Enforcing Shape Constraints on a Probability Density Estimate Using an Additive Adjustment Curve

(Short title: Density Estimation by Adjustment Curves)

Mark A. Wolters
Shanghai Center for Mathematical Sciences
Fudan University
Shanghai, China
mwolters@fudan.edu.cn

W. John Braun
Irving K. Barber School of Arts and Sciences
Univerisity of British Columbia
Kelowna, British Columbia, Canada
john.braun@ubc.ca

Key Words: kernel density estimation; quadratic programming; shape restriction, star unimodality, unimodal density estimation.

ABSTRACT

A method is proposed for shape-constrained density estimation under a variety of constraints, including but not limited to unimodality, monotonicity, symmetry, and constraints on the number of inflection points of the density or its derivative. The method involves computing an adjustment curve that is used to bring a pre-existing pilot estimate into conformance with the specified shape restrictions. The pilot estimate may be obtained using any preferred estimator, and the optimal adjustment can be computed using fast, readily-available quadratic programming routines. This makes the proposed procedure generic and easy to implement.

1 Introduction

Probability density estimation is a fundamental task in data analysis. It may be done on its own to aid understanding of a data set, or it can be used as a component of more sophisticated statistical procedures. In either situation one is faced with a choice between parametric and

nonparametric approaches to density estimation—each approach with its advantages and disadvantages.

Shape-constrained nonparametric density estimation occupies a middle ground between the parametric and nonparametric options. When some shape characteristics of the density to be estimated may be safely assumed, a shape-restricted nonparametric estimate offers the potential for better statistical performance than purely nonparametric methods along with better data-adaptivity than parametric alternatives. As an added practical benefit, the constrained estimate will exhibit the desired shape characteristics for every sample, not just on average or asymptotically. These advantages are particularly important for small to moderate sample sizes, where sampling variation is more likely to produce undesirable qualitative features in unconstrained nonparametric estimates.

The motivation for using shape constraints can be illustrated by considering different situations that would make an analyst inclined to use them. The analyst might be convinced that a quantity of interest is an aggregate of many small influences, for example, and on this basis be considering a normal model for the data. If there is concern about the presence of skew, however, a safer choice would be to use a nonparametric estimator with a constraint that ensures unimodality and limits the number of inflections in the estimate (such as the proposed *bell shape* constraint introduced later in this article). Similarly, a user with nonnegative data such as failure times might believe that the unknown density producing the observations is 2-monotone (i.e., convex and decreasing; see Balabdaoui and Wellner (2007)), but might be reluctant to commit to the memoryless property inherent in an exponential model. In this case a nonparametric estimate with the 2-monotonicity constraint directly enforced (such as the one shown later, in Figure 1) would be more suitable. Finally, an investigator may have in hand a nonparametric density estimate to be presented to non-statisticians, but one that includes potentially distracting spurious modes in the tails. A constraint requiring the estimate to have monotonic tails would rectify the problem with minimal impact (and possible improvement) on the statistical validity of the estimate.

At present the barriers to wide adoption of shape-constrained density estimation are threefold. There does not exist an estimation framework that i) can handle a variety of different constraints, ii) acts as an extension of familiar density estimators, and iii) is constructed

in a manner that yields easy optimization problems. This article describes a constraint-handling approach intended to eliminate these barriers. The new method has the following distinguishing features:

1. It operates as an adjustment to an existing density estimate (called the *pilot* estimate), so it can be used as an adjunct to various familiar estimators.

2. It can handle a number of important constraints, including but not limited to monotonicity, convexity, unimodality, symmetry, and a proposed *bell shape* constraint based on the inflection points of the density. It can also incorporate roughness penalties based on the integral of the square of the second derivative of the density.

3. All of the aforementioned constraints and penalties can be enforced using fast and readily-available quadratic programming (QP) routines.

4. While the method is best suited to univariate problems, its mathematical structure can be extended to higher dimensions. A bivariate case is demonstrated in the Examples section to follow.

The core of the method is a general construction of an adjustment to a pilot estimate, that allows various problem instances to be expressed as quadratic programs and solved using the same optimization framework.

The remainder of this section reviews existing methods for shape constrained nonparametric density estimation, and provides an overview of the proposed methodology. The new method is then described in detail, including a variety of alternatives in the construction of the adjustment and the set up of the optimization. The third section discusses how these implementation alternatives impact performance and ease-of-use. It is shown how user-facing functions can be created to make shape constrained estimation just as easy as standard nonparametric density estimation. Following that, the method's performance is demonstrated using both examples and simulations.

For concreteness, we will describe the method for the case where the pilot estimator is a kernel density estimator (KDE; see, e.g., Wand and Jones (1995); Sheather (2004)). The supplementary material accompanying the article includes a MATLAB (The Mathworks,

Inc. 2007) implementation of the method. It includes both a low-level function allowing detailed control of the method with arbitrary pilot estimators, as well as all-in-one functions for obtaining various shape-restricted kernel density estimates in one function call.

1.1 Shape-constrained density estimation

One approach to the problem of finding shape-restricted estimates is to find the nonparametric maximum likelihood estimator (NPMLE) of the density under the shape constraint of interest. This was done by Grenander Grenander (1956) for the constraint of monotonicity. Later research attempted to extend the Grenander estimator to unimodal densities in general. The common premise was to combine a nondecreasing Grenander estimate to the left of the mode with a nonincreasing one to the right. When the location of the mode is known, this estimator is the NPMLE; otherwise the NPMLE does not exist. A variety of solutions to the problem of mode location in this setting have been proposed (Wegman 1972; Bickel and Fan 1996; Birgé 1997; Reboul 2005). Like the Grenander estimator itself, all of these methods produce an estimate that is a step function.

Important recent work (Dümbgen and Rufibach 2009; Cule et al. 2010) has developed the NPMLE for a different constraint, log-concavity. Both the properties of this estimator and an algorithm for computing it have been developed in arbitrary dimension. In the univariate case, this estimator can lead to estimates with a cuspate appearance (the logarithm of the estimate is piecewise linear), making it a less attractive choice when a high degree of smoothness is desired. It does not depend on any bandwidth or smoothing parameters, however, which is particularly advantageous in higher dimensions.

Estimation approaches not based on maximum likelihood have also been proposed to construct smooth density estimates under the unimodality constraint or other simple constraints. Fougères (1997) used a monotone rearrangement to transform a multimodal density estimate into a unimodal one, though under the restrictive assumption that the final mode location is known. Cheng et al. (1999) start with a unimodal template density and then iteratively apply monotone transformations (possibly with intermediate smoothing steps) to construct a more suitable unimodal estimate. The method of rearrangements has also been used by Birke (2009) to find monotone, convex, or log-concave estimates (see references

therein for additional alternatives with these constraints).

Another branch of recent research focuses on methods that can handle shape constraints using standard nonparametric estimators that are more familiar to users. Data sharpening (shifting the data points) is one such approach that has been used for accommodating constraints in both density estimation and regression (Braun and Hall 2001; Hall and Kang 2005). Finding the optimally-sharpened data values is challenging, and alternative optimization algorithms have been proposed (Wolters 2012a,c) to improve the performance of data sharpening and expand the number of constraints that can be handled with it. Du et al. (2013), expanding on the work of Hall and Huang (2001), used weights on the data points to enforce a broad class of derivative constraints on kernel regression estimates; this strategy can also be applied to density estimation. Because techniques like shifting or re-weighting data points are so general, they can work with any estimator, and can in principle handle arbitrary constraints or high-dimensional problems. This approach to constraint handling is investigated thoroughly by Wolters (2012b) . The contribution of the present work is another means of shape adjustment which, like data sharpening or re-weighting, can be applied to any density estimator.

Shape constraints are common in regression as well as density estimation. While the method described below is potentially extensible to regression problems, this extension is not pursued here. The reader is referred to the literature (Meyer 2008; Henderson and Parmeter 2009, 2015) for recent work summarizing the developments in shape-constrained regression.

1.2 The proposed approach

Let the pilot density estimate, which does not necessarily satisfy the desired constraints, be $\hat{f}^{\circ}$. A simple option for modifying the shape of $\hat{f}^{\circ}$ is to add to it a function, $\Psi(x)$, that can annihilate any of its unwanted features or contribute any desired features that are not present. It is proposed to let $\Psi(x)$ be a linear combination of k density functions ψ_i,

$i = 1, \ldots, k$. Then the shape-adjusted estimator, $\hat{f}_{\boldsymbol{a}}$, is

$$\begin{aligned} \hat{f}_{\boldsymbol{a}}(x) &= \hat{f}^{\circ}(x) + \Psi(x) \\ &= \hat{f}^{\circ}(x) + a_1\psi_1(x) + a_2\psi_2(x) + \cdots + a_k\psi_k(x) \\ &= \hat{f}^{\circ}(x) + \boldsymbol{a}^T\boldsymbol{\psi}(x), \end{aligned} \tag{1}$$

where $\boldsymbol{a} = [a_1 \cdots a_k]^T$ are the coefficients of the combination, and $\boldsymbol{\psi}(x) = [\psi_1(x) \cdots \psi_k(x)]^T$. The coefficients can be chosen to minimize the amount of adjustment made to $\hat{f}^{\circ}$, subject to the required shape constraints on $\hat{f}_{\boldsymbol{a}}$. The functions $\Psi(x)$ and $\{\psi_i(x)\}$ will be referred to as the *adjustment curve* and the *adjustment densities*, respectively.

The number of adjustment densities (k) and the specific densities chosen for each ψ_i determine which $\Psi(x)$ curves are possible; consequently, these choices must be made appropriately for the pilot estimator and shape constraints used in a particular problem. Figure 1 provides an illustration of how adjustment densities are set up appropriately in two different scenarios. The first example in the figure is based on a draw of size $n = 20$ from a standard normal distribution. A kernel density estimate with Gaussian kernel is used as the pilot density. The pilot estimate has three modes. In this case the ψ_i are chosen to be normal densities positioned with uniform spacing over the range of the pilot estimate. This arrangement of adjustment densities is used to construct an adjustment curve that renders the final estimate unimodal.

The second example in the figure uses as its pilot an edge frequency polygon (Jones et al. 1998). The data are a random sample of size $n = 50$ from an exponential distribution. Because the pilot estimate in this case is piecewise linear, it is more appropriate to use triangular adjustment densities, arranged so that $\Psi(x)$ is also piecewise linear with slope changes at the same points as the pilot estimate. In this second example two constraints are enforced: i) the estimate must be zero for negative abscissa values, and ii) the density must be 2-monotone on the positive half-line. The optimization methodology used to obtain this second estimate is identical to that used for the first example.

The Gaussian KDE case is prototypical, so this estimator is considered for the remainder of the article. The next section provides more detail on how the problem may be set up and solved for this estimator with a variety of constraints.

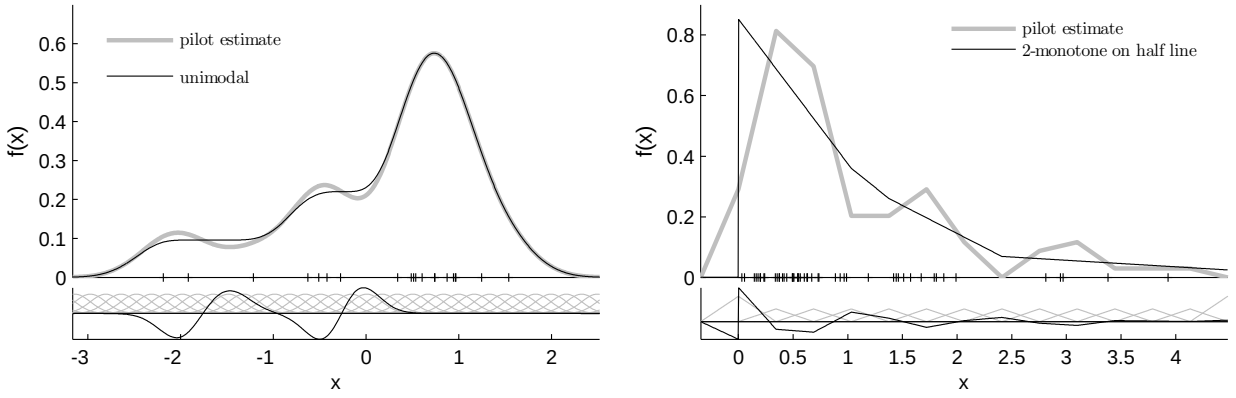

Figure 1: Illustration of the adjustment curve method. In the left panel, a kernel density estimate is rendered unimodal using normal adjustment densities. In the right panel, an edge frequency polygon is rendered 2-monotone on the positive half-line, and zero elsewhere, using triangular adjustment densities. The adjustment densities and the adjustment curve are shown beneath each plot, with vertical scaling chosen to enhance visualization.

1.3 Remark on large-sample properties

Consider the case where the sample size is n and the pilot estimator $\widehat{f}_n(x)$ is a KDE with kernel $K(\cdot)$ and bandwidth h_n. The kernel is assumed to be a bounded symmetric continuous probability density function having bounded variation (an assumption stronger than necessary, but satisfied by the kernels we are considering). Suppose the density being estimated, $f(x)$, satisfies the chosen constraints and has a uniformly continuous rth derivative on $(-\infty, \infty)$. Suppose also that $h_n \to 0$ and $nh_n^{2r+1}/\log(1/h_n) \to \infty$ as $n \to \infty$.

Silverman (1978) showed that under these conditions a KDE with an appropriately-chosen bandwidth exhibits strong uniform consistency as an estimator of not only the density itself, but also of its derivatives. That is,

$$\lim_{n\to\infty} \sup_x |\widehat{f}_n^{(r)}(x) - f^{(r)}(x)| = 0$$

with probability 1. Silverman's theorem implies that there exists N such that for any $\varepsilon > 0$,

$$\sup_x |\widehat{f}_n^{(r)}(x) - f^{(r)}(x)| < \varepsilon, \qquad \forall n > N$$

with probability 1.

Most of the shape constraints to be considered are expressed as constraints on the derivatives of $\widehat{f}_n(x)$ over intervals. The above result indicates that for n sufficiently large, the

magnitude of any violations of such constraints will be bounded by the arbitrarily small constant ε.

Consider the unimodality case, with mode m, as an example. In this case $f'(x) \geq 0$ for all $x < m$ and $f'(x) \leq 0$ for all $x > m$. Let $\varepsilon > 0$ be given. Then there exist N_1 and N_2 such that (with probability 1)

$$\widehat{f}_n'(x) \geq -\varepsilon, \qquad \forall x < m$$

for all $n > N_1$ and

$$\widehat{f}_n'(x) \leq \varepsilon, \qquad \forall x > m$$

for all $n > N_2$. Therefore, $\widehat{f}_n(x)$ will be unimodal up to the tolerance ε for all $n > \max(N_1, N_2)$, with probability 1.

Informally, we expect the pilot estimate to need less and less adjustment as the sample size grows—assuming the chosen constraints are in fact features of the density being estimated, and the pilot estimator has sufficiently good asymptotic properties. This ability to "borrow" the asymptotics of the pilot estimator is an advantage of the proposed approach: asymptotically, the adjustment can do no harm, but in finite samples it can bring improvements. The benefits of including shape restrictions in small samples are illustrated in the simulation at the end of this paper, and have also been observed elsewhere (Braun and Hall 2001).

2 Details of the new method

The new method will now be described in greater depth. The first part of this section describes how the optimal estimate in the general formulation (1) can be defined as the solution to a quadratic program. The second part describes a number of shape restrictions that can be enforced through linear inequality constraints in $\boldsymbol{a}$, as required by the QP framework. The third part addresses the important question of how to lay out the adjustment densities to ensure the existence of feasible and useful solutions.

The information in this section is oriented toward the reader who wishes to understand the method in detail, for example to adapt the technique to situations beyond those described in this article. When working with specific problem instances—specific combinations of pilot

estimator and constraint—many of the aspects discussed here can be handled internally in user-facing computer code. This point is discussed further in a later section.

2.1 Quadratic objective, linear constraints

Let the optimal vector of adjustment coefficients be $\boldsymbol{a}^*$, and take it to be the minimizer of a quadratic form,

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}}{\operatorname{argmin}} \ \boldsymbol{a}^T \mathbf{H} \boldsymbol{a} + \mathbf{v}^T \boldsymbol{a}, \tag{2}$$

with the values of $\mathbf{H}$ and $\mathbf{v}$ to be specified. The minimization is subject to three groups of constraints:

$$\sum_{i=1}^{k} a_i = 0 \tag{3}$$

$$\hat{f}^{\circ}(g_l) + \boldsymbol{a}^T \boldsymbol{\psi}(g_l) \geq 0, \quad l = 1 \ldots G \tag{4}$$

$$\mathbf{A}\boldsymbol{a} \leq \mathbf{b}. \tag{5}$$

Expressions (2) through (5) define a quadratic programming problem, since the objective function is a quadratic form in $\boldsymbol{a}$ and the constraints are linear in $\boldsymbol{a}$. QP problems may be routinely solved in most statistical computing environments, with rapid computation of solutions even when $\mathbf{H}$ and $\mathbf{A}$ are of large dimensions. Further, when the matrix $\mathbf{H}$ is positive definite (as it is using either of the objective functions discussed subsequently), the QP solver returns the globally optimal solution. For more on quadratic programming, see, e.g., Antoniou and Lu (2007) or Nocedal and Wright (1999).

Consider the constraints first. Eq (3) defines a sum constraint on the a_i to ensure that $\Psi(x)$ integrates to zero. Constraint (4) is a system of G inequalities that enforce non-negativity on $\hat{f}_{\boldsymbol{a}}$ pointwise at a vector of abscissa values $\mathbf{g} = [g_1 \cdots g_G]^T$. For practical purposes this will achieve uniform nonnegativity if $\mathbf{g}$ is taken to be an evenly-spaced grid of values extending beyond the minimum and maximum data values, with the number of gridpoints (G) sufficiently large. A default rule for setting G is given below. Together, constraints (3) and (4) ensure that the adjusted estimate remains a *bona fide* probability density.

The final constraint (5) is a general system of inequalities used to implement any operative shape constraints. Shape constraints are also enforced pointwise over a grid. This allows each constraint to be expressed as a linear system of finite dimension. For convenience, we use the same vector $\mathbf{g}$, defined above, as the constraint-checking grid for the shape constraints. We will shortly see a number of constraints that can be expressed as a linear function of $\boldsymbol{a}$; examples of how specific constraints are expressed in the form $\mathbf{A}\boldsymbol{a} \leq \mathbf{b}$ are deferred to the article's supplemental material.

Turning to the objective function, the quadratic form in (2) should quantify the amount of adjustment made to the pilot estimate. One possible measure of the amount of adjustment is the integrated squared error between $\hat{f}_{\boldsymbol{a}}$ and $\hat{f}^{\circ}$:

$$\mathrm{ISE}(\boldsymbol{a}) = \int_{-\infty}^{\infty} (\hat{f}_{\boldsymbol{a}}(x) - \hat{f}^{\circ}(x))^2 dx = \int_{-\infty}^{\infty} \boldsymbol{a}^T \boldsymbol{\psi}(x) \boldsymbol{\psi}(x)^T \boldsymbol{a} \, dx. \tag{6}$$

This integral can be approximated by evaluating $\boldsymbol{\psi}(x)$ at the points in $\mathbf{g}$ and using the trapezoidal rule. With this approximation, $\mathrm{ISE}(\boldsymbol{a})$ takes the form of the objective function in Eq (2), with $\mathbf{v} = \mathbf{0}$ and $\mathbf{H} = \sum_{l=1}^{G} c_l \boldsymbol{\psi}(g_l) \boldsymbol{\psi}(g_l)^T$ (where $c_1 = c_G = 1$ and $c_l = 2$ for $l \notin \{1, G\}$; see the article's supplementary material for more details).

An alternative is to use the sum of the squared coefficients (L_2 distance):

$$L_2(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{a}, \tag{7}$$

which is reasonable because the total adjustment is zero when $\boldsymbol{a} = \mathbf{0}$. The L_2 objective corresponds to the general form (2) with $\mathbf{v} = \mathbf{0}$ and $\mathbf{H} = \mathbf{I}$. It will be shown below that when the pilot estimator is a KDE, our preferred construction of $\Psi(x)$ is indifferent to the choice of ISE or L_2 objective. For simplicity, then, the L_2 objective is used henceforth.

An extension of objective (2) will allow a penalty on the roughness of the final estimate to be included. If, following other types of roughness-penalized estimation (Ramsay and Silverman 2005), we use the integrated square of the second derivative of $\hat{f}_{\boldsymbol{a}}$ as a measure of roughness, the objective may be written $\boldsymbol{a}^T(\mathbf{H} + \lambda \mathbf{S})\boldsymbol{a} + \mathbf{v}^T\boldsymbol{a}$. In this case $\mathbf{H}$ measures the amount of adjustment made, as before. The added term $\lambda\mathbf{S}$ and the linear term $\mathbf{v}$ (which is nonzero in this case) are used to implement the penalty. The coefficient λ is an adjustable parameter controlling the size of the penalty, and therefore the smoothness of the estimate.

The roughness penalty may be used on its own, or in addition to shape constraints. An example using this penalty is presented later, in Figure 4, and computational details are given in supplement S2, but otherwise the focus of this work is on shape constraints without this type of smoothness control.

2.2 Constraints fitting the QP framework

The pth derivative of the adjusted estimate (1) is

$$\hat{f}_{\boldsymbol{a}}^{(p)}(x) = \hat{f}^{\circ(p)}(x) + \boldsymbol{a}^T \boldsymbol{\psi}^{(p)}(x),$$

which is linear in $\boldsymbol{a}$. Any shape constraints involving only linear restrictions on $\hat{f}_{\boldsymbol{a}}$ or its derivatives will therefore be linear in $\boldsymbol{a}$ as well, and expressible in a form suitable for QP. The following constraints may be implemented in this manner.

Unimodality with mode at m. For an estimate satisfying this constraint, $\hat{f}'_{\boldsymbol{a}}(x) \geq 0$ when $x \leq m$ and $\hat{f}'_{\boldsymbol{a}}(x) \leq 0$ when $x \geq m$.

This constraint can be generalized to the case of $M > 1$ modes. Let $m_1, \ldots, m_M$ be the mode locations, and $u_1, \ldots, u_{M-1}$ be the locations of the minima between adjacent modes (consequently $m_1 < u_1 < m_2 < \cdots < u_{M-1} < m_M$). In this case $\hat{f}'_{\boldsymbol{a}}$ is constrained to be positive to the left of m_1, negative between m_1 and u_1, positive between u_1 and m_2, and so on alternating over the support of the estimate.

Monotonicity on the interval $I = (x_1, x_2)$. That is, $\hat{f}'_{\boldsymbol{a}}(x) \geq 0, x \in I$ (monotonically increasing) or $\hat{f}'_{\boldsymbol{a}}(x) \leq 0, x \in I$ (monotonically decreasing). Convexity over an interval can similarly be achieved by restricting the sign of the second derivative of $\hat{f}_{\boldsymbol{a}}$.

Nonnegative support: $\hat{f}_{\boldsymbol{a}}(x) \leq \epsilon, \forall x \leq 0$, where ϵ is a small positive number. This constraint can be used to prevent a Gaussian KDE from having appreciable probability mass on the negative half-line.

Symmetry with point of symmetry s and tolerance ϵ. An estimate is considered symmetric if $|\hat{f}_{\boldsymbol{a}}(s-d) - \hat{f}_{\boldsymbol{a}}(s+d)| \leq \epsilon$, $\forall d > 0$.

The tolerance ϵ in this definition is required to make the constraint expressible as a system of inequalities, rather than strict equalities. Because there are a finite number of adjustment densities, it may not be possible to render $\hat{f}_{\boldsymbol{a}}(x)$ perfectly symmetric for all possible choices of the reflection point s. Small values of ϵ (0.001 times the maximum height of $\hat{f}^{\circ}$, for example) are usually sufficient to make the problem feasible.

Bell shape (type 1): $\hat{f}_{\boldsymbol{a}}$ has exactly two inflection points, at v_1 and v_2. That is,

$$\begin{aligned} \hat{f}''_{\boldsymbol{a}}(x) &\geq 0, \quad x < v_1 \quad \text{or} \quad v_2 \leq x \\ \hat{f}''_{\boldsymbol{a}}(x) &\leq 0, \quad v_1 \leq x < v_2. \end{aligned}$$

Bell shape (type 3): $\hat{f}'_{\boldsymbol{a}}$ has exactly three inflection points, at v_1, v_2, and v_3, i.e.,

$$\begin{aligned} \hat{f}'''_{\boldsymbol{a}}(x) &\geq 0, \quad x < v_1 \quad \text{or} \quad v_2 \leq x < v_3 \\ \hat{f}'''_{\boldsymbol{a}}(x) &\leq 0, \quad v_1 \leq x < v_2 \quad \text{or} \quad v_3 \leq x. \end{aligned}$$

A number of comments can help to clarify this list. As previously mentioned, computer implementation of each of these constraints involves pointwise constraint checking over a set of grid points. The supplement to this article describes in more detail how this is done. Also, note that it is not difficult to apply multiple constraints from the above list simultaneously, for example to achieve an estimate that is both symmetric and unimodal.

The bell shape constraints were introduced by Wolters (2012b) as a means of obtaining unimodal estimates with a degree of smoothness and qualitative resemblance to typical parametric forms greater than that possible with a simple unimodality constraint. The type 1 restriction ensures that the estimate has minimal waves or kinks by preventing extra inflection points from appearing in the density. The type 3 restriction ensures an even higher degree of smoothness by restricting the inflections of the density's derivative. An intermediate type 2 option was also proposed, but it is of less practical interest than these two choices.

Most of the constraints listed above only satisfy the QP structure if the locations of certain points such as the mode, point of symmetry, or inflection points are known. We refer to these as *important points*. Since the locations of these points are not known in practice,

it is necessary to embed the QP solver inside an outer optimization procedure to determine their optimal values. Taking unimodality as an example, the globally optimal adjustment curve can by found by QP for any given mode location m; when m is not known beforehand, a univariate search must be conducted to choose its value, with the QP solver called at each candidate mode location.

The need to search for the best combination of the important points adds complexity to the problem and destroys any guarantee of global optimality in practical application. Nevertheless, good constrained estimates can be found as long as the number of important points is not too large. The following approach is recommended for locating the important points. Let the number of important points be r, and label the points from left to right in ordered sequence $v_1 \leq v_2 \leq \ldots \leq v_r$. Let v_0 and v_{r+1} be lower and upper bounds for the search, respectively. When $r = 1$, the best estimate may be found by performing a one-dimensional minimization of the QP objective as a function of v_1, over the interval (v_0, v_2). For $r > 1$, a good solution can be found by iteratively optimizing each v_i over (v_{i-1}, v_{i+1}), and stopping when no improvement can be made. Any one-dimensional, gradient-free minimizer can be used for the minimization step; each evaluation of the objective function during the minimization requires the quadratic program to be solved for a particular value of $(v_1, \ldots, v_r)$.

This iterative procedure has advantages over attempting to optimize $(v_1, \ldots, v_r)$ simultaneously. It naturally handles the order constraints on the v_i values, and permits a simple, numerically stable optimizer to be used for each step. In our implementation, we carry out the univariate optimizations using golden section search, which is much more efficient than a brute-force grid search over the interval. For the small r values in our constraints ($r \leq 3$), the overall procedure terminates in reasonable time (on the order of seconds).

2.3 Choosing the adjustment densities

The quality of the solution obtained by the QP solver (and the existence of a solution in the first place) depends on the particular set of adjustment densities $\{\psi_i\}_{i=1,\ldots,k}$ used to construct the adjustment curve. To perform its function well, the adjustment curve should be smooth, but still have a high degree of shape flexibility over the support of the density—enough that $\hat{f}_{\boldsymbol{a}}$ can take shapes ranging from sharp peaks to completely flat sections.

The primary means of achieving shape flexibility in the adjustment curve is to choose k sufficiently large and to let each individual ψ_i have its probability mass concentrated over a small region of the support. To avoid introducing unwanted discontinuities in $\hat{f}_{\boldsymbol{a}}$ or its derivatives, the adjustment densities should have the same degree of smoothness as the pilot estimator. If these requirements are met, the specific functional form of the adjustment densities is of little importance.

When the pilot estimator is a Gaussian KDE, a convenient way of defining the adjustment curve is to let the ith adjustment density be a $N(\mu_i, \sigma_i^2)$ density,

$$\psi_i(x) = \frac{1}{\sigma_i}\phi\left(\frac{x-\mu_i}{\sigma_i}\right), \tag{8}$$

where $\phi(\cdot)$ is the standard normal density function. Good performance of $\Psi(x)$ can then be ensured by appropriate choices of (μ_i, σ_i), $i = 1, \ldots, k$. Two options appear most natural.

Option 1: match the pilot KDE's kernel functions

In this option, $k = n$ and the ith adjustment density has parameters $\mu_i = x_i$ and $\sigma_i = h$, where h is the bandwidth parameter of the KDE. In effect, each adjustment density is assigned to one data point and serves to increase or decrease the contribution of the kernel at that point. With the $\{\psi_i\}$ matched to the Gaussian KDE in this way, $\hat{f}_{\boldsymbol{a}}(x)$ is

$$\begin{aligned}
\hat{f}_a(x) &= \hat{f}^{\circ}(x) + \sum_{i=1}^{n} a_i \psi_i(x) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x-x_i}{h}\right) + \sum_{i=1}^{n}\frac{a_i}{h}\phi\left(\frac{x-x_i}{h}\right) \\
&= \frac{1}{h}\sum_{i=1}^{n}\left(\frac{1}{n}+a_i\right)\phi\left(\frac{x-x_i}{h}\right),
\end{aligned} \tag{9}$$

which is equivalent to a variable-weight kernel estimator, with the ith point receiving weight $w_i = \frac{1}{n} + a_i$. This shows that using variable weights to enforce shape constraints on a KDE is a special case of the adjustment curve method, and that optimal weights can also be found using quadratic programming.

Figure 2 shows what the estimate looks like using this arrangement of adjustment densities. The data in the figure are a random sample of size 50 from a lognormal distribution.

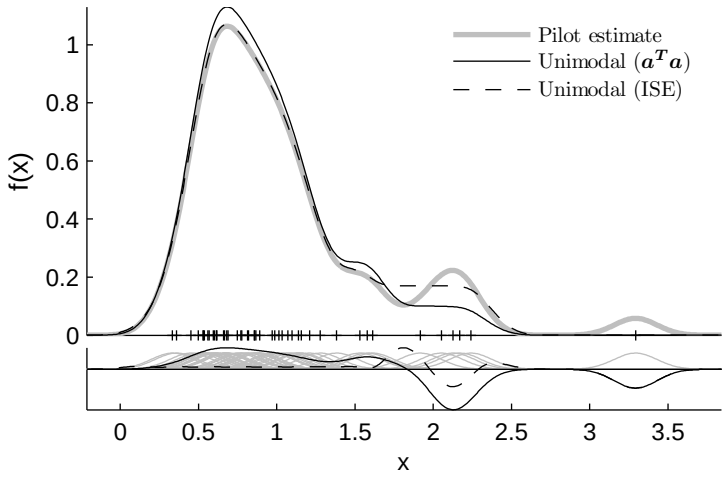

Figure 2: A unimodal KDE with adjustment densities at the data points. The pilot estimate and two unimodal estimates are shown. Underneath are found the set of adjustment densities (scaled down to fit on the plot), with the adjustment curves superimposed.

The pilot estimate (with $h = 0.75h_{SJ}$, where h_{SJ} is the plug-in bandwidth of Sheather and Jones (1991)), is trimodal with an outlying point. Optimal unimodal estimates are shown for both the ISE and L_2 objective functions.

The figure illustrates the advantages of constructing $\Psi(x)$ in this way. The weight interpretation of $\boldsymbol{a}$ is an advantage in itself. Also, the adjustment curve is able to perfectly annihilate any unwanted features of the pilot density (as with the outlying mode in this example), because the adjustment densities are equal to the kernel functions. Simplicity is another advantage, since k and $\{\mu_i\}$ are fixed by the data, and choosing the pilot bandwidth determines $\{\sigma_i\}$.

Several important disadvantages of this construction are also apparent in the estimates. First, in some circumstances it may be necessary to give points zero weight ($a_i = -\frac{1}{n}$) in order to find a feasible solution. This is the case for the outlying point in Figure 2. It is not possible for the constrained estimator to extend its right tail all the way out to this outlier. Second, this method inherits a general feature of variable-weight estimators, that a local adjustment in one region of the curve may require compensatory adjustment in a distant region. In the figure, this effect is more obvious when the $\boldsymbol{a}^T\boldsymbol{a}$ objective is used. The fact that the two objective functions produce such different estimates is also discouraging, as both options should promote selection of solutions that are close to the pilot density. Finally, the

adjustment densities may become unnecessarily concentrated in the high-density regions of the curve. This becomes increasingly inefficient as n grows, and could cause ill-conditioning of the coefficient matrices used by the QP solver.

Option 2: location-shifted, overlapping densities on a grid.

The second natural choice is to let all the adjustment densities have the same standard deviation σ, and locate them on an evenly-spaced grid. Let l and u be lower and upper bounds for the grid, selected so that (l, u) extends beyond the data in either direction (setting $l = x_{(1)} - 4h$ and $u = x_{(n)} + 4h$ would seem reasonable). Then the set of densities is fixed by specifying k and σ. As a rule of thumb, it is proposed to use

$$k = \left\lceil \frac{2(u-l)}{h} \right\rceil \qquad \text{and} \qquad \sigma = \frac{u-l}{k-1} \equiv \Delta, \tag{10}$$

where $\lceil\ \rceil$ represents the ceiling function and Δ is the grid spacing. With this rule, the adjustment densities are centered at $\mu_i = l + (i-1)\Delta$, $i = 1, \ldots, k$.

The logic behind recommendation (10) is as follows. Take l and u as given. The set of adjustment densities must be able to reproduce the pilot pdf to within some tolerance, otherwise $\Psi(x)$ would not be able to eliminate unwanted features of the density. So the grid must be dense enough that every data point is close to a grid point μ_i. The bandwidth h can be taken as a measure of closeness, so a grid spacing of approximately $\frac{h}{2}$ should be sufficient. The grid spacing is $\Delta = \frac{u-l}{k-1}$, so ideally one would choose

$$\frac{u-l}{k-1} = \frac{h}{2} \qquad \Rightarrow \qquad k = \frac{2(u-l)}{h} + 1.$$

The value suggested in (10) results by noting that $2(u-l)/h \gg 1$ and that k must be an integer.

With the values of k and Δ thus determined, we set $\sigma = \Delta$ to ensure that the $\psi_i(x)$ overlap to an appropriate degree. A trade-off exists in the choice of σ. If it is made too large, the adjustment densities will overlap too much, and the adjustment curve will be too smooth—unable to make rapid local changes of shape. If σ is too small, on the other hand, the adjustment curve (or its derivatives, which are used in the constraints) will be insufficiently smooth, and the solver might not be able to find a solution. Experience has shown that setting $\sigma = \Delta$ provides a good compromise between these two extremes.

Figure 3 demonstrates the results of this construction of $\Psi(x)$ on the lognormal-data example of Figure 2. In this case the adjustments to the pilot density are confined to those regions near the constraint violations, and the adjusted estimate does extend out to the outlying point. Also, the two different objective functions return nearly indistinguishable solutions. This is a consequence of defining the adjustment densities in this way, and the agreement between $ISE(\boldsymbol{a})$ and $L_2(\boldsymbol{a})$ improves as n or k grow (see the article's supplementary material for more on this point). Given these appealing characteristics, the grid construction for $\Psi(x)$ with rule of thumb (10) is used from this point forward.

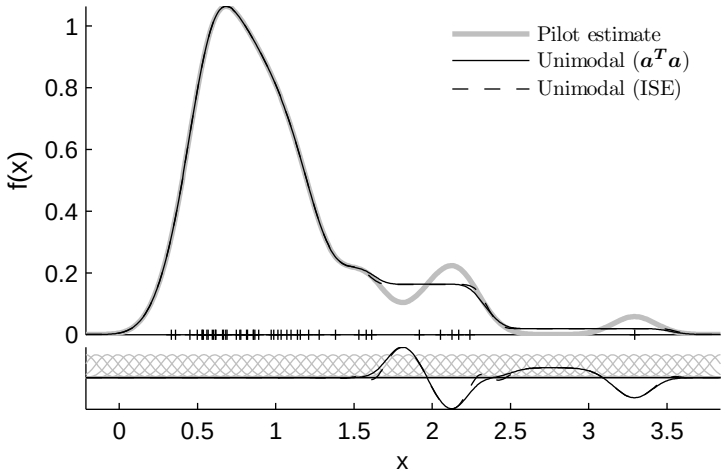

Figure 3: A unimodal KDE with adjustment densities on a grid. Compare with Figure 2. The rule of thumb (10) chose $k = 60$ for these data.

When using this rule of thumb for setting up the adjustment densities, it is also important to ensure that G, the number of constraint checking points, is sufficiently large. If G is too small, then some adjustment densities might fall between points in $\mathbf{g}$, and the corresponding a values will have no influence on the constraints inside the QP solver. This can lead to solutions with unintended constraint violations. A default setting of $G = 2k$ is recommended to avoid this problem. This default is used in all of the examples and simulations to follow.

3 User-facing estimation routines

The adjustment curve approach to constraint handling, in its general form, is conceptually simple: the adjustment is a linear combination of density functions, and the constraints are

evaluated pointwise over the support of the density, yielding a quadratic program. Most of the complexity encountered in the preceding section arises in the process of matching this general structure to a specific type of pilot density (the KDE) so that the resulting optimization problem is feasible, numerically stable, and produces desirable solutions. It is not uncommon to encounter such implementation issues when applying a general computational scheme to a specific problem. More importantly, these implementation details primarily impact optimization performance, rather than statistical performance. Once they have been worked out for a particular type of pilot estimator, they can be fixed and built into computer code so that the end user need not be aware of them.

As justification of these claims, we can review the design decisions covered in the previous section, taking the perspective of a programmer whose goal is to implement the method to add constraints to a KDE in a user-friendly way.

Constraint-checking grid. It is necessary for numerical stability that every $\psi_i(x)$ be involved in at least one constraint inequality. This is not hard to achieve for a given set of adjustment densities. Beyond this, setting the number of check points sufficiently large (say, 100 or more) will ensure that any constraint violations at intermediate points are negligible. Adding more check points will increase the size of the matrices fed to the QP solver, but will have minimal effect on the final estimate.

Objective function. The optimization is indifferent to the choice of ISE or L_2 objective, so this choice depends on user preference. As explained previously, the two options have little difference when the ψ_i are arranged on a grid.

Constraint tolerances. The symmetry and nonnegative-support constraints each include a tolerance ϵ required to ensure feasible solutions exist. Experience has shown that, in both cases, a very wide range of ϵ values (many orders of magnitude) can ensure numerical stability while having no discernable effect on the final estimate.

Placement of the adjustment densities. For the case of a KDE pilot estimator (and not for other situations), the option exists to match the adjustment densities to the pilot kernel functions ("option 1"). This may be viewed as a serendipitous result for

the KDE case, since it allows us to obtain a shape-restricted weighted KDE using the adjustment curve technology. This choice fixes all ψ_i and admits no further tuning of the adjustment densities.

The alternative of letting the ψ_i be k location-shifted copies of a normal density ("option 2") is more generally applicable and could be used with other smooth pilot estimators. In this case, it is crucial for feasibility that the densities overlap sufficiently, and that k is sufficiently large. Guidelines for satisfying these requirements were given, but a wide range of grid arrangements will suffice, with minimal impact on the final estimate.

From this summary we see that once we have settled on option 1 or option 2 for the placement of the ψ_i, all of the other settings are easily made and have little impact on the shape of the final estimate. In particular, the visually discernable effect of the algorithmic options is dwarfed by the effect of setting the estimator's true tuning parameter—the pilot bandwidth, h.

The MATLAB code accompanying this article includes individual functions to run the procedures described throughout. In addition, there are six functions that could be described as user-facing—they handle all of the algorithmic options internally and compute constrained density function values at specified points. These functions use a Gaussian KDE as the pilot estimator and handle six scenarios: unimodality, type 1 bell shape, and type 3 bell shape, each with or without the additional constraint of symmetry. These functions take only the data and the pilot bandwidth as inputs, making them as easy to use as unconstrained estimation functions. Internally, the functions set up the adjustment densities and the constraint checking grid automatically, and solve the resulting QP problem repeatedly until the best important points (mode location or inflection points) are found. The functions were used to process the examples and simulations given in the following sections.

4 Examples

In this section, two data sets from the literature are used to illustrate some characteristics of the adjustment curve method. The first data set is univariate, consisting of 57 wind

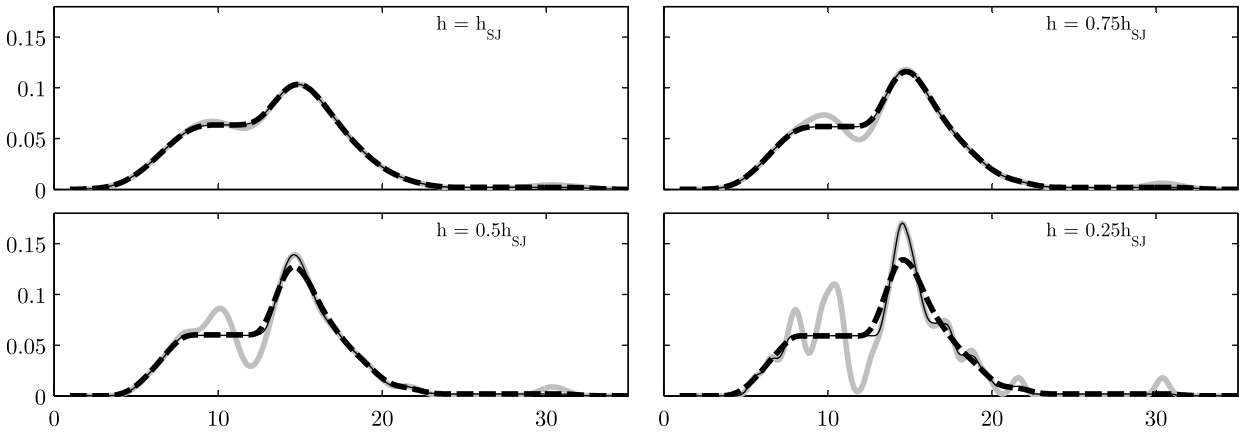

Figure 4: Unimodal estimates for the wind speed data at different bandwidths. Each plot shows the pilot estimate (grey) the unimodal estimate (solid black), and the unimodal estimate with roughness penalty (dashed). All plots have the same axis scaling.

speed measurements made at an elevation of 10 meters in Italy's Messina Strait region (Alibrandi and Ricciardi 2008). The second data set is bivariate, consisting of standardized measurements of systolic blood pressure (SBP) and concentration of low density lipoprotein (LDL) in 160 diseased patients who were part of a larger study of risk factors for heart disease in South Africa (Hastie and Tibshirani 1987; Hastie et al. 2009). All of the constrained estimates below were produced using the default settings introduced so far (Gaussian KDE pilot estimator, L_2 objective, grid arrangement of the ψ_i, and the rule of thumb choices for k and G).

4.1 Univariate example

The unconstrained KDE for the wind speed data is shown in grey in Figure 4, using four different bandwidths: 1, 0.75, 0.5, and 0.25 times the Sheather-Jones bandwidth (which equals 1.55 for these data). These bandwidths were chosen to illustrate the behavior of constrained estimates as the bandwidth is reduced. The unconstrained estimate at $h = h_{SJ}$ has three modes: a main central peak, a broad shoulder to its left, and a mode to its right, caused by an outlying point at speed 30.4. As the bandwidth is reduced, these three modes become more distinct and additional modes begin to appear.

Suppose it is reasonable to assume that the true distribution of wind speeds is unimodal. Each plot in Figure 4 also shows (as a solid black line) the best unimodal estimate achieved

using adjustment curves. In all cases the mode of the best constrained estimate was located at the same point as the highest mode in the pilot estimate (this need not be the case generally, as the mode is selected to optimize the objective). These estimates achieve unimodality by flattening out the density across any constraint violations. The estimate looks increasingly like a step function as h gets smaller and the number of constraint violations grow. This illustrates how $\hat{f}_{\boldsymbol{a}}$ does not necessarily inherit the smoothness of the pilot KDE, because the adjustment curve operates over the whole line, and not just at the data points.

Smoother unimodal estimates can be obtained by adding a roughness penalty to the objective function, as shown in Figure 4. The resulting penalized unimodal estimates are also shown in the figure. The penalty causes the estimates to be less sensitive to pilot bandwidth choice, and prevents the step-like appearance from arising as the bandwidth is reduced. Nevertheless, the value of roughness-penalized estimation is arguable in this situation, because the estimates require the selection of both a bandwidth and a smoothing parameter. For the plots in this example, the smoothing parameter λ was chosen by cross-validation (Wasserman 2006, p. 127).

A better alternative for improving the smoothness of the estimates is to use a different shape constraint that more accurately captures the desired qualitative features of the estimate. The bell-shaped constraints are an example of more restrictive criteria that should produce smoother estimates. Figure 5 shows, for the same four pilot bandwidths, the bell-shaped estimates of type 1 and 3. At each bandwidth the best choices of inflection points for each estimate were found using the algorithm described earlier. These shape-restricted estimates have a high degree of smoothness built in to them by definition, so they cannot have plateaus or steps in them. Only bandwidth selection is required.

We note in passing that the problem of selecting a bandwidth that is (in some sense) optimal for a shape-restricted KDE is difficult, and is not addressed here. Fortunately, there is some justification for simply using a bandwidth that is suitable for the pilot KDE, and performing the shape adjustment after bandwidth selection. These matters are discussed in more detail elsewhere (Wolters 2012a,b). For the wind speed data, then, the simplest approach would be to use a standard bandwidth choice such as h_{SJ}, and to subsequently apply the bell-shaped constraint. This would produce the estimates in the upper left of

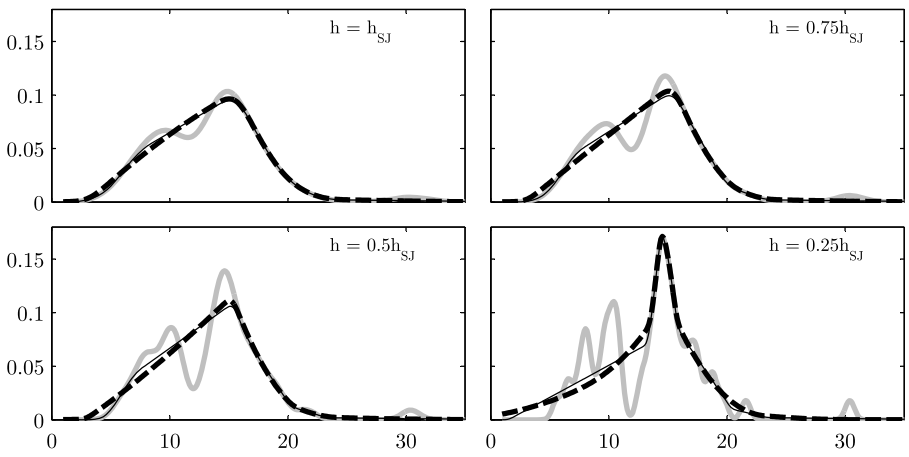

Figure 5: Bell shaped estimates for the wind speed data. Each plot shows the pilot estimate (grey), the type 1 bell shaped estimate (black), and the type 3 bell shaped estimate (dashed). All plots have the same axis scaling.

Figure 5 as the final estimate.

4.2 Bivariate example

The estimator $\hat{f}_{\boldsymbol{a}}$ is in principle easily extended to higher dimensions. If a d-dimensional estimate is required, one only needs to define the k adjustment densities to be d-variate functions. The constrained estimator remains linear in $\boldsymbol{a}$, and $\boldsymbol{a}$ is still a $k \times 1$ vector. Practical implementation of the method in d dimensions involves two significant complications, however.

The first difficulty is the potential explosion of the required number of adjustment densities and constraint-checking points as d increases. The size of the system of inequalities in the QP problem will quickly become unmanageable if the univariate strategy of putting the ψ_i and $\mathbf{g}_i$ on a grid is extended to d-dimensional rectangular meshes. A second, more fundamental problem is the inability to express higher-dimensional shape constraints as linear inequalities in $\boldsymbol{a}$. Simple univariate constraints like unimodality or bell shape do not translate easily to higher dimensions, and more complex restrictions (for example, unimodality of all conditional distributions) are difficult to express mathematically without assuming that a large number of important points are pre-specified.

Despite these difficulties, some progress can be made. The heart disease data is bivariate,

and for $d = 2$ it is still possible to put the adjustment densities and constraint-checking points on a mesh without exceeding the capacities of a typical personal computer. Also, one multivariate constraint that can be implemented using QP is *star unimodality*. This constraint specifies that the density is decreasing along all rays emanating from the mode location $\mathbf{m}$ (Klemelä 2009). When $\mathbf{m}$ is taken as known, the directional derivative of $\hat{f}_{\boldsymbol{a}}(x)$ along the ray from $\mathbf{m}$ to $\mathbf{g}_i$ can be expressed as a function that is linear in $\boldsymbol{a}$ (see the supplementary material). The constraint can be implemented by establishing a set of constraint-enforcement points $\{\mathbf{g}_i\}$, and requiring the directional derivative to be negative at all elements of the set.

Figure 6 shows the star unimodal estimator. The adjustment surface was constructed using a 20×20 grid of independent bivariate normal distributions, with component standard deviations equal to the grid spacing. The constraint was enforced at a 35×35 grid of points. The kernel function for the pilot KDE was an uncorrelated bivariate normal density with covariance matrix $h^2\mathbf{I}$. The bandwidth was set to $h = 0.23$, which maximized a pseudo-likelihood criterion defined and motivated by Wolters (2012a). Applying the constraint does improve the qualitative smoothness of the estimate noticeably. The adjusted estimate has one visible violation of the constraint (noted by an arrow in the figure). Increasing the density of the grid would eliminate such artifacts, at the cost of longer run time. The estimate in Figure 6 was obtained in approximately 30 seconds on a laptop computer.

This bivariate example has been provided only as a proof of concept, which may be of interest given the dearth of shape-constrained density estimators available for higher dimensions (the log-concave NPMLE being a notable exception). We defer more detailed consideration of the bivariate case to future work. The simulations of the next section will focus on one-dimensional estimation.

5 A simulation study

A simulation study was performed to observe how the addition of different shape constraints influences the quality of estimation afforded by the univariate Gaussian KDE. Data sets for the simulation were drawn from the t distribution with 3 degrees of freedom, with two sample sizes, $n = 50$ or $n = 200$. At each sample size, 260 data vectors were drawn and used to

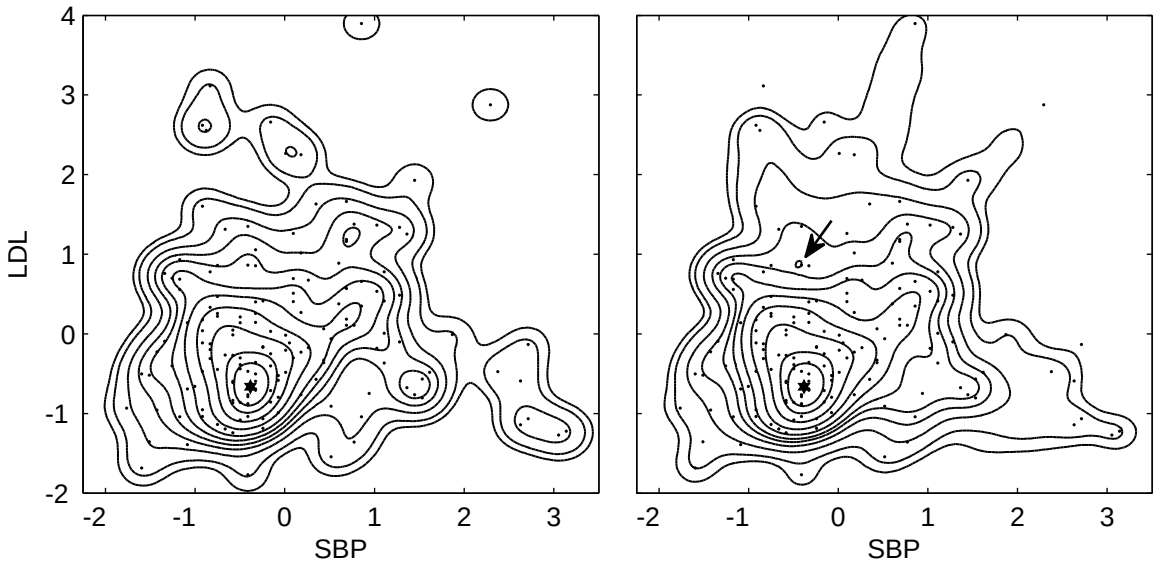

Figure 6: Pilot density estimate (left) and star unimodal estimate (right) for the heart disease data. The bandwidth used was $h = 0.23$, and the highest mode of the pilot density (labelled by a star) was used as the mode for the adjusted estimate.

produce five different estimates: 1) no constraint, 2) unimodal, 3) unimodal and symmetric, 4) type 1 bell shape, and 5) type 1 bell shape and symmetric around zero.

Each estimate was computed at 10 different pilot bandwidths, evenly spaced between 0.2 and 0.8. In total, 13000 estimates were calculated for each sample size (all combinations of 260 data sets, five constraints, and 10 bandwidths).

Note that unimodality, bell shape, symmetry, and symmetry around zero are all true characteristics of the t densities, so each of the constraints introduces valid auxiliary information that should enhance estimation performance. The main goal of the study was to observe whether the different constraints, which include different amounts of auxiliary information, produce appreciable differences in mean estimation quality and bandwidth sensitivity.

The results of the study are summarized in Figure 7, which shows the mean value of the integrated squared error (ISE) between the estimates and the truth, as a function of h, for each constraint and both sample sizes. The horizontal dashed line on each plot shows the mean value of the appropriate distance when each unconstrained KDE was computed with an *oracle* bandwidth selector—the bandwidth that actually minimizes the distance between the estimate and the truth. Performance with the oracle bandwidth represents the best

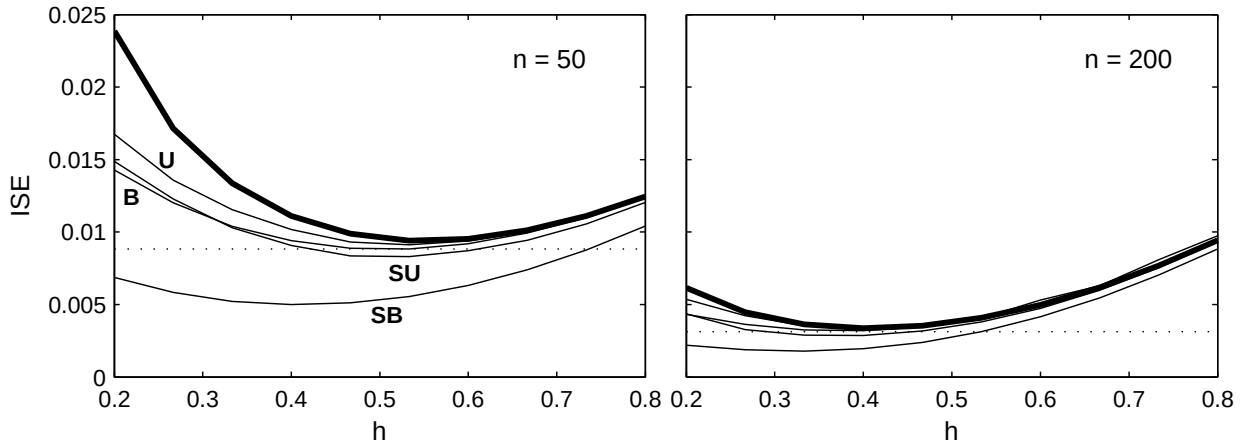

Figure 7: Statistical performance of constrained estimates using $\hat{f}_{\boldsymbol{a}}$, at two sample sizes. The thick line is the result for the pilot estimator. Labels on the other four lines indicate the operative constraints: U for unimodality, B for bell shape, and S for symmetry. The dotted horizontal lines give the performance of the unconstrained estimator with oracle bandwidth. Both plots have the same axis scaling. The lines in the $n = 200$ plot are in the same relative positions as the $n = 50$ case.

possible performance of an unconstrained KDE, and provides a useful reference point.

The results suggest that adding constraints does improve performance and reduce bandwidth sensitivity. The constraints involving more qualitative information yield greater improvements. The symmetric and bell shaped estimator performed particularly well, likely because the correct point of symmetry (zero) was supplied to this estimator. It should also be noted that the optimal bandwidth is largest for the unconstrained estimate, and becomes smaller as better constrained estimators are used.

The benefits of adding shape constraints are greatest for the smaller sample size. This behaviour is expected whenever the pilot estimator is consistent and the constraints are valid. Constraint violations should get smaller as n increases, leaving less opportunity for improvement.

The simulation runs also provide information on typical run times required to obtain constrained estimates. Figure 8 plots the median run time as a function of h and the constraint type. The run times in the plot reflect the combined effects of two factors: the size of the system of inequalities necessary to enforce the constraints, and the repetitions required to find the best inflection or mode points. The system of inequalities becomes larger

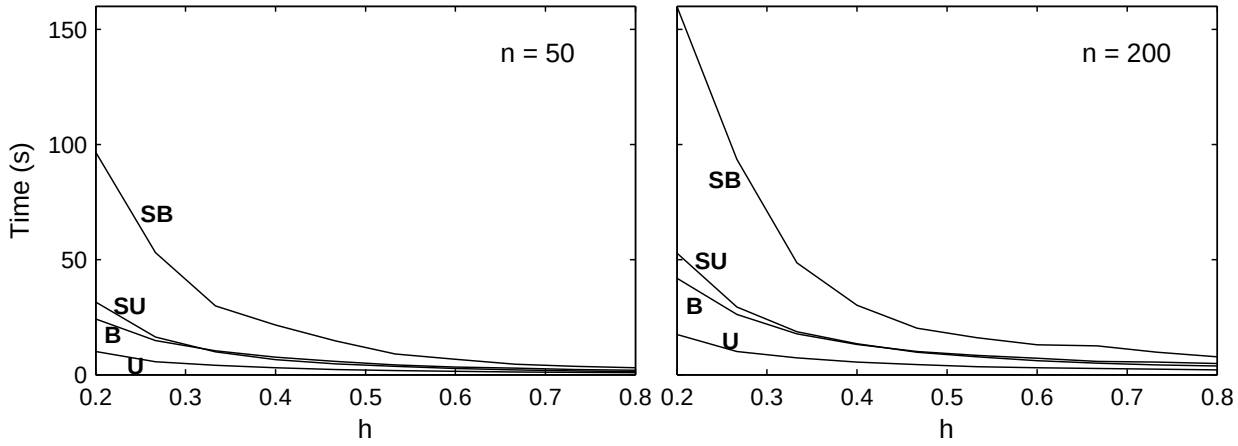

Figure 8: Median run times for the adjustment curve estimates, for two sample sizes. Line labels are the same as in Figure 7. The unconstrained estimate is not shown because its run times are too close to zero (typically 0.007s).

as h gets smaller (a consequence of the rule of thumb for choosing k and G), and also becomes larger when the symmetry constraint is added. Figure 7 suggests optimal bandwidths fall in the range $(0.4, 0.6)$. In this range, estimates can be obtained quickly: unimodal estimation is nearly instantaneous, and the most difficult case of symmetry and bell shape can be solved in 30 seconds or less. These reasonable computation times are a result of the golden section line search used in the algorithm for finding the important points.

6 Extensions and future work

The method of adjustment curves has some attractive features, foremost of which is the ability to use fast and reliable quadratic programming routines to obtain many types of constrained estimates. In addition, the construction of the adjustment curve can be varied for different purposes, and is not coupled to the form of the estimator, as it is for other constraint-handling approaches like data sharpening or weighted estimation. This offers potentially greater flexibility in determining the constrained estimator's characteristics, and opens up a number of avenues for refinement or expansion of the method. Several such ideas are summarized here.

- The method of constructing the adjustment curve is open to alteration. One obvious

change is to use adjustment densities that are compactly supported. Alternatively, it may be possible to define a_i to be the height of the adjustment curve at point μ_i, and to define $\psi(x)$ as a curve that interpolates these points; or to use a spline function, rather than a linear combination of densities, as the adjustment curve. Such changes might simplify the construction of $\Psi(x)$ or improve numerical performance.

- Two options for placement of the adjustment densities were proposed in this chapter: putting them at the data locations, or putting them on a grid. An adaptive or hybrid method of locating the adjustment densities could be proposed, that combines the advantages of both options by putting more densities in data-rich regions, and a regular grid of densities in data-poor regions.

- Many shape constraints are actually restrictions on the derivatives of the estimate. It may be possible to apply the adjustment curve to the appropriate derivative of the KDE rather than to the KDE itself. This approach could be expected to give better numerical stability and smoother density estimates.

- It should be possible to adapt the adjustment curve approach to constrained nonparametric regression problems. Because of its additive structure, it may be easier to find an optimal adjustment than to constrain the regression estimator directly.

The question of bandwidth selection for shape-restricted KDEs was briefly touched upon at the end of Section 4.1. Further exploration of improvements to bandwidth selection is certainly warranted. Work on bandwidths for density derivative estimation (e.g. Henderson and Parmeter 2012) may be relevant, since most of our constraints are based on derivatives of the density. It is also possible that enforcing constraints may produce a side benefit of improved bandwidth selection when using data-driven selection procedures such as least-squares cross validation. This question has already been explored for constraints handled by data sharpening (Wolters 2012a); similar work could be carried out for the adjustment curve method.

This article has focused solely on problems for which QP can be used to find solutions. It is important to note that $\hat{f}_{\boldsymbol{a}}$ is not limited to only these cases. As long as a sufficiently effective optimizer is available, other constraints could be satisfied by this adjustment method.

For example, a particle swarm optimizer has been proposed (Wolters 2012c) for general constrained estimation problems.

7 Supporting Information

S1 Code. A MATLAB implementation. Functions and scripts for constrained estimation using the methods of this article (ZIP).

The code is included as an attachment here:

S2 Appendix. Sample quadratic programs. A document describing in detail for several examples how the objective function and constraints are expressed in the form of Eq (2) through (5) (PDF).

The file is included as an attachment here:

References

Alibrandi, U. and Ricciardi, G. (2008), "Efficient evaluation of the pdf of a random variable through the kernel density maximum entropy approach," *International Journal for Numerical Methods in Engineering*, 75, 1511–1548.

Antoniou, A. and Lu, W. (2007), *Practical optimization: algorithms and engineering applications*, Springer-Verlag New York Inc.

Balabdaoui, F. and Wellner, J. (2007), "Estimation of a k-monotone density: limit distribution theory and the spline connection," *The Annals of Statistics*, 35, 2536–2564.

Bickel, P. J. and Fan, J. (1996), "Some Problems on the Estimation of Unimodal Densities," *Statistica Sinica*, 6, 23–45.

Birgé, L. (1997), "Estimation of Unimodal Densities Without Smoothness Assumptions," *The Annals of Statistics*, 25, 970–981.

Birke, M. (2009), "Shape Constrained Kernel Density Estimation," *Journal of Statistical Planning and Inference*, 139, 2851–2862.

Braun, W. J. and Hall, P. (2001), "Data Sharpening for Nonparametric Inference Subject to Constraints," *Journal of Computational and Graphical Statistics*, 10, 786–806.

Cheng, M.-Y., Gasser, T., and Hall, P. (1999), "Nonparametric Density Estimation under Unimodality and Monotonicity Constraints," *Journal of Computational and Graphical Statistics*, 8, 1–21.

Cule, M., Samworth, R., and Stewart, M. (2010), "Maximum likelihood estimation of a multi-dimensional log-concave density," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 545–607.

Du, P., Parmeter, C. F., and Racine, J. S. (2013), "Nonparametric Kernel Regression with Multiple Predictors and Multiple Shape Constraints," *Statistica Sinica*, 23, 1347–1371.

Dümbgen, L. and Rufibach, K. (2009), "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency," *Bernoulli*, 15, 40–68.

Fougères, A.-L. (1997), "Estimation de densites unimodales," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 25, 375–387.

Grenander, U. (1956), "On the Theory of Mortality Measurement, Part II," *Skandinavisk Aktuarietidskrift*, 39, 125–153.

Hall, P. and Huang, L.-S. (2001), "Nonparametric Kernel Regression Subject to Monotonicity Constraints," *The Annals of Statistics*, 29, pp. 624–647.

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

Hastie, T. and Tibshirani, R. (1987), "Non-Parametric Logistic and Proportional Odds Regression," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36, 260–267.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Science+Business Media.

Henderson, D. J. and Parmeter, C. F. (2009), "Imposing Economic Constraints in Nonparametric Regression: Survey, Implementation, and Extension," *Advances in Econometrics*, 25, 433–469.

Henderson, D. J. and Parmeter, C. F. (2012), "Canonical Higher-Order Kernels for Density Derivative Estimation," *Statistics and Probability Letters*, 82, 1383–1387.

Henderson, D. J. and Parmeter, C. F. (2015), *Applied Nonparametric Econometrics*, New York: Cambridge University Press.

Jones, M., Samiuddin, M., Al-Harbey, A., and Maatouk, T. (1998), "The edge frequency polygon," *Biometrika*, 85, 235–239.

Klemelä, J. (2009), *Smoothing of Multivariate Data*, Wiley.

Meyer, M. (2008), "Inference using shape-restricted regression splines," *The Annals of Applied Statistics*, 2, 1013–1033.

Nocedal, J. and Wright, S. J. (1999), *Numerical Optimization*, Springer.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis*, 2nd ed., Springer.

Reboul, L. (2005), "Estimation of a Function under Shape Restrictions. Applications to Reliability," *The Annals of Statistics*, 33, pp. 1330–1356.

Sheather, S. J. (2004), "Density Estimation," *Statistical Science*, 19, 588–597.

Sheather, S. J. and Jones, M. C. (1991), "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society, Series B*, 53, 683–690.

Silverman, B. W. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, 6, pp. 177–184.

The Mathworks, Inc. (2007), *MATLAB Version 7.4.0*, Natick, Massachusetts.

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

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer.

Wegman, E. J. (1972), "Nonparametric Probability Density Estimation: I. A Summary of Available Methods," *Technometrics*, 14, 533–546.

Wolters, M. A. (2012a), "A Greedy Algorithm for Unimodal Kernel Density Estimation by Data Sharpening," *Journal of Statistical Software*, 47, 1–26.

— (2012b), "Methods for Shape-Constrained Kernel Density Estimation," PhD thesis, University of Western Ontario.

— (2012c), "A particle swarm algorithm with broad applicability in shape-constrained estimation," *Computational Statistics & Data Analysis*, 56, 2965–2975.