Semiparametric Bivariate Density Estimation with Irregularly Truncated Data

Chad M. Schafer*

September 5, 2006

Abstract

This work develops an estimator for the bivariate density given a sample of data truncated to a non-rectangular region. Such inference problems occur in various fields; the motivating application here was a problem in astronomy. The approach is semiparametric, combining a nonparametric local likelihood density estimator with a simple parametric form to account for the dependence of the two random variables. Large sample theory for M-estimators is utilized to approximate the distribution for the estimator. A method is described for approximating the integrated mean squared error of the estimator; smoothing

*Chad Schafer is Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (email: cschafer@stat.cmu.edu). Research supported by NSF Grants #0434343 and #0240019. The author acknowledges Chris Genovese and Larry Wasserman for many helpful discussions. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

parameters can be selected to minimize this quantity. Results are described from the analysis of data from the measurements of quasars. A Fortran implementation is available, along with an R wrapper function.

Keywords: Truncated data; semiparametric method; bivariate density estimation; local modelling; M-estimator; boundary effect.

1 Introduction

Given the truncated data shown in the scatter plot of Figure 1, how could one estimate the bivariate density over (a) the irregular observable region, and (b) the entire region? It is clear that a priori assumptions must be made in order to infer the nature of the density. Answering question (b) requires strong assumptions, which in some cases (e.g. a physical model) may be warranted, but question (a) is itself challenging because of significant boundary effects. Scott and Wand (1991) demonstrated the feasibility of using kernel density estimators for high-dimensional problems, but standard kernel methods are inadequate with such truncation. We present a semiparametric approach which allows one to place minimal assumptions on the form of the bivariate density and greatly diminish artifacts due to truncation and an irregular boundary.

This inference problem appears in astronomy. Petrosian (1992) gives an excellent overview of the motivation and history; we will summarize this background here. The measured redshift of an astronomical object (e.g. galaxy or quasar) is a proxy for its distance from the observer since the increase in the wavelength of light is related to how far it has traveled. It is often of interest to determine the distribution of some other characteristic of the object under study as a function of redshift. This leads to data sets such as that shown in Figure 1. Each of the 11,242 dots shown in the scatter plot represents one quasar observed by the Sloan Digital Sky Survey (SDSS)

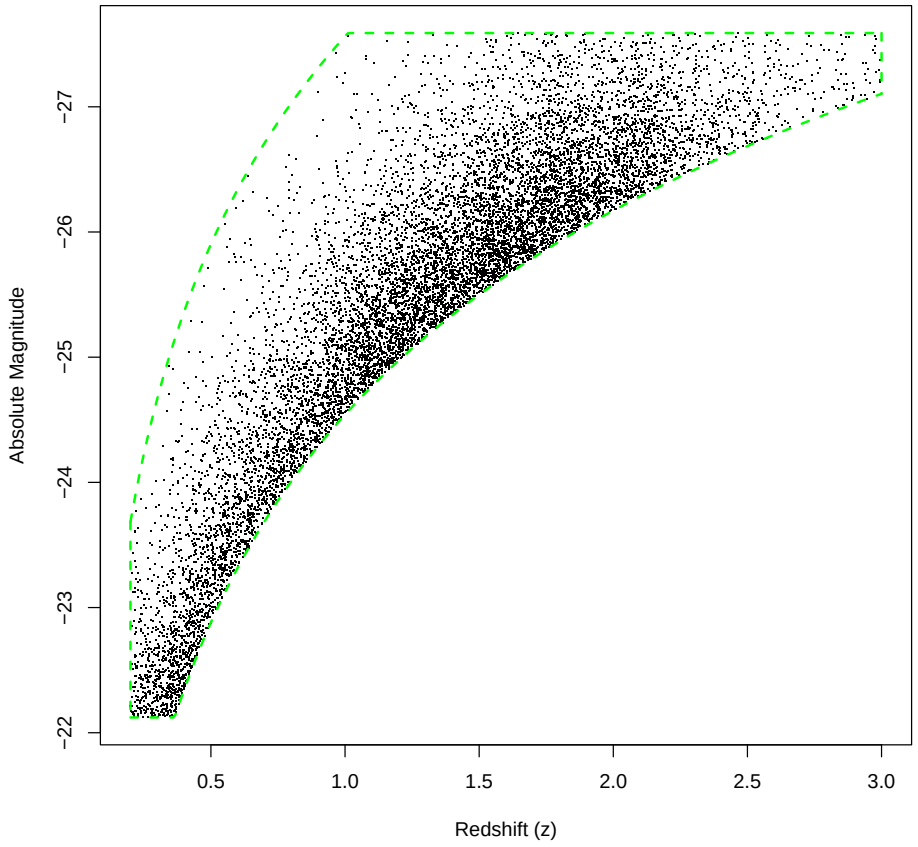

Figure 1: Quasar data from the Sloan Digital Sky Survey. There are 11,242 quasars in this sample after truncation to the region indicated by the dashed line. The vertical axis is inverted because brightness increases as absolute magnitude decreases.

(Schneider, 2003). (We are currently using Data Release 2.)

Along with the redshift of each quasar, the survey measured the apparent luminosities, i.e. the observed brightness of each object. The apparent luminosity is translated into an absolute luminosity (L), the brightness of the object at the source. (This transformation requires that one assume a particular cosmology and assign values to unknown cosmological parameters. We will not address this problem here, but due to this fact these data are not useful in determining the values of these phys-

ical constants.) As a final step, the data are often expressed in terms of absolute magnitude $M = -2.5\log(L) + k$ where k is an unimportant constant. This is the quantity shown on the vertical axis of the plot. The brightest quasars possess the smallest absolute magnitude, so the vertical axis is customarily plotted "backwards." The *luminosity function* is the distribution of absolute luminosity (absolute magnitude) as a function of redshift. Understanding how the luminosity function evolves with redshift is of primary interest. In statistical terms, the luminosity function is the conditional distribution of Y (the variable on the vertical axis) given X (on the horizontal axis). We seek to estimate the bivariate density as a means to estimating these key conditional distributions.

The challenge is that quasars will not be observable if their apparent luminosity lies outside of some range. When transforming this truncation bound to a bound on absolute magnitude, we arrive at the irregular region region traced by the dashed line in Figure 1. The larger gap in the lower right portion of the plot represents quasars that are too dim, the smaller gap in the upper left of the plot would include those that are too bright; these are not distinguishable from other nearby objects. This truncation bias would also appear when observing other astronomical objects such as galaxies.

Lynden-Bell (1971) introduced in the astronomy literature the nonparametric maximum likelihood (NPMLE) estimator for the case of one-sided truncation of absolute magnitude and Woodroofe (1985) derived some of the asymptotic properties of this estimator. Efron and Petrosian (1999) extended the NPMLE to the case of double truncation of absolute magnitude. Each of these papers assumes that absolute magnitude and redshift are statistically independent (and, hence, that the luminosity function does not evolve with redshift.) The density estimate (or distribution function estimate) which results from a NPMLE procedure places all of the probability

on observed data values, but even smoothing this estimate may not be sufficient to remove artifacts: An estimate can suffer from what Woodroofe (1992) referred to as "large jumps," where lone data points can greatly influence the estimator. Efron and Petrosian (1999) also developed a permutation test for independence of the two variables. Independence of absolute magnitude and redshift is a strong assumption, and evidence suggests that it is not justified; see Boyle et al. (2000). Parametric maximum likelihood is feasible with the irregular truncation, and Boyle et al. (2000) utilize this approach.

A semiparametric approach was proposed by Efron and Tibshirani (1996) and implemented for quasar data by Efron and Petrosian (1999), but in this analysis they did not exploit the nonparametric portion of the estimator, likely due to issues with the truncation. (Efron and Tibshirani (1996) also includes a nice application to similarly truncated galaxy data, although they also do not consider the effect of the truncation.) These *special exponential family* models estimate a density $f(\cdot)$ using models of the form

$$\log(f(x)) = \mathbf{f}(x) + \sum_{k=1}^{p} t_k(x)\,\beta_k \tag{1}$$

where the $t_k(\cdot)$ are the assumed sufficient statistics, $\mathbf{f}(\cdot)$ is estimated nonparametrically, and the β_k are estimated using maximum likelihood, once the nonparametric portion is fixed. See also Hjort and Glad (1995), who consider "correcting" an initial parametric density estimate by following it with a nonparametric fit.

The key advantage of our approach is that it allows one to avoid assuming that the two variables are independent, but also avoid imposing a tight parametric form on the bivariate density. We fit a variant on the special exponential family by decomposing the bivariate density $h(x, y)$ into

$$\log h(x, y) = \mathbf{f}(x) + \mathbf{g}(y) + \mathbf{h}(x, y, \theta) \tag{2}$$

where $\mathbf{h}(x, y, \theta)$ will take any form linear in the parameters; it is intended to model the dependence between the two random variables. For example, there may be a physical, parametric model for the evolution of the luminosity function which could be incorporated into $\mathbf{h}(x, y, \theta)$. Alternatively, one could use $\mathbf{h}(x, y, \theta) = \theta xy$ as a first-order approximation to the dependence. The functions $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are estimated nonparametrically, with bandwidth parameters to control the amount of smoothness in the estimate. We develop a single criterion, related to the local likelihood function and adjusted to account for truncation, which is maximized to obtain the estimate. The approach allows us to avoid the question of which portion (nonparametric or parametric) to fit first: Although an iterative fitting algorithm is developed, there is a unique estimate which maximizes the criterion regardless of the starting point.

We implemented the method as a Fortran subroutine with R wrapper. It is available for download, along with documentation, from

`http://www.stat.cmu.edu/~cschafer/BivTrunc`

The paper is organized as follows. Section 2 describes our approach in detail. Section 3 uses asymptotic properties of M-estimators to obtain approximations to the distribution and standard error of the estimator. Also, this section gives a description of how the integrated mean squared error can be approximated using cross-validation in a computationally tractable manner; the bandwidths can then be chosen to minimize this quantity. Section 4 presents some results from simulations and from the analysis of the quasar data and Section 5 is a brief discussion.

2 The Model

In this section we will describe the method in detail, starting the development with a heuristic description for the case when the two random variables are assumed in-

dependent. Once the details for the independence case are established, incorporating dependence is a simple extension.

Our approach originates in the following, naive method. Let $h(x, y)$ denote the joint density of random variables X and Y with respect to Lebesgue measure over some rectangular region. For the moment we assume X and Y are independent; write $h(x, y) = f(x)g(y)$ where $f(\cdot)$ is the density of X and $g(\cdot)$ is the density of Y. Let $\mathcal{A}$ denote the region outside of which of the data are truncated and let $\mathcal{A}(x, \cdot) \equiv \{y : (x, y) \in \mathcal{A}\}$ denote the cross-section of $\mathcal{A}$ at $X = x$. Let (X^*, Y^*) have the same distribution as (X, Y) conditional on $(X, Y) \in \mathcal{A}$. The available data allow for estimation of

$$f^*(x) \equiv f(x) \int_{\mathcal{A}(x,\cdot)} g(y)\, dy \Big/ \int_{\mathcal{A}} f(u)\, g(v)\, du\, dv \tag{3}$$

for all x, since it is the marginal density of X^*. Assuming for the moment that $g(\cdot)$ were known, it is possible to turn an estimator for $f^*(\cdot)$ into an estimator for $f(\cdot)$ using

$$\widehat{f}(x) \propto \widehat{f^*}(x) \Big/ \left(\int_{\mathcal{A}(x,\cdot)} g(y)\, dy \right) \tag{4}$$

and then normalizing to get a final estimate of the density. Starting with an initial guess at $g(\cdot)$, we could iterate between assuming $g(\cdot)$ is known, and estimating $f(\cdot)$, and vice versa. But, the variance of this estimator could be huge: Consider the behavior of $\widehat{f}(x)$ for x where $\int_{\mathcal{A}(x,\cdot)} g(y)dy$ is small. Also, it is possible to construct examples where this algorithm will converge to different estimates starting from different initial guesses.

The fundamental problem is that a well-chosen estimator (i.e., well-chosen smoothing parameters) for $f^*(\cdot)$ does not necessarily lead to a good estimator for $f(\cdot)$. In fact, if the region $\mathcal{A}$ has sharp corners, no amount of smoothing of $\widehat{f^*}(\cdot)$ will produce

a smooth $\widehat{f}(\cdot)$. Instead, we write

$$\log(f^*(x)) = \left[\log(f(x)) - \log\left(\int_{\mathcal{A}} h(u,v)\,du\,dv\right)\right] + \log\left(\int_{\mathcal{A}(x,\cdot)} g(y)\,dy\right) \tag{5}$$

and include

$$\log\left(\int_{\mathcal{A}(x,\cdot)} g(y)\,dy\right) \tag{6}$$

as an offset term in local polynomial models for $\log(f^*(x))$ and consider estimators that allow one to control smoothing in terms of $f(\cdot)$, not $f^*(\cdot)$.

2.1 Local Likelihood Density Estimation

We utilize the *local likelihood density estimator* developed independently by Loader (1996) and Hjort and Jones (1996). In the simple case of estimating $f^*(\cdot)$, the density of X^*, we seek the function $\mathbf{a}_u^*(\cdot)$ which maximizes

$$\mathcal{L}_u(\mathbf{a}_u^*, \mathbf{X}) \equiv \sum_{j=1}^{n} K^*(X_j, u, \lambda)\,\mathbf{a}_u^*(X_j) - \left[n\int_{\mathcal{X}} K^*(x,u,\lambda)\exp(\mathbf{a}_u^*(x))\,dx\right], \tag{7}$$

where K^* is a kernel function, λ is a smoothing parameter, and data $\mathbf{X} \equiv (X_1, X_2, \ldots, X_n)$ have density $f^*(\cdot)$ with respect to Lebesgue measure. The kernel function is written as $K^*(x,u,\lambda)$ because it will not, in general, be a simple function of $(x-u)/\lambda$; see Remark 2 below. Equation (7) is the *localized log likelihood* at u (equation (6) in Loader (1996)). Typically, $\mathcal{L}(\mathbf{a}_u^*, \mathbf{X})$ is maximized over degree p polynomials expanded around u:

$$\mathbf{a}_u^*(x) = a_{u0}^* + a_{u1}^*(x-u) + \cdots + a_{up}^*(x-u)^p\,. \tag{8}$$

The use of the notation $\mathbf{a}_u^*$ is intended to simultaneously identify the function, and the vector of coefficients $(a_{u0}^*, a_{u1}^*, \ldots, a_{up}^*)$ of this expansion.

Remark 1. The usual maximum likelihood estimate for a density can be found by maximizing

$$\sum_{j=1}^{n} \log f(X_j) - n\int f(x)\,dx \tag{9}$$

over all nonnegative functions f (possibly restricted to some smooth class). Note how $\mathcal{L}_u(\mathbf{a}_u^*, \mathbf{X})$ localizes this around u by introducing the kernel weighting. The form for localized likelihood is also motivated by noting that, just as for the standard likelihood function, the expected value of $\mathcal{L}_u(\mathbf{a}_u^*, \mathbf{X})$ is maximized by setting $\mathbf{a}_u^*(x) = \log f^*(x)$ for all x.

Remark 2. Our seemingly unusual choice for the form of the kernel function K^* comes from a requirement that, for all x,

$$\int K^*(x, u, \lambda)\, du = 1. \tag{10}$$

Although this function could take any nonnegative form (subject to this constraint), we will utilize

$$K^*(x, u, \lambda) = K\left(\frac{x-u}{\lambda}\right) \Big/ \left(\int_{\mathcal{X}} K\left(\frac{x-u}{\lambda}\right) du\right) \tag{11}$$

where $K(\cdot)$ is a "usual" kernel function, in our case the tricube kernel:

$$K(u) = \left(1 - |u|^3\right)^3 1_{\{|u|\leq 1\}}. \tag{12}$$

This requirement is not the same as the common assumption that the kernel integrates to one, because here, due to the boundary, the normalization term varies with x.

Let $\widehat{\mathbf{a}}_u^*$ denote the $\mathbf{a}_u^*$ which maximizes $\mathcal{L}_u(\mathbf{a}_u^*, \mathbf{X})$. The standard approach is to create an estimator for $f^*(u)$ via $\widehat{f^*}(u) \equiv \exp(\widehat{\mathbf{a}}_u^*(u)) = \exp(a_{u0}^*)$, but we consider an alternative method, motivated as follows: When fitting a local model around u, we obtain useful information about the value of density not only at u, but also for x near u. We use as our estimator

$$\widehat{f^*}(x) \equiv \int_{\mathcal{X}} K^*(x, u, \lambda) \exp(\widehat{\mathbf{a}}_u^*(x))\, du. \tag{13}$$

In equation (13) we obtain $\widehat{f}^*(x)$ by taking a weighted average of the individual $\widehat{\mathbf{a}}^*_u(x)$ for u near x. This will lead to significant analytical and computational advantages, and this is why it is necessary to require the kernel integrate to one in the way stated above: For fixed x, the quantity $K(x, u, \lambda)$ is the weight placed on the estimate of $f^*(x)$ from the local model fit around u. We can also allow λ to vary with u, so that

$$\widehat{f}^*(x) = \int_{\mathcal{X}} K^*(x, u, \lambda_u) \exp(\widehat{\mathbf{a}}^*_u(x))\, du. \tag{14}$$

Now we can look to choose λ_u so that $\widehat{\mathbf{a}}^*_u(x)$ is close to $\log f^*(x)$ for x close to u.

Define

$$\begin{aligned} \mathcal{L}(\mathbf{a}^*, \mathbf{X}) &\equiv \int_{\mathcal{X}} \mathcal{L}_u(\mathbf{a}^*_u, \mathbf{X})\, du \\ &= \int_{\mathcal{X}} \sum_{j=1}^{n} K^*(X_j, u, \lambda)\, \mathbf{a}^*_u(X_j)\, du - n \int_{\mathcal{X}} \left[\int_{\mathcal{X}} K^*(x, u, \lambda) \exp(\mathbf{a}^*_u(x))\, du \right] dx. \end{aligned} \tag{15}$$

The localized likelihood is maximized for each $u \in \mathcal{X}$, so we are finding the family of functions $\widehat{\mathbf{a}}^* \equiv \{\widehat{\mathbf{a}}^*_u : u \in \mathcal{X}\}$ that maximizes $\int_{\mathcal{X}} \mathcal{L}_u(\mathbf{a}^*_u, \mathbf{X})\, du$. As before, this has the property that the expected value of $\mathcal{L}(\mathbf{a}^*_u, \mathbf{X})$ is maximized by setting $\mathbf{a}^*_u(x) = \log f^*(x)$ for all u and all x, and hence $\mathcal{L}(\mathbf{a}^*, \mathbf{X})$ is maximized by setting $\widehat{f}^*(x)$ equal to $f^*(x)$ for all x.

2.2 The Offset Version

We now return to the goal of estimating $f(\cdot)$, the marginal density for X. We could also use the local likelihood approach; we model

$$\log f(x) - \log \left(\int_{\mathcal{A}} f(u)\, g(v)\, du\, dv \right) \approx \mathbf{a}_u(x) \equiv a_{u0} + a_{u1}(x - u) + \cdots + a_{up}(x - u)^p \tag{16}$$

for x near u. Recalling equation (5), our model for $f^*(\cdot)$ becomes

$$\log f^*(x) \approx \mathbf{a}_u(x) + \log \left(\int_{\mathcal{A}(x,\cdot)} g(y)\, dy \right) \tag{17}$$

for x near u. Note the following:

1. Instead of the above, we could define the offset term in equation (17) as

$$\log\left(\int_{\mathcal{A}(u,\cdot)} g(y)\,dy\right), \tag{18}$$

but then our estimate $\widehat{f}(x)$ would have the undesirable form found in equation (4): A scaled version of $\widehat{f^*}(x)$. The chosen version allows for smoothing of this term via the choice of λ_u.

2. Finding for each u the $\mathbf{a}_u$ which maximizes $\mathcal{L}_u(\mathbf{a}_u + \text{OFFSET}, \mathbf{X})$ leads to

$$\int_{\mathcal{X}} K^*(x, u, \lambda_u) \exp(\widehat{\mathbf{a}}_u(x))\,du \tag{19}$$

as an estimator for

$$f(x) \Big/ \left(\int_{\mathcal{A}} h(u, v)\,du\,dv\right). \tag{20}$$

3. In practice we will set up a grid of values of u at which local models are fit. The optimization is not difficult, since $\mathcal{L}_u(\mathbf{a}_u + \text{OFFSET}, \mathbf{X})$ is almost surely strictly concave as a function of $a_{u0}, a_{u1}, \ldots, a_{up}$.

2.3 The Iterative Algorithm

The above derivation assumed that $g(\cdot)$ were known. Instead, imagine that $g(\cdot)$ is estimated in a similar manner. Let $g^*(\cdot)$ denote the density for the observable Y^* and consider models of the form

$$\log g^*(y) \approx \mathbf{b}_v(y) + \log\left(\int_{\mathcal{A}(\cdot,y)} f(x)\,dx\right) - \log\left(\int_{\mathcal{A}} h(u, v)\,du\,dv\right) \tag{21}$$

where

$$\mathbf{b}_v(y) \equiv b_{u0} + b_{u1}(y - v) + \cdots + b_{up}(y - v)^p\,. \tag{22}$$

Now it becomes convenient to use

$$\log\left(\int_{\mathcal{A}(\cdot,y)} f(x)\,dx\right) - \log\left(\int_{\mathcal{A}} h(u,v)\,du\,dv\right) \tag{23}$$

as the offset, since it will lead directly to an estimate of $f(\cdot)$, and we found an estimator for this quantity as

$$\log\left(\int_{\mathcal{A}(\cdot,y)} \int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\widehat{\mathbf{a}}_u(x))\,du\,dx\right) \tag{24}$$

using equation (19). Thus, for each $v \in \mathcal{Y}$, the local likelihood $\mathcal{L}_v(\mathbf{b}_v + \text{OFFSET}, \mathbf{Y})$, where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, is maximized over offset polynomial forms

$$\mathbf{b}_v(y) + \log\left(\int_{\mathcal{A}(\cdot,y)} \int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\widehat{\mathbf{a}}_u(x))\,du\,dx\right) \tag{25}$$

where now we use the estimate of the offset to stress that the algorithm will proceed by alternating between estimating $f(\cdot)$ and $g(\cdot)$. But note the following:

$$\begin{aligned}
\mathcal{L}(\mathbf{b},\mathbf{Y}) \;\equiv\; & \int_{\mathcal{Y}} \mathcal{L}_v(\mathbf{b}_v + \text{OFFSET}, \mathbf{Y})\,dv = \int_{\mathcal{Y}} \sum_{j=1}^{n} K^*(Y_j, v, \lambda_v)\,\mathbf{b}_v(Y_j)\,dv \\
-\; n & \int_{\mathcal{Y}} \int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp(\mathbf{b}_v(y)) \left(\int_{\mathcal{A}(\cdot,y)} \int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\widehat{\mathbf{a}}_u(x))\,du\,dx\right) dy\,dv \\
= \; & \int_{\mathcal{Y}} \sum_{j=1}^{n} K^*(Y_j, v, \lambda_v)\,\mathbf{b}_v(Y_j)\,dv \\
-\; n & \int_{\mathcal{A}} \left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp(\mathbf{b}_v(y))\,dv\right] \left[\int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\widehat{\mathbf{a}}_u(x))\,du\right] dy\,dx. \qquad (26)
\end{aligned}$$

As before, we find $\widehat{\mathbf{b}} \equiv \{\widehat{\mathbf{b}}_v(\cdot)\colon v \in \mathcal{Y}\}$ which maximizes $\mathcal{L}(\mathbf{b},\mathbf{Y})$ and now set

$$\widehat{g}(y) \equiv \int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp\left(\widehat{\mathbf{b}}_v(y)\right) dv. \tag{27}$$

Returning to the first step in the process, we will use $\widehat{g}(\cdot)$ in place of the unknown $g(\cdot)$, and we see that

$$\begin{aligned}
\mathcal{L}(\mathbf{a},\mathbf{X}) \;\equiv\; & \int_{\mathcal{X}} \mathcal{L}_u(\mathbf{a}_u + \text{OFFSET}, \mathbf{X})\,du = \int_{\mathcal{X}} \sum_{j=1}^{n} K^*(X_j, u, \lambda_u)\,\mathbf{a}_u(X_j)\,du \\
-\; n & \int_{\mathcal{A}} \left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp(\widehat{\mathbf{b}}_v(y))dv\right] \left[\int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\mathbf{a}_u(x))\,du\right] dy\,dx. \qquad (28)
\end{aligned}$$

This leads to the following critical result:

Theorem 1. *Assume we place a restriction on the parameters of the local models that*

$$\int_{\mathcal{X}} a_{u0}\, du = 0. \tag{29}$$

Then the iterative algorithm will either diverge or converge to a unique $\widehat{\mathbf{a}}$ *and* $\widehat{\mathbf{b}}$*, and hence a unique estimate of* $f(\cdot)$ *and* $g(\cdot)$*.*

Proof. Consider the quantity

$$\begin{aligned}\mathcal{L}^*(\mathbf{a},\mathbf{b},\mathbf{X},\mathbf{Y}) \;\equiv\; & \int_{\mathcal{X}} \sum_{j=1}^{n} K^*(X_j,u,\lambda_u)\,\mathbf{a}_u(X_j)\,du \;+\; \int_{\mathcal{Y}} \sum_{j=1}^{n} K^*(Y_j,v,\lambda_v)\,\mathbf{b}_v(Y_j)\,dv \\ -\; n\int_{\mathcal{A}} & \left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp(\mathbf{b}_v(y))\,dv\right]\left[\int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\mathbf{a}_u(x))\,du\right] dy\,dx \\ & - \left(\int_{\mathcal{X}} a_{u0}\,du\right)^2. \end{aligned} \tag{30}$$

Maximizing $\mathcal{L}^*$ with $\mathbf{b}$ fixed is equivalent to maximizing $\mathcal{L}(\mathbf{a},\mathbf{X})$ with fixed offset term, under the restriction of equation (29). Maximizing $\mathcal{L}^*$ with $\mathbf{a}$ fixed is equivalent to maximizing $\mathcal{L}(\mathbf{b},\mathbf{Y})$ with fixed offset term. Since $\mathcal{L}^*$ is almost surely strictly concave as a function of the model coefficients a_{uk} and b_{vk}, the algorithm must converge to a unique point regardless of the starting value. □

Remark 3. Equation (29) is an identifiability condition; note that adding a constant onto a_{u0} for all u can be compensated for by subtracting the same constant off of all the b_{u0}. When the maximization is performed holding $\widehat{\mathbf{b}}$ fixed, it is not important to consider this restriction: The maximization is performed without the condition, and then the a_{u0} are normalized afterwards.

Remark 4. Note that

$$\left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v)\exp\left(\widehat{\mathbf{b}}_v(y)\right)dv\right]\left[\int_{\mathcal{X}} K^*(x,u,\lambda_u)\exp(\widehat{\mathbf{a}}_u(x))\,du\right] \tag{31}$$

is our final estimator for

$$f(x)\, g(y) \Big/ \int_{\mathcal{A}} h(u,v)\, du\, dv. \tag{32}$$

(See equations (13) and (27).) We can now normalize to get estimates of h, f, and g.

2.4 Removing the Independence Assumption

When we remove the assumption that X and Y are independent, we write the logarithm of the joint density $h(x,y)$ as

$$\log h(x,y) = \mathbf{f}(x) + \mathbf{g}(y) + \mathbf{h}(x,y,\theta)\,. \tag{33}$$

The term $\mathbf{h}(x,y,\theta)$ is the parametric portion of the estimator; here we use it to model the dependency between X and Y. Here, we focus on a simple form for the dependence: $\mathbf{h}(x,y,\theta) = \theta xy$.

We update our criterion of equation (30) to be

$$\begin{aligned}
&\mathcal{L}^*(\mathbf{a},\mathbf{b},\theta,\mathbf{X},\mathbf{Y}) \;\equiv \\
&\int_{\mathcal{X}} \sum_{j=1}^{n} K^*(X_j,u,\lambda_u)\, \mathbf{a}_u(X_j)\, du \;+ \int_{\mathcal{Y}} \sum_{j=1}^{n} K^*(Y_j,v,\lambda_v)\, \mathbf{b}_v(Y_j)\, dv \;+ \sum_{j=1}^{n} \mathbf{h}(X_j,Y_j,\theta) \\
&- n \int_{\mathcal{A}} \exp(\mathbf{h}(x,y,\theta)) \left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v) \exp(\mathbf{b}_v(y))\, dv \right] \\
&\qquad\qquad \left[\int_{\mathcal{X}} K^*(x,u,\lambda_u) \exp(\mathbf{a}_u(x))\, du \right] dy\, dx - \left(\int_{\mathcal{X}} a_{u0}\, du \right)^2,
\end{aligned} \tag{34}$$

and retain the identifiability constraint of equation (29). We assume that $\exp(\mathbf{h}(x,y,\theta))$ is strictly concave as a function of θ, e.g. $\mathbf{h}(x,y,\theta)$ is a linear function of θ. The algorithm is updated in a simple manner: Now there is a third step in which a search for the value of θ which maximizes $\mathcal{L}^*$ is sought while holding $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ fixed. Note that the offset terms become

$$\log\left(\int_{\mathcal{A}(x,\cdot)} \exp(\mathbf{h}(x,y,\theta)) \int_{\mathcal{Y}} K^*(y,v,\lambda_v) \exp(\widehat{\mathbf{b}}_v(y)) dv\, dx \right) \tag{35}$$

when finding $\widehat{\mathbf{a}}$ and

$$\log\left(\int_{\mathcal{A}(\cdot,y)} \exp(\mathbf{h}(x,y,\theta)) \int_{\mathcal{X}} K^*(x,u,\lambda_u) \exp(\widehat{\mathbf{a}}_u(x))\, du\, dx\right) \tag{36}$$

when finding $\widehat{\mathbf{b}}$. This will converge to a unique final combination $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\theta})$ which maximizes $\mathcal{L}^*$, with

$$\exp\left(\mathbf{h}(x,y,\widehat{\theta})\right) \left[\int_{\mathcal{Y}} K^*(y,v,\lambda_v) \exp(\widehat{\mathbf{b}}_v(y))dv\right] \left[\int_{\mathcal{X}} K^*(x,u,\lambda_u) \exp(\widehat{\mathbf{a}}_u(x))\, du\right] \tag{37}$$

being our estimator for

$$h(x,y) \Big/ \int_{\mathcal{A}} h(u,v)\, du\, dv. \tag{38}$$

3 Inference and Bandwidth Selection

In this section we will derive some of the key tools for statistical inference and bandwidth selection for use when applying the method of the previous section. Emphasis is placed on computational feasibility. As a first step, consider that the above derivation fits a local model at each $u \in \mathcal{X}$ when $\widehat{\mathbf{b}}$ and $\widehat{\theta}$ were held constant, and at each $v \in \mathcal{Y}$ when $\widehat{\mathbf{a}}$ and $\widehat{\theta}$ were held constant. In practice, we must consider a grid of values $u_1, u_2, \ldots, u_g \in \mathcal{X}$ and $v_1, v_2, \ldots, v_g \in \mathcal{Y}$ at which these respective models are fit. Now we denote the parameters such that the local model around $u_i \in \mathcal{X}$ is

$$\mathbf{a}_i(x) \equiv a_{i0} + a_{i1}(x - u_i) + \cdots + a_{ip}(x - u_i)^p + \text{offset} \tag{39}$$

and the local model around $v_i \in \mathcal{Y}$ is

$$\mathbf{b}_i(y) \equiv b_{i0} + b_{i1}(y - v_i) + \cdots + b_{ip}(y - v_i)^p + \text{offset}. \tag{40}$$

We will here focus on the case $\mathbf{h}(x,y,\theta) = \theta xy$, although extension to other cases is not difficult. Thus, we have a list of $2(p+1)g + 1$ parameters to estimate:

$$\boldsymbol{\beta} = [\, a_{10} \;\; \cdots \;\; a_{1p} \;\; \cdots \;\; a_{g0} \;\; a_{g1} \;\; \cdots \;\; a_{gp} \;\; b_{10} \;\; \cdots \;\; b_{1p} \;\; \cdots \;\; b_{g0} \;\; b_{g1} \;\; \cdots \;\; b_{gp} \;\; \theta \,]^T \tag{41}$$

and our new "discrete grid" criterion function is

$$\mathcal{L}_n^*(\boldsymbol{\beta}) = n^{-1} \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}) \tag{42}$$

where

$$\begin{aligned}
\ell_j(\boldsymbol{\beta}) \;\equiv\; & \sum_{i=1}^{g} K^*(X_j, u_i, \lambda_{u_i})\, \mathbf{a}_i(X_j) \;+ \sum_{i=1}^{g} K^*(Y_j, v_i, \lambda_{v_i})\, \mathbf{b}_i(Y_j) \;+ \mathbf{h}(X_j, Y_j, \theta) \\
& - \int_{\mathcal{A}} \exp(\mathbf{h}(x, y, \theta)) \left[\sum_{i=1}^{g} K^*(y, v_i, \lambda_{v_i}) \exp(\mathbf{b}_i(y)) \right] \left[\sum_{i=1}^{g} K^*(x, u_i, \lambda_{u_i}) \exp(\mathbf{a}_i(x)) \right] dy\, dx \\
& - \left(\sum_{i=1}^{g} a_{i0} \right)^2
\end{aligned} \tag{43}$$

and now we assume that

$$\sum_{i=1}^{g} K^*(x, u_i, \lambda_{u_i}) = \sum_{i=1}^{g} K^*(y, v_i, \lambda_{v_i}) = 1 \tag{44}$$

for each x and y.

3.1 Approximating the Distribution of the Estimator

We will utilize standard results from the theory of M-estimators. Let $\mathbf{V}_{x,y}$ be the matrix such that

$$\mathbf{V}_{x,y}\boldsymbol{\beta} = \begin{bmatrix}
\mathbf{a}_1(x) + \mathbf{b}_1(y) + \mathbf{h}(x, y, \theta) \\
\mathbf{a}_1(x) + \mathbf{b}_2(y) + \mathbf{h}(x, y, \theta) \\
\vdots \\
\mathbf{a}_1(x) + \mathbf{b}_g(y) + \mathbf{h}(x, y, \theta) \\
\mathbf{a}_2(x) + \mathbf{b}_1(y) + \mathbf{h}(x, y, \theta) \\
\mathbf{a}_2(x) + \mathbf{b}_2(y) + \mathbf{h}(x, y, \theta) \\
\vdots \\
\mathbf{a}_2(x) + \mathbf{b}_g(y) + \mathbf{h}(x, y, \theta) \\
\vdots \\
\mathbf{a}_g(x) + \mathbf{b}_g(y) + \mathbf{h}(x, y, \theta)
\end{bmatrix}. \tag{45}$$

Define

$$\mathbf{K}_{x,y} = \begin{bmatrix} K^*(x, u_1, \lambda_{u_1})\, K^*(y, v_1, \lambda_{v_1}) \\ K^*(x, u_1, \lambda_{u_1})\, K^*(y, v_2, \lambda_{v_2}) \\ \vdots \\ K^*(x, u_1, \lambda_{u_1})\, K^*\big(y, v_g, \lambda_{v_g}\big) \\ K^*(x, u_2, \lambda_{u_2})\, K^*(y, v_1, \lambda_{v_1}) \\ K^*(x, u_2, \lambda_{u_2})\, K^*(y, v_2, \lambda_{v_2}) \\ \vdots \\ K^*(x, u_2, \lambda_{u_2})\, K^*\big(y, v_g, \lambda_{v_g}\big) \\ \vdots \\ K^*\big(x, u_g, \lambda_{u_g}\big)\, K^*\big(y, v_g, \lambda_{v_g}\big) \end{bmatrix} \tag{46}$$

and

$$\mathbf{Z}_{x,y}(\boldsymbol{\beta}) \equiv \mathrm{diag}(\exp(\mathbf{V}_{\mathrm{x,y}}\boldsymbol{\beta}))\, \mathrm{diag}(\mathbf{K}_{\mathrm{x,y}}) \tag{47}$$

so that

$$\exp(\mathbf{V}_{x,y}\boldsymbol{\beta})^T \mathbf{K}_{x,y} = \mathbf{e}^T \mathbf{Z}_{x,y}(\boldsymbol{\beta})\, \mathbf{e} \tag{48}$$

where and $\mathbf{e}$ is a vector filled with ones. Define $\mathbf{m}_j \equiv \mathbf{V}^T_{X_j,Y_j}\mathbf{K}_{X_j,Y_j}$ and let $\mathbf{c}$ denote a vector the same length as $\boldsymbol{\beta}$, with ones in the positions corresponding to the parameters $a_{10}, a_{20}, \ldots a_{g0}$, and zeros everywhere else; we can now write

$$\ell_j(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \mathbf{m}_j - \int_{\mathcal{A}} \mathbf{e}^T \mathbf{Z}_{x,y}(\boldsymbol{\beta})\, \mathbf{e}\, dx\, dy - \left(\boldsymbol{\beta}^T \mathbf{c}\right)^2. \tag{49}$$

It follows that

$$\frac{d\ell_j}{d\boldsymbol{\beta}} = \mathbf{m}_j - \int_{\mathcal{A}} \mathbf{V}^T_{x,y} \mathbf{Z}_{x,y}(\boldsymbol{\beta})\, \mathbf{e}\, dx\, dy - 2\left(\boldsymbol{\beta}^T \mathbf{c}\right) \mathbf{c}. \tag{50}$$

and

$$\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} = -\int_{\mathcal{A}} \mathbf{V}^T_{x,y} \mathbf{Z}_{x,y}(\boldsymbol{\beta})\, \mathbf{V}_{x,y}\, dx\, dy - 2\mathbf{c}\mathbf{c}^T = \mathrm{E}\left[\frac{d\ell_j}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\right]. \tag{51}$$

All of this notation is needed for the following results.

Theorem 2. *There is a unique parameter vector which maximizes* $\mathrm{E}(\ell_j(\boldsymbol{\beta}))$.

Proof. Since $\mathrm{E}(\ell_j(\boldsymbol{\beta}))$ is strictly concave as a function of $\boldsymbol{\beta}$, there can be at most one maximal $\boldsymbol{\beta}$. Let α denote a scalar. For any parameter vector $\boldsymbol{\beta}$,

$$\begin{aligned}\lim_{\alpha\to\infty}\mathrm{E}(\ell_j(\alpha\boldsymbol{\beta})) &= \lim_{\alpha\to\infty}\left[\alpha\boldsymbol{\beta}^T\mathrm{E}(\mathbf{m}_j) - \int_{\mathcal{A}}\mathbf{e}^T\mathbf{Z}_{x,y}(\alpha\boldsymbol{\beta})\,\mathbf{e}\,dx\,dy - \left(\alpha\boldsymbol{\beta}^T\mathbf{c}\right)^2\right] \\ &= \lim_{\alpha\to\infty}\left[\boldsymbol{\beta}^T\mathrm{E}(\mathbf{m}_j) - \int_{\mathcal{A}}\left(\mathbf{V}_{x,y}\boldsymbol{\beta}\right)^T\mathbf{K}_{x,y}\mathbf{e}^T\mathbf{Z}_{x,y}(\alpha\boldsymbol{\beta})\,\mathbf{e}\,dx\,dy - 2\alpha\boldsymbol{\beta}^T\mathbf{c}\right] \\ &= -\infty. \end{aligned} \tag{52}$$

The conclusion is that in any direction $\boldsymbol{\beta}$, $\mathrm{E}(\ell_j(\alpha\boldsymbol{\beta}))$ eventually decreases for sufficiently large α. Thus, there must be a single, unique maximal $\boldsymbol{\beta}$. □

Theorem 3. *Let* $\boldsymbol{\beta}_\lambda$ *denote the parameter vector which maximizes* $\mathrm{E}(\ell_j(\boldsymbol{\beta}))$ *and let* $\widehat{\boldsymbol{\beta}}_n$ *denote the parameter vector which maximizes* $\mathcal{L}_n^*(\boldsymbol{\beta})$. *As* $n\to\infty$, $\widehat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_\lambda$ *and*

$$\begin{aligned}&\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_\lambda\right) \xrightarrow{\mathcal{D}} \\ &\quad \mathrm{N}\left(\mathbf{0},\ \left(\left.\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}_\lambda}\right)^{-1}\mathrm{E}\left[\left(\left.\frac{d\ell_j}{d\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_\lambda}\right)\left(\left.\frac{d\ell_j}{d\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_\lambda}\right)^T\right]\left(\left.\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}_\lambda}^T\right)^{-1}\right).\end{aligned} \tag{53}$$

Proof. We will utilize results in Haberman (1989). The almost sure consistency is given by Theorem 5.1. Note that his conditions 1, 2, and 6 are trivially true since the parameter space is $(2(p+1)g+1)$-dimensional Euclidean space, clearly closed. His condition 5 is implied by condition 3 which requires that $\mathrm{E}(\ell_j(\boldsymbol{\beta}))$ be finite for all $\boldsymbol{\beta}$; this is clearly true since the random variables are bounded by the truncation.

The central limit result is Theorem 6.1 in Haberman (1989). Condition 7 requires that the Hessian matrix

$$\left.\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}_\lambda} \tag{54}$$

exist and be nonsingular. For any parameter vector $\boldsymbol{\beta}_1 \neq \mathbf{0}$, we note that

$$\boldsymbol{\beta}_1^T \left(\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}_\lambda} \right) \boldsymbol{\beta}_1 = -\int_{\mathcal{A}} (\mathbf{V}_{x,y}\boldsymbol{\beta}_1)^T \mathbf{Z}_{x,y}(\boldsymbol{\beta}_\lambda) (\mathbf{V}_{x,y}\boldsymbol{\beta}_1)\, dx\, dy - 2\boldsymbol{\beta}_1^T \mathbf{c}\mathbf{c}^T \boldsymbol{\beta}_1 < 0 \quad (55)$$

since

$$\boldsymbol{\beta}_1^T \mathbf{c}\mathbf{c}^T \boldsymbol{\beta}_1 > 0 \quad (56)$$

and $\mathbf{Z}_{x,y}(\boldsymbol{\beta})$ is a diagonal matrix with all positive entries along the diagonal. Thus, the matrix is negative definite, and must be invertible. Condition 10 requires that

$$\boldsymbol{\beta}_1^T \mathrm{E}\left(\frac{d\ell_j(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}_1} \right) < \infty \quad (57)$$

for every $\boldsymbol{\beta}_1$. This is again true since the criterion is continuous and the random variables are all bounded. Conditions 7 and 10 imply conditions 8 and 9. □

Remark 5. In our code we approximate

$$\mathrm{E}\left[\left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}_\lambda} \right) \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}_\lambda} \right)^T \right] \approx n^{-1} \sum_{j=1}^{n} \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right)^T \quad (58)$$

and

$$\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}_0} \approx \frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\widehat{\boldsymbol{\beta}}_n}. \quad (59)$$

Finally, we apply the Δ-method to get asymptotic normality of our estimator of $h(x, y)$. Let $h_{\boldsymbol{\beta}}(x, y)$ denote our estimator for

$$h(x, y) \Big/ \left(\int_{\mathcal{A}} h(u, v)\, du\, dv \right), \quad (60)$$

using parameter vector $\boldsymbol{\beta}$, as previously discussed. Note that

$$h_{\widehat{\boldsymbol{\beta}}}(x, y) \equiv \mathbf{e}^T \mathbf{Z}_{x,y}(\widehat{\boldsymbol{\beta}})\, \mathbf{e}. \quad (61)$$

We have that

$$\frac{dh_{\boldsymbol{\beta}}(x, y)}{d\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}_\lambda} = \mathbf{V}_{x,y}^T \mathbf{Z}_{x,y}(\boldsymbol{\beta}_\lambda)\, \mathbf{e}, \quad (62)$$

so we can approximate

$$h_{\widehat{\boldsymbol{\beta}}}(x,y) \sim \mathrm{N}\Big(h_{\boldsymbol{\beta}_\lambda}(x,y)\,,\; n^{-1}\mathbf{e}^T\mathbf{Z}_{x,y}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_{x,y}\boldsymbol{\Sigma}\,\mathbf{V}_{x,y}^T\mathbf{Z}_{x,y}(\widehat{\boldsymbol{\beta}}_n)\,\mathbf{e}\Big) \tag{63}$$

and

$$\mathrm{Cov}(h_{\widehat{\boldsymbol{\beta}}}(x,y)\,,\; h_{\widehat{\boldsymbol{\beta}}}(x',y')) = n^{-1}\mathbf{e}^T\mathbf{Z}_{x,y}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_{x,y}\,\boldsymbol{\Sigma}\,\mathbf{V}_{x',y'}^T\mathbf{Z}_{x',y'}(\widehat{\boldsymbol{\beta}}_n)\,\mathbf{e} \tag{64}$$

where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}_n$.

3.2 Bandwidth Selection

Standard bandwidth selection techniques can be employed with this estimator. Following Rudemo (1982), the integrated mean squared error of the estimator

$$\mathrm{E}\left[\int_{\mathcal{A}}\left(h_{\widehat{\boldsymbol{\beta}}}(x,y) - h(x,y)\Big/\left(\int_{\mathcal{A}}h(u,v)\,du\,dv\right)\right)^2 dx\,dy\right] \tag{65}$$

can be approximated (up to an unimportant constant) using the least squares cross-validation estimator

$$\int_{\mathcal{A}} h^2_{\widehat{\boldsymbol{\beta}}}(x,y)\,dx\,dy - \frac{2}{n}\sum_{j=1}^{n} h_{\widehat{\boldsymbol{\beta}}_{(-j)}}(X_j,Y_j) \tag{66}$$

where $h_{\widehat{\boldsymbol{\beta}}_{(-j)}}(X_j,Y_j)$ denotes the estimate found when omitting the j^{th} data pair from the analysis. Alternatively, the likelihood cross-validation score is calculated as

$$n^{-1}\sum_{j=1}^{n}\log\Big(h_{\widehat{\boldsymbol{\beta}}_{(-j)}}(X_j,Y_j)\Big)\,. \tag{67}$$

Maximizing the likelihood cross-validation approximates minimizing the Kullback-Leibler divergence between the estimate and the true h; see Bowman (1984).

Exact computation of these leave-one-out estimates is computationally prohibitive, so instead we make the following approximation. Let $\mathcal{L}^*_{(-j)}(\boldsymbol{\beta})$ denote the criterion

evaluated with the j^{th} data pair removed. Then

$$\begin{aligned}\mathcal{L}^*_{(-j)}(\widehat{\boldsymbol{\beta}}_n + \boldsymbol{\beta}') &\approx \mathcal{L}^*_{(-j)}(\widehat{\boldsymbol{\beta}}_n) + \boldsymbol{\beta}'^T \left(\frac{d\mathcal{L}^*_{(-j)}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) + \boldsymbol{\beta}'^T \left(\frac{d\ell_j}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) \boldsymbol{\beta}'/2 \\ &= \mathcal{L}^*_{(-j)}(\widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\beta}'^T \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) / (n-1) + \boldsymbol{\beta}'^T \left(\frac{d\ell_j}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) \boldsymbol{\beta}'/2 \quad (68)\end{aligned}$$

since

$$\mathbf{0} = \left(\frac{n}{n-1} \right) \left(\frac{d\mathcal{L}^*_n(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) = \left(\frac{d\mathcal{L}^*_{(-j)}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) + \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) / (n-1) \,. \quad (69)$$

We find that this is maximized by setting

$$\boldsymbol{\beta}' = \left(\frac{d\ell_j}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right)^{-1} \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) / (n-1) = \left(\frac{d\mathrm{E}(\ell_j)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right)^{-1} \left(\frac{d\ell_j}{d\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}_n} \right) / (n-1) \quad (70)$$

so we can approximate the leave-one-out estimate $\widehat{\boldsymbol{\beta}}_{(-j)} \approx \widehat{\boldsymbol{\beta}}_n + \boldsymbol{\beta}'$. Computationally, this is not difficult because we already found the necessary quantities when deriving the asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}$ (see Remark 5).

Remark 6. We can only estimate the integrated squared error over $\mathcal{A}$ (equation 66) since there is no available information outside of $\mathcal{A}$, **but** we stress that the cross-validation estimator we use estimates the expected integrated squared error in estimating

$$h(x,y) \Big/ \left(\int_{\mathcal{A}} h(u,v)\, du\, dv \right) \quad (71)$$

not for estimating $h_{\boldsymbol{\beta}_\lambda}(x,y)$. Of course, it is this former quantity we truly want to estimate.

4 Results

In this section the method is applied to some simulated data sets, with the estimates compared to the known truth, and to the SDSS quasar sample of Figure 1.

4.1 Simulations

In these simulations, the observable region $\mathcal{A}$ is defined to be the subset of the unit square $[0,1] \times [0,1]$ where $y \geq x - 0.2$ and $(x-0.6)^2 + (y-0.4)^2 \geq 0.07$. Three distributions are considered for the data. In **Case I**, X and Y are independent, normal random variables, each with mean 0.5 and standard deviation 0.2. In **Case II**, X and Y are normal, each with mean 0.5 and SD 0.2, but the correlation between X and Y is 0.5. In **Case III**, X is normal with mean 0.5 and SD 0.2, and the conditional distribution of Y given X is normal with mean $\sqrt{X}$ and SD 0.2. In all simulations, a sample consisting of 10,000 (X, Y) pairs (that fall into the observable region) is generated. In all analyses, $p = 1$ and $g = 50$.

For illustrative purposes, Figure 2 shows one data set simulated under Case II, along with the true density (solid contours) and the estimate (dashed contours). In each case our method was applied to each of fifty simulated data sets. The bandwidth used was chosen by minimizing the least squares cross-validation criterion. Tables 3, 4, and 5 summarize the results. The tables compare the true density (in column 2) at a collection of (x, y) values with the average estimate over the 50 simulations (in column 3). Also, the nominal standard error for each of these estimates is averaged in quadrature over the 50 simulations (column 4) and compared with the standard deviation of the 50 estimates. If the asymptotic theory for approximating the SE is appropriate, these quantities should be close. Finally, column 5 shows the squared error in the estimate averaged over the 50 simulations. If column 5 is significantly larger than column 4, this is a sign of bias in the estimator. The most noticeable problems occur in Case III, which is not surprising given that this is the one of the three cases where the true density cannot be written in a form with $\mathbf{h}(x, y, \theta) = \theta xy$. All estimates in the table are for the density normalized to one over $\mathcal{A}$.

4.2 SDSS Quasar Sample

Figures 6 through 10 show results from applying this approach to the SDSS sample of 11,242 quasars. For this analysis, the redshift range of the observable region is 0.2 to 3.0, the lower truncation limit for absolute magnitude is

$$\max\left(-29.28 - 2.43\log\left(1+\mathrm{z}-\sqrt{1+\mathrm{z}}\right) + 0.61\log(1+\mathrm{z}), -27.59\right) \qquad (72)$$

and the upper truncation limit is

$$\min\left(-26.26 - 2.43\log\left(1+\mathrm{z}-\sqrt{1+\mathrm{z}}\right) + 0.61\log(1+\mathrm{z}), -22.12\right), \qquad (73)$$

where z denotes redshift. The analysis was performed using $p = 2$ and $g = 100$, and the optimal global bandwidth is 0.15 using either the least squares cross-validation or the likelihood cross-validation criteria. (This bandwidth is stated on the scale of the data after it has been transformed to lie in the unit square.)

Figure 6 depicts the estimate of the bivariate density superimposed on the data with dashed truncation region. An evident problem is that the estimate is under-smoothed outside of the observable region. This is not surprising given that the bandwidth is chosen to limit the integrated mean squared error only over $\mathcal{A}$. Figure 7 shows the estimate of the marginal distribution for redshift. Figure 8 shows the estimate of the marginal distribution for absolute magnitude. As is customary, this is plotted on a logarithmic scale. Figures 9 and 10 depict the estimates of the conditional density of absolute magnitude (the luminosity function) for redshift of 0.5 and 1.5. Error bars are one standard error limits. The shaded region represents the range over which data are not observable for that redshift. Care should be taken when interpreting the estimate within this region.

5 Discussion

The effect of the irregular boundary is significant when trying to estimate the density, thus this method is useful even if one believes it is unrealistic to estimate the density over regions where data are unobserved. The shape of the estimate shown in Figure 6, when restricted to the observable region, would have been difficult to obtain had the density not also been estimated over the unobservable region.

The nonparametric portion of the estimator is novel in that local models are fit, but then smoothed together, making double use of the bandwidth. This construction led to a simple, single criterion function which, when maximized, gives the estimate. This criterion is intuitive in that it is strongly related to the usual likelihood equation.

The results of Section 3 allow the user to vary the bandwidths over the different local models. A wide class of models is available as these bandwidths are varied. A simple approach was presented for estimating the integrated mean squared error of the estimator. This is perhaps the greatest asset of this approach: One can minimize (an estimate of) the integrated mean squared error over a wide class of models. The results from the analysis of observed quasars show that careful selection of the bandwidths is required, or else there will be undersmoothing in the truncated regions. Varying bandwidths will likely address this problem; future work will focus on searching the space of possible bandwidths in a computationally feasible manner.

References

Bowman, A. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.

Boyle, B., Shanks, T., Croom, S., Smith, R., Miller, L., Loaring, N., and Heymans, C. (2000), "The 2df QSO Redshift Survey - I. The optical luminosity function of quasi-stellar objects," *Month. Not. Royal Astron. Soc.*, 317, 1014–1022.

Efron, B. and Petrosian, V. (1999), "Nonparametric Methods for Doubly Truncated Data," *J. Am. Stat. Assoc.*, 94, 824–834.

Efron, B. and Tibshirani, R. (1996), "Using Specially Designed Exponential Families for Density Estimation," *Ann. Stat.*, 24, 2431–2461.

Haberman, S. (1989), "Concavity and Estimation," *Ann. Stat.*, 17, 1631–1661.

Hjort, N. and Glad, I. (1995), "Nonparametric Density Estimation with a Parametric Start," *Ann. Stat.*, 23, 882–904.

Hjort, N. and Jones, M. (1996), "Locally Parametric Nonparametric Density Estimation," *Ann. Stat.*, 24, 1619–1647.

Loader, C. (1996), "Local Likelihood Density Estimation," *Ann. Stat.*, 24, 1602–1618.

Lynden-Bell, D. (1971), "A Method of Allowing for known Observational Selection in Small Samples Applied to 3CR Quasars," *Month. Not. Royal Astron. Soc.*, 155, 95–118.

Petrosian, V. (1992), "Luminosity Function of Flux-Limited Samples," in *Statistical Challenges in Modern Astronomy*, eds. Feigelson, E. and Babu, G., New York: Spring-Verlag, pp. 173–194.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scan. J. of Stat.*, 9, 65–78.

Schneider, D. (2003), "The Sloan Digital Sky Survey Quasar Catalog II. First Data Release." *The Astronomical Journal*, 126, 2579–2593.

Scott, D. and Wand, M. (1991), "Feasibility of Multivariate Density Estimates," *Biometrika*, 78, 197–205.

Woodroofe, M. (1985), "Estimating a Distribution Function with Truncated Data," *Ann. Stat.*, 13, 163–177.

— (1992), "Luminosity Function of Flux-Limited Samples: Discussion," in *Statistical Challenges in Modern Astronomy*, eds. Feigelson, E. and Babu, G., New York: Spring-Verlag, pp. 196–200.

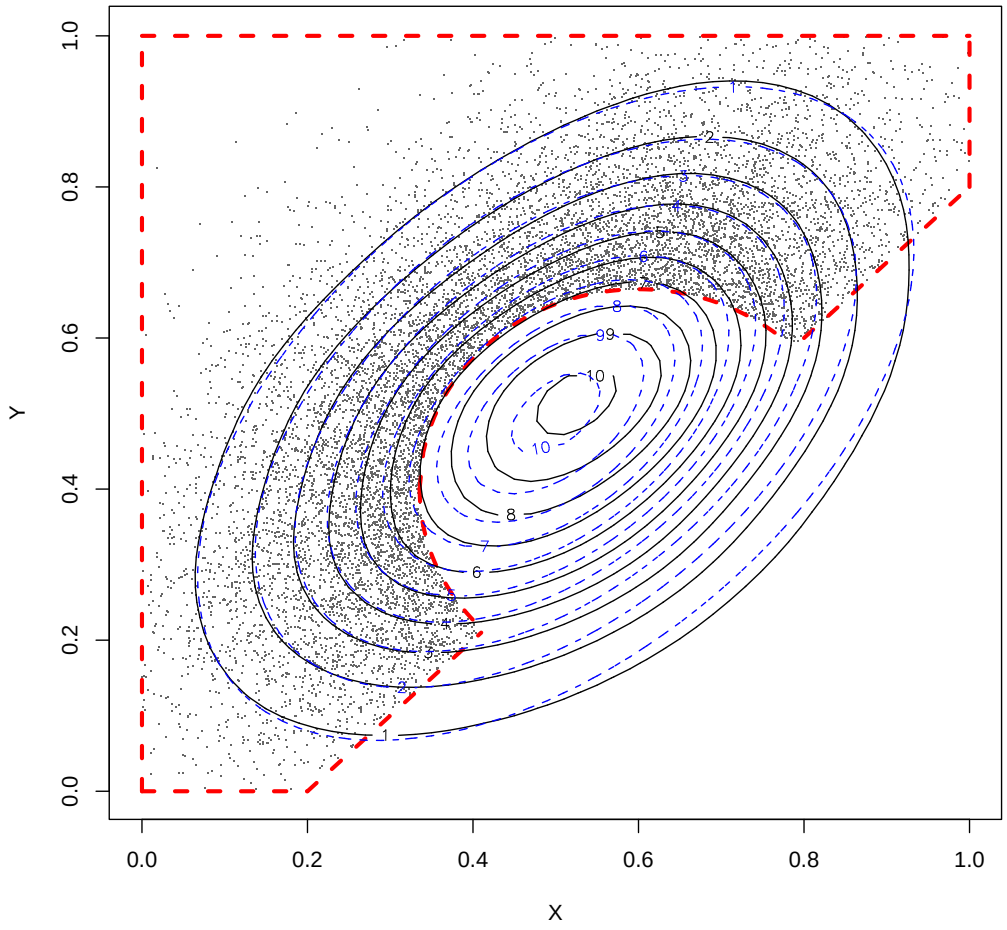

Figure 2: One simulated data set consisting of $10,000$ pairs. The outer dashed line gives the truncation region. The data set depicted is simulated under Case II, with the true density shown by the dashed contours. The estimate, using the described method on this simulated data set, is shown by the solid contours.

Estimand	True Value	Average Estimate	RMS of Nominal SE	SD of Estimates	RMSE of Estimates
bivariate density					
(0.225, 0.225)	1.46	1.44	0.0579	0.0559	0.0604
(0.475, 0.225)	3.73	3.64	0.246	0.254	0.258
(0.725, 0.225)	1.99	1.95	0.281	0.262	0.282
(0.225, 0.475)	3.73	3.76	0.102	0.0952	0.107
(0.475, 0.475)	9.51	9.55	0.460	0.455	0.457
(0.725, 0.475)	5.09	5.11	0.394	0.403	0.391
(0.225, 0.725)	1.99	2.00	0.056	0.0573	0.0554
(0.475, 0.725)	5.09	5.09	0.106	0.117	0.105
(0.725, 0.725)	2.73	2.73	0.0903	0.0874	0.0896
X **marginal**					
0.225	1.88	1.86	0.0277	0.0302	0.035
0.475	4.81	4.73	0.171	0.172	0.187
0.725	2.57	2.53	0.175	0.172	0.178
Y **marginal**					
0.225	1.88	1.81	0.136	0.132	0.155
0.475	4.81	4.73	0.201	0.212	0.212
0.725	2.57	2.52	0.034	0.0341	0.0624

Figure 3: Simulations results for Case I.

Estimand	True Value	Average Estimate	RMS of Nominal SE	SD of Estimates	RMSE of Estimates
bivariate density					
(0.225, 0.225)	2.94	2.93	0.0732	0.0866	0.0741
(0.475, 0.225)	3.26	3.33	0.249	0.234	0.256
(0.725, 0.225)	0.451	0.486	0.0841	0.0733	0.0902
(0.225, 0.475)	3.26	3.31	0.0903	0.0893	0.100
(0.475, 0.475)	10.3	10.5	0.479	0.532	0.511
(0.725, 0.475)	4.02	4.19	0.324	0.335	0.365
(0.225, 0.725)	0.451	0.458	0.0223	0.0227	0.0232
(0.475, 0.725)	4.02	4.01	0.103	0.105	0.102
(0.725, 0.725)	4.46	4.46	0.118	0.114	0.117
X **marginal**					
0.225	1.75	1.73	0.0282	0.0296	0.0374
0.475	4.47	4.53	0.171	0.178	0.178
0.725	2.39	2.43	0.114	0.109	0.118
Y **marginal**					
0.225	1.75	1.73	0.0834	0.0747	0.0849
0.475	4.47	4.52	0.179	0.201	0.186
0.725	2.39	2.35	0.0311	0.034	0.055

Figure 4: Simulations results for Case II.

Estimand	True Value	Average Estimate	RMS of Nominal SE	SD of Estimates	RMSE of Estimates
bivariate density					
(0.225, 0.225)	1.08	1.09	0.0476	0.0516	0.0499
(0.475, 0.225)	0.405	0.498	0.0442	0.043	0.103
(0.725, 0.225)	0.0237	0.0171	0.00312	0.00297	0.00729
(0.225, 0.475)	2.34	2.27	0.0787	0.0771	0.109
(0.475, 0.475)	3.37	3.58	0.209	0.198	0.295
(0.725, 0.475)	0.544	0.424	0.0445	0.0421	0.129
(0.225, 0.725)	1.07	1.1	0.0315	0.0399	0.0474
(0.475, 0.725)	5.89	6.05	0.12	0.124	0.204
(0.725, 0.725)	2.62	2.48	0.0841	0.0761	0.164
X **marginal**					
0.225	1.17	1.16	0.021	0.025	0.0233
0.475	3.00	2.91	0.067	0.0697	0.110
0.725	1.60	1.20	0.0321	0.0314	0.407
Y **marginal**					
0.225	0.72	0.426	0.0211	0.0201	0.294
0.475	1.41	1.57	0.0675	0.0647	0.175
0.725	1.90	2.41	0.0367	0.033	0.507

Figure 5: Simulations results for Case III.

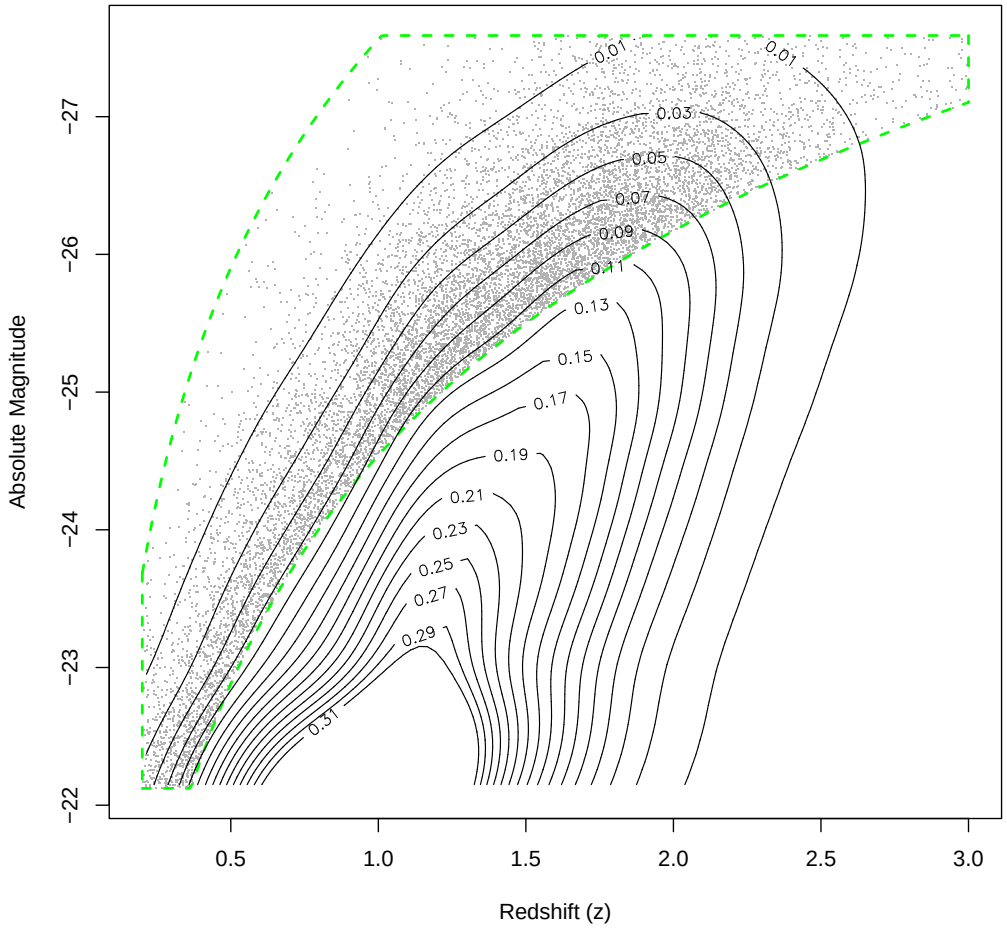

Figure 6: Estimate of bivariate density from the analysis of the SDSS quasar sample.

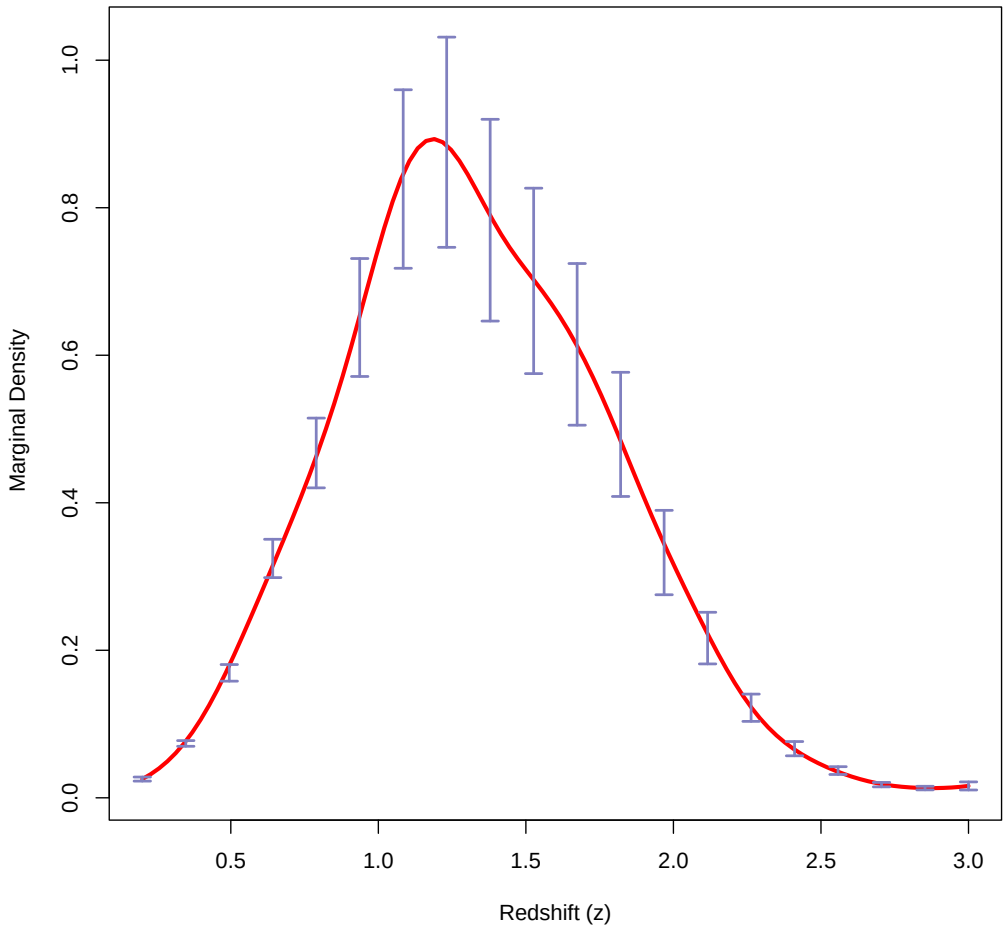

Figure 7: Estimate of marginal distribution for redshift. Error bars represent one standard error.

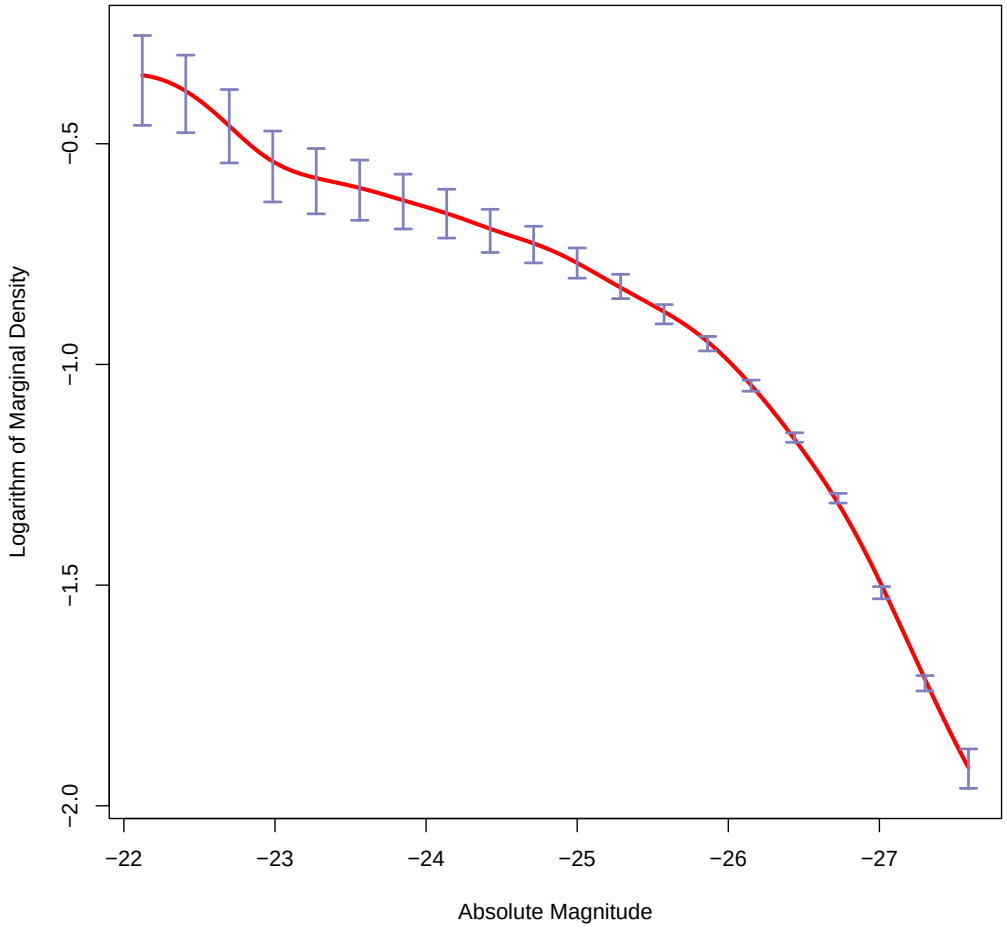

Figure 8: Estimate of log marginal density for absolute magnitude.

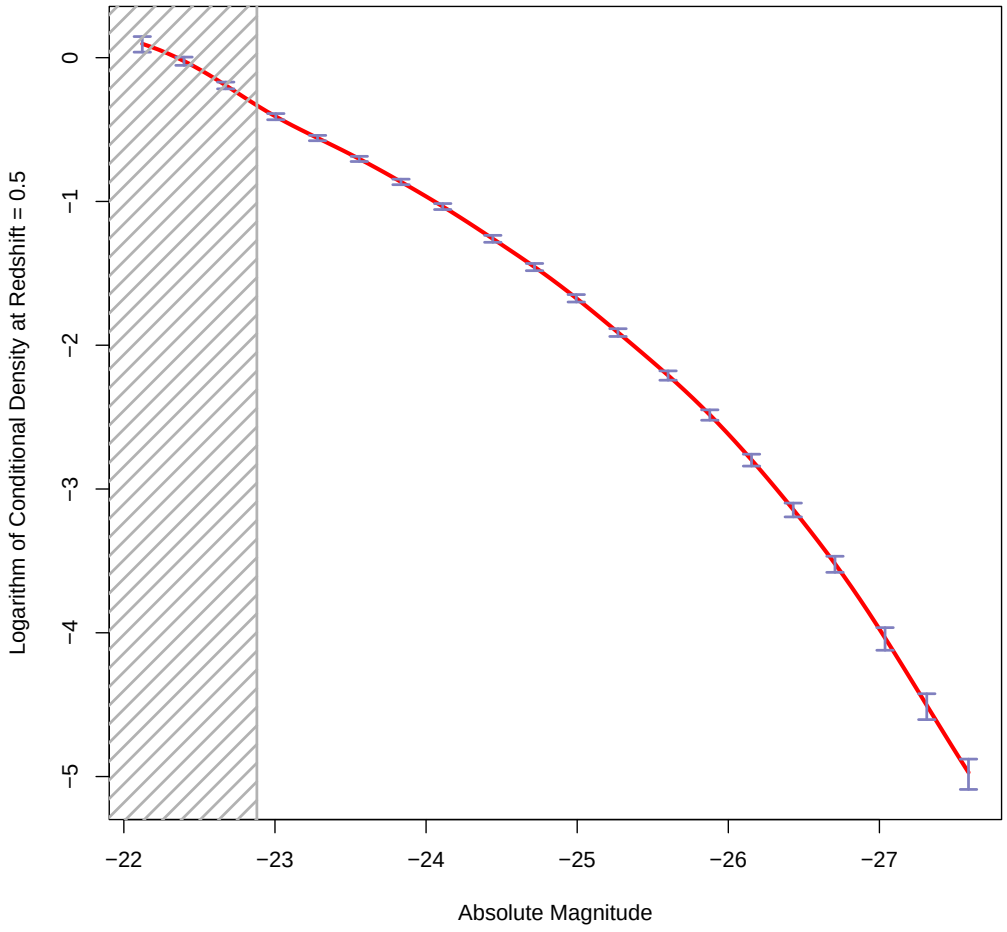

Figure 9: Estimate of the log conditional density (the luminosity function) at redshift of 0.5. The shaded region represents the range over which quasars are unobservable at this redshift.

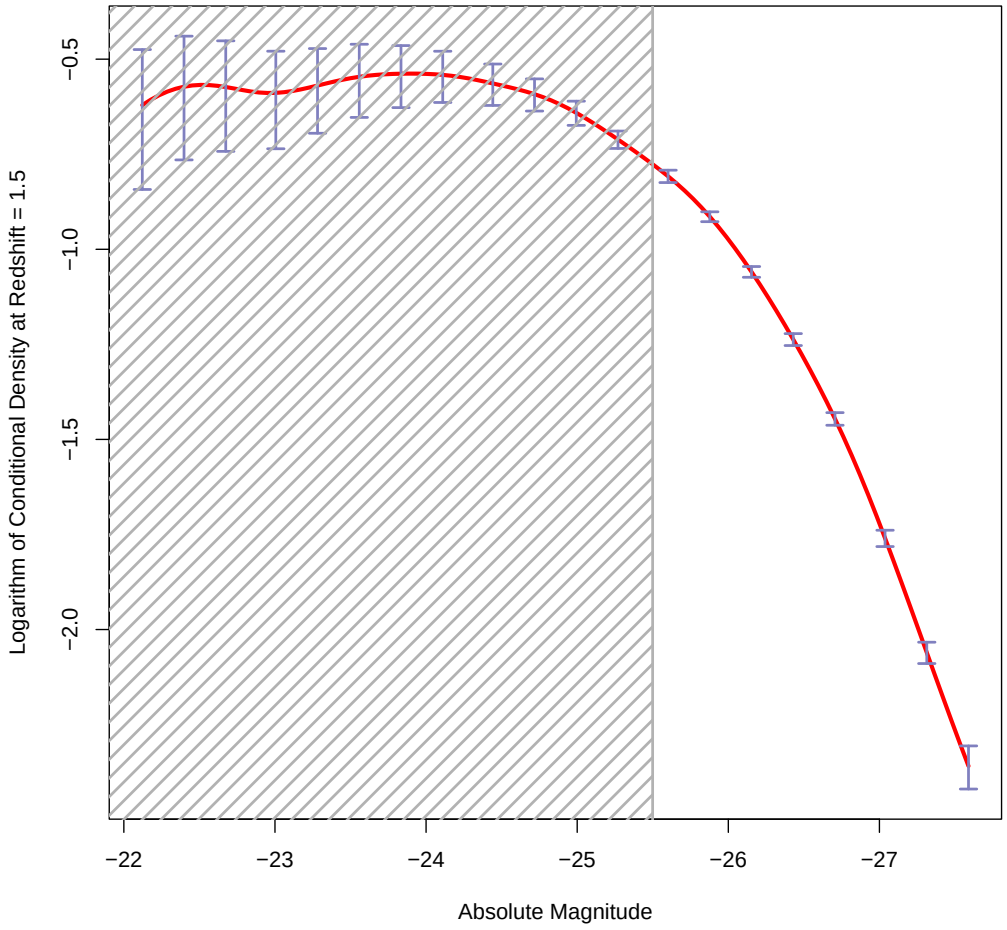

Figure 10: Estimate of the log conditional density (the luminosity function) at redshift of 1.5.