Unimodal Density Estimation Using Bernstein Polynomials

Bradley C. Turnbull and Sujit K. Ghosh

Department of Statistics, North Carolina State University

NC State Department of Statistics Technical Report 2650

Abstract

The estimation of probability density functions is one of the fundamental aspects of any statistical inference. Many data analyses are based on an assumed family of parametric models, which are known to be unimodal (e.g., exponential family, etc.). Often a histogram suggests the unimodality of the underlying density function. Parametric assumptions, however, may not be adequate for many inferential problems. This paper presents a flexible class of mixture of Beta densities that are constrained to be unimodal. We show that the estimation of the mixing weights, and the number of mixing components, can be accomplished using a weighted least squares criteria subject to a set of linear inequality constraints. We efficiently compute the number of mixing components and associated mixing weights of the beta mixture using quadratic programming techniques. Three criterion for selecting the number of mixing weights are presented and compared in a small simulation study. More extensive simulation studies are conducted to demonstrate the performance of the density estimates in terms of popular functional norms (e.g., L_p norms). The true underlying densities are allowed to be unimodal symmetric and skewed, with finite, infinite or semi-finite supports. Code for an R function is provided which allows the user to input a data set and returns the estimated density, distribution, quantile, and random sample generating functions.

1 Introduction

Statistical inference is typically based on an assumed family of unimodal parametric models. Nonparametric density estimation is a popular alternative when that parametric assumption is not appropriate for modeling the density of the underlying population. The kernel method, developed by Parzen (1962), is one of the most popular methods of nonparametric density estimation. It is defined as the weighted average of kernel functions centered at the observed

values. This average is taken with respect to the empirical cumulative distribution function (ECDF), $F_n(\cdot)$, and is dependent on a smoothing or bandwidth parameter.

If one believes the underlying population's density is unimodal, there are two major advantages to including a unimodality constraint in the density estimate. First, incorporating extra information about the shape of the density should improve the overall accuracy of the estimate. Second, extraneous modes, which may hinder the usefulness of the density estimate as a visual aid and exploratory tool, will be eliminated (Wolters, 2012).

1.1 Unimodal Density Estimation

Silverman (1981) developed a bandwidth test for unimodality stemming from a nonparametric density estimate. Unfortunately, this test cannot be used to form the basis for a unimodal density estimate. The density estimate constructed by the test is smoothed in a global manner that is influenced solely by the features of the density located around the mode (Cheng et al., 1999). This can result in considerable over-smoothing in some places, and under-smoothing in others. The bandwidth test is also sensitive to clusters of data located away from the center of the distribution and therefore requires a large bandwidth value in order to produce a unimodal density estimate. As the sample size increases, the bandwidth may even diverge to infinity if the data are sampled from a heavy tailed density, similar to the Student's t distribution with small degrees of freedom.

Other approaches to unimodal density estimation extend from the estimation of monotone densities, which are simply special cases of unimodal densities with the mode located on a boundary of the density's support. Grenander (1956) proposed the nonparametric maximum likelihood estimator which is the derivative of the least concave majorant of the ECDF for nonincreasing densities, and the derivative of the greatest convex minorant of the ECDF for nondecreasing densities. Later research focussed on extending the Grenander estimator to any unimodal density. The typical approach is to combine a nondecreasing Grenander estimate to the left of the mode and a nonincreasing estimate to the right. The key element of this approach is determining the location of the mode. Wegman (1972) proposed specifying a modal interval, while Bickel and Fan (1996) used a consistent point estimate of the mode location, and Birgé (1997) selected the mode that minimized the distance between

the distribution estimate and ECDF. Since all these estimates are based on the nonsmooth Grenander estimate, they each produce a step-function density estimate. Bickel and Fan (1996), however, did present methods for smoothing the estimated density.

Other novel approaches include, Fougéres (1997) who used a monotone rearrangement, as suggested by Hardy et al. (1952), to transform a multimodal density estimate into unimodal form. This, however, requires the assumption that the location of the mode is known. Cheng et al. (1999) developed a unique method which treats a general unimodal density as a transformation of some known, but subjective, unimodal template. They presented a recursive algorithm for estimating the transformation and showed how to adjust the technique for density estimation under the monotonicity constraint. The algorithm produces a sequence of successive step function approximations of the true density, which require some form of smoothing in order to make the method an appealing estimate of a smooth density.

Recent unimodal density estimation research has been focussed on utilizing data sharpening techniques, introduced by Choi and Hall (1999) and Choi et al. (2000), to implement unimodal constraints on standard nonparametric density estimators. Data sharpening involves shifting data points in a controlled manner before executing estimation techniques. The goal is to shift the data as little as necessary in order to bestow the estimator with some desired characteristics. Data sharpening is an attractive approach to density estimation as it can be applied to any nonparametric estimator with any shape constraint (Wolters, 2012). Braun and Hall (2001) showed that data sharpening can improve the performance of numerous estimators, including unimodal kernel density estimators.

Braun and Hall (2001) and Hall and Kang (2005) measured the closeness of the sharpened data and original data set using a L_α distance, for $1 \leq \alpha \leq 2$. They obtained the sharpened data vector that minimized the L_α norm using sequential quadratic programming (SQP) techniques. Hall and Huang (2002) also used SQP methods to perform unimodal density estimation by reweighting, or tilting, the empirical distribution. There are numerous issues with using SQP for unimodal density estimation, including the requirement that the location of the mode must be explicitly defined, and when $\alpha = 1$ the L_1 norm is not strictly convex so solutions may not be unique. The biggest issue, however, is that the constraint functions may not always be convex functions of the sharpened data set, so the SQP could improperly

converge to local optima, or in some cases may not converge at all (Wolters, 2012). Wolters (2012) attempted to remedy these issues by proposing a greedy algorithm which always converges to a sensible solution, does not require the location of the mode to be pre-specified, and requires less computing time than SQP, but like SQP, the algorithm is sensitive to its starting values.

1.2 Density Estimation with Bernstein Polynomials

Bernstein polynomials were first studied by Bernstein in 1912, who developed them as a probabilistic proof of the Weierstrass Approximation Theorem. He showed that any continuous function, $f(x)$, on a closed interval $[a, b]$ can be uniformly approximated using Bernstein polynomials by,

$$B_m(x, f) = \sum_{k=1}^{m} f\left(a + \frac{k-1}{m-1}(b-a)\right) \binom{m-1}{k-1} \left(\frac{x-a}{b-a}\right)^{k-1} \left(\frac{b-x}{b-a}\right)^{m-k}, \qquad (1)$$

for $a \leq x \leq b$. The Bernstein-Weierstrass Approximation Theorem assures that as the degree of the polynomial increases to infinity the Bernstein polynomial approximation converges uniformly to the true function, i.e. $\|B_m(\cdot\,, f) - f(\cdot)\|_\infty \equiv \sup_{a \leq x \leq b} |B_m(x, f) - f(x)| \to 0$, as $m \to \infty$ (Lorentz, 1986).

Bernstein polynomials are an attractive approach to density estimation as they are the simplest example of a polynomial approximation with a probabilistic interpretation. They also naturally lead to estimators with acceptable behavior near the boundaries (Leblanc, 2010). Vitale (1975) was the first to propose using Bernstein polynomials to produce smooth density estimates. Babu et al. (2002) investigated the asymptotic properties of using Bernstein polynomials to approximate bounded and continuous density functions. Kakizawa (2004) demonstrated that Bernstein polynomials can be used as a nonparametric prior for continuous densities, and Leblanc (2010) focussed on a bias reduction approach using a Bernstein-based estimator. Petrone (1999) performed nonparametric density estimation in a fully Bayesian setting using Bernstein polynomials. The asymptotical properties of this method were further investigated by Ghosal (2001) and Petrone and Wasserman (2002).

This paper presents a method of unimodal density estimation using Bernstein polynomials. A density estimate is obtained using quadratic programming techniques to minimize a scaled squared distance between the Bernstein distribution function estimate constrained to

unimodality and the ECDF of the data. Multiple approaches for selecting the degree of the polynomial are presented as well, along with a small simulation comparing the effectiveness of each approach. The performance of our proposed method when estimating densities of various supports and levels of skewness is assessed in a Monte Carlo simulation. We also show how our method handles data that is contaminated with outliers. A small section of this paper is dedicated to describing R code we provide for easy implementation our method. Finally, we apply our method to two real data sets and discuss its performance compared to the traditional approach of assuming the data follow some unimodal parametric distribution.

2 Methodology

We begin by assuming $X_i \overset{iid}{\sim} f(\cdot)$, for $i = 1, 2, \ldots, n$, and it is known that $f(\cdot)$ is a unimodal continuous density function. In other words, we assume that there exists an $x^* \in \mathbb{R}$, such that $f(x)$ is non-decreasing on $(-\infty, x^*)$ and $f(x)$ is non-increasing on (x^*, ∞). Our goal is to construct an estimate of $f(\cdot)$, $\hat{f}(\cdot)$, which satisfies the following conditions:

(i) $\hat{f}(x) \geq 0$, for all $x \in \mathbb{R}$,

(ii) $\int \hat{f}(x)\, dx = 1$,

(iii) $\hat{f}(x)$ is unimodal, i.e. there exists an $\hat{x}^* \in \mathbb{R}$, such that $\hat{f}(x)$ is non-decreasing on $(-\infty, \hat{x}^*)$ and $\hat{f}(x)$ is non-increasing on $(\hat{x}^*, \infty)$.

It is also desirable that the proposed density estimate, $\hat{f}(\cdot)$, satisfies a set of asymptotic properties (e.g. consistency). Without loss of generality, we first consider densities with support $[0, 1]$, and then extend our methodology to more general supports.

We begin by considering a Bernstein polynomial of order $m - 1$ to estimate $f(x)$,

$$B_m(x, f) = \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right) \binom{m-1}{k-1} x^{k-1}(1-x)^{m-k}. \tag{2}$$

However, it is clear that $B_m(x, f)$, defined in (2), is not a proper density function. We therefore consider a re-scaled version of $B_m(x, f)$,

$$f_m(x) = \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right) m \binom{m-1}{k-1} x^{k-1}(1-x)^{m-k} \Big/ \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right), \tag{3}$$

which is a proper density function for any value m. This motivates us to consider the following class of density estimates,

$$f_m(x, \boldsymbol{\omega}) = \sum_{k=1}^{m} \omega_k \, f_b(x; k, m-k+1)\,, \tag{4}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)^T$ is a vector of weights of size m, and $f_b(\cdot)$ is the Beta density function with shape parameters k and $m-k+1$. Notice that $f_m(x, \boldsymbol{\omega})$ will obtain all the desired properties of $\hat{f}(\cdot)$ if the vector of weights, $\boldsymbol{\omega}$, satisfies the following constraints:

(a) $\omega_k \geq 0$, for $k = 1, \ldots, m$,

(b) $\sum_{k=1}^{m} \omega_k = 1$, and

(c) $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_{k^*} \geq \omega_{k^*+1} \geq \cdots \geq \omega_m$, for some $k^* \in \{1, 2, \ldots, m\}$, i.e. the ω_k's are non-decreasing for $k \leq k^*$ and non-increasing for $k \geq k^*$.

Constraint (a) implies property (i), as $\binom{m-1}{k-1} x^{k-1}(1-x)^{m-k}$ is non-negative for all values of x in $[0, 1]$. Integrating $f_m(x, \boldsymbol{\omega})$ over the interval $[0, 1]$ shows that property (ii) holds when constraint (b) is satisfied. Finally, Carnicer and Peña (1993) showed that a Bernstein polynomial basis is "optimally" shape-preserving, such that the unimodal shape constraint on the weights, (c), will result in a unimodal density estimate, $f_m(x, \boldsymbol{\omega})$, satisfying property (iii).

Our next goal is to estimate $\boldsymbol{\omega}$ and k^* for a given m, and then select m by some information theoretic and related criteria. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$ be the ECDF for some observed independent and identically distributed (i.i.d.) data $x_1, \ldots, x_n$ from $f(\cdot)$. One obvious estimation method is to simply set $\tilde{\omega}_k = F_n\left(\frac{k}{m-1}\right) - F_n\left(\frac{k-1}{m-1}\right)$ (Leblanc, 2010) and let $\hat{k}^* = \underset{1 \leq k \leq m-1}{\operatorname{argmax}} \, \tilde{\omega}_k$. Using the results established by Babu et al. (2002) it follows that $f_m(x, \tilde{\boldsymbol{\omega}})$ is consistent, with respect to the sup-norm, L_∞, for $f(x)$ as $m \to \infty$ and $n \to \infty$, such that $2 \leq m \leq (n/\log n)$.

It can also be shown that $\hat{x}^*_m = \underset{x \in [0,1]}{\operatorname{argmax}} \, f_m(x, \tilde{\boldsymbol{\omega}})$ is consistent for x^* as $m \to \infty$ and $n \to \infty$, satisfying the above conditions, using the continuous and monotonicity properties of $f(x)$ on $[0, x^*)$ and $(x^*, 1]$. Although the above estimate, $(\tilde{\boldsymbol{\omega}}, \hat{k}^*)$, is consistent for estimating $f_m(x, \boldsymbol{\omega})$, given by (4), $\tilde{\boldsymbol{\omega}}$ is not guaranteed to satisfy condition (c), and hence $f_m(x, \tilde{\boldsymbol{\omega}})$ is not necessarily a unimodal density satisfying property (iii).

It is well known that the Anderson-Darling test (Anderson and Darling, 1954) is one of the most powerful tests for detecting the functional form of a given density (Stephens, 1974). Motivated by this result, we propose the following criteria to estimate $\boldsymbol{\omega}$: find $\hat{\boldsymbol{\omega}}$ that minimizes

$$\sum_{i=1}^{n} \frac{n[F_n(x_i) - \tilde{F}_m(x_i, \boldsymbol{\omega})]^2}{(F_n(x_i) + \varepsilon_n)(1 + \varepsilon_n - F_n(x_i))}, \tag{5}$$

subject to the restrictions (a)-(c) with $k^* = \hat{k}^*$, and

$$\tilde{F}_n(x, \boldsymbol{\omega}) = \int_0^x f_m(u, \boldsymbol{\omega}) du = \sum_{k=1}^{m} \omega_k \, F_b\left(x; k, m - k + 1\right),$$

where $F_b(\cdot, k, m - k + 1)$ is the cdf of a Beta distribution with shape parameters k and $m - k + 1$. The small nudge factor, $\varepsilon_n = \frac{3}{8n}$, is added to avoid numerical instabilities, following the second-order corrections suggested by Anscombe and Aumann (1963).

It can be easily shown that the optimization problem in (5) can be solved by quadratic programming techniques subject to linear inequality constraints. More specifically, (5) can be written as: minimize

$$-\boldsymbol{b}\,\boldsymbol{\omega} + \frac{1}{2}\boldsymbol{\omega}^T A \boldsymbol{\omega} + c, \tag{6}$$

subject to $R\boldsymbol{\omega} \geq \boldsymbol{d}$, where $A, \boldsymbol{b}, c, R$, and $\boldsymbol{d}$ are given in the Appendix.

Our methodolgy can be extended to more general supports of $[a, b]$ using the simple linear transformation, $u = (x - a)/(b - a)$. When no information is known about the support of $f(\cdot)$, $a = x_{(1)} - s/\sqrt{n}$ and $b = x_{(n)} + s/\sqrt{n}$ provide reasonable bounds for the estimated density, where $x_{(1)}$ and $x_{(n)}$ are the first and last order statistics of the data, and s is the sample standard deviation. These bounds are motivated by the fact that $\Pr\left[X_{(1)} - \frac{s}{\sqrt{n}} \leq X_{n+1} \leq X_{(n)} - \frac{s}{\sqrt{n}}\right] \geq \Pr\left[X_{(1)} \leq X_{n+1} \leq X_{(n)}\right] = \frac{n-1}{n+1}$, with justification provided in the Appendix.

2.1 Methods for selecting the number of weights

2.1.1 Condition Number (CN)

Before solving the minimization problem, the optimal number of weights, m, must be selected. The square matrix A, from expression (6), must be positive definite, and the size of the matrix depends on m. If too many weights are selected the matrix no longer remains

positive definite due to numerical instabilities, and the minimization procedure cannot be completed. Ideally we would like to include a sufficient number of weights to properly estimate the density, while keeping A positive definite.

We present a novel procedure for selecting the number of weights by examining the condition number of A. Since A is a normal matrix the condition number is evaluated by, $\mathrm{CN}(m) = \left|\frac{\lambda_{max}(A)}{\lambda_{min}(A)}\right|$, where $\lambda_{max}(A)$ and $\lambda_{min}(A)$ are the maximum and minimum eigenvalues of A respectively. We select m by including the largest number of weights possible while still bounding $\log_{10}\mathrm{CN}(m)$ by $\sqrt{n}$.

2.1.2 AIC and BIC

We also present more traditional methods for selecting the number of weights, specifically, the Akaike information criterion (AIC) and Bayesian information criteria (BIC). In the realm of density estimation, these criteria are given by,

$$\mathrm{AIC}(m) = -2\sum_{i=1}^{n}\log[\hat{f}_m(x_i, \widehat{\boldsymbol{\omega}}_m)] + 2(m-1), \tag{7}$$

$$\mathrm{BIC}(m) = -2\sum_{i=1}^{n}\log[\hat{f}_m(x_i, \widehat{\boldsymbol{\omega}}_m)] + \log(n)(m-1), \tag{8}$$

where m is the number of weights, $\hat{f}_m$ is the estimated density using m weights, $\widehat{\boldsymbol{\omega}}_m$ is the vector of estimated weights, and x_i for $i = 1, \ldots, n$ are the observations. The degrees of freedom for the estimated density is $m-1$ since our procedure estimates the weights under one dimension constraint.

A variation of Theorem 3.1 in Babu et al. (2002) shows that $\hat{f}_m$ will converge uniformly to f for $2 \leq m \leq (n/\log n)$ as $n \to \infty$, so we implement both methods by estimating the density with m weights ranging from 2 to $\lceil n/\log n \rceil$. The AIC or BIC is calculated for each fit, then the density estimate with the lowest AIC or BIC is selected as the best estimate.

2.2 Comparison of criterion for selecting the number of weights

We compare these three methods of selecting m by performing a small simulation study. We generate data from a mixture of Beta densities with a fixed vector of weights, which follows the required format of our density estimation model. We then generate density estimates using the AIC, BIC, and CN criterion to select m. Our goal is to find the method

which provides the best density estimate in terms of a low root mean integrated squared error (RMISE) and selects the true number of weights of the Beta mixture. The RMISE is approximated as follows,

$$\text{RMISE} = \text{E}\Big[\|\hat{f}_{\widehat{m}} - f\|_2\Big] \approx \frac{1}{N}\sum_{\ell=1}^{N}\sqrt{d\sum_{j=1}^{J}\Big[\hat{f}_{\widehat{m}}^{(\ell)}(x_j) - f(x_j)\Big]^2}, \tag{9}$$

where $J = 100$, $\Pr[x_1 \leq X \leq x_J] = 0.999$, and $d = (x_j - x_{j-1}) = \frac{x_J - x_1}{J-1}$ for all $j \geq 2$. In the above expression, $\hat{f}_{\widehat{m}}^{(\ell)}$ denotes the estimated density at the ℓ^{th} sample for $\ell = 1, \ldots, N$ with $\widehat{m}$ chosen by one of the three criteria.

We examine a symmetric and right-skewed beta mixture each with 7 weights (i.e. $m = 7$). The weight vectors are fixed to be $\boldsymbol{\omega}_1 = \{0.05, 0.1, 0.2, 0.3, 0.2, 0.1, 0.05\}$ for the symmetric distribution, and $\boldsymbol{\omega}_2 = \{0.1, 0.4, 0.25, 0.15, 0.05, 0.03, 0.02\}$ for the right-skewed distribution. We generate $N = 1000$ Monte Carlo (MC) samples of size $n = 15, 30, 100$, and 500 for each beta mixture. Table 1 displays the MC estimated RMISE values as well as the average estimate of m across the MC samples. Statistically significant lower RMISE values, based on a paired t-test, are in bold.

	Symmetric				Right-Skewed			
$n =$	15	30	100	500	15	30	100	500
	RMISE							
CN	**3.631**	2.718	1.772	0.999	3.943	**2.807**	1.914	1.051
	(0.049)	(0.029)	(0.018)	(0.008)	(0.049)	(0.0340)	(0.020)	(0.009)
AIC	3.673	**2.542**	1.347	**0.749**	3.944	3.080	**1.871**	**0.712**
	(0.047)	(0.033)	(0.017)	(0.007)	(0.046)	(0.028)	(0.025)	(0.010)
BIC	3.680	**2.548**	**1.293**	**0.752**	3.943	3.102	2.302	0.773
	(0.047)	(0.033)	(0.017)	(0.006)	(0.045)	(0.026)	(0.023)	(0.014)
	Average Number of Weights							
CN	3.882	5.934	12.235	24.223	3.799	5.744	11.867	24.082
	(0.010)	(0.008)	(0.014)	(0.046)	(0.013)	(0.014)	(0.016)	(0.043)
AIC	3.030	3.125	3.411	3.977	3.031	3.223	4.429	5.226
	(0.006)	(0.012)	(0.026)	(0.034)	(0.007)	(0.022)	(0.037)	(0.022)
BIC	3.010	3.046	3.085	3.203	3.019	3.034	3.571	4.883
	(0.003)	(0.007)	(0.011)	(0.018)	(0.005)	(0.008)	(0.027)	(0.018)

Table 1: RMISE values $\times 10$ and average estimate of m for 1000 MC samples with MC standard errors displayed in parentheses

The AIC method has the lowest average RMISE in most situations. Unfortunately, none

of the methods have average number of weight estimates that are close to 7. The CN method appears to select the number of weights at around $\sqrt{n}$, which is much larger than 7 for the cases of $n = 100$ and 500. Both the AIC and BIC under estimate m for all sample sizes with the AIC having slightly larger estimates. Overall it appears the CN criterion is a better option for samples with fewer observations while the AIC performs the best for larger samples of size 100 and 500.

2.3 R function: umd

To enable users to easily implement our method we provide R code for the function `umd`. This function has six inputs:

- `data`, a vector of univariate data
- `fix.lower`, a known lower bound for the support of the data. For example, lifetime data typically has a support with a lower bound of 0 as negative time values are usually nonsensical. If no value is given, the function defaults to the value $x_{(1)} - s/\sqrt{n}$ as outlined in Section 2. To avoid using bounds which are extremely distant from the provided data, the function uses $x_{(1)} - s/\sqrt{n}$ as the lower bound of the support if it is larger than the given `fix.lower` value.
- `fix.upper`, a known upper bound for the support of the data. If no value is given, the function defaults to the value $x_{(n)} + s/\sqrt{n}$. To avoid using bounds which are extremely distant from the provided data, the function uses $x_{(n)} + s/\sqrt{n}$ as the upper bound of the support if it is less than the given `fix.upper` value.
- `crit`, the type of criterion to use for selecting the number of weights. The user can input either `'AIC'`, `'BIC'`, or `'CN'`, corresponding to each of the three criterion presented in Section 2.1. If no value is given, the function defaults to `'CN'`.
- `m`, the number of weights the user wants the beta mixture to contain. In general, it is more practical for the user to leave this parameter empty and allow the AIC, BIC, or CN criterion select the optimal number of weights. The function will calculate the optimal number of weights regardless if this parameter is given by the user.

- `warning`, a boolean parameter with default value `TRUE`. If set to `TRUE`, the `umd` function will return a warning if the user sets the `m` parameter to a value that is not equal to the optimal number of weights selected by the given criterion. No warnings will be returned if this parameter is set to `FALSE`.

The `umd` function returns an object with the four typical distribution functions, the number of weights (`m.hat`), and a vector of the estimated weights (`weights`). The four distribution functions returned are `dumd`, `pumd`, `qumd`, and `rumd`, corresponding to the density, distribution, quantile, and random sample generator functions available for most distributions in R. Note that a user can easily obtain an estimate of the mode of the population distribution by inputting the returned estimated density function into a one-dimensional optimization function. Our code is available from the first author upon request.

3 Simulation Study

We demonstrate the performance of our density estimation technique through multiple simulation studies. We first assess its performance estimating densities with three common supports: $(-\infty, \infty)$, $[0, \infty)$, and $[-1, 1]$. Densities with left skewed, right skewed, and symmetric shapes are chosen for each support with the exception of a symmetric density for the $[0, \infty)$ case. In this section we present the results using the CN criterion to select m; we include the results for the AIC method in the Appendix.

3.1 Density estimation with support $(-\infty, \infty)$

On the $(-\infty, \infty)$ support we select a normal density with $\mu = 0$ and $\sigma^2 = 4$. We skew this normal distribution by multiplying the density function by the scaled cdf, $F^*(cx)$. In other words the true density is given by,

$$f(x) = \frac{1}{8\pi} \exp\left\{-\frac{x^2}{8}\right\} \int_{-\infty}^{cx} \exp\left\{-\frac{t^2}{8}\right\} dt. \tag{10}$$

We set $c = 0, 2.2$, and -2.2 resulting in symmetric, left skewed, and right skewed distributions with skewness factors of 0, 0.5, and -0.5 respectively.

3.2 Density estimation with support $[0, \infty)$

A gamma distribution with $\alpha = 16$, $\beta = 0.5$, mean $\alpha\beta = 8$, and skewness $= 0.5$ is used for a right skewed distribution on the $[0, \infty)$ support. A Beta distribution, with $\alpha = 10$ and

$\beta = 5.5$, scaled by 25 is selected to simulate a left skewed density on that same support. The resulting expression for the Beta density function is,

$$f(x) = \frac{1}{25}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{x}{25}\right)^{(\alpha-1)}\left(1-\frac{x}{25}\right)^{(\beta-1)},\ x \in (0,25).$$

This distribution has a skewness factor of approximately -0.5 as well. When estimating each of these distributions we assume it is known that the data cannot have values less than 0, so the `fix.lower` parameter is set to 0.

3.3 Density estimation with support $[-1, 1]$

We select the less traditional half circle density on the $[-1, 1]$ support. We create this proper density by rescaling the half circle function. This density is skewed by multiplying by the scaled cdf, resulting in the following density function,

$$f(x) = \frac{4}{\pi}\sqrt{1-x^2}\left[\frac{1}{2}\arcsin(cx) + \frac{cx}{2}\sqrt{1-(cx)^2} + \pi/4\right],\ x \in [-1,1]. \tag{11}$$

Symmetric and skewed distributions are obtained by setting $c = 0, 1$, and -1. The skewed distributions have skewness factors of approximately 0.4 and -0.4. The probability integral transformation is used to generate random samples from the symmetric distribution. We use rejection sampling with proposal density proportional to $2\sqrt{1-x^2}$ to sample from the skewed distributions. We assume the true densities' support of -1 to 1 is known when applying our density estimation method, so the `fix.lower` and `fix.upper` parameters are set to -1 and 1 respectively.

3.4 Evaluation method

We evaluate the performance of the density estimates using the RMISE, outlined in equation (9), and the estimated expected value of two other popular functional norms:

$$\mathrm{ML}_1\mathrm{E} = \mathrm{E}\left[\|\hat{f}_{\widehat{m}} - f\|_1\right] \approx \frac{1}{N}\sum_{\ell=1}^{N}\left[d\sum_{j=1}^{J}\left|\hat{f}_{\widehat{m}}^{(\ell)}(x_j) - f(x_j)\right|\right],$$

$$\mathrm{ML}_\infty\mathrm{E} = \mathrm{E}\left[\|\hat{f}_{\widehat{m}} - f\|_\infty\right] \approx \frac{1}{N}\sum_{\ell=1}^{N}\max_{x_1 \le x_j \le x_J}\left|\hat{f}_{\widehat{m}}^{(\ell)}(x_j) - f(x_j)\right|,$$

where $J = 100$, and x_1 and x_J are chosen such that based on the true density of X, $\Pr[x_1 \le X \le x_J] = 0.999$, and $d = (x_j - x_{j-1}) = \frac{x_J - x_1}{J-1}$ for all $j \ge 2$.

3.5 Results

We generate $N = 1000$ MC samples of size $n = 15, 30$, and 100. Table 3 displays the results of the average expected norm values across the MC samples, the MC standard errors are displayed in parentheses. All values are multiplied by 10.

		Left-Skewed			Symmetric			Right-Skewed		
		ML_1E	RMISE	$ML_\infty E$	ML_1E	RMISE	$ML_\infty E$	ML_1E	RMISE	$ML_\infty E$
$(-\infty, \infty)$	$n = 15$	3.683	1.622	1.296	3.532	1.299	0.878	3.671	1.715	1.591
		(0.038)	(0.017)	(0.017)	(0.037)	(0.014)	(0.012)	(0.038)	(0.019)	(0.023)
	$n = 30$	2.501	1.085	0.846	2.437	0.870	0.540	2.509	1.119	0.934
		(0.029)	(0.013)	(0.011)	(0.028)	(0.010)	(0.007)	(0.028)	(0.013)	(0.014)
	$n = 100$	1.425	0.606	0.450	1.412	0.494	0.297	1.451	0.622	0.472
		(0.016)	(0.007)	(0.006)	(0.016)	(0.006)	(0.004)	(0.017)	(0.008)	(0.006)
$[0, \infty)$	$n = 15$	3.553	1.316	0.882	-	-	-	3.508	1.347	1.020
		(0.041)	(0.015)	(0.012)	-	-	-	(0.036)	(0.015)	(0.014)
	$n = 30$	2.335	0.848	0.545	-	-	-	2.466	0.912	0.625
		(0.028)	(0.010)	(0.007)	-	-	-	(0.028)	(0.011)	(0.009)
	$n = 100$	1.468	0.528	0.325	-	-	-	1.451	0.520	0.332
		(0.016)	(0.006)	(0.004)	-	-	-	(0.017)	(0.006)	(0.004)
$[-1, 1]$	$n = 15$	3.168	2.826	5.508	2.961	2.554	4.690	2.911	2.820	7.029
		(0.042)	(0.035)	(0.072)	(0.046)	(0.035)	(0.058)	(0.041)	(0.038)	(0.102)
	$n = 30$	2.249	2.038	4.098	2.253	1.939	3.641	2.260	2.127	5.274
		(0.027)	(0.023)	(0.060)	(0.030)	(0.023)	(0.041)	(0.029)	(0.026)	(0.080)
	$n = 100$	1.502	1.353	2.793	1.393	1.209	2.466	1.481	1.374	3.425
		(0.016)	(0.014)	(0.038)	(0.017)	(0.013)	(0.027)	(0.016)	(0.014)	(0.049)

Table 2: Estimated expected value of functional norms $\times 10$ using the CN criterion across 1000 MC samples for left-skewed, symmetric, and right-skewed distributions on supports of $(-\infty, \infty)$, $[0, \infty)$, and $[-1, 1]$ with samples of size 15, 30, and 100. MC standard errors are displayed in parentheses.

Numerically it can be seen that our method performs best when estimating symmetric densities, and the half circle density is the most difficult to estimate. As expected the estimations improve as the sample size, n, increases.

Figure 5 displays plots of the average density estimates over the 1000 MC samples for each sample size. Visually, our method appears to accurately estimate each density. It does struggle for small samples of size $n = 15$, but the estimates drastically improve for larger samples of $n = 50$ and $n = 100$. It can also be seen that for the larger samples our method always accurately captures the location of the true mode.

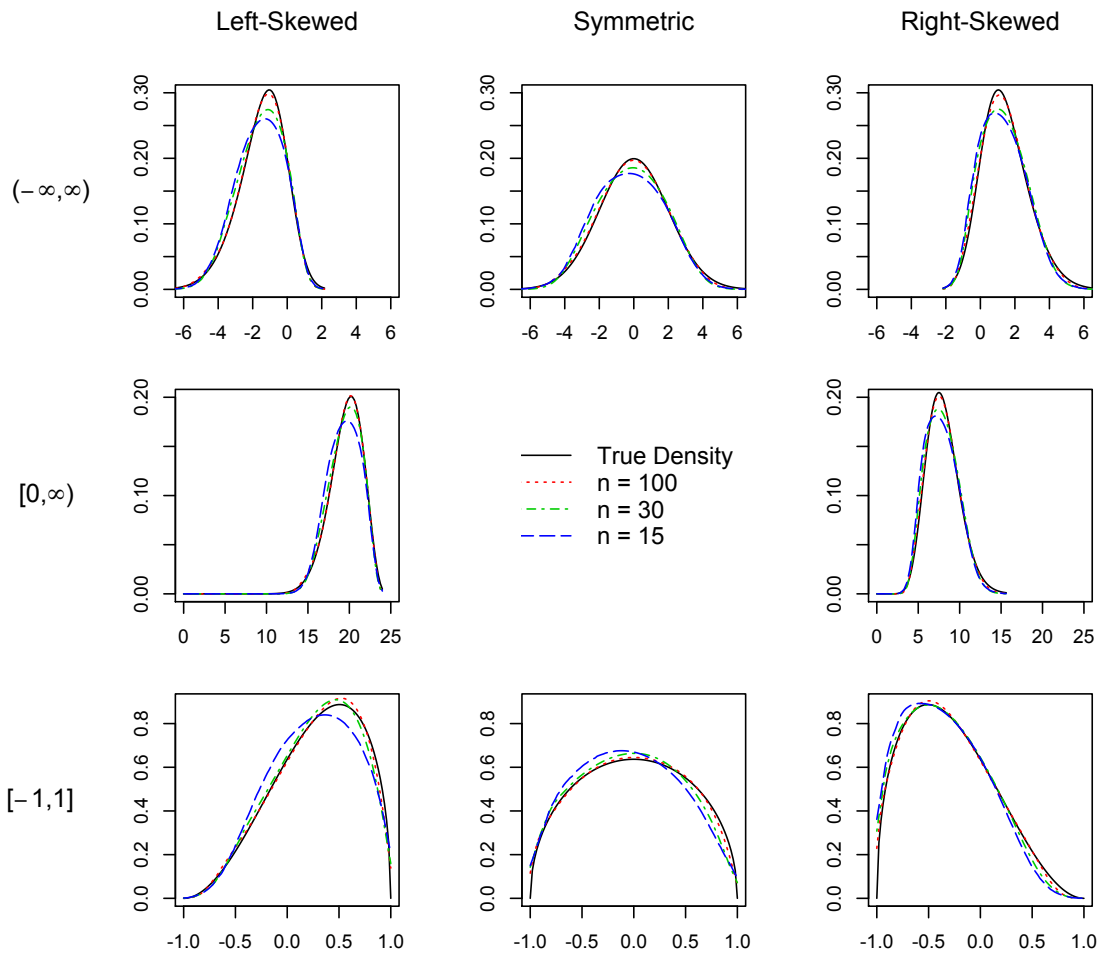

Figure 1: Plots of average estimated densities using the CN criterion across 1000 MC samples for samples of size 15 (dashed line), 30 (dot-dash line), and 100 (dotted line) along with the true density curve (solid line) for each support and type of skewness.

3.6 Outliers

One concern is how well our method performs when the data is contaminated with outliers. We explore this by generating data from a $\mathcal{N}(0,1)$ density mixed with a $\mathcal{N}(4,0.5)$ density. We set the mixture proportion to 95%, resulting in the following normal mixture,

$$p\,\mathcal{N}(0,1) + (1-p)\,\mathcal{N}(4,0.5), \tag{12}$$

where $p \sim$ bernoulli(0.95).

We generate 1000 MC samples from the mixture distribution with sizes $n = 15, 30$, and 100. Figure 2 displays the average density estimates across the 1000 MC samples. Assuming

the data is truly from a $\mathcal{N}(0,1)$ distribution, the 5% contamination does not completely compromise the density estimate. For the $n = 100$ sample, the density estimate still properly locates the mode at 0 and there is only a slightly heavier tail on the side of the contaminating distribution.

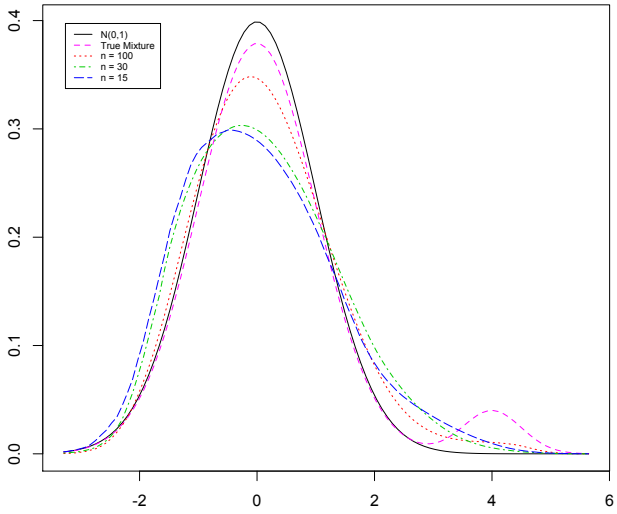

Figure 2: Plots of the average estimated densities using the CN criterion for 1000 MC samples of size 15 (long dashed line), 30 (dot-dash line), and 100 (dotted line) along with the true mixture distribution (short dashed line) and standard normal distribution (solid line)

4 Real Data Examples

4.1 Suicide Data

We first demonstrate the application of method to a data set presented in Silverman (1986) of the duration in days of psychiatric treatments for 86 control group patients from a study of suicide risks. This data set is chosen because it appears to have an exponential distribution shape with the mode on the boundary of the support, which is a special case of unimodal distributions. Since these data are measured in number of days we estimate the density with a fixed minimum of 0 for the support (i.e. `fix.lower` $= 0$), as values of negative days are not practical. The AIC method selects 7 as the optimal number of weights whereas the CN criterion selects 10. Regardless of this difference in number of weights, both estimates are almost identical so only the CN criterion estimate is reported.

The data has an exponential shape, so we compare our density estimate to simply using an exponential distribution with scale parameter equal to the mle of the data, 122. Figure 3 displays each of the density estimates on top of a histogram of the data. Both density estimates appear to be reasonable estimates of the data, and it can be seen that our estimate does not have any difficulty estimating a distribution with the mode on the boundary. A Kolmogorov-Smirnov test is conducted comparing the ECDF of the data to each of the estimated distribution functions; in both cases the test failed to disprove the null hypothesis. Our method, however, resulted in a test statistic of 0.055 with p-value, 0.957, while the exponential distribution had a test statistic of 0.0954 and p-value, 0.414, indicating that our method provides a superior density estimate of the data.

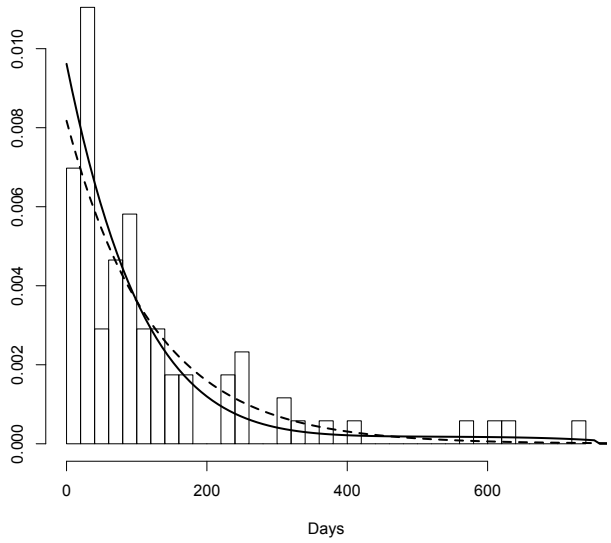

Figure 3: Histogram of duration in days of psychiatric treatments for 86 control group patients, CN criterion unimodal Bernstein polynomial density estimate (solid line), and exponential distribution with sample mean (dashed line).

4.2 S&P 500 Log Returns Data Set

We also explore a data set of the Standard and Poor's 500 daily log returns for the years 2008 to 2010. This data was acquired from the Yahoo Finance webpage (`http://finance.yahoo.com/q?s=%5EGSPC`, accessed: 11/15/2012). The S&P 500 is one of the most commonly followed stock market indices and is based on common stock prices of 500 top publicly

traded American companies, which are determined by S&P. It is considered one of the best indicators of the state of the market and economy (Markowitz and Nilufer, 1996). General finance models assume that returns are normally distributed or follow some other popular unimodal parametric distribution (Fergusson and Platen, 2006). These parametric assumptions, however, may not always be appropriate, resulting in inaccurate estimates of the underlying density of the log returns (Fergusson and Platen, 2006).

We apply our density estimation method to the S&P 500 data using both the AIC and CN criterion to select the number of weights. This data set features 756 log return values in total. The AIC and CN criterion selected 55 and 19 weights respectively. We also investigated a normal distribution with mean and standard deviation parameters set equal to the corresponding sample statistics of the data. A Kolmogorov-Smirnov test was carried out for the AIC, CN, and normal distribution, but the null hypothesis was rejected in all three cases. We remedy this issue by manually increasing the number of weights until the Kolmogorov-Smirnov test returns a p-value greater than 0.5. We find that 117 is the smallest m value that satisfies this condition resulting in a p-value of 0.502. Figure 4 displays a histogram of the data, our density estimates for each criterion and m = 117, and the normal distribution. It can be seen that the CN criterion density estimate does not properly capture the extremely peaked shape of the histogram. While the AIC density estimate captures the general shape of the histogram it has slightly thinner tails than the histogram indicates. The density estimate with 117 weights, however, both captures the peaked nature of the histogram and has properly weighted tails. This shows that in certain cases users may find it useful to manually explore alternate values of $\widehat{m}$ other than those selected by the AIC, BIC, and CN criterion. Overall, our method provides a superior density estimate than simply assuming the log returns follow a normal distribution.

5 Discussion

We introduce a unimodal density estimation technique using Bernstein polynomials presented as a mixture of Beta kernels. The weight values for the Beta mixture are estimated using quadratic programming techniques. We also present three different criterion for determining the number of mixture weights. R code and clearly defined instructions are provided upon

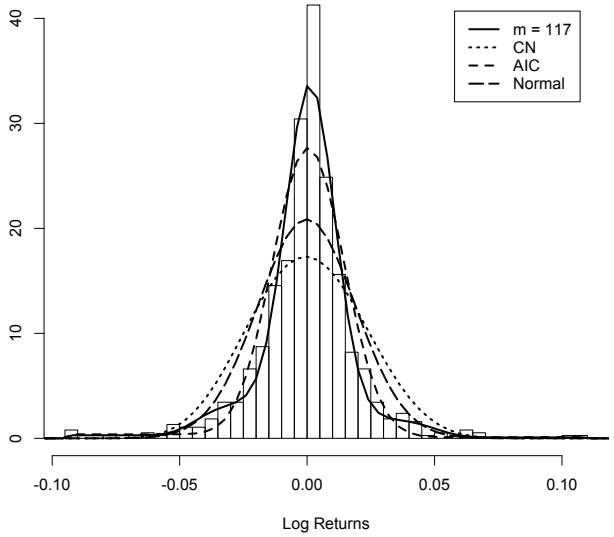

Figure 4: Histogram of 756 S&P 500 daily log return values from 2008 to 2010 along with the Bernstein polynomial density estimate with $\widehat{m} = 117$ weights (solid line), AIC Bernstein polynomial density estimate (short dashed line), CN criterion Bernstein polynomial density estimate (dotted line), and normal distribution with sample mean and sample standard deviation (long dashed line).

request for easy implementation of our methodology. Simulation studies and two real data examples demonstrate the effectiveness of our approach and provide insight as to when each criterion for selecting the number of weights is most useful.

One limitation of our approach is that our Beta mixture may be too constricting to properly estimate the most oddly shaped distributions. In order to account for the odd shapes in the underlying distribution, the number of weights must increase, resulting in a larger number of parameters that must be estimated. Further studies on the selection of the number of weights are needed. One possibility, which we utilize in the S&P 500 data example, is the use of a Kolmogorov-Smirnov or some other nonparametric goodness of fit test. One could determine the number of weights by ensuring that the test statistic or p-value achieve a specific value.

A future area of interest is expanding this methodology to multivariate unimodal distributions and log-concave densities. The log-concave densities are necessarily unimodal but the reverse is not true in general.

References

Anderson, T. and Darling, D. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, 49(268):765–769.

Anscombe, F. and Aumann, R. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205.

Babu, G., Canty, A., and Chaubey, Y. (2002). Application of bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105:377–392.

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*, 3:970–981.

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics*, 10:786–806.

Carnicer, J. M. and Peña, J. M. (1993). Shape preserving representations and optimality of the bernstein basis. *Adv. Comput. Math.*, pages 173–196.

Cheng, M. Y., Gasser, T., and Hall, P. (1999). Nonparametric density estimation under unimodality and monotonicity constraints. *Journal of Computation and Graphical Statistics*, 8:1–21.

Choi, E. and Hall, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika*, 86:941–947.

Choi, E., Hall, P., and Rousson, V. (2000). Data sharpening for bias reduction in nonparametric regression. *The Annals of Statistics*, 28:1339–1355.

Fergusson, K. and Platen, E. (2006). On the distributional characterization of daily log-returns of a world stock index. *Applied Mathematical Finance*, 13(1):19–38.

Fougéres, A. (1997). Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387.

Ghosal, S. (2001). Convergence rates for density estimation with bernstein polynomials. *The Annals of Statistics*, 29:1264–1280.

Grenander, U. (1956). On the theory of mortality measurement. *Skand. Aktuarietidskr*, 39:125–153.

Hall, P. and Huang, L.-S. (2002). Unimodal density estimation using kernel methods. *Statistica Sinica*, 12:965–990.

Hall, P. and Kang, K.-H. (2005). Unimodal kernel density estimation by data sharpening. *Statistica Sinica*, 15:73–98.

Hardy, G., Littlewood, J., and Pólya, G. (1952). *Inequalities*. Cambridge University Press, Cambridge.

Kakizawa, Y. (2004). Bernstein polynomial probability density estimation. *Nonparametric Statistics*, 16:709–729.

Leblanc, A. (2010). A bias-reduced approach to density estimation using bernstein polynomials. *Journal of Nonparametric Statistics*, 22:459–475.

Lorentz, G. (1986). *Bernstein Polynomials*. New York: Chelsea Publishing.

Markowitz, H. M. and Nilufer (1996). The likelihood of various stock market return distributions, part 1: Principles of inference. *Journal of Risk and Uncertainty*, 13:207–219.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Petrone, S. (1999). Bayesian density estimation using bernstein polynomial posteriors. *Canadian Journal of Statistics*, 27:105–126.

Petrone, S. and Wasserman, L. (2002). Consistency of bernstein polynomial posteriors. *Journal of the Royal Statistical Society, Series B*, 64:79–100.

Silverman, B. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society*, 43:97–99.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Stephens, M. (1974). Components of goodness-of-fit statistics. *Annales De L Institut Henri Poincare Section B-Calcul des Probabilites et Statistique*, 10:37–54.

Vitale, R. (1975). A bernstein polynomial approach to density function estimation. *Statistical Inference and Related Topics*, 2:87–99.

Wegman, E. (1972). Nonparametric probability density estimation: I. a summary of available methods. *Technometrics*, 14(3):533–546.

Wolters, M. A. (2012). A greedy algorithm for unimodal kernel density estimation by data sharpening. *Journal of Statistical Software*, 47.

Appendix

I. *Details of the decomposition of expression (5) and linear contraints*:

We begin by expressing the minimization problem in (5) as,

$$(\boldsymbol{f} - B\boldsymbol{\omega})^T L (\boldsymbol{f} - B\boldsymbol{\omega}), \tag{13}$$

where $B_{n\times m} = ((F_b(x_i, j, m-j+1)))$, $\boldsymbol{f}_{n\times 1} = (F_n(x_i))$, and
$L_{n\times n} = \text{Diag}\left(n\left[\left(F_n(x_i) + \frac{3}{8n}\right)\left(1 + \frac{3}{8n} - F_n(x_i)\right)\right]^{-1}\right)$, for $1 \le i \le n$ and $1 \le j \le m$.

Expression (13) expands to $\boldsymbol{f}^T L\boldsymbol{f} - 2\boldsymbol{f}^T LB\boldsymbol{\omega} + \boldsymbol{\omega}^T B^T LB\boldsymbol{\omega}$ and now takes the exact form of (6), with $\boldsymbol{b} = 2\boldsymbol{f}^T LB$, $A = B^T LB$, and $c = \boldsymbol{f}^T L\boldsymbol{f}$.

The constraints on $\boldsymbol{\omega}$ can be expressed in matrix form as follows:

(a) $\mathbf{1}^T_{1\times m}\,\boldsymbol{\omega}_{m\times 1} = 1$,

(b) $\boldsymbol{I}_{m\times m}\,\boldsymbol{\omega}_{m\times 1} \ge \mathbf{0}_{m\times 1}$,

(c) $C_{(m-1)\times m}\boldsymbol{\omega}_{m\times 1} \geq \mathbf{0}_{(m-1)\times 1}$, where $C_{(m-1)\times m} = ((\, c_{ij}\,))$, such that,

$$c_{ij} = \begin{cases} -1, & i = j < \hat{k}^* \text{ or } i+1 = j \geq \hat{k}^*, \\ 1, & i+1 = j < \hat{k}^* \text{ or } i = j \geq \hat{k}^*, \\ 0, & j < i \text{ or } j \geq i+1, \end{cases}$$

for $1 \leq i \leq (m-1)$ and $1 \leq j \leq m$. These constraints can then be combined into a single matrix and vector, $R^T = \left(\mathbf{1}_{m\times 1},\ \boldsymbol{I}_{m\times m},\ C^T_{m\times(m-1)}\right)_{m\times 2m}$ and $\boldsymbol{d} = (1,\, 0,\, 0,\, \ldots, 0\,)_{2m\times 1}$ with the constraint for the first element of $\boldsymbol{d}$ set strictly to equality.

II. *Proof that* $Pr\left[X_{(1)} - \frac{s}{\sqrt{n}} \leq X_{n+1} \leq X_{(n)} - \frac{s}{\sqrt{n}}\right] \geq Pr\left[X_{(1)} \leq X_{n+1} \leq X_{(n)}\right] = \frac{n-1}{n+1}$ *for* $X_1, \ldots, X_n, X_{n+1} \overset{i.i.d.}{\sim} G$*, where* G *is any distribution function.*

We begin by showing that $\Pr\left[X_{(1)} \leq X_{n+1} \leq X_{(n)}\right] = \frac{n-1}{n+1}$. We condition on $X_{(1)}$ and $X_{(n)}$, giving us,

$$\mathrm{E}\big[\Pr\left(X_{(1)} \leq X_{n+1} \leq X_{(n)} \mid X_{(1)}, X_{(n)}\right)\big],$$

which reduces to,

$$\mathrm{E}\big[\Pr\left(X_{n+1} \leq X_{(n)} \mid X_{(n)}\right) - \Pr\left(X_{n+1} \leq X_{(1)} \mid X_{(1)}\right)\big] = \mathrm{E}\big[\, G(X_{(n)}) - G(X_{(1)})\big].$$

It is known that $G(X_{(j)})$ follows a Beta distribution with shape parameters j and $n-j+1$. With this knowledge we can easily complete the remaining steps of the proof,

$$\mathrm{E}\big[\, G(X_{(n)}) - G(X_{(1)})\big] = \mathrm{E}\big[\, G(X_{(n)})\big] - \mathrm{E}\big[\, G(X_{(1)})\big] = \frac{n}{n+1} - \frac{1}{n+1} = \frac{n-1}{n+1}.$$

Finally, since the interval $\left(X_{(1)}, X_{(n)}\right)$ is contained in $\left(X_{(1)} - \frac{s}{\sqrt{n}}, X_{(n)} + \frac{s}{\sqrt{n}}\right)$, we can conclude that $\Pr\left[X_{(1)} - \frac{s}{\sqrt{n}} \leq X_{n+1} \leq X_{(n)} - \frac{s}{\sqrt{n}}\right] \geq \frac{n-1}{n+1}$.

III. *Additional Tables and Figures for Section 3.5*

		Left-Skewed			Symmetric			Right-Skewed		
		ML_1E	RMISE	$ML_\infty E$	ML_1E	RMISE	$ML_\infty E$	ML_1E	RMISE	$ML_\infty E$
$(-\infty, \infty)$	$n = 15$	4.640	2.063	1.720	4.460	1.660	1.188	4.385	2.067	2.028
		(0.034)	(0.015)	(0.017)	(0.035)	(0.013)	(0.013)	(0.033)	(0.015)	(0.020)
	$n = 30$	3.414	1.485	1.193	3.140	1.126	0.720	3.463	1.578	1.436
		(0.037)	(0.016)	(0.017)	(0.034)	(0.013)	(0.011)	(0.033)	(0.016)	(0.021)
	$n = 100$	1.591	0.687	0.514	1.745	0.608	0.344	1.647	0.712	0.533
		(0.021)	(0.009)	(0.006)	(0.023)	(0.008)	(0.004)	(0.022)	(0.010)	(0.007)
$[0, \infty)$	$n = 15$	4.467	1.671	1.160	-	-	-	4.333	1.680	1.341
		(0.033)	(0.012)	(0.012)	-	-	-	(0.032)	(0.012)	(0.013)
	$n = 30$	3.127	1.143	0.778	-	-	-	3.455	1.300	0.976
		(0.035)	(0.013)	(0.012)	-	-	-	(0.035)	(0.014)	(0.015)
	$n = 100$	1.432	0.514	0.314	-	-	-	1.581	0.567	0.351
		(0.018)	(0.007)	(0.004)	-	-	-	(0.021)	(0.007)	(0.004)
$[-1, 1]$	$n = 15$	3.464	3.164	6.732	2.993	2.696	5.541	3.034	3.038	7.988
		(0.032)	(0.027)	(0.062)	(0.036)	(0.028)	(0.050)	(0.028)	(0.025)	(0.081)
	$n = 30$	2.762	2.539	6.199	2.421	2.157	4.680	2.568	2.565	7.325
		(0.020)	(0.017)	(0.074)	(0.022)	(0.019)	(0.051)	(0.017)	(0.018)	(0.084)
	$n = 100$	1.435	1.343	3.723	1.175	1.060	2.614	1.443	1.396	4.201
		(0.020)	(0.019)	(0.072)	(0.020)	(0.016)	(0.036)	(0.021)	(0.021)	(0.082)

Table 3: Estimated expected value of functional norms $\times 10$ using the AIC across 1000 MC samples for left-skewed, symmetric, and right-skewed distributions on supports of $(-\infty, \infty)$, $[0, \infty)$, and $[-1, 1]$ with samples of size 15, 30, and 100. MC standard errors are displayed in parentheses.

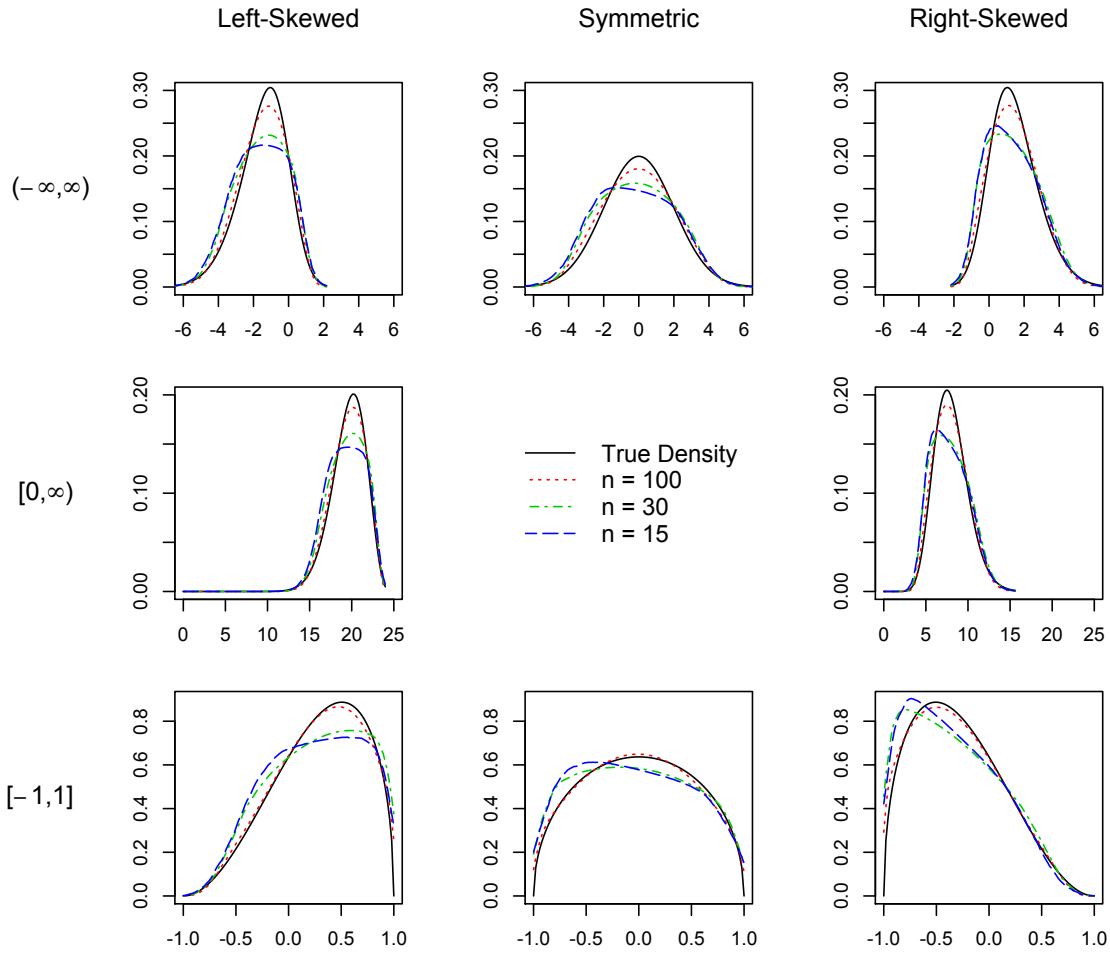

Figure 5: Plots of average estimated densities using the AIC criterion across 1000 MC samples for samples of size 15 (dashed line), 30 (dot-dash line), and 100 (dotted line) along with the true density curve (solid line) for each support and type of skewness.