Methods for estimating a time series of densities

Thilaksha Silva
Department of Econometrics & Business Statistics
Monash University, Clayton VIC 3800, Australia
Email: ??

Rob J Hyndman
Department of Econometrics & Business Statistics
Monash University, Clayton VIC 3800, Australia
Email: Rob.Hyndman@monash.edu
Corresponding author

23 April 2019

JEL classification: C13, C14, C32

Methods for estimating a time series of densities

Abstract

We consider the problem of estimating a time series of density functions. A data set comprising many observations is recorded at each time period, and the associated probability density function is to be estimated for each time period. It is assumed that the densities change slowly over time and that neighbouring densities are similar but not identical.

We consider several methods for estimating a time series of densities: (1) A logspline approach applied to each data set separately, where each estimated density has common knots but different coefficients; (2) A conditional logspline approach with a full splines framework to account for the time variation; (3) A conditional logspline approach applied to all data simultaneously with common knots and kernel weights to account for the time variation; (4) A conditional kernel approach with kernel weights to account for the time variation.

In the full splines framework, we allow the degree of the splines to change from 2 to 10, and select the optimal value based on Bayesian information criterion (BIC). For the conditional kernel estimator, we compute two new bandwidth selection approaches that explicitly account for the discrete nature of the time conditioning.

We apply our methods to two simulated examples (unimodal and bimodal) and four real examples. The four data sets comprise UK and Australian income and age data over many years with thousands of observations per year. Probability integral transforms and proper scoring rules such as logarithmic score, quadratic score and spherical score are used to evaluate the density estimates.

Keywords: density estimation, functional data, kernel estimation, time series, splines, probability scoring, density evaluation, income distribution

1 Introduction

We consider methods for nonparametric estimation of a time series of density functions, allowing for smooth changes over time. This is of particular interest when studying probability distributions based on samples collected regularly, such as with income distributions.

Consider a time series of univariate density functions $\{f_t(y)\}_{t=1}^{T}$ with a similar structure using data $\{y_{t,j}\}$ for $t = 1, \ldots, T$ and $j = 1, \ldots, n_t$, where n_t is the number of observations at time t and $\{y_{t,j}\} \in \mathbb{R}$. We assume that the observations are independent of each other, but that the densities $f_t(y)$ are smooth in both y and t.

Previous work on estimating a sequence of density functions include Kneip & Utikal (2001), Park & Marron (1990), and Wand, Marron & Ruppert (1991). However, they ignore the time ordering of the densities and estimate the densities independently for each year, without accounting for smoothness in the t direction. We propose methods to nonparametrically estimate a time series of density functions using either a logspline approach or a kernel approach, taking account of the time ordering of the densities.

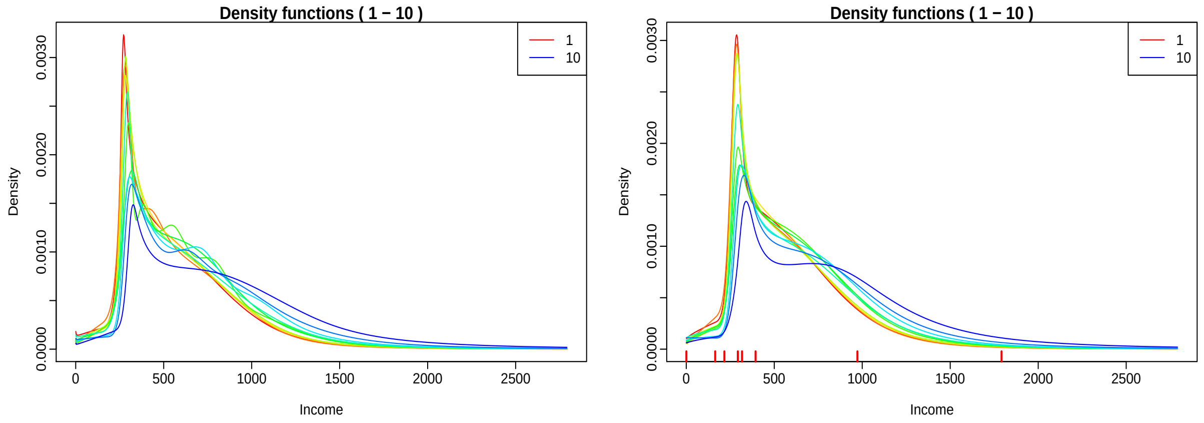
The logspline model was proposed by Kooperberg & Stone (1991), Kooperberg & Stone (1992), and Stone et al. (1997), and is given by

$$\hat{f}_t(y; \boldsymbol{\theta}_t) = \exp\left(b_t(y; \boldsymbol{\theta}_t) - c_t(\boldsymbol{\theta}_t)\right) \tag{1}$$

where $\boldsymbol{\theta}_t = \{\theta_{t,k}\}_{k=1}^{p_t} \in \Theta$ is a vector of parameters, $b_t(y; \boldsymbol{\theta}_t) = \sum\limits_{k=1}^{p_t} \theta_{t,k} B_{t,k}(y)$, $c_t(\boldsymbol{\theta}_t) = \log\left\{\int_{\mathbb{R}} \exp\left(b_t(y; \boldsymbol{\theta}_t)\right) dy\right\}$, $B_{t,k}(y)$ is the kth basis function at time t, $p_t = K_t - 1$ and K_t is the number of knots at time t.

In (1), we could treat every data set separately and obtain a sequence of densities that are estimated independently. However, this is inefficient and ignores the time ordering of the densities. Instead, we consider estimating a time series of density functions using the logspline density approach with the same number and locations of knots at each time $t = 1, 2, \ldots, T$, where each estimated density has common knots but different parameters.

Figure 1 shows Australian income data between 1994/1995–2007/2008. The right plot shows estimates with common knots (denoted using red tick marks), with a consistent structure over time. This method with common knots across different density functions produces meaningful parameters and provides more information about the multiple densities than a situation with different knots.



(a) *Density functions with different knots*

(b) *Density functions with common knots*

Figure 1: *Australian income distributions: 1994/1995–2007/2008. Left: densities estimated with different knots. Right: densities estimated with common knots (snown in red). The curves are ordered chronologically according to the colours of the rainbow, with the oldest years shown in red and the most recent years in blue.*

We will also consider variants of this approach involving an adaptive polynomial framework and kernel weights. We will also include a conditional kernel estimation method for comparison.

Thus we consider and compare four possible approaches to this problem, described in the following four sections. In Section 2 we propose a logspline approach applied to each data set separately, where each estimated density has common knots but different parameters. A new conditional logspline approach is proposed in Section 3 allowing flexible adjustment of the smoothness of conditional density functions via adaptive polynomials. A third method is proposed in Section 4, combining logspline density estimation with kernel weights. In Section 5, we propose the conditional kernel estimator with two new bandwidth selection approaches.

Section 6 compares these methods on six data sets. Then in Section 7, we compare the methods using proper scoring rules, adapted to the situation involving a sequence of densities. We conclude by summarising the main findings and results in Section 8.

2 Logspline with common knots

Let $K(\geq 3)$ be the number of knots and let $\{k'_k\}_{k=1}^{K}$ be a knot sequence with $-\infty < k'_1 < \cdots < k'_K < \infty$. The knot sequence $\{k'_k\}$ is the same for each time series. Let S denote the collection of twice continuously differentiable functions s on $\mathbb{R}$ such that the restriction of s to each of the intervals $(-\infty, k'_1], [k'_1, k'_2], \ldots, [k'_{K-1}, k'_K], [k'_K, \infty)$ is a cubic polynomial. The space S is $(K+4)$-dimensional and the functions in this space are referred to as cubic splines, which have knots at $k'_1, \ldots, k'_K$. Let S_0 denotes the K-dimensional subspace of S consisting of functions $s \in S$ such that s is linear on $(-\infty, k'_1]$ and $[k'_K, \infty)$. Set $p = K - 1$. Then S_0 has a basis of the form $B_0(y) \equiv 1, B_1(y), \ldots, B_p(y)$, and they can be chosen such that $B_1(y)$ is linear with a negative slope on $(-\infty, k'_1]$; $B_2(y), \ldots, B_p(y)$ are constant on $(-\infty, k'_1]$; $B_p(y)$ is linear with a positive slope on $[k'_K, \infty)$; and $B_1(y), \ldots, B_{p-1}(y)$ are constant on $[k'_K, \infty)$.

Let $f_t(y; \boldsymbol{\theta}_t)$ be the probability density function of data $\{y_{t,j}\}_{j=1}^{n_t}$ given $\boldsymbol{\theta}_t = \{\theta_{t,k}\}_{k=1}^{p} \in \Theta$. The logspline model at time t is:

$$f_t(y; \boldsymbol{\theta}_t) = \exp\left(b_t(y; \boldsymbol{\theta}_t) - c_t(\boldsymbol{\theta}_t)\right) \tag{2}$$

where $b_t(y; \boldsymbol{\theta}_t) = \sum\limits_{k=1}^{p} \theta_{t,k} B_k(y)$ and $c_t(\boldsymbol{\theta}_t) = \log\left\{\int_{\mathbb{R}} \exp\left(b_t(y; \boldsymbol{\theta}_t)\right)\, dy\right\}$.

The two differences in (1) and (2) are the number of knots (K_t v. K) and the basis functions ($B_{t,k}(y)$ v. $B_k(y)$). Since the usual logspline density estimation approach treats every data set separately in (1), the number of knots is different in each time series $t = 1, 2, \ldots, T$. Hence, the basis functions are different. However, the logspline model in (2) has a fixed number and fixed locations of knots in each time series t; hence, the basis functions are unchanging over time.

The optimal parameters at time t are estimated maximising the log-likelihood function of the logspline density function at time t is given by:

$$L_t(\boldsymbol{\theta}_t) = \sum_{j=1}^{n_t} \log f_t(y_{t,j}; \boldsymbol{\theta}_t) \tag{3}$$

Stone (1990) showed that the log-likelihood of a univariate logspline density function was a strictly concave function on Θ, and that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}_t$ was unique if it existed.

2.1 Common knots selection

By using the ideas of the knot placement for a single data set for logspline density estimation, we developed a new algorithm to select knots for multiple data sets in an optimal manner. This algorithm consists of four steps. Firstly, the initial number of knots is selected using $2(n^*)^{1/5}$ where n^* depends on the number of observations of each data set $\{n_t\}_{t=1}^{T}$. We propose n^* as the average number of observations, $\bar{n}$. This choice is theoretically supported using a large sample inference for a logspline density estimator with common knots as $\left\|\hat{f}_t(y) - f_t(y)\right\|_2 = O\left(\bar{n}^{-\xi}\right)$ where $f_t(y) = f_t(y; \boldsymbol{\theta}_t)$ is the density function at time t, $\hat{f}_t(y) = f_t(y; \hat{\boldsymbol{\theta}}_t)$ is the logspline density estimate of $f_t(y)$, $\xi = \frac{d}{(2d+1)}$, and $d(> 0.5)$ depends on the smoothness condition of $f_t(y)$ (Stone

1990). Hence, the number of initial knots is taken as $2\bar{n}^{1/5}$. The power $1/5$ is used based on a recommendation by Kooperberg & Stone (1991), and the multiple 2 is somewhat arbitrary and mainly as a result of experience.

Initial knots are placed in the second step of the algorithm. Originally, we placed two boundary knots at the first- and last-order statistics of the pooled data ($\min(y)$ and $\max(y)$). However, because of the linear tail restrictions on the basis functions, we found that placing the boundary knots beyond the first- and last-order statistics works relatively well in practice. We place $k'_1 = \min(y) - \lambda$ and $k'_K = \max(y) + \lambda$ where $\lambda = (\max(y) - \min(y))/4$. If a log transformation is considered for the density estimation and $k'_1 = \min(y) - \lambda$ is non-positive, we consider $k'_1 = \min(y)$. In this paper, we use a close approximation of the Chebyshev procedure (De Boor 1978) to place the initial knots K_{int}, which splits the semicircle into $N - 1$ arcs of equal length to select N points over an interval $[\min(y), \max(y)]$.

After obtaining the number and locations of the initial knots, the next step of the algorithm involves the stepwise addition of knots. The BIC criterion is used for the knot addition: $-2 \log L(\cdot) + \tau p$ where $L(\cdot)$ is the log-likelihood function in (3), p is the number of parameters in the model, and the penalty parameter $\tau = \log n$ where n is the sample size. We calculate the BIC value for each data set and for each knot interval $(k'_1, k'_2), \dots, (k'_{K-1}, k'_K)$. For each knot interval, the average BIC value over all data sets is then calculated. The potential knot with the smallest average BIC is added to the model. Knots are added until the maximum number of knots $3\bar{n}^{1/5}$ is reached and the multiple 3 is chosen based on experience. In the final step, the least important knots are removed successively during stepwise knot deletion where the BIC is also used to measure importance. This procedure is repeated until there are five knots left.

2.2 Constrained optimisation and initial parameters

The splines are constructed in such a way that any linear combination of the spline functions in two tails is linear. For T time series, the sufficient conditions for feasible parameters are:

$$\theta_{1,1} < 0, \dots, \theta_{t,1} < 0, \dots, \theta_{T,1} < 0 \quad \text{and} \quad \theta_{1,p} < 0, \dots, \theta_{t,p} < 0, \dots, \theta_{T,p} < 0. \tag{4}$$

For some data sets (e.g., income data), the logspline estimation process could be applied to log data instead of original data. After density estimation has been performed on the log data, it is a common practice that there will be a spike at the left end. Silverman (1986) discussed an approach of using reflection as a remedy for this spike, while Wand, Marron & Ruppert (1991) indicated another possibility by modifying the family of transformations to ensure that less mass is spread below zero. We found another possibility for removing the spike by constraining the first parameter of the log transformed model. When a log transformation is considered for the density estimation, the sufficient conditions for feasible parameters on the log space are:

$$\eta_{1,1} < \frac{1}{D_1^{(1)}(z)}, \ldots, \eta_{t,1} < \frac{1}{D_1^{(1)}(z)}, \ldots, \eta_{T,1} < \frac{1}{D_1^{(1)}(z)} \quad \text{and} \quad \eta_{1,p} < 0, \ldots, \eta_{t,p} < 0, \ldots, \eta_{T,p} < 0 \tag{5}$$

where $D_1^{(1)}(z)$ is the first derivative of the first basis function for Z. As we use fixed basis functions for each time period, the threshold $\frac{1}{D_1^{(1)}(z)}$ for the first parameter is the same for each t. Constraining the first parameter vector with this positive threshold results that the left end of the back-transformed density function goes to zero, with no spike.

It is well known that optimisation algorithms converge to the solution quickly if the initial parameters are close to the optimal values. In finding the initial parameters, we follow Jin (1990) and Kooperberg (1991). The idea behind constructing initial parameters is minimising the L_2 distance from the score function to its logspline approximation relative to the (empirical) density. We found the initial parameters solving the linear equation: $U\hat{\boldsymbol{\eta}}^* = V$ where U is a $p \times p$ matrix with elements $u_{k,j} = \sum_{t=1}^{T} \sum_{i=1}^{n_t} D_k^{(1)}(z_{t,i}) D_j^{(1)}(z_{t,i})$, V is a column vector of length p with elements $v_j = -\sum_{t=1}^{T} \sum_{i=1}^{n_t} D_j^{(2)}(z_{t,i})$ and $k, j = 1, \ldots, p$.

When original data are used for the density estimation, and if the initial parameters produced by this method do not satisfy the tail constraints in (4), the constraints are imposed directly. In addition, when the log data are used, and if the initial parameters do not satisfy the tail constraints in (5), the constraints are imposed directly.

3 Conditional logspline with adaptive polynomial framework

In Section 2, we proposed a method using logspline density estimation to estimate the densities independently for every year, ignoring all other years. Under normal circumstances, it is reasonable to assume that these univariate densities are related to the ones in nearby years, and it is natural to account the nearby years to the density estimation of a particular year. This is the notion behind the 'conditional density estimation'. Masse & Truong (1999) proposed a spline-based conditional logspline density estimation method which applies to bivariate data where both response and conditioning variables are continuous. In this section, we propose a modified version of the conditional method with two features: conditional density estimation with the discrete support of the conditioning variable, time T'; optimising spline degree—we not only choose the optimal knots but also optimise the spline degree of the density estimate.

The concept of selecting the optimal spline degree for B-spline basis functions has also been used in the R package crs (Racine, Nie & Ripley 2014).

Let $f(y|t)$ be the conditional density function of Y given $T' = t$. $f(y|t)$ is modelled using sets of basis functions on Y and T'. We generate a set of B-spline basis functions on Y as in Section 2 and specify the candidate polynomial degrees of the basis functions from 2 to 10. In contrast, the basis functions in the T' dimension are treated as piecewise cubic polynomials with a fixed degree value 3, and they are generated as follows. Let $J(\geq 3)$ be the number of knots in the T' dimension and let $-\infty < k''_1 < \cdots < k''_J < \infty$ be a knot sequence. Set $q = J - 1$. Let B-spline basis functions in the T' dimension be $G_0(t) \equiv 1, G_1(t), \ldots, G_q(t)$. The boundary basis functions on T' are the same as the ones on Y so that $G_1(t)$ is linear with a negative slope on $(-\infty, k''_1]$; $G_2(t), \ldots, G_q(t)$ are constant on $(-\infty, k''_1]$; $G_q(t)$ is linear with a positive slope on $[k''_J, \infty)$; and $G_1(t), \ldots, G_{q-1}(t)$ are constant on $[k''_J, \infty)$.

The univariate logspline model is extended to the conditional logspline model as:

$$\log f(y|t;\boldsymbol{\alpha}) = \sum_{k=0}^{p} \theta_k(t) B_k(y), \tag{6}$$

$\theta_0(t) = -\log\left\{\int_{\mathbb{R}} \exp\left(\sum_{k=1}^{p}\sum_{j=0}^{q} \alpha_{j,k} G_j(t) B_k(y)\right) dy\right\}$ and $\theta_k(t) = \sum_{j=0}^{q} \alpha_{j,k} G_j(t), k = 1, \ldots, p$ where $\boldsymbol{\alpha} = \{\alpha_{j,k}\} \in \mathcal{A}$ are parameters for $j = 0, \ldots, q$ and $k = 1, \ldots, p$, and $\theta_0(t)$ is the normalising constant so that $\int_{\mathbb{R}} f(y|t;\boldsymbol{\alpha})\, dy = 1$.

The log-likelihood function of the conditional logspline model with a discrete conditioning variable is defined by: $L(\boldsymbol{\alpha}) = \sum_{t=1}^{T}\sum_{i=1}^{n_t} \log f(y_{t,i}|t;\boldsymbol{\alpha}) = \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(\sum_{k=1}^{p}\sum_{j=0}^{q} \alpha_{j,k} G_j(t) B_k(y_{t,i}) - a(\boldsymbol{\alpha})\right)$ where $a(\boldsymbol{\alpha}) = \log\left\{\int_{\mathbb{R}} \exp\left(\sum_{k=1}^{p}\sum_{j=0}^{q} \alpha_{j,k} G_j(t) B_k(y)\right)\ dy\right\}$.

3.1 Knots and spline degree selection

We choose the fixed number of knots in the T' dimension, $J = 5$. In the Y dimension, we follow Section 2.1 to find the initial number and locations of the knots. Having all initial knots plus the first added knot, the optimal spline degree is found by minimising BIC. After each knot addition, a search is made for the best spline degree from the candidate spline degrees from 2 to 10. Knots are added until the maximum number of knots $3\bar{n}^{1/5}$.

3.2 Constrained optimisation and initial parameters

A vector of parameters in the conditional logspline model $\{\alpha_{j,k}\}$ for $j = 0, \dots, q$ and $k = 1. \dots, p$ is said to be feasible if:

$$\alpha_{0,1} < 0, \dots, \alpha_{q,1} < 0 \text{ and } \alpha_{0,p} < 0, \dots, \alpha_{q,p} < 0. \tag{7}$$

In Section 2.2, we proposed a constrained optimisation in univariate logspline model. Following the idea into the conditional logspline model, when a log transformation is considered for the conditional density estimation, the sufficient conditions for feasible parameters on the log space are:

$$\phi_{0,1} < \frac{1}{D_1^{(1)}(z)}, \dots, \phi_{q,1} < \frac{1}{D_1^{(1)}(z)} \text{ and } \phi_{0,p} < 0, \dots, \phi_{q,p} < 0 \tag{8}$$

where $\boldsymbol{\phi} = \{\phi_{j,k}\} \in \Phi$ is a set of parameters on log space for $j = 0, \dots, q$ and $k = 1, \dots, p$.

Masse & Truong (1999) modified Kooperberg (1991)'s approach into the conditional situation by minimising the L_2 distance from the score function to its conditional logspline approximation relative to the (empirical) bivariate density. We adapted Masse & Truong (1999)'s approach in finding the initial parameters when the conditional density has a discrete support on the conditioning variable. When original data are used for the density estimation, and if the initial parameters produced by this method do not satisfy the tail constraints in (7), the constraints are imposed directly. When log data are used, and if the initial parameters do not satisfy the tail constraints in (8), the constraints are imposed directly.

4 Conditional logspline with kernel weights

This section proposes an alternative method for estimating conditional densities combining two density estimation approaches: logspline and kernel. Our suggested method integrates the logspline approach with kernel weights to account for time variations in estimating the conditional density of Y conditional on $T' = t$. This weight function $w(t)$ was proposed by Hyndman, Bashtannyk & Grunwald (1996) for conditional kernel density estimation with continuous response and conditioning variables. It accounts for the neighbouring years in the estimate of the density of a particular year such that it gets the most weight, while the years before and after the specific year get less weight. This results in a series of estimated densities that change slowly over time, with neighbouring densities that are similar but not identical.

The proposed conditional logspline with kernel weights model is:

$$\log f(y|t;\boldsymbol{\theta},h) = \sum_{s=1}^{T}\sum_{k=1}^{p} w_s(t)\theta_{s,k}B_k(y) - r(\boldsymbol{\theta},h) \tag{9}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t=1}^{T}$, $w_s(t) = \frac{K\left(\frac{t-T_s'}{h}\right)}{\sum\limits_{s=1}^{T} K\left(\frac{t-T_s'}{h}\right)}$, $r(\boldsymbol{\theta},h) = \log\left\{\int_{\mathbb{R}} \exp\left(\sum\limits_{s=1}^{T}\sum\limits_{k=1}^{p} w_s(t)\theta_{s,k}B_k(y)\right)\, dy\right\}$, h is the bandwidth in T' dimension, and $K(\cdot)$ is a kernel function satisfying the properties:

$$\sum K(v) = 1, \sum vK(v) = 0,\ \sigma_K^2 = \sum v^2K(v) < \infty,\ R(K) = \sum K^2(v) < \infty \text{ and } G(K) = \sum v^2K^2(v) < \infty. \tag{10}$$

4.1 Optimisation methods

The conditional logspline model with kernel weights is parameterised on the data-dependent knots and the bandwidth. Two optimisation methods of the model involves finding the optimal parameters and the optimal bandwidth. Firstly, we calculate the log-likelihood function $L_t(\boldsymbol{\theta}_t)$ for each t as in (3) and find the optimal parameters for each time period by maximising the corresponding log-likelihood function. Secondly, the leave-one-out maximum likelihood estimation is used for finding the optimal bandwidth $\hat{h}$.

The log-likelihood function of the conditional logspline with kernel weights density function at time t is given by:

$$L_t(h) = \sum_{i=1}^{n_t} \log f(y_{t,i}|t;\boldsymbol{\theta},h) \tag{11}$$

where $\log f(y|t;\boldsymbol{\theta},h)$ is defined as in (9). This suggests that the optimal bandwidth $\hat{h}$ could be found by maximising the sum of the log-likelihood functions over all T data sets:

$$\hat{h} = \arg\max_{h}\ l(h) \quad \text{where} \quad l(h) = \sum_{t=1}^{T} L_t(h). \tag{12}$$

We found that the maximum likelihood estimation has a trivial maximum at $h = 0$. This implies that there is no difference between the conditional density estimates and the univariate density estimates. Therefore, the leave-one-out maximum likelihood estimation is invoked by replacing $\log f(y_{t,i}|t;\boldsymbol{\theta},h)$ in (11) by the leave-one-out logspline density estimation model at time t:

$$\log f_{-t}(y|t;\boldsymbol{\theta},h) = \sum_{s(\neq t)=1}^{T}\sum_{k=1}^{p} w_s(t)\theta_{s,k}B_k(y) - r_{-t}(\boldsymbol{\theta},h) \tag{13}$$

where $r_{-t}(\boldsymbol{\theta}, h) = \log \left\{ \int_{\mathbb{R}} \exp \left(\sum_{s(\neq t)=1}^{T} \sum_{k=1}^{p} w_s(t) \theta_{s,k} B_k(y) \right) dy \right\}$. Using leave-one-out log-likelihood functions over all T data sets, $\{L_t(h)\}_{t=1}^{T}$, the optimal bandwidth $\hat{h}$ could be found by maximising the sum of the leave-one-out log-likelihood functions as in (12).

5 Conditional kernel estimation

Bandwidth selection is the most important aspect of the kernel density estimation. After the introduction of conditional kernel density, Bashtannyk & Hyndman (2001) proposed several bandwidth selection approaches ranging from fast rule-of-thumb approaches to slow hi-tech approaches with continuous response and conditioning variables. In this section, we propose two bandwidth selection approaches for the conditional kernel estimation with the discrete conditioning variable.

The conditional kernel density estimator of Y given $T' = t$ is:

$$\hat{f}(y|t) = \frac{1}{h_y} \sum_{s=1}^{T} \sum_{j=1}^{n_s} w_s(t) K\left(\frac{y - Y_{s,j}}{h_y} \right) \tag{14}$$

where $\{Y_{s,1}, \cdots, Y_{s,n_s}\}$ is a sample of observations at time s, n_s is the number of observations at time s, $w_s(t) = \frac{K\left(\frac{t-T'_s}{h_t}\right)}{\sum_{s=1}^{T} n_s K\left(\frac{t-T'_s}{h_t}\right)}$, $K(\cdot)$ is a kernel function satisfying (10), and h_t and h_y are bandwidths for conditioning variable T' and response variable Y respectively.

The asymptotic bias and variance of the density estimator $\hat{f}(y|t)$ are calculated in then appendix. Adding squared bias to the variance gives the asymptotic mean square error for the estimator:

$$\begin{aligned} \text{AMSE}\hat{f}(y|t) = \frac{h_y^4 \sigma_K^4}{4} \left\{ \frac{\partial^2 f(y|t)}{\partial y^2} \right\}^2 + \frac{R(K) f(y|t)}{n h_t h_y m_t} \left[R(K) - h_y f(y|t) \right] + O\left(h_y^6\right) + O\left(h_t^4\right) \\ + O\left(h_y^2 h_t^2\right) + O\left(\frac{h_y^2}{n}\right) + O\left(\frac{h_t}{n}\right) + O\left(\frac{h_y h_t}{n}\right) + O\left(\frac{h_y}{n h_t}\right). \end{aligned} \tag{15}$$

This estimator is consistent provided that $h_t \to 0$, $h_y \to 0$, $n \to \infty$, $nh_t \to \infty$ as $n \to \infty$ and $nh_t h_y \to \infty$ as $n \to \infty$. The IMSE, derived by taking the integral with respect to y and summing over all values of $t = 1, 2, \ldots, T$, is in the form:

$$\text{IMSE}\hat{f}(y|t) \approx \frac{A}{n h_t h_y} - \frac{B}{n h_t} + D h_y^4 \tag{16}$$

where the constants A, B and D are given by: $A = \sum_{t=1}^{T} R^2(K)$, $B = \sum_{t=1}^{T} \left(\int_{\mathbb{R}} R(K) f^2(y|t)\, dy\right)$, and $D = \sum_{t=1}^{T} \left(\int_{\mathbb{R}} \frac{\sigma_K^4 m_t}{4} \left\{\frac{\partial^2 f(y|t)}{\partial y^2}\right\}^2 dy\right)$. Differentiating (16) with respect to h_y and setting the derivatives to zero leads to the approximate optimal bandwidth of h_y:

$$\hat{h}_y = \left\{\frac{A}{4nD}\right\}^{1/5} \hat{h}_t^{-1/5}. \tag{17}$$

Discrete uniform reference rule (Discrete URR)

Assuming the conditional distribution is normal with linear mean $\delta_1 + \delta_2 t$ and linear standard deviation $\delta_3 + \delta_4 t$, the conditional distribution could be written as $[Y|T' = t] \sim \mathcal{N}(\delta_1 + \delta_2 t, (\delta_3 + \delta_4 t)^2)$. Based on the assumption that the conditional standard deviation is constant ($\delta_4 = 0$), the optimal bandwidth $\hat{h}_y$ of h_y is:

$$\hat{h}_y = \left\{\frac{8\sqrt{\pi} R^2(K) T \delta_3^5}{3n\sigma_K^4}\right\}^{1/5} \hat{h}_t^{-1/5} \tag{18}$$

where $\hat{h}_t$ is the optimal bandwidth of h_t.

We have observed that $\hat{h}_t$ could not be found by differentiating (16) with respect to h_t. Alternatively, we use $\hat{h}_t$ expression of Bashtannyk & Hyndman (2001) URR method. They assumed continuous conditioning variable is uniformly distributed over the sample space. Our conditioning variable is non-random discrete and uniformly distributed if approximately the same number of observations exist in each year. Therefore, assuming that the conditioning variable is uniformly distributed, the optimal bandwidth in the T' dimension is found from Bashtannyk & Hyndman (2001), p286:

$$\hat{h}_t = \left\{\frac{4\sqrt{\pi} R^2(K) T \delta_3^5}{3n\sigma_K^4 \delta_2^5}\right\}^{1/6}. \tag{19}$$

A plug-in bandwidth selection approach

Following Hall & Marron (1987) and Wand, Marron & Ruppert (1991), a new plug-in bandwidth selection approach is developed in this section. This approach plugs a value for the integrated squared derivatives of the conditional density where the unknown conditional density estimator is considered normal.

The integrated squared r derivative of conditional kernel density estimator of $f(y|t)$ is:

$$\int_{\mathbb{R}} \left(\hat{f}^{(r)}(y|t)\right)^2 dy = \frac{\frac{1}{\left(h_y^{2r+1}\right)^2} \sum\limits_{s_2=1}^{T} \sum\limits_{s_1=1}^{T} \sum\limits_{j=1}^{n_{s_2}} \sum\limits_{i=1_{i\neq j}}^{n_{s_1}} K\left(\frac{t-T'_{s_1}}{h_t}\right) K\left(\frac{t-T'_{s_2}}{h_t}\right) \int_{\mathbb{R}} K^{(r)}\left(\frac{y-Y_{s_1,i}}{h_y}\right) K^{(r)}\left(\frac{y-Y_{s_2,j}}{h_y}\right) dy}{\left(\sum\limits_{s=1}^{T} n_s K\left(\frac{t-T'_s}{h_t}\right)\right)^2} \tag{20}$$

where $K^{(r)}(\cdot)$ is the r derivative of the kernel function.

Following Hall & Marron (1987) idea in univariate density estimation, we propose the integrated squared second derivative of the conditional kernel density estimator of $f(y|t)$:

$$\int_{\mathbb{R}} \left(\hat{f}^{(2)}(y|t)\right)^2 \ dy = \frac{\frac{1}{h_y^5} \sum\limits_{s_2=1}^{T} \sum\limits_{s_1=1}^{T} \sum\limits_{j=1}^{n_{s_2}} \sum\limits_{i=1_{i\neq j}}^{n_{s_1}} K\left(\frac{t-T'_{s_1}}{h_t}\right) K\left(\frac{t-T'_{s_2}}{h_t}\right) K^{(4)}\left(\frac{Y_{s_1,i}-Y_{s_2,j}}{h_y}\right)}{\left(\sum\limits_{s=1}^{T} n_s K\left(\frac{t-T'_s}{h_t}\right)\right)^2} \tag{21}$$

We choose the Gaussian $\mathcal{N}(0, h_y^2)$ kernel function in the Y dimension. The fourth derivative of the kernel function for $(Y_{s_1,i} - Y_{s_2,j})$ is given by:

$$\frac{1}{h_y^5} K^{(4)}\left(\frac{Y_{s_1,i} - Y_{s_2,j}}{h_y}\right) = \frac{1}{\sqrt{2\pi} h_y^5} \exp\left\{-\frac{z}{2}\right\}\left(z^2 - 6z + 3\right). \tag{22}$$

where $z = \left(\frac{Y_{s_1,i}-Y_{s_2,j}}{h_y}\right)^2$. Following Wand, Marron & Ruppert (1991) idea in univariate density estimation, h_y is defined as $h_y = \sqrt{2} h'_y$ where the pilot bandwidth $h'_y = \hat{\sigma}_Y \left(\frac{84\sqrt{\pi}}{5\bar{n}^2}\right)^{1/13}$, $\hat{\sigma}_Y$ is the standard deviation of pooled Y data and $\bar{n}$ is the average number of observations.

Substituting (22) into (21), we get the integrated squared second derivative of the conditional density kernel density estimator of $f(y|t)$:

$$\int_{\mathbb{R}} \left(\hat{f}^{(2)}(y|t)\right)^2 \ dy = \frac{\frac{1}{\sqrt{2\pi} {h_y}^5} \left[\sum\limits_{s_2=1}^{T} \sum\limits_{s_1=1}^{T} \sum\limits_{j=1}^{n_{s_2}} \sum\limits_{i=1}^{n_{s_1}} K\left(\frac{t-T'_{s_1}}{h_t}\right) K\left(\frac{t-T'_{s_2}}{h_t}\right) \mathrm{Y} - 3 \sum\limits_{s=1}^{T} n_s \left(K\left(\frac{t-T'_s}{h_t}\right)\right)^2\right]}{\left(\sum\limits_{s=1}^{T} n_s K\left(\frac{t-T'_s}{h_t}\right)\right)^2} \tag{23}$$

where $\mathrm{Y} = \exp\left\{-\frac{z}{2}\right\}\left(z^2 - 6z + 3\right)$, $z = \left(\frac{Y_{s_1,i}-Y_{s_2,j}}{h_y}\right)^2$, $h_y = \sqrt{2} h'_y$, $h'_y = \hat{\sigma}_Y \left(\frac{84\sqrt{\pi}}{5\bar{n}^2}\right)^{1/13}$, $\hat{\sigma}_Y$ is the standard deviation of pooled Y data, $\bar{n}$ is the average number of observations and h_t is obtained from (19). With the integrated square of the second derivative of the density estimator of $f(y|t)$ in (23), the optimal bandwidth in the Y dimension, $\hat{h}_y$, is obtained from (17).

6 Application

This section presents the results obtained from four estimation methods discussed in Sections 2-5. We apply our methods to two simulated data sets (unimodal and bimodal) and four real data sets. In this paper, the density estimation of all six data sets is performed on log-transformed data, and then back-transformed to obtain the densities on the original scale.

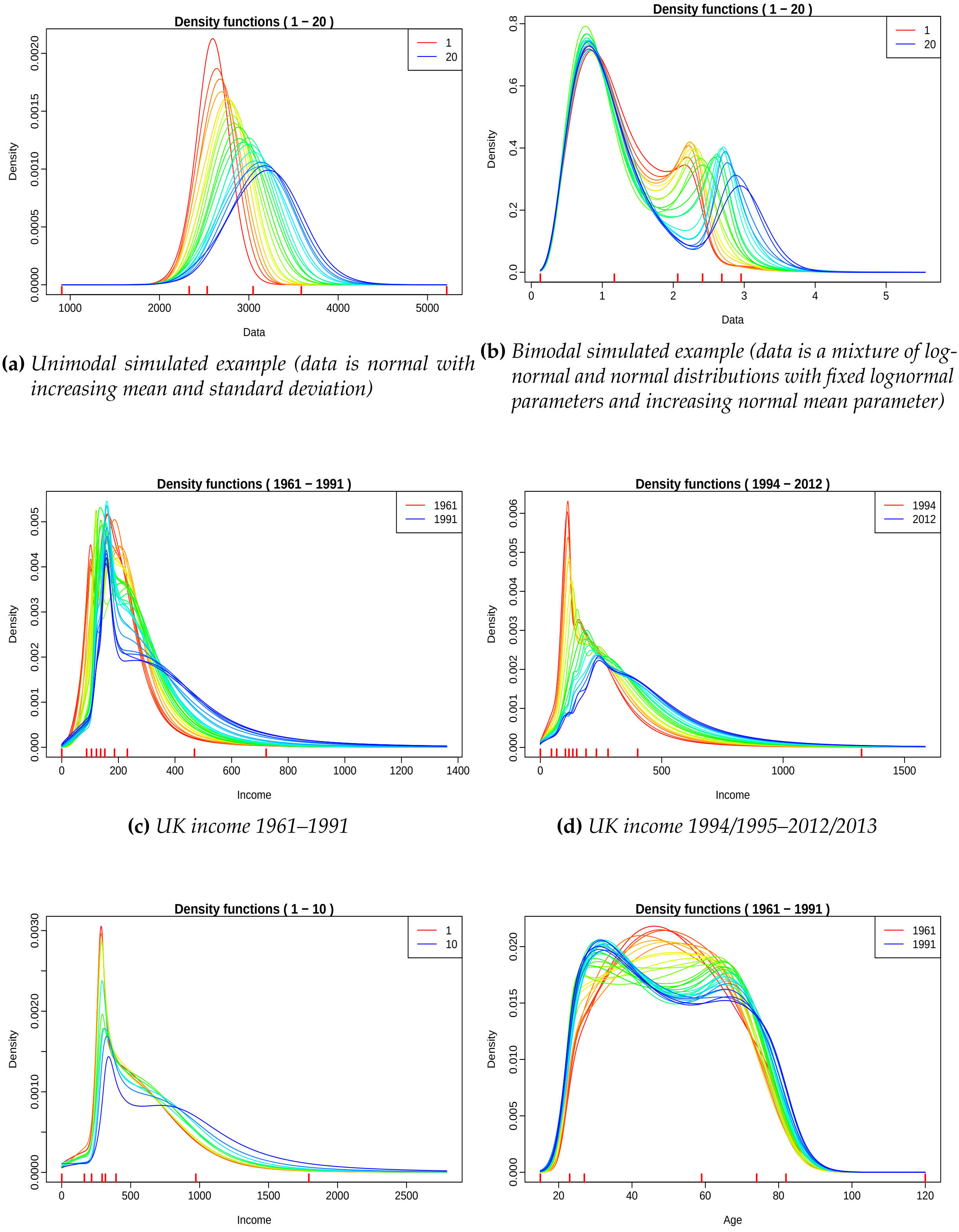
6.1 Logspline with common knots

Application	Number of series	Average number of observations	Initial knots	After stepwise knot addition	Optimal knots
Unimodal example	20	2000	9	9	6
Bimodal example	20	2000	9	11	7
UK income 1961–1991	31	6066	11	13	11
UK income 1994/1995–2012/2013	19	24725	15	18	13
Australian income 1994/1995–2007/2008	10	9432	12	14	9
UK age 1961–1991	31	6100	11	11	7

Table 1: *Results of the logspline with common knots method.*

Figure 2 shows the behaviour of the density functions over time. For a clear view, graphs based on income data are drawn up to a limited value on the income axis. The two plots in the first row of Figure 2 depict data from a known series of densities. The estimated densities evidently show the changing behaviour in the simulated data. The data for Figure 2(c) consist of UK income data for 31 years between 1961 and 1991. The sharp peak is due to the UK national old-age pension, which caused many people to have nearly identical incomes. The height and location of the spike are accurately estimated by the method, which places at least two initial knots near the spike.

The data for Figure 2(d) comprise UK income data for a period of 19 years between 1994/1995 and 2012/2013. Over time, the height of the mode consistently decreases, and the range of the income values increases. The data for Figure 2(e) contain Australian income data for 10 years between 1994/1995 and 2007/2008. The series of densities lookalike in Figure 2(c) and the sharp peak decreases over time. The data for Figure 2(f) consists of UK age data over 31 years between 1961 and 1991. Densities of age are visibly different from the densities of income. Age densities



(a) *Unimodal simulated example (data is normal with increasing mean and standard deviation)*

(b) *Bimodal simulated example (data is a mixture of log-normal and normal distributions with fixed lognormal parameters and increasing normal mean parameter)*

(c) *UK income 1961–1991*

(d) *UK income 1994/1995–2012/2013*

(e) *Australian income 1994/1995–2007/2008*

(f) *UK age 1961–1991*

Figure 2: *A series of density functions estimated using the logspline with common knots method. The positions of the knots are displayed at the bottom. The oldest years are shown in red and the most recent years are in blue. The curves are ordered chronologically according to the colours of the rainbow.*

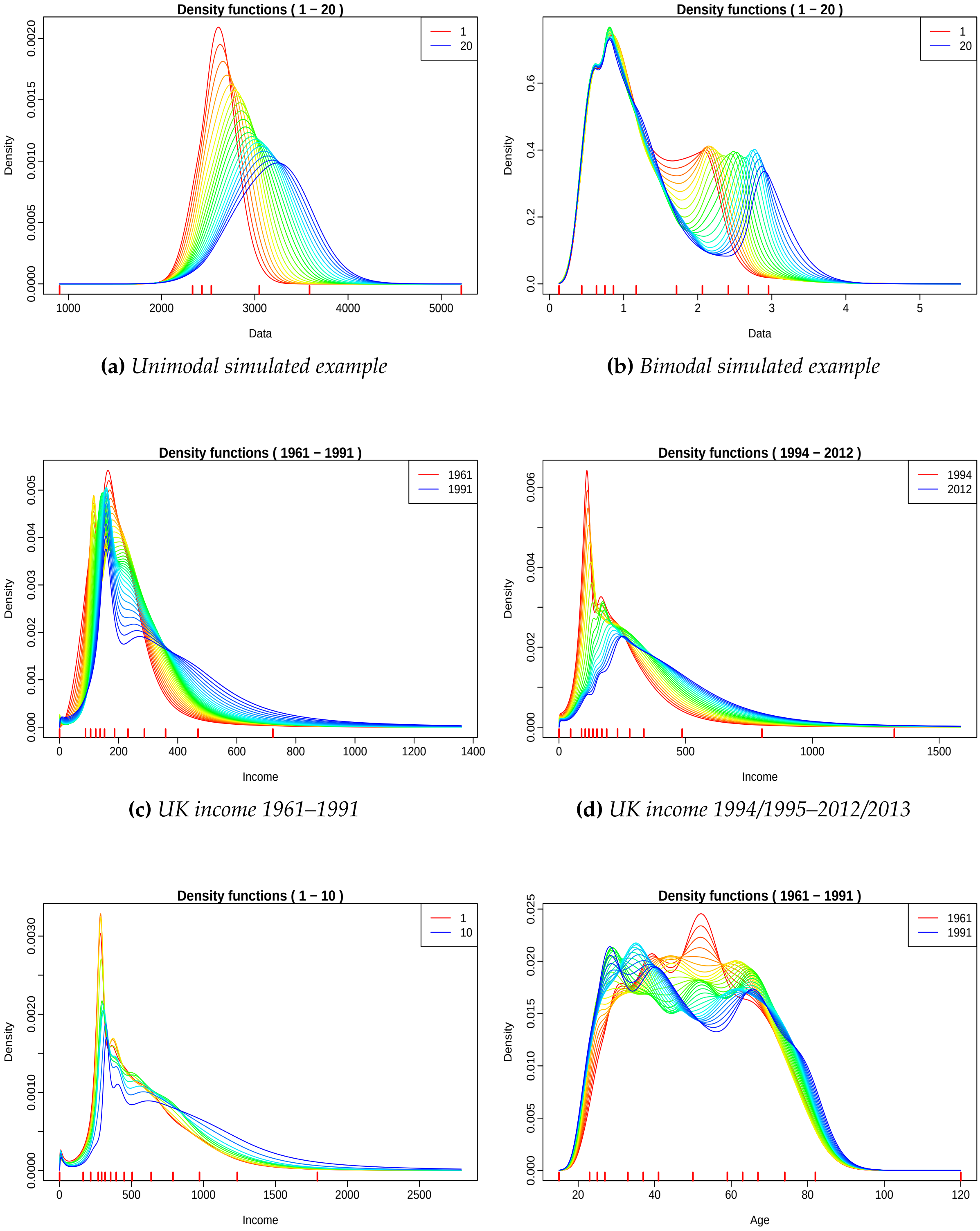
at early years are unimodal, and then the distribution tends to have two modes. The summary of the results is presented in Table 1.

6.2 Conditional logspline with adaptive polynomial framework

Application	Initial knots	After stepwise knot addition	Optimal knots	Optimal degree
Unimodal example	9	10	7	3
Bimodal example	9	12	12	2
UK income 1961–1991	11	13	13	5
UK income 1994/1995–2012/2013	15	18	15	5
Australian income 1994/1995–2007/2008	12	16	16	5
UK age 1961–1991	11	14	13	6

Table 2: *Results of the conditional logspline with adaptive polynomial framework method.*

Figure 3 displays the conditional density plots that were estimated using a conditional logspline approach with an adaptive polynomial framework. Conditional logspline density plots in Figure 3 are looking smoother than the corresponding sequential logspline density plots in Figure 2. The summary of the knots in the Y dimension, along with the optimal spline degree, is shown in Table 2. The optimal degree of each example shows that most density estimates can plausibly have a higher spline degree, resulting in relatively wiggly density functions.



(a) *Unimodal simulated example*

(b) *Bimodal simulated example*

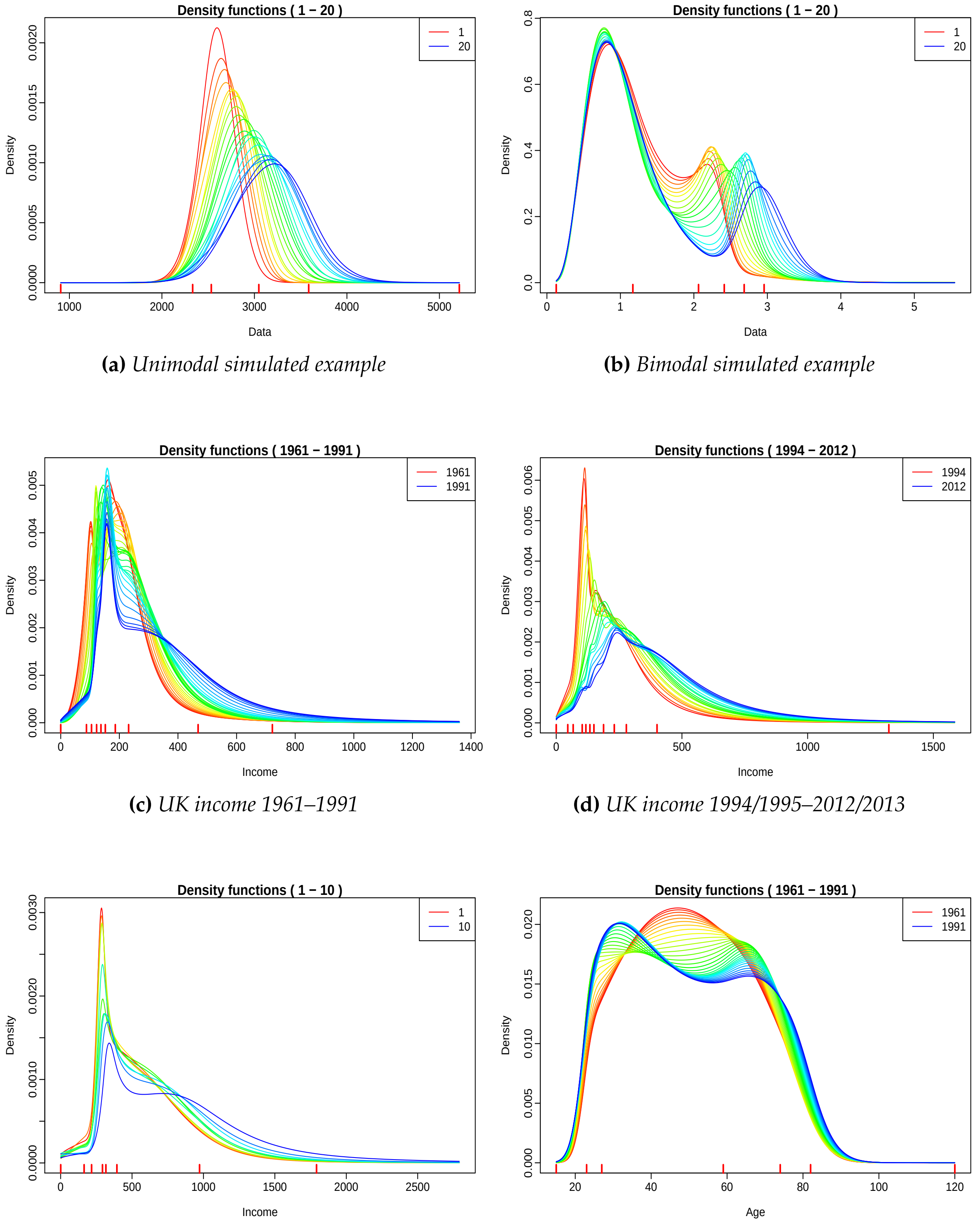
(c) *UK income 1961–1991*

(d) *UK income 1994/1995–2012/2013*

(e) *Australian income 1994/1995–2007/2008*

(f) *UK age 1961–1991*

Figure 3: *A series of conditional density functions estimated using the conditional logspline with adaptive polynomial framework method. The positions of the knots are displayed at the bottom. The oldest years are shown in red and the most recent years are in blue. The curves are ordered chronologically according to the colours of the rainbow.*



(a) *Unimodal simulated example*

(b) *Bimodal simulated example*

(c) *UK income 1961–1991*

(d) *UK income 1994/1995–2012/2013*

(e) *Australian income 1994/1995–2007/2008*

(f) *UK age 1961–1991*

Figure 4: *A series of conditional density functions estimated using the conditional logspline with kernel weights method. The positions of the knots are displayed at the bottom. The oldest years are shown in red and the most recent years are in blue. The curves are ordered chronologically according to the colours of the rainbow.*

6.3 Conditional logspline with kernel weights

Application	Initial knots	After stepwise knot addition	Optimal knots	Optimal bandwidth in the T' dimension
Unimodal example	9	9	6	0.32
Bimodal example	9	11	7	0.99
UK income 1961–1991	11	13	11	0.99
UK income 1994/1995–2012/2013	15	18	13	0.31
Australian income 1994/1995–2007/2008	12	14	9	0.10
UK age 1961–1991	11	11	7	2.09

Table 3: *Results of the conditional logspline with kernel weights method.*

Combining the logspline density estimation method with kernel weights, we obtain conditional density functions in Figure 4. The density functions are smoothly changing over time and look smoother compared to the corresponding plots in Figure 3. The summary of the results of this method with an optimal bandwidth in the T' dimension is shown in Table 3. As this method considers the results from the logspline with common knots method, the summary of the number of knots in the Y dimension is the same as in Table 1.

6.4 Conditional kernel estimation

We implemented two existing bandwidth selection approaches in Bashtannyk & Hyndman (2001)—the URR and the bootstrap approach—with our continuous response variable Y and the discrete conditioning variable T'. The existing approaches assume that both response and conditioning variables are continuous random variables. The discreteness in the conditioning variable conflicts with the theory behind the existing approaches, although it works in practice. Also, we implemented two bandwidth selection approaches proposed in this paper.

A careful examination of the conditional kernel plots (not included in this paper) reveals that the conditional densities from the bootstrap method differ from the conditional densities obtained using the other three bandwidth selection methods. This may be because the other three methods are related. Both the URR and the discrete URR assumes that the underlying conditional density is normal. The plug-in approach plugs in a value for the second derivative of the conditional density; however, the pilot distribution is also considered normal. A summary of the bandwidths

Application	h_t	h_y
Unimodal example		
URR	1.45	0.0151
Bootstrap	0.94	0.0189
Discrete URR	1.45	0.0176
Plug-in	1.45	0.0182
Bimodal example		
URR	13.2	0.0536
Bootstrap	5.27	0.0778
Discrete URR	13.2	0.0622
Plug-in	13.2	0.0554
UK income 1961–1991		
URR	3.67	0.0530
Bootstrap	1.84	0.0688
Discrete URR	3.67	0.0612
Plug-in	3.67	0.0544
UK income 1994/1995–2012/2013		
URR	1.54	0.0594
Bootstrap	0.85	0.0772
Discrete URR	1.54	0.0690
Plug-in	1.54	0.0586
Australian income 1994/1995–2007/2008		
URR	1.59	0.0661
Bootstrap	0.64	0.0958
Discrete URR	1.59	0.0775
Plug-in	1.59	0.0625
UK age 1961–1991		
URR	49.4	0.0209
Bootstrap	19.8	0.0417
Discrete URR	49.4	0.0241
Plug-in	49.4	0.0200

Table 4: *Results of the conditional kernel method.*

of four bandwidth selection methods is presented in Table 4. The bandwidths in the T' and Y dimensions are shown in columns h_t and h_y respectively.

6.5 Computational efficiency

Table 5 shows the computational time for each density estimation method and for each example, using an HP L1906 computer with a Intel(R) Core(TM) i5-2400 CPU @ 3.10 GHz processor, 4GB of RAM and 4-core parallel processing. Each example is a large data set with total number of observations of at least 40,000.

URR and discrete URR are the least time-consuming methods. It is not surprising as both URR and discrete URR are reference rule methods. Two logspline models—logspline with common knots method and conditional logspline with kernel weights method—have much less

Method	Ex. 1 ($\bar{n}$=2000) (T=20)	Ex. 2 ($\bar{n}$=2000) (T=20)	Ex. 3 ($\bar{n}$=6066) (T=31)	Ex. 4 ($\bar{n}$=24725) (T=19)	Ex. 5 ($\bar{n}$=9432) (T=10)	Ex. 6 ($\bar{n}$=6066) (T=31)
Univariate log.	0:52	1:82	7:48	58:13	5:72	4:59
Con. log.-spline	73:46	69:68	121:65	531:93	74:04	176:33
Con. log.-kernel	0:52	1:85	7:66	58:58	5:77	4:82
Con. kernel						
URR	0:8	0:9	0:24	0:58	0:13	0:34
Bootstrap	46:38	33:72	144:29	391:86	58:39	96:04
Discrete URR	0:8	0:9	0:26	1:04	0:15	0:35
Plug-in	3:86	3:86	125:25	574:24	15:72	125:72

Table 5: *Mean computation time for different methods. Times are measured in minutes:seconds. $\bar{n}$ and T denote the average number of observations and the number of data sets respectively. The abbreviations Ex., Con and log. denote the example, Conditional and logspline respectively.*

computational time. The bootstrap method and plug-in method are time consuming, but the most tedious task is the conditional logspline with adaptive polynomial framework method.

The best performing strategy with good computational time seems to be the conditional logspline with kernel weights method.

7 Evaluation

After the estimated densities have been obtained, it is necessary to assess how close they are to the actual densities. In the usual density evaluation, a series of densities over a certain period is considered to have one observation for every time period. Nonetheless, in this paper, we have thousands of observations for each time period. Therefore, we propose the following method for evaluating the reliability of density estimates.

Let $f_t(y)$ and $\hat{f}_t(y)$ be the true density and estimated density at time t respectively. Each density has thousands of observations. Given that the estimates should be evaluated by assessing whether the estimated densities are correct, we test whether the collection of estimated densities is equal to the collection of true densities: $\{f_t(y)\}_{t=1}^T = \{\hat{f}_t(y)\}_{t=1}^T$. It is difficult to test this null hypothesis, as the true density $\hat{f}_t(y)$ is never obtained for real data sets. However, the properties of the following modified probability integral transformation provide the solution to this problem:

$$\left\{ \{u_{t,j}\}_{j=1}^{n_t} = \left\{ \int_{-\infty}^{y_{t,j}} \hat{f}_{t,j}(v) \;\, dv \right\}_{j=1}^{n_t} \right\}_{t=1}^{T} \tag{24}$$

Rosenblatt (1952) showed that the distribution of PIT is uniform(0,1) for the true generating density. In this situation, we assess whether the distribution of PITs under the null hypothesis is

Application	UPIT
Unimodal example	20/20
Bimodal example	20/20
UK income 1961–1991	31/31
UK income 1994/1995–2012/2013	19/19
Australian income 1994/1995–2007/2008	10/10
UK age 1961–1991	29/31

Table 6: *Evaluation results of the logspline with common knots method.*

uniform(0,1). Specifically, if the realisations at time t, $\{y_{t,j}\}_{j=1}^{n_t}$, come from the estimated densities at time t, $\hat{f}_t(y)$, then the series of PITs at time t, $\{u_{t,1}, \ldots, u_{t,n_t}\}$, is uniform(0,1). If the series of PITs deviates from the uniform(0,1), then the estimated density at time t, $\hat{f}_t(y)$, is not identical to the true density, $f_t(y)$.

A summary of the evaluation results of the logspline with common knots method is presented in Table 6. The number of uniformly distributed PIT histograms out of the total number of histograms is stated under the column 'UPIT'. Kolmogorov–Smirnov test is used for testing uniformity. As the insignificant p-values (0.04153, 0.04621) in the sixth example are close to the significance level 0.05, we conclude that all univariate estimated densities obtained from logspline with common knots method are not significantly different from the corresponding series of true densities.

In this paper, we compare six conditional density estimation methods (two conditional logspline methods in Sections 3 and 4, and four bandwidth selection methods in the conditional kernel estimation in Section 5) using four proper scoring rules: LS, QS, SS and CRPS. Proper scoring rules assess calibration and sharpness simultaneously, and the method producing the lowest proper scoring rule score is preferred (Gneiting, Balabdaoui & Raftery 2007; Gneiting & Katzfuss 2014; Gneiting & Raftery 2007). A summary of the evaluation results of these methods for six examples is presented in Tables 7–12. Blue-coloured bold values and black-coloured bold values denote the minimum and next minimum values of proper scoring rules respectively. As shown in Tables 7–12, two conditional logspline methods outperform four bandwidth selection methods in conditional kernel estimation, and the bootstrap method performs better than the other three bandwidth selection methods in conditional kernel density estimation. However, it can be seen that four types of scores are extremely close. Therefore, the statistical significance of the scores is evaluated to test the equal estimation performance.

The significance of the scores is statistically tested using the Diebold–Mariano test (Gneiting & Katzfuss 2014, p.137). This is a paired t-test that can be used to test for equal performance of two competing reliable density estimation methods. By performing paired t-tests, we can conclude

Method	UPIT(/20)	LS	QS	SS	CRPS
Conditional logspline-spline	19	**7.09**	**-0.000984**	**-0.0312**	**168**
Conditional logspline-kernel	20	**7.09**	**-0.000984**	**-0.0312**	**168**
Conditional kernel					
URR	19	7.12	-0.000957	-0.0308	172
Bootstrap	19	7.12	-0.000958	-0.0308	173
Discrete URR	18	7.12	-0.000953	-0.0307	173
Plug-in	18	7.12	-0.000952	-0.0307	173

Table 7: *Evaluation results of six conditional methods for unimodal simulated example.*

Method	UPIT(/20)	LS	QS	SS	CRPS
Conditional logspline-spline	20	**0.941**	**-0.456**	**-0.675**	**0.432**
Conditional logspline-kernel	20	**0.941**	**-0.457**	**-0.676**	**0.430**
Conditional kernel					
URR	3	0.999	-0.436	-0.660	0.438
Bootstrap	12	0.996	-0.437	-0.661	0.440
Discrete URR	3	1.00	-0.435	-0.660	0.438
Plug-in	3	0.999	-0.436	-0.660	0.438

Table 8: *Evaluation results of six conditional methods for bimodal simulated example.*

Method	UPIT(/31)	LS	QS	SS	CRPS
Conditional logspline-spline	14	**6.18**	**-0.00272**	**-0.0519**	**78.9**
Conditional logspline-kernel	26	**6.18**	**-0.00272**	**-0.0520**	**78.4**
Conditional kernel					
URR	8	6.21	-0.00266	-0.0515	80.3
Bootstrap	18	6.21	-0.00267	-0.0515	80.7
Discrete URR	8	6.21	-0.00265	-0.0514	80.4
Plug-in	8	6.21	-0.00266	-0.0515	80.3

Table 9: *Evaluation results of six conditional methods for real example 1: UK income 1961–1991.*

Method	UPIT(/19)	LS	QS	SS	CRPS
Conditional logspline-spline	11	**6.60**	**-0.00186**	**-0.0428**	**127**
Conditional logspline-kernel	19	**6.60**	**-0.00186**	**-0.0429**	**126**
Conditional kernel					
URR	6	6.62	-0.00183	-0.0425	130
Bootstrap	10	6.62	-0.00183	-0.0426	130
Discrete URR	6	6.62	-0.00182	-0.0425	130
Plug-in	6	6.62	-0.00183	-0.0425	130

Table 10: *Evaluation results of six conditional methods for real example 2: UK income 1994/1995–2012/2013.*

Method	UPIT(/10)	LS	QS	SS	CRPS
Conditional logspline-spline	5	**7.15**	**-0.00105**	**-0.0322**	**207**
Conditional logspline-kernel	10	**7.15**	**-0.00105**	**-0.0322**	**207**
Conditional kernel					
URR	2	7.18	-0.00101	-0.0317	210
Bootstrap	5	7.18	-0.00100	-0.0316	213
Discrete URR	2	7.18	-0.00100	-0.0316	211
Plug-in	2	7.17	-0.00101	-0.0318	210

Table 11: *Evaluation results of six conditional methods for real example 3: Australian income 1994/1995–2007/2008.*

Method	UPIT(/31)	LS	QS	SS	CRPS
Conditional logspline-spline	20	**4.16**	**-0.0164**	**-0.128**	**9.73**
Conditional logspline-kernel	22	**4.16**	**-0.0164**	**-0.128**	**9.72**
Conditional kernel					
URR	0	4.18	-0.0161	-0.127	9.88
Bootstrap	1	4.18	-0.0160	-0.127	9.89
Discrete URR	1	4.18	-0.0161	-0.127	9.86
Plug-in	0	4.19	-0.0162	-0.127	9.90

Table 12: *Evaluation results of six conditional methods for real example 4: UK age 1961–1991.*

that two conditional logspline methods do not differ in performance in producing sharp density estimates; however, they perform better than four bandwidth selection methods in conditional kernel estimation.

8 Concluding remarks

We considered several methods for estimating a time series of densities. In the logspline with common knots method, a logspline approach was applied to each data set separately, where each estimated density had common knots but different coefficients. We found that the PITs of the logspline density estimates in each example were uniformly distributed, thus concluding that the logspline with common knots method is a reliable density estimation method.

We also proposed four conditional density estimation methods: (1)~The spline-based conditional logspline method allowed the degree of the splines to change from 2 to 10, and we selected the optimal value based on the BIC; (2)~In conditional logspline with kernel weights method, a logspline approach was applied to all data simultaneously, with common knots and kernel weights to account for the time variation. An optimal bandwidth was selected by maximising the sum of the leave-one-out log-likelihood functions over all T data sets; (3) and (4)~We proposed two new bandwidth selection methods for conditional kernel estimation that explicitly accounted for the discrete nature of the time conditioning.

The proposed four conditional density estimation methods were compared with conditional kernel estimation using two existing bandwidth selection methods: URR and bootstrap. The relative accuracy of these six conditional density estimates was measured using proper scoring rules by assessing both calibration and sharpness. Testing for equal estimation performance, we concluded that the two proposed conditional logspline methods outperformed the four conditional kernel density estimation methods. Both conditional logspline methods were reliable and also produced equally sharp density estimates.

9 Supplementary documents

An R package was developed with the implementations mentioned in this paper. It estimates a time series of density functions using logspline approach and can be accessed by https://github.com/ThilakshaSilva/densityEst.

10 Appendix

10.1 Lemma

Let T' be a discrete variable with probability mass function $m_t = \frac{n_t}{n}$, n_t be the number of observations at time t, n be the total number of observations, $q(y|t)$ be an at least twice continuously differentiable function at time t and defined on the sample space of Y, $K(\cdot)$ be a kernel function satisfying the discrete version of (10), and h_t be a constant. Then as $h_t \to 0$:

$$\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = q(y|t)m_t + O\left(h_t^2\right), \tag{25}$$

$$\mathrm{E}\left[\frac{1}{h_t^2}K^2\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = \frac{q(y|t)m_t R(K)}{h_t} + O\left(h_t\right) \tag{26}$$

and:

$$\mathrm{var}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = q^2(y|t)m_t\left[\frac{R(K)}{h_t} - m_t\right] + O\left(h_t\right). \tag{27}$$

Proof:

$$\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = \frac{1}{h_t}\sum_s K\left(\frac{t-s}{h_t}\right)q(y|s)m_s$$

If roughly the same number of observations is taken for each data set, $n_s = n_t$ and hence, $m_s = m_t$. Then, we get: $\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = \sum_u K(u)q(y|(t-uh_t))m_t$ where $u = \frac{t-s}{h_t}$.

With different s values, $q(y|t)$ could be estimated as: $q(y|t) = \frac{1}{h_t}\sum_{s=1}^{T} K\left(\frac{t-s}{h_t}\right)q(y|s)$.

Then we could write: $q(y|s) = q(y|(t - uh_t)) = q(y|t) + O\left(h_t\right)$ and

$$\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t - T'}{h_t}\right)q(y|T')\right] = \sum_u K(u)\left[q(y|t) + uh_t q'(y|t) + O\left(h_t^2\right)\right]m_t.$$

Therefore, the expectation becomes: $\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = q(y|t)m_t + O\left(h_t^2\right).$

Using a similar argument, we obtain (26). Then using (25) and (26), (27) is derived as follows: $\mathrm{var}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right] = \mathrm{E}\left[\frac{1}{h_t^2}K^2\left(\frac{t-T'}{h_t}\right)q^2(y|T')\right] - \left\{\mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'}{h_t}\right)q(y|T')\right]\right\}^2.$

10.2 Derivations

The asymptotic bias and the variance of the conditional density estimator $\hat{f}(y|t)$ are derived as:

$$\mathrm{E}[\hat{f}(y|t)] - f(y|t) = \frac{h_y^2\sigma_K^2}{2}\frac{\partial^2 f(y|t)}{\partial y^2} + O\left(h_y^4\right) + O\left(h_t^2\right) + O\left(h_y^2h_t^2\right) + O\left(\frac{h_y^2}{n}\right) + O\left(\frac{h_t}{n}\right) + O\left(\frac{h_y^2h_t}{n}\right) + O\Bigg(\frac{h_y^2}{nh_t} \tag{28}$$

$$\mathrm{var}[\hat{f}(y|t)] = \frac{R(K)f(y|t)}{nh_th_ym_t}\left[R(K) - h_yf(y|t)\right] + O\left(\frac{h_y^2}{n}\right) + O\left(\frac{h_t}{n}\right) + O\left(\frac{h_yh_t}{n}\right) + O\left(\frac{h_y}{nh_t}\right). \tag{29}$$

Proof:

The conditional density estimator can be expressed as the ratio of two random variables: $\hat{f}(y|t) = \frac{\frac{1}{n}\sum\limits_{s=1}^{T}\sum\limits_{j=1}^{n_s} q_1(T'_s, Y_{s,j})}{\frac{1}{n}\sum\limits_{s=1}^{T}\sum\limits_{j=1}^{n_s} q_2(T'_s)}$ where $q_1(T'_s, Y_{s,j})$ and $q_2(T'_s)$ are two random variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, and with the covariance σ_{12}^2. Following Lemma 2 in Hyndman, Bashtannyk & Grunwald (1996), pp.333–335, the expected value and the variance of the estimator $\hat{f}(y|t)$ can be taken as:

$$\mathrm{E}[\hat{f}(y|t)] \approx \frac{\mu_1}{\mu_2} + \frac{1}{n}\left(\frac{\mu_1\sigma_2^2}{\mu_2^3} - \frac{\sigma_{12}^2}{\mu_2^2}\right) \tag{30}$$

$$\mathrm{var}[\hat{f}(y|t)] \approx \frac{1}{n\mu_2^2}\left(\sigma_1^2 + \frac{\mu_1^2\sigma_2^2}{\mu_2^2} - \frac{2\mu_1\sigma_{12}^2}{\mu_2}\right). \tag{31}$$

<u>Finding μ_2 and σ_2^2</u> \ By applying Lemma in Section 10.1, we obtain: $\mu_2 = \mathrm{E}\left[\frac{1}{h_t}K\left(\frac{t-T'_s}{h_t}\right)\right] = m_t + O\left(h_t^2\right)$ and \ $\sigma_2^2 = \mathrm{var}\left[\frac{1}{h_t}K\left(\frac{t-T'_s}{h_t}\right)\right] = m_t\left[\frac{R(K)}{h_t} - m_t\right] + O\left(h_t\right).$

<u>Finding μ_1</u> \ Conditioning on T'_s, the expectation: $\mathrm{E}\left[\frac{1}{h_th_y}K\left(\frac{t-T'_s}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)|T'_s\right] = \frac{1}{h_t}K\left(\frac{t-T'_s}{h_t}\right)\mathrm{E}\left[\frac{1}{h_y}K\left(\frac{y-Y_{s,j}}{h_y}\right)|T'_s\right].$

Then applying Lemma 1 in Hyndman, Bashtannyk & Grunwald (1996), pp.331–332, gives:

$$\frac{1}{h_t}K\left(\frac{t-T_s'}{h_t}\right)\mathrm{E}\left[\frac{1}{h_y}K\left(\frac{y-Y_{s,j}}{h_y}\right)|T_s'\right]=\frac{1}{h_t}K\left(\frac{t-T_s'}{h_t}\right)\left[f(y|T_s')+\frac{h_y^2\sigma_K^2}{2}\frac{\partial^2 f(y|T_s')}{\partial y^2}+O\left(h_y^4\right)\right]. \tag{32}$$

Applying Lemma in Section 10.1 gives the unconditional expectation: \ $\mu_1 = \mathrm{E}\left[\mathrm{E}\left[\frac{1}{h_t h_y}K\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)|T_s'\right]\right]=m_t\left[f(y|t)+\frac{h_y^2\sigma_K^2}{2}\frac{\partial^2 f(y|t)}{\partial y^2}\right]+O\left(h_y^4\right)+O\left(h_t^2\right).$

Finding σ_1^2 \ $\sigma_1^2 = \mathrm{var}\left[\frac{1}{h_t h_y}K\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)\right]$ can be re-written as: $\sigma_1^2 = V_1 + V_2$ where \ $V_1 = \mathrm{var}\left[\mathrm{E}\left[\frac{1}{h_t h_y}K\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)|T_s'\right]\right]$ and $V_2 = \mathrm{E}\left[\mathrm{var}\left[\frac{1}{h_t h_y}K\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)|T_s'\right]\right]$.

To find V_1, get the variance of (32): $V_1 = f^2(y|t)m_t\left[\frac{R(K)}{h_t}-m_t\right]+O\left(h_y^2\right)+O\left(h_t\right)+O\left(\frac{h_y^2}{h_t}\right).$

To find V_2, firstly, apply Lemma 1 in Hyndman, Bashtannyk & Grunwald (1996), pp.331–332, for the variance conditioning on T_s'. Then the expectation of the result gives V_2:

$$V_2=\frac{m_t R(K)}{h_t}\left[f(y|t)\left[\frac{R(K)}{h_y}-f(y|t)\right]+\frac{h_y G(K)}{2}\frac{\partial^2 f(y|t)}{\partial y^2}\right]+O\left(h_t\right)+O\left(\frac{h_y^2}{h_t}\right).$$

The unconditional variance σ_1^2 is the sum of V_1 and V_2.

Finding σ_{12}^2 \ Conditioning on T_s' and applying Lemma 1 in Hyndman, Bashtannyk & Grunwald (1996), pp.331–332, gives:

$$\mathrm{E}\left[\frac{1}{h_t^2 h_y}K^2\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)|T_s'\right]=\frac{1}{h_t^2}K^2\left(\frac{t-T_s'}{h_t}\right).\left[f(y|T_s')+\frac{h_y^2\sigma_K^2}{2}\frac{\partial^2 f(y|T_s')}{\partial y^2}+O\left(h_y^4\right)\right].$$

Then applying Lemma in Section 10.1 gives: \

$$\mathrm{E}\left[\frac{1}{h_t^2 h_y}K^2\left(\frac{t-T_s'}{h_t}\right)K\left(\frac{y-Y_{s,j}}{h_y}\right)\right]=\frac{m_t R(K)}{h_t}\left[f(y|t)+\frac{h_y^2\sigma_K^2}{2}\frac{\partial^2 f(y|t)}{\partial y^2}\right]+O\left(h_t\right)+O\left(\frac{h_y^4}{h_t}\right).$$

Subtracting $\mu_1\mu_2$ from the above expression gives the covariance: \

$$\sigma_{12}^2=m_t f(y|t)\left[\frac{R(K)}{h_t}-m_t\right]+O\left(h_y^2\right)+O\left(h_t\right)+O\left(h_y^2 h_t^2\right)+O\left(\frac{h_y^2}{h_t}\right).$$

Finding the asymptotic bias of the conditional density estimator $\hat{f}(y|t)$ \ Using the result $1/(a+b)=1/a-b/a^2+O(b)$, we obtain $\frac{\mu_1}{\mu_2}$, $\frac{\mu_1\sigma_2^2}{\mu_2^3}$ and $\frac{\sigma_{12}^2}{\mu_2^2}$. Hence, from (30), we obtain (28).

Finding the asymptotic variance of the conditional density estimator $\hat{f}(y|t)$ \ We obtain $\frac{\sigma_1^2}{\mu_2^2}$, $\frac{\mu_1^2\sigma_2^2}{\mu_2^4}$ and $\frac{\mu_1\sigma_{12}^2}{\mu_2^3}$. Hence, from (31), we obtain (29).

References

Bashtannyk, DM & RJ Hyndman (May 2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis* **36**(3), 279–298.

De Boor, C (1978). *A practical guide to splines*. Springer, New York.

Gneiting, T, F Balabdaoui & AE Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of Royal Statistical Society* **69**(Part 2), 243–268.

Gneiting, T & M Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Applications* **1**, 125–151.

Gneiting, T & AE Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Hall, P & JS Marron (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters* **6**, 109–115.

Hyndman, RJ, DM Bashtannyk & GK Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* **5**(4), 315–336.

Jin, K (1990). "Empirical smoothing parameter selection in adaptive estimation". PhD thesis. University of California, Berkeley.

Kneip, A & KJ Utikal (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96**(454), 519–542.

Kooperberg, C (1991). "Smoothing images, curves and densities". PhD thesis. University of California, Berkeley.

Kooperberg, C & CJ Stone (Nov. 1991). A study of logspline density estimation. *Computational Statistics & Data Analysis* **12**, 327–347.

Kooperberg, C & CJ Stone (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**(4), 301–328.

Masse, BR & YK Truong (1999). Conditional logspline density estimation. *The Canadian Journal of Statistics* **27**(4), 819–832.

Park, BU & JS Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of American Statistical Association* **85**(409), 66–72.

Racine, J, Z Nie & BD Ripley (2014). *crs: Categorical Regression Splines*. R package version 0.15-24. http://cran.r-project.org/package=crs.

Rosenblatt, M (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23**(3), 470–472.

Silverman, BW (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.

Stone, CJ (1990). Large-sample inference for log-spline models. *The Annals of Statistics* **18**(2), 717–741.

Stone, CJ, MH Hansen, C Kooperberg & YK Truong (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics* **25**(4), 1371–1425.

Wand, MP, JS Marron & D Ruppert (1991). Transformations in density estimation. *Journal of the American Statistical Association* **86**(414), 343–353.