1}}|$ sufficiently large, we shall have $\sum_{l \leq j+1} w_{i_l} K_{i_l}^{\mathcal{D}} > 0$ on $(\overline{\cup_{l=1}^{j+1} \mathbf{O}_{i_l}})_\epsilon$. At the last step, $w_{i_k}$ should also be chosen to satisfy $\sum_{l=1}^{k} w_{i_l} \neq 0$.

Next, we set $w_j = 0$ for each $j \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$. Then the normalized weights $p_i = w_i / \sum_{l=1}^{n} w_l$, for each $i \in \{1, \ldots, n\}$ gives a set of weights satisfying the requisite property.

### 7.2 Proof of Theorem 2

Let $p_0 = (1/n, \ldots, 1/n)'$ be the vector of unconstrained weights. From Li, Liu, and Li (2017), we have $D(p) = (Ap - Ap_0)'(Ap - Ap_0) = (p - p_0)' A' A (p - p_0)$ where $A$ is the matrix of kernel weights with $(i, j)$th entry $K_i(\mathbf{x}_j)$. Let $p$ be the vector of constraint weights minimizing $D(p)$. Define vectors $q = Ap$ with the $j^{th}$ element $q_j = A_{[j,]} p$ and $q_0 = Ap_0$ with $j^{th}$ element $q_{0j} = A_{[j,]} p_0$, where $A_{[j,]}$ is the $j^{th}$ row of $K^{\mathcal{D}}$.

For the proofs that follow we first establish some technical lemmas. Lemma 1 partitions the weights into two sets, one where individual weights provide identical $q_j$ and have similar orders, and one where they are a set of fixed constants (that may differ), and is used exclusively in Theorem 2 (iii). Lemma 2 establishes that there exist a set of weights that guarantee that the constraint is satisfied with probability approaching one in the limit. These weights do not have to equal the optimal

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

weights, but are used to bound the distance metric evaluated at the optimal weights. Lemma 2 is used in Theorem 2 (ii) and (iii). Lemma 3 establishes that with probability approaching one in the limit the unrestricted estimate $\tilde{f}$ satisfies $\tilde{f}^{\mathcal{D}}(\mathbf{x}) > \frac{1}{3} f^{\mathcal{D}}(\mathbf{x})$ for any point $\mathbf{x}$ of certain distance away from $\mathcal{X}_0$. It is used to determine the suitable distance away from $\mathcal{X}_0$ for a point $\mathbf{x}$ to have $\hat{f}^{\mathcal{D}}(\mathbf{x}|\hat{p}) > 0$, which is then used to derive the orders of $Z_1$ and $Z_2$ in Theorem 2 (iii).

**Lemma 1.** *If $\mathcal{A}$ and $\mathcal{B}$ are complementary subsets of the integers $1, \ldots, n$ and if $p_i$, for $i \in \mathcal{A}$ are fixed, then the values of $p_j$ for $j \in \mathcal{B}$ that minimize $D(p)$ are those that provide identical $q_j = A_{[j,]}p = \sum_{i=1}^{n}(p_i - 1/n)K_i(\mathbf{x}_j)$, and are uniquely determined if a constraint like $\sum_{i=1}^{n} p_i = 1$ is enforced.*

*Proof.* Proof of Lemma 1.

We find the optimal $q_j$s by minimizing $\sum\limits_{j\in\mathcal{B}} (q_j - q_{0j})^2$ subject to $\sum\limits_{j\in\mathcal{B}} p_j = -\sum\limits_{i\in\mathcal{A}} p_i$. Without loss of generality, assume $n \in \mathcal{B}$. By incorporating the constraint directly into our distance metric we obtain

$$\min_{q_j, j\in\mathcal{B}\setminus\{n\}} \left[ \sum_{j\in\mathcal{B}\setminus\{n\}} (q_j - q_{0j})^2 - \left( \sum_{i\in\mathcal{A}} q_i + \sum_{j\in\mathcal{B}\setminus\{n\}} q_j - q_{0j} \right)^2 \right],$$

which yields the first order conditions, for $j \in \mathcal{B}\setminus\{n\}$, $q_j = -\sum\limits_{i\in\mathcal{A}} q_i - \sum\limits_{k\in\mathcal{B}\setminus\{n\}} q_k$. Since $\sum\limits_{k\in\mathcal{B}\setminus\{n\}} q_k$ is fixed we see that all $q_j$, $j \in \mathcal{B}\setminus\{n\}$, are identical. Moreover, the excluded $q_n$ is also equal to $-\left(\sum\limits_{j\in\mathcal{A}} q_j + \sum\limits_{k\in\mathcal{B}\setminus\{n\}} q_k\right)$ which proves Lemma 1. □

We remark here that in both Hall and Huang (2002) and Du, Parmeter, and Racine (2013), for the regression setting, it was assumed that the weights sum to one, $\sum\limits_{i=1}^{n} p_i = 1$. This was done primarily so that the weights could be interpreted as probability weights, consistent with the power-divergence metric originally proposed to enforce the constraints, as in Hall and Huang (2002). However, the work of Du, Parmeter, and Racine (2013) demonstrated in the context of an $L_2$-norm, that the weights did not need to be nonnegative nor sum to one for the constraints to be successfully implemented. In the density setting we need $\sum\limits_{i=1}^{n} a_i = 0$ in order to ensure that the constrained density is in fact a proper probability density function. When the weights sum to zero the effect is that even with the constraints imposed, the estimated density integrates to one. We also mention that for the case of constraining the log-density, such as log-concavity, $\sum\limits_{i=1}^{n} a_i = 0$ will no longer ensure that the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

constrained density will integrate to one. This implies that some level of rescaling will be necessary to construct a log-concave density which is a proper probability density function. However, having the constraint weights sum to zero, though arbitrary, does ensure that the integrated difference between the log constrained density and the log unconstrained density is 0. Additionally, Lemma 1 would continue to hold for any finite value for which the weights were constrained to sum to. Thus, whether we are working in the pure density setting, or the log-density setting, we keep the restriction that $\sum_{i=1}^{n} a_i = 0$ and work with this in Lemma 1.

**Lemma 2.** *For each $\delta > 0$ there exists a $\tilde{p} = \tilde{p}(\delta)$ satisfying*

$$P\left\{\hat{f}^{\mathcal{D}}(\mathbf{x}|\tilde{p}) > 0 \quad \forall \mathbf{x} \in \mathcal{J}\right\} > 1 - \delta, \tag{7.1}$$

*for all sufficiently large $n$. Moreover, this yields $D(\hat{p}) = O_p\left(n^{-1}h^{2|\mathbf{d}|+2r+1}\right)$.*

*Proof.* Proof of Lemma 2.

Choose any $\mathbf{x}_0 \in \mathcal{X}_0$. Consider a function $L(\mathbf{x}) = \sum_{j=1}^{r} L_j(x_j)$, where each component $L_j$ is a fixed and compactly supported function with appropriate derivatives to be specified later. Define

$$\tilde{p}_i = n^{-1}\left\{1 + \Delta + h^{|\mathbf{d}|+r} f(\mathbf{X}_i)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right)\right\}, \tag{7.2}$$

where $\Delta$ is a constant defined by $\sum_{i=1}^{n} \tilde{p}_i = 1$. Note that the compactness of $L$ is component-wise such that for some constant $C_1 > 0$, $L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right) \neq 0$ requires that $|x_{0j} - X_{ij}| \leq C_1 h$ for at least one $1 \leq j \leq r$. This is different from the compactness of $K$ (and its derivatives) which indicates that there exists a constant $C_2 > 0$ such that $K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \neq 0$ requires that $|x_j - X_{ij}| \leq C_2 h$ for all $j = 1, \ldots, r$. Now

$$\begin{aligned}
\hat{f}^{\mathcal{D}}(\mathbf{x}|\tilde{p}) =& \sum_{i=1}^{n} \tilde{p}_i K_i^{\mathcal{D}}(\mathbf{x}) \\
=& (1+\Delta) n^{-1} \sum_{i=1}^{n} K_i^{\mathcal{D}}(\mathbf{x}) + n^{-1} \sum_{i=1}^{n} \left[h^{|\mathbf{d}|+r} f(\mathbf{X}_i)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right) K_i^{\mathcal{D}}(\mathbf{x})\right] \\
=& (1+\Delta)\tilde{f}^{\mathcal{D}}(\mathbf{x}) + h^{|\mathbf{d}|+r} B(\mathbf{x}),
\end{aligned} \tag{7.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

where

$$B(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} f(\mathbf{X}_i)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right) K_i^{\mathcal{D}}(\mathbf{x}).$$

The compactness of $L(\cdot)$ gives us

$$n^{-1} \sum_{i=1}^{n} f(\mathbf{X}_i)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right) = O_p\left(h\right)$$

and we have $\sum_{i=1}^{n} \tilde{p}_i = 1 + \Delta + h^{|\mathbf{d}|+r} O_p\left(h\right)$, which implies $\Delta = O_p\left(h^{|\mathbf{d}|+r+1}\right)$.

Since a higher order derivative $(\mathbf{j})$ implies a higher order for $K^{(\mathbf{j})}(\mathbf{x})$, we have $B(\mathbf{x}) = O_p\left(\tilde{B}(\mathbf{x})\right)$ with

$$\tilde{B}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} f(\mathbf{X}_i)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_i}{h}\right) K_i^{(\mathbf{d})}(\mathbf{x}),$$

where only the highest order derivative is included. We can then bound $\tilde{B}(\mathbf{x})$ as follows.

Without loss of generality, we have

$$\begin{aligned}
\tilde{B}(\mathbf{x}) =& h^{-(|\mathbf{d}|+r)} n^{-1} \sum_{i=1}^{n} f^{-1}(\mathbf{X}_i) K^{(\mathbf{d})}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) L\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \\
=& (-h)^{|\mathbf{d}|} \int f^{-1}(\mathbf{x} - h\mathbf{y}) K^{(\mathbf{d})}(\mathbf{y}) L\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right) d\mathbf{y} + O_p(1), \\
=& O_p\left((-h)^{|\mathbf{d}|} \int K(\mathbf{y}) L^{(\mathbf{d})}\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right) d\mathbf{y}\right) + O_p(1), \qquad (7.4)
\end{aligned}$$

where the last equality follows from the definition of the integral, integration by parts and the compactness of $L$ and $K$.

Combining (7.3) and (7.4) gives us

$$\hat{f}^{\mathcal{D}}(\mathbf{x}|\tilde{p}) = \tilde{f}^{\mathcal{D}}(\mathbf{x}) + (-1)^{|\mathbf{d}|} h^r \int K(\mathbf{y}) L^{(\mathbf{d})}\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right) d\mathbf{y} + B_2(\mathbf{x}), \qquad (7.5)$$

where $B_2(\mathbf{x}) = o_p(h^r)$ uniformly in $\mathbf{x} \in \mathcal{J}$. Particularly, we require $nh^{r-1} \to \infty$ here to claim $h^{|\mathbf{d}|+r} B_1(\mathbf{x}) = o_p(h^r)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

For a subset $S \subset \mathcal{J}$, define

$$B(S, \delta) = \left\{ \mathbf{x} \in \mathcal{J} : \inf_{\mathbf{y} \in S} |\mathbf{x} - \mathbf{y}| \leq \delta \right\}.$$

Consider any point $\mathbf{x} \in \mathcal{J} \setminus B\left(\mathcal{X}_0, Ch^{r/2}\right)$. Let $\mathbf{y}_0 \in \mathcal{X}_0$ be the point such that $|\mathbf{x} - \mathbf{y}_0| = \inf_{\mathbf{y} \in \mathcal{X}_0} |\mathbf{x} - \mathbf{y}|$. The existence of such a $\mathbf{y}_0$ is guaranteed by the fact that $\mathcal{X}_0$ is a closed subset. Since $\partial f^{[\mathbf{d}]}/\partial \mathbf{x}$ is continuous, we must have $\frac{\partial f^{\mathcal{D}}}{\partial \mathbf{x}}(\mathbf{y}_0) = \mathbf{0}$. Hence, the Taylor expansion at $\mathbf{y}_0$ gives $f^{\mathcal{D}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{y}_0)^T \frac{\partial^2 f^{\mathcal{D}}(\mathbf{y}_0)}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x} - \mathbf{y}_0) + O_p\left(h^{3r/2}\right)$. Since $\tilde{f}$ is a consistent estimate of $f$, for sufficiently large $n$ and $C > 0$ we have

$$P\left\{ \tilde{f}^{\mathcal{D}}(\mathbf{x}) > 3h^r, \quad \forall \mathbf{x} \in \mathcal{J} \setminus B\left(\mathcal{X}_0, Ch^{r/2}\right) \right\} > 1 - \frac{\delta}{3}. \tag{7.6}$$

In (7.5), when $\mathbf{x}$ is in the neighbourhood of $\mathbf{x}_0$, $\tilde{f}^{[d]}(\mathbf{x}) = O_p\left(\left(nh^{2|\mathbf{d}|+r}\right)^{-1/2}\right) = O_p(h^r)$ if we require $\left(nh^{2|\mathbf{d}|+3r}\right)^{-1} = O(1)$. Hence, for such an $\mathbf{x}$, we need to have a positive dominating second term in (7.5) in order to claim

$$P\{\tilde{f}^{\mathcal{D}}(\mathbf{x}) + (-1)^{|\mathbf{d}|} h^r \int K(\mathbf{y}) L^{(\mathbf{d})}\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right) d\mathbf{y} > 2h^r \quad \forall \mathbf{x} \in B(\mathcal{X}_0, Ch^{r/2})\} > 1 - \frac{1}{3}\delta. \tag{7.7}$$

Given (7.6), we must also have

$$(-1)^{|\mathbf{d}|} \int K(\mathbf{y}) L^{(\mathbf{d})}\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right) d\mathbf{y} > -1 \quad \forall \mathbf{x} \in \mathcal{J} \setminus B(\mathcal{X}_0, Ch^{r/2}). \tag{7.8}$$

To guarantee (7.7) and (7.8), given both $C$ and $\delta$, we may choose $L$ such that each component of $L^{(\mathbf{d})}$ is a positive or negative constant with sufficiently large absolute value on a sufficiently wide interval containing the origin and returning sufficiently slowly to 0 on either side of the interval where it is constant. Note that our restrictions on $\mathcal{X}_0$ play a role here. A point in the neighbourhood of $\mathcal{X}_0$ will have at least one coordinate falling in the neighbourhood of the corresponding coordinate of the chosen point $\mathbf{x}_0$ in (7.2). This is critical for guaranteeing that $L^{(\mathbf{d})}\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h} + \mathbf{y}\right)$ in (7.7) has at least one non-zero component for any $\mathbf{x} \in B(\mathcal{X}_0, Ch^{r/2})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Lastly, for all $C, \delta > 0$, we have,

$$P\left\{B_2(\mathbf{x}) > -h^r, \quad \forall \mathbf{x} \in \mathcal{J}\right\} > 1 - \frac{\delta}{3}.$$

The derivation in (7.5) with the above set probabilities yields

$$P\left\{\hat{f}^{\mathcal{D}}(\mathbf{x}|\tilde{p}) > 0, \quad \forall \mathbf{x} \in \mathcal{J}\right\} > 1 - \delta, \tag{7.9}$$

which establishes (7.1).

Next, we will show that $D(\tilde{p}) = O_p\left(n^{-1}h^{2|\mathbf{d}|+2r+1}\right)$. This follows from $\Delta = O_p\left(h^{|\mathbf{d}|+r+1}\right)$ and

$$n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} f(\mathbf{X}_j)^{-2} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_j}{h}\right)^2 K_j^2(\mathbf{X}_i) = O_p\left(h\right)$$

for each $i$, which follows from the compactness of $L(\cdot)$ and $K(\cdot)$. Replacing $p_i$ with $\tilde{p}_i$ yields

$$\begin{aligned}
&\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}(\tilde{p}_j - 1/n)K_j(\mathbf{X}_i)\right\}^2 \\
=&\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\Delta + h^{|\mathbf{d}|+r} f(\mathbf{X}_j)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_j}{h}\right)\right\} K_j(\mathbf{X}_i)\right]^2 \\
\leq& 2\sum_{i=1}^{n}\left[\Delta^2\left\{\frac{1}{n}\sum_{j=1}^{n} K_j(\mathbf{X}_i)\right\}^2 + h^{2|\mathbf{d}|+2r}\left\{\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}_j)^{-1} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_j}{h}\right) K_j(\mathbf{X}_i)\right\}^2\right] \\
=&2\Delta^2\sum_{i=1}^{n}\tilde{f}^2(\mathbf{X}_i) + 2nh^{2|\mathbf{d}|+2r}\cdot n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} f(\mathbf{X}_j)^{-2} L\left(\frac{\mathbf{x}_0 - \mathbf{X}_j}{h}\right)^2 K_j^2(\mathbf{X}_i) \\
=&O_p\left(nh^{2|\mathbf{d}|+2r+2}\right) + O_p\left(nh^{2|\mathbf{d}|+2r+1}\right) \\
=&O_p\left(nh^{2|\mathbf{d}|+2r+1}\right).
\end{aligned}$$

Since $\hat{p}$ minimizes $D(p)$ subject to non-negativity of $\hat{f}^{\mathcal{D}}$ on $\mathcal{J}$, this implies that for all sufficiently large $n$,

$$P\left\{D(\hat{p}) \leq D(\tilde{p})\right\} > 1 - \delta.$$

Hence $D(\hat{p}) = O_p\left(n^{-1}h^{2|\mathbf{d}|+2r+1}\right)$. □

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

**Lemma 3.** *For each $\delta > 0$, there exists $C = C(\delta)$ such that, for all sufficiently large $n$,*

$$P\left\{\tilde{f}^{\mathcal{D}}(\mathbf{x}) > \frac{1}{3}f^{[\mathbf{d}]}(\mathbf{x}), \forall \mathbf{x} \text{ such that } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq Ch^{r/2}\right\} > 1 - \delta. \tag{7.10}$$

*Proof.* Proof of Lemma 3.

Let

$$C_n = \sup\left\{M > 0 : \tilde{g}^{\mathcal{D}}(\mathbf{x}) \leq \frac{1}{3}f^{[\mathbf{d}]}(\mathbf{x}), \forall \mathbf{x} \text{ such that } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \leq M\right\}.$$

Then, for any $\mathbf{x}$ such that $\inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq C_n$, we have

$$\tilde{f}^{\mathcal{D}}(\mathbf{x}) > \frac{1}{3}f^{[\mathbf{d}]}(\mathbf{x}), \forall \mathbf{x} \text{ such that } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq C_n.$$

Therefore, we only need to show that $C_n = O_p\left(h^{r/2}\right)$.

Define the stochastic processes $\gamma = \tilde{f}^{\mathcal{D}} - f^{\mathcal{D}}$. For an integer $j \geq 0$, let

$$\mathcal{J}_j = \left\{\mathbf{x} \in \mathcal{C}(\mathcal{J} \setminus \mathcal{X}_0) : jh^{r/2} \leq \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x_0}| \leq (j+1)h^{r/2}\right\},$$

where $\mathcal{C}(S)$ is the closure of a subset $S \subset \mathcal{J}$. Let $J = \max\{j : \mathcal{J}_j \neq \emptyset\}$. Define $\lambda_j(\mathbf{x}) = \sqrt{n}h^{|\mathbf{d}|+\frac{r}{2}}\gamma(\mathbf{x})$ on $\mathcal{J}_j$ for each $0 \leq j \leq J$. We want to show that

$$\sup_j P\left\{\sup_{\mathbf{x} \in \mathcal{J}_j} |\lambda_j(\mathbf{x})| > v\right\} \leq C_1(1+v)^{-2} \tag{7.11}$$

for all $v \in [0, C_2 h^{-r/2}]$, where the constants $C_1, C_2 > 0$ do not depend on $n$.

For this purpose, we need to use a strong approximation result found in Komlós, Major, and Tusnády (1975), generally referred to as the Hungarian embedding. Let $z_i$, $i = 1, \ldots, n$ be a sequence of independent and identically distributed random variables with $E[z_i] = 0$ and $E[z_i] = \sigma_z^2$, and let $S_n = \sum_{i=1}^n z_i$. Then the Hungarian embedding result says that there exists a sequence $T_n$ with the same distribution as $S_n$ and a sequence of a standard Brownian bridge $\mathbb{B}_n(t)$ such that

$$\sup_{0 \leq t \leq 1} \left|T_{\lfloor nt \rfloor} - \sigma_z\sqrt{n}\mathbb{B}_n(t)\right| = O_p(\log n),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

where $\lfloor x \rfloor$ signifies the integer part of $x$.

Again, we only need to prove (7.11) for the highest order derivative $\mathbf{d}$. To economize on notation, we now use $\gamma = \tilde{f}^{(\mathbf{d})} - f^{(\mathbf{d})}$ such that

$$\gamma(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left\{K_i^{(\mathbf{d})}(\mathbf{x}) - f^{(\mathbf{d})}(\mathbf{x})\right\}. \tag{7.12}$$

Note that $\mathrm{Var}(\gamma(\mathbf{x})) = \left(nh^{2|\mathbf{d}|+r}\right)^{-1} C_K f(\mathbf{x}) + o\left((nh^{2|\mathbf{d}|+r})^{-1}\right)$, where $C_K$ is a constant depending only on $K(\cdot)$. Also, the fact that $K(\cdot)$ is compactly supported indicates that the sum in (7.12) essentially has only $\lfloor nC_{\mathbf{x},h} \rfloor$ terms, where $C_{\mathbf{x},h} = P\{\mathbf{X} \in B(\mathbf{x}, Ch)\}$ for some constant $C$. Now applying the Hungarian embedding result to each $\sqrt{n}\lambda_j(\mathbf{x})$ yields

$$\sup_{\mathbf{x}\in\mathcal{J}_j}|\lambda_j(\mathbf{x})| \le \sup_{\mathbf{x}\in\mathcal{J}_j}\left|\sqrt{C_{k,K}f(\mathbf{x})}\mathbb{B}_n(C_{\mathbf{x},h})\right| + O_p\left(\frac{\log n}{\sqrt{n}}\right).$$

Then (7.11) follows from the modulus of continuity for a Brownian bridge; see e.g., Chapter 1 of Csörgő and Révész (1981).

Therefore, letting $w_j \equiv (3h^r)^{-1}\inf_{\mathbf{x}\in\mathcal{J}_j} f^{\mathcal{D}}(\mathbf{x})$ and $v_j \equiv \min\left(w_j, C_2 h^{-r/2}\right)$, we have

$$\begin{aligned} P\left\{|\gamma(\mathbf{t})| > \frac{1}{3}f^{\mathcal{D}}(\mathbf{t}) \text{ for some } \mathbf{t} \in \cup_{j=j_0}^{J}\mathcal{J}_j\right\} &\le \sum_{j=j_0}^{J} P\left\{|\gamma(\mathbf{t})| > \frac{1}{3}f^{[\mathbf{d}]}(\mathbf{t}) \text{ for some } \mathbf{t}\in\mathcal{J}_j\right\} \\ &\le \sum_{j=j_0}^{J} P\left\{\sup_{\mathbf{x}\in\mathcal{J}_j}|\lambda_j(\mathbf{x})| > v\right\} \\ &\le \sum_{j=j_0}^{J} C_1(1+v_j)^{-2}. \end{aligned} \tag{7.13}$$

For any $\mathbf{x} \in \mathcal{J}_j$, let $\mathbf{x}_0 = \arg\min_{\mathbf{y}_0\in\mathcal{X}_0}|\mathbf{x} - y_0| \in \vartheta\mathcal{X}_0 \setminus \vartheta\mathcal{J}$ and $e = (\mathbf{x} - \mathbf{x}_0)/|\mathbf{x} - \mathbf{x}_0|$ be a unit vector. Then there exists a constant $0 \le u \le 1$ such that $\mathbf{x} = \mathbf{x}_0 + (j+u)h^{r/2}e$. The conditions imposed on $f$ in the theorem imply that

$$f^{\mathcal{D}}(\mathbf{x}_0 + (j+u)h^{r/2}e) = \frac{1}{2}(j+u)^2 h e^T \frac{\partial^2 f^{\mathcal{D}}}{\partial\mathbf{x}\partial\mathbf{x}^T}(\mathbf{x}_0)e + O_p\left(h^{3r/2}\right).$$

Hence, $w_j \ge C_3 j^2$ and $v_j \ge \min(C_3 j^2, C_4 h^{-r/2})$, where $C_3, C_4 > 0$ do not depend on $n$. Since

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

$\sum_{j=j_0}^{J}\left(1+C_3j^2\right)^{-2}\to 0$ as $j_0\to\infty$ and $\sum_{j=j_0}^{J}\left(1+C_4h^{-r/2}\right)^{-2}=O_p\left(Jh^r\right)=O_p\left(h^{r/2}\right)$, it now follows from (7.13) that

$$\lim_{C\to\infty}\limsup_{n\to\infty}P\left\{\tilde{f}^{\mathcal{D}}(\mathbf{x})>\frac{1}{3}f^{\mathcal{D}}(\mathbf{x}),\ \text{for some } \mathbf{x} \text{ such that } \inf_{\mathbf{x}_0\in\mathcal{X}_0}|\mathbf{x}-\mathbf{x}_0|\geq Ch^{r/2}\right\}=0.$$

Applying the results to $\gamma$ we obtain (7.10). □

*Proof.* Proof of Theorem 2.

We prove each part of Theorem 2 in turn.

1. For $n$ sufficiently large the uniform weights will automatically satisfy the constraints. In this case the constrained estimator will be equivalent to the unconstrained estimator.

2. By the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\hat{f}^{(\mathbf{j})}(\mathbf{x}|\hat{p})-\tilde{f}^{(\mathbf{j})}(\mathbf{x})\right|\leq&\left|n^{-1}\sum_{i=1}^{n}(n\hat{p}_i-1)^2\right|^{1/2}\left|n^{-1}\sum_{i=1}^{n}K_i^{(\mathbf{j})}(\mathbf{x})^2\right|^{1/2}\\ \leq&O_p\left(n^{-1}h^{2|\mathbf{d}|+2r+1}\right)^{1/2}O_p\left(h^{-(r+2|\mathbf{j}|)}\right)^{1/2}\\ =&O_p\left(h^{j_{\mathbf{j}}}\right),\end{aligned}\tag{7.14}$$

   where $j_{\mathbf{j}}=|\mathbf{d}|+\frac{r+1}{2}-|\mathbf{j}|$. The last inequality follows from Lemma 2 and the fact that $K(\cdot)$ is compactly supported and $\sum_{i=1}^{n}(n\hat{p}_i-1)^2\asymp D(\hat{p})$. Note that (7.14) indicates that when $\mathbf{j}$ is of lower order than $\mathbf{d}$,

$$\sup_{\mathbf{x}\in\mathcal{J}}\left|\hat{f}^{(\mathbf{j})}(\mathbf{x}|\hat{p})-\tilde{f}^{(\mathbf{j})}(\mathbf{x})\right|\leq\sup_{\mathbf{x}\in\mathcal{J}}\left|\hat{f}^{(\mathbf{d})}(\mathbf{x}|\hat{p})-\tilde{f}^{(\mathbf{d})}(\mathbf{x})\right|=O_p\left(h^{j_{\mathbf{d}}}\right)=O_p\left(h^{\frac{r+1}{2}}\right).\tag{7.15}$$

   We used Lemma 2 to establish the final order of the above difference. In particular, $\sup_{\mathbf{x}\in\mathcal{J}}|\hat{f}(\mathbf{x}|\hat{p})-\tilde{f}(\mathbf{x})|=O_p\left(h^{|\mathbf{d}|+\frac{r+1}{2}}\right)$. This proves part (ii) of Theorem 2.

3. Since $K(\cdot)$ is compactly supported, there exists a constant $M$ such that

$$K_i(\mathbf{x})=0\quad\text{if } |\mathbf{x}-\mathbf{X}_i|\geq Mh.\tag{7.16}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

Let $Z_2$ be a random variable such that

$$Z_2 h^{\frac{r+1}{4}} = \inf\left\{ z > 0 : \hat{p}_i = n^{-1}(1+\Theta_i), \text{ such that } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x}_0 - \mathbf{X}_i| \geq z \right\},$$

where $\Theta_i$ is a random variable such that $q = \sum_{j=1}^{n} (\hat{p}_j - 1/n) K_j(\mathbf{x}_i)$ does not depend on $i$ (see Lemma 1). This infinitum is well defined given Lemma 1. Given (7.16), any $\mathbf{x}$ such that $\inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq Mh + Z_2 h^{\frac{r+1}{4}}$ implies $\hat{p}_i = n^{-1}(1+\Theta_i)$, $\forall i$ such that $K_i(\mathbf{x}) \neq 0$. Let $\Theta = \sup_{\{i: K_i(\mathbf{x}) \neq 0\}} |\Theta_i|$. Then Lemma 3 yields

$$P\Bigg\{ \hat{f}^{\mathcal{D}}(\mathbf{x}|\hat{p}) > \frac{1}{3}(1+\Theta) f^{\mathcal{D}}(\mathbf{x}) > 0, \forall \mathbf{x} \text{ such that } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq \max\left( Ch^{r/2}, Mh + Z_2 h^{\frac{r+1}{4}} \right) \Bigg\} > 1 - \delta. \tag{7.17}$$

Using (7.15) along with the convergence rate of $\tilde{f}^{\mathcal{D}}$ gives us $\hat{f}^{\mathcal{D}}(\mathbf{x}|\hat{p}) = f^{\mathcal{D}}(\mathbf{x}) + O_p\left(h^{\frac{r+1}{2}}\right)$ for any $\mathbf{x} \in \mathcal{J}$. In particular, let $\mathbf{x} = \mathbf{x}_0 + \left(Ch^{r/2} + Z_2 h^{\frac{r+1}{4}}\right) e \in \mathcal{J} \setminus \mathcal{X}_0$, where $\mathbf{x}_0 = \arg\min_{\mathbf{y}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{y}_0| \in \vartheta\mathcal{X}_0 \setminus \vartheta\mathcal{J}$ and $\mathbf{e} = (\mathbf{x} - \mathbf{x}_0)/||\mathbf{x} - \mathbf{x}_0||$ is a vector of unit length. Using $Z(h) = Ch^{r/2} + Z_2 h^{\frac{r+1}{4}}$ and a Taylor expansion gives

$$\begin{aligned} \hat{f}^{\mathcal{D}}\left(\mathbf{x}_0 + Z(h)\mathbf{e}|\hat{p}\right) =& \frac{1}{2} Z(h)^2 \mathbf{e}^T \frac{\partial^2 f^{\mathcal{D}}}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0)\mathbf{e} + O_p\left(h^{\frac{r+1}{2}}\right) + o_p\left(Z(h)^2\right) \\ =& O_p\left(h^r\right) + O_p\left(Z_2^2 h^{\frac{r+1}{2}}\right) + O_p\left(h^{\frac{r+1}{2}}\right). \end{aligned}$$

Clearly, we should have $Z_2 = O_p(1)$ in order to guarantee (7.17). Hence, $\hat{p}_i$ is equal to $n^{-1}(1+\Theta_i)$ for some random variable $\Theta_i$ and for all $i$ such that $|\mathbf{X}_i - \mathbf{x}_0| \geq Z_2 h$, where $Z_2 = O_p(1)$. Now we can define a random variable $Z_1$ such that

$$Z_1 h^{\frac{r+1}{4}} = \inf\Bigg\{ z \geq 0 : \forall \mathbf{x} \in \mathcal{J} \text{ for which } \inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq z, \; K_i(\mathbf{x}) = 0 \text{ whenever } \hat{p}_i \neq n^{-1}(1+\Theta_i) \Bigg\}.$$

Clearly $Z_1 = O_p(1)$ and $1 - \Theta \leq \hat{f}(\mathbf{x}|\hat{p})/\tilde{f}(\mathbf{x}) \leq 1 + \Theta$ for all values $\mathbf{x} \in \mathcal{J}$ such that $\inf_{\mathbf{x}_0 \in \mathcal{X}_0} |\mathbf{x} - \mathbf{x}_0| \geq Z_1 h^{\frac{r+1}{4}}$. This establishes the second part of Theorem 2 (iii).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

The order of $\Theta$ is derived as follows. From Lemma 1 we have that the weights $\hat{p}_i = n^{-1}(1+\Theta_i)$ are such that $q = \sum_{j=1}^{n}(\hat{p}_j - 1/n)K_j(\mathbf{x}_i)$ does not depend on $i$ for indices $i$ that lie a distance $O\left(h^{r/2}\right)$ from $\mathbf{x}_0$. Then $\Theta_i$ of such indices $i$ will have the same order. Let there be $N$ such indices, and let $\mathcal{A}$ be the set of remaining indices. Then $N = O_p(n)$, $N_1 \equiv |\mathcal{A}| = O_p\left(nh^{r/2}\right)$ and $\sum_{i=1}^{n}\hat{p}_i = 1$ deliver

$$\begin{aligned}
|\Theta| &\leq N^{-1}\sum_{i\in\mathcal{A}}|n\hat{p}_i - 1| \\
&\leq N^{-1}\left\{\sum_{i\in\mathcal{A}}1\right\}^{1/2}\left\{\sum_{i\in\mathcal{A}}(n\hat{p}_i-1)^2\right\}^{1/2} \\
&\leq N^{-1}N_1^{1/2}\left\{\sum_{i=1}^{n}(n\hat{p}_i-1)^2\right\}^{1/2} \\
&= O_p\left(h^{|\mathbf{d}|+r+1}\right).
\end{aligned}$$

This establishes the first part of Theorem 2 (iii). □

The proof of Theorem 4 is divided into two steps: the first step uses the Karush-Kuhn-Tucker conditions to obtain a closed form for $D(\hat{p})$ and the second step applies standard asymptotic results for multivariate kernel density estimation and their derivatives to obtain the asymptotic distribution of $D(\hat{p})$. Some tedious details are skipped in the second step for the sake of brevity.

*Proof.* Proof of Theorem 4.

The Karush-Kuhn-Tucker conditions say that there exist constants $\lambda_1$ and $\lambda_{2j}, j = 1, \ldots, N$ such that

$$2(\hat{p}_i - n^{-1}) + \lambda_1 - \sum_{j=1}^{N}\lambda_{2j}\psi_i(\mathbf{x}_j) = 0, i = 1, \ldots, n, \tag{7.18}$$

$$\lambda_{2j}\sum_{i=1}^{n}\hat{p}_i\psi_i(\mathbf{x}_j) = 0, j = 1, \ldots, N, \tag{7.19}$$

$$\sum_{i=1}\hat{p}_i = 1, \tag{7.20}$$

$$\sum_{i=1}^{n}\hat{p}_i\psi_i(\mathbf{x}_j) \geq 0, \lambda_{2j} \geq 0, j = 1, \ldots, N, \tag{7.21}$$

where (7.18) is the stationary condition, (7.19) is the complementary slackness condition, and (7.20)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

and (7.21) are feasibility conditions.

Let $\bar{\psi}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\psi_i(\mathbf{x})$. Solving (7.18) for $\hat{p}_i$ we have

$$\hat{p}_i = n^{-1} - \frac{\lambda_1}{2} + \frac{1}{2}\sum_{j=1}^{N}\lambda_{2j}\psi_i(\mathbf{x}_j). \tag{7.22}$$

Summing up (7.22) and using (7.20) we can obtain $\lambda_1 = \sum_{j=1}^{N}\lambda_{2j}\bar{\psi}(\mathbf{x}_j)$. Substitution of this into (7.22) yields

$$\hat{p}_i = \frac{1}{n} - \frac{1}{2}\sum_{j=1}^{N}\lambda_{2j}(\psi_i(\mathbf{x}_j) - \bar{\psi}(\mathbf{x}_j)). \tag{7.23}$$

In light of (7.19), we can permute $(\lambda_{21},\ldots,\lambda_{2N})$ to obtain $(\delta_1,\ldots,\delta_N)$ such that $\delta_j > 0$ for $j = 1,\ldots,M$ and $\delta_j = 0$ for $j = M+1,\ldots,N$, where $M$ is the number of nonzero $\lambda_{2j}$'s. Permute $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ accordingly to obtain $(\mathbf{x}_1^*,\ldots,\mathbf{x}_N^*)$, and (7.23) becomes

$$\hat{p}_i = \frac{1}{n} - \frac{1}{2}\sum_{j=1}^{M}\delta_j(\psi_i(\mathbf{x}_j^*) - \bar{\psi}(\mathbf{x}_j^*)). \tag{7.24}$$

For $j = 1,\ldots,M$, (7.19) implies $\sum_{i=1}^{n}\hat{p}_i\psi_i(\mathbf{x}_j^*) = 0$, which, after plugging in (7.24), yields

$$\sum_{k=1}^{M}\delta_k\sigma_{jk}^{(n)} = \frac{2}{n}\bar{\psi}(\mathbf{x}_j^*), j = 1,\ldots,M, \tag{7.25}$$

where $\sigma_{jk}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\mathbf{x}_j^*)\psi_i(\mathbf{x}_k^*) - \bar{\psi}(\mathbf{x}_j^*)\bar{\psi}(\mathbf{x}_k^*)$.

Let $\Sigma_n$ be the matrix with element $(j,k)$ equal to $\sigma_{jk}^{(n)}$,$\boldsymbol{\delta} = (\delta_1,\ldots,\delta_M)^T$, $\mathbf{x}^* = (\mathbf{x}_1^*,\ldots,\mathbf{x}_M^*)^T$, and for a function $f$, denote $\mathbf{f}(\mathbf{x}^*) = (f(\mathbf{x}_1^*),\ldots,f(\mathbf{x}_M^*))^T$. Then (7.25) gives $\boldsymbol{\delta} = \frac{2}{n}\Sigma_n^{-1}\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$. Combining this with (7.24) gives

$$D(\hat{p}) = \sum_{i=1}^{n}\left(\frac{1}{n} - \hat{p}_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left[\bar{\boldsymbol{\psi}}(\mathbf{x}^*)^T\Sigma_n^{-1}(\boldsymbol{\psi}_i(\mathbf{x}^*) - \bar{\boldsymbol{\psi}}(\mathbf{x}^*))\right]^2 \equiv \frac{1}{n^2}\sum_{i=1}^{n}T_i^2, \tag{7.26}$$

where $T_i = \bar{\boldsymbol{\psi}}(\mathbf{x}^*)^T\Sigma_n^{-1}(\boldsymbol{\psi}_i(\mathbf{x}^*) - \bar{\boldsymbol{\psi}}(\mathbf{x}^*))$.

We now investigate the asymptotic behaviour of the $T_i$'s. For the sake of brevity we only provide a sketch of the proof here. The details we skip are essentially straightforward and tedious extensions of the existing local asymptotic results in Singh (1987) and Chacón, Duong, and Wand (2011). We

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

will use the empirical process notation for part of our argument. Let $E_n$ denote the empirical mean such that $E_n[k(\mathbf{X})] = \frac{1}{n}\sum_{i=1}^n k(\mathbf{X}_i)$ for any function $k$, where $\mathbf{X}$ is the random vector following the distribution of $\mathbf{X}_i$.

First note that under Assumptions A1(i)-(iv), $\bar{\psi}(\mathbf{x}) = E_n[K_X^{\mathcal{D}}(\mathbf{x})]$ is a consistent estimate of $f^{\mathcal{D}}(\mathbf{x})$ with

$$\mathrm{Cov}(\bar{\psi}(\mathbf{t}_1), \bar{\psi}(\mathbf{t}_2)) = \frac{1}{nh^{2|\mathbf{d}|+r}}\left[\int K^{(\mathbf{d})}(\mathbf{y})K^{(\mathbf{d})}\left(\mathbf{y} + \frac{\mathbf{t}_2 - \mathbf{t}_1}{h}\right)d\mathbf{y} + o(1)\right], \tag{7.27}$$

and $\sigma_{jk}^{(n)} = E_n[\psi_X(\mathbf{x}_j^*)\psi_X(\mathbf{x}_k^*)] - E_n[\psi_X(\mathbf{x}_j^*)]E_n[\psi_X(\mathbf{x}_k^*)]$ is a consistent estimate of $\mathrm{Cov}(\psi_X(\mathbf{x}_j^*), \psi_X(\mathbf{x}_k^*))$. Under Assumption A2(ii), when $j \neq k$, $\mathbf{x}_j^*$ and $\mathbf{x}_k^*$ can be considered as sufficiently far away from each other to claim $\mathrm{Cov}(\psi_X(\mathbf{x}_j^*), \psi_X(\mathbf{x}_k^*)) = 0$. Whence, $\mathrm{Var}(\psi_X(\mathbf{x}_j^*)) = h^{-2|\mathbf{d}|+r}\sigma^2_{K^{(\mathbf{d})}}$ and $\Sigma_n = h^{-2|\mathbf{d}|+r}(\sigma^2_{K^{(\mathbf{d})}} I_M + o_p(1))$, where $I_M$ is the identity matrix of dimension $M$. Together with the consistency of $\bar{\psi}(\mathbf{x})$, we thus have

$$\bar{\boldsymbol{\psi}}(\mathbf{x}^*) = \mathbf{f}^{\mathcal{D}}(\mathbf{x}^*) + o_p(1) \text{ and } \Sigma_n^{-1} = h^{2|\mathbf{d}|+r}(\sigma^{-2}_{K^{(\mathbf{d})}} I_M + o_p(1)).$$

By Slutsky's Theorem, the asymptotic distribution of $T_i$ is equivalent to that of the random variable

$$W_i = \sigma^{-2}_{K^{(\mathbf{d})}} h^{2|\mathbf{d}|+r}\left[\mathbf{f}^{\mathcal{D}}(\mathbf{x}^*)\right]^T\left(\boldsymbol{\psi}_i(\mathbf{x}^*) - \mathbf{f}^{\mathcal{D}}(\mathbf{x}^*)\right). \tag{7.28}$$

Thus, the asymptotic distribution of $D(\hat{p})$ is equivalent to that of $\frac{1}{n^2}\sum_{i=1}^n W_i^2$. Under Assumptions A1(i)-(iv), $h^{|\mathbf{d}|+r/2}(\boldsymbol{\psi}_i(\mathbf{x}^*) - \mathbf{f}^{\mathcal{D}}(\mathbf{x}^*))$ are independent and identically distributed random vectors that are asymptotically normal with zero mean and covariance $\sigma^2_{K^{(\mathbf{d})}} I_M$. Hence $W_i$ is asymptotically equivalent to $\left\{\sigma^{-1}_{K^{(\mathbf{d})}} h^{|\mathbf{d}|+r/2}\sum_{j=1}^M f^{\mathcal{D}}(\mathbf{x}_j^*)\right\} Z_i$, where $Z_i$'s are i.i.d. standard normal random variables. Thus

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

Note that $M$ may increase with $n$, so Assumption A2(i) ensures that $\frac{1}{n}\sum_{j=1}^M f^{\mathcal{D}}(\mathbf{x}_j^*) = O(1)$. □

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

## 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-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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.

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.

Dümbgen, L., and K. Rufibach. 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

———. 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.

———. 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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. 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0112

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.

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.