



City Research Online

City St George's, University of London

Citation: Bagkavos, D., Guillen, M. & Nielsen, J. P. (2024). Nonparametric conditional survival function estimation and plug-in bandwidth selection with multiple covariates. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 33(4), pp. 1225-1257. doi: 10.1007/s11749-024-00945-7

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/33387/>

Link to published version: <https://doi.org/10.1007/s11749-024-00945-7>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

Nonparametric conditional survival function estimation and plug-in bandwidth selection with multiple covariates

Dimitrios Bagkavos · Montserrat Guillen · Jens P. Nielsen

Received: date / Revised: date

Abstract The present research provides two methodological advances, simulation evidence and a real data analysis, all contributing to the area of local linear survival function estimation and bandwidth selection. The first contribution is the development of a double smoothed local linear survival function estimator which admits an arbitrary number of covariates and the analytic establishment of its asymptotic properties. The second contribution is the efficient implementation of the estimator in practice. This is achieved by developing an automatic plug-in smoothing parameter selector which optimizes the estimator's performance in all coordinate directions. The traditional problem of vectorization of higher-order derivatives which lead to increasingly intractable matrix algebraic expressions is addressed here by introducing an alternative vectorization that exploits the analytic relationships between the functionals involved. This yields simpler, tractable and efficient in terms of computing time expressions which greatly facilitate the implementation of the rule in practice. The analytic study of the rule's rate of convergence shows that in contrast to the traditional cross validation approach, the proposed bandwidth selector is functional even for a large number of covariates. The benefits of all methodological advances are illustrated with the analysis of a motivating real-world dataset on credit risk.

Keywords Bandwidth selection · Censoring · Local linear estimation · Kernel smoothing

MSC Classification 62G05, 62N02, 62P10, 62G3

1 Introduction.

The present research considers the local polynomial modeling of the conditional survival function in the randomly right censored data setting in the presence of an arbitrary number ($d \geq 1$) of covariates. The proposed estimator is based on the double smoothing approach, i.e., it is obtained as the intercept estimator of a local least squares approximation to a kernel smoothed version of the empirical distribution function and embodies covariate effects via multidimensional hyperplane fits in small regions of the input space. This scheme allows the estimator to effectively capture the nonlinear structure of the underlying curve, without imposing any specific relationship among the covariates. Precise mathematical formulation and justification is provided in Section 2.1: establishing the asymptotic unbiasedness of the Kaplan–Meier when conditioning on the covariate values prompts its use as the response in a nonparametric regression model for the underlying conditional cumulative distribution function (c.d.f.) with the sample points used as the design. The correspondence between the local hyperplane coefficients and the derivatives in a Taylor expansion of the regression function at a nearby point, naturally yields the proposed smooth estimator of the conditional c.d.f. and subsequently of the conditional survival function. As discussed in Section 2.2, the benefit of the approach lies on that the local polynomial smoothing of randomly right censored data automatically reinstates the kernel mass falling outside the region of estimation back in. As a consequence, the resulting estimator automatically adjusts at the boundary without the negative side effects caused by manual methods such as the use of boundary kernels.

Dimitrios Bagkavos
Department of Mathematics, University of Ioannina, Ioannina, 45110, Greece
E-mail: dimitrios.bagkavos@gmail.com

Montserrat Guillen
Department of Econometrics, Statistics and Applied Economics, Universitat de Barcelona, Catalonia, Spain

Jens P. Nielsen
Bayes Business School, City, University of London, UK.

The present approach can be thought as the extension to the conditional survival function setting of the density estimators discussed in Fan et al. (1996) and Fan and Yim (2004). It is also related to the conditional survival function estimators discussed in Kim et al. (2010) and Spierdijk (2008) and to the unconditional local linear density estimator of Cheng (1997) and Cheng et al. (1997). However, all conditional estimators in the aforementioned articles are developed and studied analytically assuming a single covariate. Even though one can incorporate multiple covariates as a single variable via the indexing technique, e.g. Strzalkowska-Kominiak and Cao (2013), inclusion of covariates in the functional form of the estimator as separate variables serves the much more realistic need in practice to quantify the impact of each covariate on the overall survival separately from each other, and thus enhance the inference drawn. At the same time though, it raises multiple methodological issues, including plug-in bandwidth selection in multiple dimensions which is known to be a challenging problem, e.g. Wand and Jones (1994), Yang and Tschernig (1999), Chacón and Duong (2018), as well as selection of the optimal combination of variables to include in the model. Due to the complexity of both issues, the former is addressed in detail here while the latter is treated in separate research which is already underway. It should be noted that the literature is rather thin on bandwidth selection for conditional c.d.f. estimators under censoring. Plug-in bandwidth selection methodology for conditional survival function estimation based on the estimator of Beran (1981) was developed recently in Peláez et al. (2021a,b, 2022a,b) in the credit scoring setting (Dirick et al. 2003; Allen and Rose 2006; Beran and Djaidja 2007; Cao et al. 2009). Bandwidth selection is also discussed in Peláez et al. (2024); their bandwidth rule is employed on a nonparametric cure model estimator of the conditional survival function which is based on the approach of López-Cheda (2018); López-Cheda et al. (2017a,b) – see also Amico and Van Keilegom (2018) and the references therein. To the best of our knowledge there is no other plug-in bandwidth selection method for the survival function in the right censored data setting in the literature. However, see Meira-Machado et al. (2015) for a comparative study of conditional survival function estimators and the references therein for an account of important contributions on the subject up to that point.

The contributions of the present article start with analytical quantification, in Theorem 1 below, of the asymptotic properties of the proposed conditional survival function estimator. Theorem 1 is the basis for developing a data driven, consistent plug-in bandwidth selector so as to facilitate the implementation of the proposed estimator. As can be immediately seen by Theorem 1, implementation of the rule requires the estimation of a – fast growing as the number of covariates increases – number of functionals. This issue is addressed here by exploiting the analytic relationship between the functionals appearing in the analytic expression of the rule, hence – in contrast to CV based approaches – allowing its effective deployment in practice even for a large number of covariates. The analytical results established herein include, in Theorem 2 below, its uniqueness as a minimizer of the MISE and quantification, in Theorem 3, of the rate of convergence of its data-driven version to the *ideal* MISE bandwidth, thus verifying its more stable performance over CV based bandwidths.

The practical performance of the proposed bandwidth rule in comparison to its ideal version is discussed via distributional data simulations in Section 4.1. The inference of the proposed methodology (estimator and bandwidth selector) is illustrated with the analysis of a motivating credit data set in Section 4.2. Section 5 summarizes the conclusions of the research and provides directions for future work. Assumptions, auxiliary lemmas and proofs of the results established herein are contained in Section 6 while trivial calculations, implementation details and additional numerical examples are deferred to the Supplementary Material.

2 Local linear conditional c.d.f. estimation.

2.1 Notation and motivation of the proposed estimator.

Let $T > 0$ be a random variable representing survival time from a probability density function (p.d.f.) f_T and c.d.f. F_T . T is (possibly) censored on the right by the random variable $U > 0$ which comes from c.d.f. G . In the sequel, matrices are denoted in bold, vectors in italic bold and univariate quantities in plain typeface. The covariates that affect survival at each given time point and are typically denoted as a d -variate vector $\mathbf{X}$, assumed to be coming from a d -variate p.d.f. $f_{\mathbf{X}}$ and c.d.f. $F_{\mathbf{X}}$. Given $\mathbf{X} = \mathbf{x} = (x_1, \dots, x_d)^\top$, the c.d.f. of the observed duration $Y = \min(T, U)$ is F . The conditional distribution functions of Y and U are $F(\cdot|\mathbf{X})$ and $G(\cdot|\mathbf{X})$ respectively while $F_T(\cdot|\mathbf{X})$ is the conditional distribution function of T . The objective is to estimate $F_T(y|\mathbf{X} = \mathbf{x}) = P(T \leq y|\mathbf{X} = \mathbf{x})$ based on an i.i.d. random sample $(\mathbf{X}_i, Y_i, \delta_i), i = 1, \dots, n$ from the population $(\mathbf{X}, Y, \delta)$, $\delta = \mathbf{1}_{\{T \leq U\}}$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function of the event $\{\cdot\}$, and for fixed i , $\mathbf{X}_i = (X_{i1}, \dots, X_{id})^\top$. It is assumed that $y \in [0, T_Y]$ where $T_Y = \sup\{y : 1 - F_T(y) > \varepsilon\}$ for a small $\varepsilon > 0$. For fixed i , set

$$\mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} = \prod_{j=1}^d k \left(\frac{X_{ij} - x_j}{h_j} \right), \quad W(y) = \int_{-\infty}^y w(u) du.$$

That is, $\mathbf{K}$ denotes a d -variate *kernel* function which is used to smooth the data in the x coordinate directions, $\mathbf{H} = \text{diag}\{h_1, \dots, h_d\}$ is a diagonal, positive definite $d \times d$ *bandwidth* matrix, denoting the amount of smoothing applied to the data in each direction via the univariate kernel function k , which is usually taken to be a symmetric p.d.f.. The kernel function k can be different than the kernel function w (also usually taken to be a symmetric p.d.f.) which is used to smooth the data in the y axis, using bandwidth $b > 0$. Assumptions A1-A3 and A6 in Section 6.1 summarize the properties of the kernel functions and the bandwidths utilized here; note that these are standard assumptions satisfied by all kernels and bandwidths used in practice. Denote with $\hat{F}_T(y)$ the well-known the Kaplan–Meier estimate of the unconditional c.d.f. $F_T(y)$, introduced in Kaplan and Meier (1958) and defined by

$$\hat{F}_T(y) = \begin{cases} 0, & \text{if } 0 \leq y \leq Z_1, \\ 1 - \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1} \right)^{\Lambda_i}, & \text{if } Z_{k-1} < y \leq Z_k, k = 2, \dots, n, \\ 1, & \text{if } y > Z_n, \end{cases} \quad (1)$$

where $(Z_i, \Lambda_i), i = 1, \dots, n$ denote the ordered Y_i 's, along with their corresponding δ_i 's. By Lemma 1, for a fixed $i \in \{1, \dots, n\}$ and for $0 < \beta < 1/2$

$$\mathbb{E} \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] = F_T(y|\mathbf{x}) \{1 + O(b^2) + O_p(n^{-\beta})\}, \quad (2)$$

where $1 - \hat{G}(y)$ is the Kaplan–Meier estimator of $1 - G$. Estimator $1 - \hat{G}(y)$ results by inverting in (1) the intuitive role played by the Y_i and δ_i and slightly altering the estimator's definition in the case where $y > Z_n$ so as to avoid division by zero when the estimator is used in the denominator, see Marron and Padgett (1987). Specifically, $1 - \hat{G}(y)$ is given by

$$1 - \hat{G}(y) = \begin{cases} 1, & \text{if } 0 \leq y \leq Z_1, \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2} \right)^{1-\Lambda_i}, & \text{if } Z_{k-1} < y \leq Z_k, k = 2, \dots, n, \\ \prod_{i=1}^n \left(\frac{n-i+1}{n-i+2} \right)^{1-\Lambda_i}, & \text{if } Z_n < y. \end{cases}$$

Also, by Theorem 2.1, Chen and Lo (1997), for $0 < \beta < 1/2$,

$$\sup_{y \in [0, T_Y]} |\hat{F}_T(y) - F_T(y)| = o(n^{-\beta}) \text{ a.s..} \quad (3)$$

By (2) and (3) it is reasonable to set

$$m(y|\mathbf{X}_i = \mathbf{x}) = \mathbb{E} \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] \simeq F_T(y|\mathbf{X}_i = \mathbf{x}),$$

and regard the data $\{W \{(Y_i - y)b^{-1}\} \delta_i (1 - \hat{G}(Y_i))^{-1}, \mathbf{X}_i, \delta_i = 1\}, i = 1, \dots, n$, as of coming from the nonparametric regression

$$\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} = F_T(y|\mathbf{X}_i) + \varepsilon_i(y), i = 1, \dots, n, \quad (4)$$

where the $\varepsilon_i(y)$'s denote conditionally mean zero error terms. This motivates a local linear approach in estimating $F_T(y|\mathbf{x})$: in parallel to Fan et al. (1996), a multivariate Taylor expansion of $F_T(y|\mathbf{X} = \mathbf{z})$ around $\mathbf{x} = (x_1, \dots, x_d)^\top$ reveals that each coefficient of the expansion corresponds to a vector of partial derivatives of the conditional distribution function. Hence, provided that $F_T(y|\mathbf{z})$ is a smooth function, it can be accurately approximated locally by d -dimensional hyperplanes with coefficients the scalar β_0 and the d -variate vector $\beta_1 = (\beta_{1,1}, \dots, \beta_{1,d})^\top$, i.e.

$$\mathbb{E} \left[\frac{W \{(Y - y)b^{-1}\} \delta}{1 - \hat{G}(Y)} \middle| \mathbf{X} = \mathbf{z} \right] = F_T(y|\mathbf{z}) \simeq F_T(y|\mathbf{x}) + \dot{F}_T(y|\mathbf{x})^\top (\mathbf{z} - \mathbf{x}) \equiv \beta_0 + \beta_1^\top (\mathbf{z} - \mathbf{x}), \quad (5)$$

where for any conditional p.d.f. or c.d.f. function $g(y|\mathbf{x})$, $\dot{g}(y|\mathbf{x}) = (\partial g(y|\mathbf{x})/\partial x_1, \dots, \partial g(y|\mathbf{x})/\partial x_d)^\top$ denotes its gradient vector and $\ddot{g}(y|\mathbf{x}) = (\partial^2 g(y|\mathbf{x})/\partial x_i \partial x_j)_{d \times d}, i, j = 1, \dots, d$ its Hessian matrix. In an entirely obvious manner $\dot{G}(\mathbf{x})$ and $\ddot{G}(\mathbf{x})$ denote the gradient vector and Hessian matrix respectively of any real function G . By (5) it is immediately seen that estimation of β_0 is equivalent to estimation of $F_T(y|\mathbf{x})$. Combining (4) and (5),

data-driven estimators of β_0 and β_1 are obtained by solving the kernel weighted least squares optimization problem

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} - \beta_0 - \beta_1^\top (\mathbf{X}_i - \mathbf{x}) \right]^2 \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \}. \quad (6)$$

In particular the solution $\hat{\beta}_0 \equiv \hat{F}_L(y|\mathbf{x})$ of (6) is

$$\hat{\beta}_0 \equiv \hat{F}_L(y|\mathbf{x}) = e_1^\top (\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^\top \mathbf{W}_x \mathbf{Y}, \quad (7)$$

where $e_1^\top = (1, 0)$ and

$$\mathbf{X}_x = \begin{pmatrix} 1 & \mathbf{X}_1 - \mathbf{x} \\ \vdots & \vdots \\ 1 & \mathbf{X}_n - \mathbf{x} \end{pmatrix}, \quad \mathbf{W}_x = \begin{pmatrix} \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_1 - \mathbf{x}) \} & & 0 \\ & \ddots & \\ 0 & & \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_n - \mathbf{x}) \} \end{pmatrix},$$

$$\mathbf{Y} = \left(\frac{W \{(Y_1 - y)b^{-1}\} \delta_1}{1 - \hat{G}(Y_1)} \dots \frac{W \{(Y_n - y)b^{-1}\} \delta_n}{1 - \hat{G}(Y_n)} \right)^\top.$$

Note that $\hat{F}_L(y|\mathbf{x})$ is continuous in both $\mathbf{x}$ and y . For $d = 1$ estimator $\hat{F}_L(y|\mathbf{x})$ was discussed in Kim et al. (2010), see also Buch-Kromann and Nielsen (2012). The wide popularity of $\hat{F}_L(y|\mathbf{x})$ lies on its flexibility to capture the nonlinear dynamics between the covariates, time and survival in the general case where $d > 1$ and subsequently on its ability to evaluate the effect on survival of each covariate, separately from each other.

2.2 Automatic boundary adjustment and asymptotic properties of the estimator.

The automatic adjustment of $\hat{F}_L$ in the boundary is even more necessary in the multidimensional setting. The reason is that as d increases, so does the ratio of the volume of the boundary region over the volume of the estimated curve, see Müller and Stadtmüller (1999) and the detailed discussion in Appendix A.1. Further, for $d > 1$ not only the distance to the boundary but also the local shape of the boundary plays a role. In particular, curved boundaries make the asymptotic analysis conceptually more complicated compared with the one-dimensional case. As shown in Appendix A.1 the asymptotic boundary behaviour of $\hat{F}_L$ is stable on $\mathcal{D} = [-1, c_1] \times \dots \times [-1, c_d]$, where $\mathbf{c} = (c_1, \dots, c_d)^\top$ is a vector of constants and after assuming without loss of generality that $\mathbf{K}$ is supported in the compact set $T = [-1, 1]^d$. Consequently, this particular choice of $\mathcal{D}$ ensures the validity the estimator's asymptotic properties. Since the computation of $\hat{F}_L(y|\mathbf{x})$ can be demanding for $d > 1$, in practice it is asymptotically equivalent to use the following computationally friendly version of (7) (established in Appendix A.1)

$$\tilde{F}_L(y|\mathbf{x}) = \frac{1}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^n \mathbf{K}_{(c)} \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)}, \quad (8)$$

where for $\mathbf{x} = \mathbf{J} + \mathbf{H}\mathbf{c}$, $\mathbf{J} = (J_1, \dots, J_d)^\top$ and $\mathbf{1}_{\{\mathbf{u} \in T_x\}}$ being shorthand notation for $(\mathbf{1}_{[-c_1, \infty)}, \dots, \mathbf{1}_{[-c_d, \infty)})$,

$$\mathbf{K}_{(c)}(\mathbf{x}) = \begin{cases} \mathbf{K}_c(\mathbf{x}) & \text{if } \mathbf{x} \in S_b, \\ \mathbf{K}(\mathbf{x}) & \text{if } \mathbf{x} \in S_{\text{int}} \end{cases}, \quad \mathbf{K}_c(\mathbf{u}) = \frac{S_{2,c} - S_{1,c} \mathbf{u} |\mathbf{H}|^{-1}}{S_{2,c} S_{0,c} - S_{1,c}^2} \mathbf{K}(\mathbf{u}) \mathbf{1}_{\{\mathbf{u} \in T_x\}},$$

with $S_b = S_{b,l} \cup S_{b,r}$ denoting the boundary region, $S_{b,l}$ and $S_{b,r}$ being respectively the left and right boundary regions, S_{int} is the interior of the estimation region, $T_x = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z} = \mathbf{H}^{-1}(\mathbf{x} - \mathbf{y}), \mathbf{y} \in (\mathbf{x} - \mathbf{H}T) \cap S\}$ (see also Appendix A.1),

$$S_{0,c} = \int_{\mathcal{D}} \mathbf{K}(\mathbf{u}) d\mathbf{u}, \quad S_{1,c} = \int_{\mathcal{D}} \mathbf{u} \mathbf{K}(\mathbf{u}) d\mathbf{u}, \quad S_{2,c} = \int_{\mathcal{D}} \mathbf{u} \mathbf{u}^\top \mathbf{K}(\mathbf{u}) d\mathbf{u},$$

with $\int$ being shorthand for $\int \dots \int$, $d\mathbf{u}$ for $du_1 \dots du_d$. $\mathbf{K}_c$ is typically referred to as the *equivalent* kernel and by construction restores the mass of $\mathbf{K}$ falling outside the region of estimation back in so as to automatically address edge-point effects of $\hat{F}_L(y|\mathbf{x})$. Set

$$A_{1,1} = \int \mathbf{u} \mathbf{K}^*(\mathbf{u}) W(z) d\mathbf{u} dz, \quad (9)$$

where $\mathbf{K}^*(\mathbf{u}) = e_1^\top (\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x)^{-1} (1 (\mathbf{u} - \mathbf{x})^\top)^\top \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{u} - \mathbf{x}) \}$ is the so-called *effective* kernel (see also Appendix A.1) and let

$$p(y|\mathbf{x}) = f_T(y|\mathbf{x})(1 - G(y))^{-1}, \quad P(z|\mathbf{x}) = \int_0^z p(y|\mathbf{x}) dy,$$

and $P(\mathbf{x}) = P(T_Y|\mathbf{x})$. The next theorem summarizes the asymptotic properties of $\hat{F}_L(y|\mathbf{x})$.

Theorem 1 *Under assumptions A1, A4 and A6, for $0 < \beta < 1/2$,*

$$\begin{aligned} \text{MISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} &= \mu_2(\mathbf{K}_{(c)}) \int \left[\frac{\text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\}}{f_{\mathbf{X}}(\mathbf{x})} \right. \\ &\quad \left. + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{tr} \left\{ \mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H} \right\} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{tr} \left\{ \mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H} \right\} \right]^2 dy + \\ &\quad \frac{R(\mathbf{K}_{(c)})}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int [P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1} - F_T^2(y|\mathbf{x})] dy + O(\mathbf{H}^\top \mathbf{H}) + O(\mathbf{H}^\top \mathbf{H}b^2) + O(b^2) + o(n^{-\beta}). \end{aligned} \quad (10)$$

Remark 1 In Section 3 it is shown that when $\hat{F}_L(y|\mathbf{x})$ employs d covariates, its optimal rate of convergence is achieved by applying bandwidth proportional to $n^{-1/(d+4)}$ to each covariate and proportional to $n^{-2/(5+d)}$ in the duration domain. With these bandwidths, (10) suggests a MISE rate of convergence of order $n^{-2/(d+4)}$.

3 Plug in bandwidth selection.

Let $\varepsilon_{y,b} = [0, T_Y] \times S_b$ and $\varepsilon_{y,o} = [0, T_Y] \times S_{\text{int}}$ and so that $\text{supp}(F_T) = \varepsilon_{y,b} \cup \varepsilon_{y,o}$. The Lebesgue measure of both $S_{b,1}$ and $S_{b,r}$ equals $h_1 \cdots h_d T_Y$ which by assumptions A5 and A6 shrinks to zero as $n \rightarrow \infty$. Therefore the Lebesgue measure of $\varepsilon_{y,b}$, which equals $2h_1 \cdots h_d$, will also shrink to zero asymptotically; hence the corresponding integral above equals zero. By this, Theorem 1 and noting also that in the interior $\mathbf{K}_c \equiv \mathbf{K}$, the MISE of $\hat{F}_L(\mathbf{H}, b)$ is asymptotically equal to the leading term of (10) with $\mathbf{K}_{(c)}$ replaced by $\mathbf{K}$. Note that (10) can be written in compact form (see Appendix A.4) as

$$\text{MISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} = \text{AMISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} + O(b^2) + O(\mathbf{H}^\top \mathbf{H}) + o(n^{-\beta}), \quad (11)$$

where

$$\text{AMISE} \left\{ \hat{F}_L(h_1, \dots, h_d, b) \right\} = \left(\sum_{i=1}^d h_i^2 \right)^2 m_1 + \left(\sum_{i=1}^d h_i^2 \right) b^2 m_2 + b^4 m_3 - \frac{bm_4}{n \prod_{i=1}^d h_i} + \frac{m_5}{n \prod_{i=1}^d h_i}, \quad (12)$$

with

$$m_1 = \mu_2^2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left\{ \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial x_i} \frac{\partial F_T(y|\mathbf{x})}{\partial x_j} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \sum_{i=1}^d \frac{\partial^2 F_T(y|\mathbf{x})}{\partial x_i^2} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \frac{\partial^2 f_{\mathbf{X}}(\mathbf{x})}{\partial x_i^2} \right\}^2 dy,$$

$$m_2 = m_1 m_3, \quad m_3 = \frac{\mu_2^2(w)}{4} \int_{\varepsilon_{y,o}} \left\{ \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\}^2 dy,$$

$$m_4 = 2R(\mathbf{K})A_{1,1}f_{\mathbf{X}}^{-1}(\mathbf{x}) \int_{\varepsilon_{y,o}} p(y|\mathbf{x}) dy \mathbf{x} = 2R(\mathbf{K})A_{1,1}f_{\mathbf{X}}^{-1}(\mathbf{x}) \int_{\varepsilon_{y,o}} \frac{f_T(y|\mathbf{x})}{1 - G(y)} dy,$$

$$m_5 = \frac{R(\mathbf{K})}{f_{\mathbf{X}}(\mathbf{x})} \int_{\varepsilon_{y,o}} \{P(y|\mathbf{x}) - F_T^2(y|\mathbf{x})\}^2 dy.$$

Under assumptions A5 and A6, (11) is well approximated by (12). When $d > 1$ equal number of different bandwidths in the x direction are needed, say $h_1, \dots, h_d$, which by symmetry will all be of the same order. Set

$$h_{\text{opt},i} = \left(\frac{m_{5,i}}{4nm_{1,i}} \right)^{1/(4+d)}, \quad b_{\text{opt}} = \left(\frac{m_4}{2nm_2} \right)^{2/(5+d)} \quad i = 1, \dots, d,$$

where $m_{1,i}, \dots, m_{5,i}$ are the univariate versions of $m_1, \dots, m_5$ with the obvious adjustment that the random variable X is replaced by the i th covariate $X_i, i = 1, \dots, d$. Now

$$\sum_{i=1}^d h_{\text{opt},i}^2 = n^{-2/(4+d)} \sum_{i=1}^d \left(\frac{m_{5,i}}{4m_{1,i}} \right)^{2/(4+d)} \equiv vn^{-2/(4+d)}.$$

Also, set $s = (m_4/(2m_2))^{2/(5+d)}$, so that $b_{\text{opt}} = sn^{-2/(5+d)}$. With this choice of bandwidths (12) becomes

$$\text{AMISE}\{\hat{F}_L(h_{\text{opt},1}, \dots, h_{\text{opt},d}, b_{\text{opt}})\} = \frac{v^2 m_1}{n^{4/(4+d)}} + \frac{vs^2 m_2}{n^{6/(4+d)}} + \frac{s^4 m_3}{n^{8/(4+d)}} - \frac{sm_4}{n^{6/(4+d)} \prod_{i=1}^d v_i} + \frac{m_5}{n^{4/(4+d)} \prod_{i=1}^d v_i} \rightarrow 0, \quad (13)$$

as $n \rightarrow \infty$, indicating that this is a sensible bandwidth. The asymptotic properties i.e. the optimality and uniqueness of the proposed bandwidths are summarized in the next theorem.

Theorem 2 *Under assumptions A1–A6, for fixed b , the $\text{AMISE}\{\hat{F}_L(\mathbf{H}, b)\}$ is minimized by a unique vector which is given by*

$$\mathbf{H}_{\text{opt}} = \text{diag}\{h_{\text{opt},1}, \dots, h_{\text{opt},d}\}, \quad (14)$$

$$b_{\text{opt}} = \{m_4/(2m_2)\}^{2/(5+d)} n^{-2/(5+d)}. \quad (15)$$

Implementation of (14) and (15) in practice requires consistent data-driven estimates of $m_1 - m_5$ which in turn depend on consistent estimates of $f_{\mathbf{X}}(\mathbf{x})$, $f_T(y|\mathbf{x})$, $F_T(y|\mathbf{x})$ and their corresponding gradient vectors and Hessian matrices. As this amounts to a substantial estimation effort, the plan is to estimate $f_{\mathbf{X}}(\mathbf{x})$ and $f_T(y|\mathbf{x})$ and then utilize the analytic relationships between these and the remaining functionals that need to be estimated so as to drastically reduce the required effort. Starting with $f_{\mathbf{X}}(\mathbf{x})$, this is estimated by the classical multivariate kernel estimate $\hat{f}_{\mathbf{X}}(\mathbf{x})$, given by

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = n^{-1} |\hat{\mathbf{H}}_l|^{-1} \sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\}, \quad (16)$$

where $\hat{\mathbf{H}}_l$ is the plug-in AMISE optimal diagonal bandwidth matrix of Duong and Hazelton (2003), see Appendix A.3. Consequently $\dot{\hat{f}}_{\mathbf{X}}$ and $\ddot{\hat{f}}_{\mathbf{X}}(\mathbf{x})$ can be consistently estimated by

$$\dot{\hat{f}}_{\mathbf{X}}(\mathbf{x}) = \left(\frac{\partial \hat{f}_{\mathbf{X}}(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial \hat{f}_{\mathbf{X}}(\mathbf{x})}{\partial x_d} \right)^\top, \quad \ddot{\hat{f}}_{\mathbf{X}}(\mathbf{x}) = \left(\frac{\partial^2 \hat{f}_{\mathbf{X}}(\mathbf{x})}{\partial x_i \partial x_j} \right)_{d \times d}, \quad i, j = 1, \dots, d.$$

Estimation of $f_T(y|\mathbf{x})$ is based on the conditional density estimator for censored data of Kim et al. (2010),

$$\hat{f}_T(y|\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\} \left\{ 1 - \hat{G}(Y_i) \right\}^{-1} w_{\tilde{b}}(Y_i - y) \delta_i}{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\}}, \quad (17)$$

with $\tilde{b} = 3.9\hat{\sigma}n^{-1/3}$ where $\hat{\sigma}$ is the minimum between the observed sample standard deviation (in the y coordinate direction) and the sample interquartile range divided by 1.359. Then, estimates of $F_T(y|\mathbf{x})$ and $\partial^2 F_T(y|\mathbf{x})/\partial y^2$, say $\hat{F}_T(y|\mathbf{x})$ and $\partial^2 \hat{F}_T(y|\mathbf{x})/\partial y^2$ respectively are obtained on the basis of the relationships

$$F_T(y|\mathbf{x}) = \int_{-\infty}^y f_T(u|\mathbf{x}) du \quad \text{and} \quad \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} = \frac{\partial f_T(y|\mathbf{x})}{\partial y}.$$

This yields,

$$\begin{aligned} \hat{F}_T(y|\mathbf{x}) &= \frac{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\} \left\{ 1 - \hat{G}(Y_i) \right\}^{-1} W \left\{ (Y_i - y) \tilde{b}^{-1} \right\} \delta_i}{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\}}, \\ \frac{\partial^2 \hat{F}_T(y|\mathbf{x})}{\partial y^2} &= \tilde{b}^{-1} \frac{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\} \left\{ 1 - \hat{G}(Y_i) \right\}^{-1} w'_{\tilde{b}}(Y_i - y) \delta_i}{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1} (\mathbf{X}_i - \mathbf{x}) \right\}}. \end{aligned}$$

Sensible estimates of $\dot{F}_T(y|\mathbf{x})$ and $\ddot{F}_T(y|\mathbf{x})$, say $\dot{\hat{F}}_T(y|\mathbf{x})$ and $\ddot{\hat{F}}_T(y|\mathbf{x})$ are obtained by multivariate numerical differentiation using the Richardson extrapolation method which yields very accurate approximations. The estimators obtained are denoted by

$$\dot{\hat{F}}_T(y|\mathbf{x}) = \left(\frac{\partial \hat{F}_T(y|\mathbf{x})}{\partial x_i}, \dots, \frac{\partial \hat{F}_T(y|\mathbf{x})}{\partial x_d} \right)^\top, \quad \ddot{\hat{F}}_T(y|\mathbf{x}) = \left(\frac{\partial^2 \hat{F}_T(y|\mathbf{x})}{\partial x_i \partial x_j} \right)_{d \times d}, \quad i, j = 1, \dots, d, \quad (18)$$

and their implementation is discussed in detail in Appendix A.6. Finally, consistent estimates of $p(y|\mathbf{x})$ and $P(y|\mathbf{x})$, say $\hat{p}(y|\mathbf{x})$ and $\hat{P}(y|\mathbf{x})$ are defined by

$$\hat{p}(y|\mathbf{x}) = \hat{f}_T(y|\mathbf{x})(1 - \hat{G}(y))^{-1},$$

and

$$\hat{P}(\mathbf{x}) = \int_0^{T_Y} \hat{p}(y|\mathbf{x}) dy = \frac{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \right\} \left\{ 1 - \hat{G}(Y_i) \right\}^{-2} \left[W \left\{ (Y_i - T_Y) \tilde{b}^{-1} \right\} - W(Y_i \tilde{b}^{-1}) \right] \delta_i}{\sum_{i=1}^n \mathbf{K} \left\{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \right\}}.$$

Substituting the above consistent estimates in place of the unknown quantities in $m_1 - m_5$, yields their practically useful versions, say $\hat{m}_1 - \hat{m}_5$. Their closed form expressions for $d = 2$ are given in Appendix A.6; see also Appendix A.5 for the general case $d > 1$.

Denote with $\hat{\mathbf{H}}_{\text{opt}}$ the bandwidth matrix $\mathbf{H}_{\text{opt}}$ with the unknown quantities $m_1 - m_5$ replaced by $\hat{m}_1 - \hat{m}_5$. Also $\hat{b}_{\text{opt}}$ denotes b_{opt} with $m_2 - m_4$ replaced by $\hat{m}_2 - \hat{m}_4$. The next theorem, in which O_p holds element-wise, establishes the consistency of $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$. Further it reveals that implementation of the bandwidth rule via the strategy of utilizing the analytic relationships between the functionals involved in its definition, does not impair its rate of convergence in comparison to using optimal bandwidths in all estimated functionals.

Theorem 3 *Under assumptions A1–A6, for fixed b , for $d \geq 2$, the relative rate of convergence of $\hat{\mathbf{H}}_{\text{opt}}$ to $\mathbf{H}_{\text{opt}}$ is*

$$\text{vech}(\hat{\mathbf{H}}_{\text{opt}} - \mathbf{H}_{\text{opt}}) = O_p \left(\mathbf{I}_d n^{-\frac{2}{d+6}} \right) \text{vech}(\mathbf{H}_{\text{opt}}), \quad \hat{b}_{\text{opt}} - b_{\text{opt}} = n^{-2(5+d)} b_{\text{opt}}. \quad (19)$$

Remark 2 Theorem 3 shows that the proposed practically useful scheme of implementing the estimate's MISE optimal bandwidth does not impair its consistency.

4 Numerical examples.

4.1 Finite sample performance of the proposed bandwidth rule.

The practical performance of the proposed bandwidth rule is investigated here via the MISE comparison of $1 - \hat{F}_L(y|\mathbf{x})$, calculated twice: once using the ideal bandwidths (14)–(15) and once using their data driven version, obtained as discussed in the previous section. We consider four different sample sizes, $n = 50, 100, 150, 250$ and three different censoring levels, 15%, 30% and 50%. The general form of the underlying survival function considered in the simulations is

$$S(y|\mathbf{x}) = e^{-A(\mathbf{x})y^2}, \quad (20)$$

which corresponds to $Y | \mathbf{X} = \mathbf{x} \sim \text{Weibull}(2, A(\mathbf{x})^{-1/2})$, with $A(\mathbf{x})$ being a polynomial and $\delta | \mathbf{X} = \mathbf{x} \sim \text{Weibull}(2, B(\mathbf{x})^{-1/2})$, where the polynomial $B(\mathbf{x})$ controls the amount of censoring. The probability that an observation will be censored is given by $P(\delta = 0 | \mathbf{X} = \mathbf{x}) = B(\mathbf{x})\{A(\mathbf{x}) + B(\mathbf{x})\}^{-1}$. Prior to the present simulation, (20) was used as the basis of the numerical examples presented in Peláez et al. (2021b, 2022a). The MISE comparisons presented here consider three realizations of (20). One model (model 1), denoted by $S_1(y|\mathbf{x})$, employs two independent from each other covariates, X_1 and X_2 , both assumed to follow the $U(0, 1)$ distribution and is given by

$$S_1(y|\mathbf{x}) = e^{-A_1(\mathbf{x})y^2}, \quad \text{where } A_1(\mathbf{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2.$$

This survival function corresponds to

$$Y | \mathbf{X} = \mathbf{x} \sim \text{Weibull}(2, A_1(\mathbf{x})^{-1/2d}) \quad \text{and} \quad \delta | \mathbf{X} = \mathbf{x} \sim \text{Weibull}(2, B_1(\mathbf{x})^{-1/2}),$$

where $B_1(\mathbf{x}) = \beta_0 + \beta_{11}x_1 + \beta_{21}x_1^2 + \beta_{12}x_2 + \beta_{22}x_2^2$. By choosing $\alpha_0 = 1, \alpha_1 = 5, \alpha_2 = 5$ and $\beta_0 = 10, \beta_{21} = 10, \beta_{22} = 20$, then, for fixed $(x_1, x_2) = (0.3, 0.6)$ and for $(\beta_{11}, \beta_{12}) = (-41.86, 10), (-39.76, 10), (-25, 15)$, the censoring conditional probability is respectively 0.15, 0.3 and 0.5. The second model (model 2), employs covariate X_1 (the same as model 1), is denoted by $S_2(y|x_1)$, and is given by

$$S_2(y|x_1) = e^{-A_2(x_1)y^2}, \quad \text{where } A_2(x_1) = \alpha_{2,0} + \alpha_{2,1}x_1.$$

This survival function corresponds to

$$Y | X_1 = x_1 \sim \text{Weibull}(2, A_2(x_1)^{-1/2}) \quad \text{and} \quad \delta | X_1 = x_1 \sim \text{Weibull}(2, B_2(x_1)^{-1/2}),$$

where $B_2(x_1) = \beta_{2,0} + \beta_{2,1}x_1 + \beta_{2,2}x_1^2$. By choosing $\alpha_{2,0} = 1, \alpha_{2,1} = 5$ and $\beta_{2,0} = 10, \beta_{2,2} = 20$, then for $\beta_{2,1} = -37.86, -35.76, -13.28$ and for fixed $x_1 = 0.3$, the censoring conditional probability is respectively 0.15, 0.3 and 0.5. Model 2 corresponds to the first model considered in Peláez et al. (2021b, 2022a) but it uses a different covariate level and different $\beta_{2,1}$ parameter values so as to obtain the required levels of censoring. Model 3 employs covariate X_2 (the same as model 1), is denoted by $S_3(y|x_2)$ and is given by

$$S_3(y|x_2) = e^{-A_3(x_2)y^2}, \text{ where } A_3(x_2) = \alpha_{3,0} + \alpha_{3,1}x_2.$$

This survival function corresponds to

$$Y | X_2 = x_2 \sim \text{Weibull}(2, A_3(x_2)^{-1/2}) \text{ and } \delta | X_2 = x_2 \sim \text{Weibull}(2, B_3(x_2)^{-1/2}),$$

where $B_3(x_2) = \beta_{3,0} + \beta_{3,1}x_2 + \beta_{3,2}x_2^2$. By choosing $\alpha_{3,0} = 1, \alpha_{3,1} = 5$ and $\beta_{3,0} = 10, \beta_{3,2} = 20$, then for $\beta_{3,1} = -28, -26.88, -24.5$ and for fixed $x_2 = 0.6$, the censoring conditional probability is respectively 0.15, 0.3 and 0.5. Model 3 corresponds also to the first model considered in the simulation studies of Peláez et al. (2021b, 2022a) but it uses different $\beta_{2,1}$ parameter values so as to obtain the required levels of censoring. The reason for implementing the MISE comparison based on these three models is that as in Peláez et al. (2021b, 2022a) for model 2 there, it can also be proved here that all models considered are close to Cox proportional hazards models. Hence the present comparison can also serve (at least approximately) as a performance comparison of the proposed methodology with the proportional hazards technique.

The corresponding local linear estimators of $S_i(y|\cdot), i = 1, 2, 3$ using ideal bandwidths are calculated as $1 - \hat{F}_L(y|\cdot)$ with the c.d.f. estimated by (7) using the bandwidths (14)–(15). The resulting estimators are denoted respectively by $\hat{S}_{1,id}(y|x_1, x_2), \hat{S}_{2,id}(y|x_1)$ and $\hat{S}_{3,id}(y|x_2)$. In an entirely obvious manner, $\hat{S}_{1,p}(y|x_1, x_2), \hat{S}_{2,p}(y|x_1)$ and $\hat{S}_{3,p}(y|x_2)$ denote the estimators of $S_i(y|\cdot), i = 1, 2, 3$ obtained via (7) with bandwidths $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$. The MISE figures reported on Table 1 are calculated as follows. For each sample size and amount of censoring, 1000 random samples $(Y_i, \delta_i, \mathbf{X}_i), i = 1, \dots, n$ are generated, based on the distribution parameters of each model. Let $A_{j,\cdot}$ denote any of the estimators $\hat{S}_{j,id}, \hat{S}_{j,p}, j = 1, 2, 3$ calculated using the k th, $k = 1, \dots, 1000$, sample. Both estimators are calculated on 50 equidistant grid points in the time domain, spanning from 0 to the 95th percentile of the Y variable. The MISE of each estimator is then approximated as the sample average of the 1000 realizations of the integrated square error of $A_{j,\cdot}$, i.e.,

$$1000^{-1} \sum_{k=1}^{1000} \int [A_{j,\cdot}(y|\mathbf{x}) - S_i(y|\mathbf{x})]^2 dy.$$

In the above MISE calculation, for each sample size and censoring level combination, in each of the 1000 iterations, the same sample is used for both $\hat{S}_{j,id}$ and $\hat{S}_{j,p}, j = 1, 2, 3$. The extended Simpson's rule is used to approximate the integral.

The MISE figures in Table 1 indicate that as the sample size grows, the percentage difference in MISE between the ideal bandwidth and its data driven counterpart decreases. The discrepancy between the two bandwidths diminishes considerably for $n = 150$ where the MISE corresponding to the data driven bandwidth is approximately 15% higher than the ideal (across all sample sizes and levels of censoring). For $n = 250$ this difference drops even further to 7%, which indicates the consistency of the proposed bandwidth rule.

4.2 Real data analysis.

Estimator $\hat{F}_L(y|\mathbf{x})$ and the proposed bandwidth selector are illustrated in this section in the analysis of a real world data set.

The example is based on the German Credit dataset which describes the profile of 1000 credit recipients of a German bank. The data set is comprised of 20 variables and it is publicly available at [http://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)). The information contained in the data set includes the duration of each credit (in months, variable: 'Duration'), the classification of each applicant as good or bad credit risk, and the variable 'Creditability' which contains the censoring indicator for each record: $\delta = 1$ denotes bad/defaulted credits and $\delta = 0$ denotes good credits. There are 300 applicants classified as bad credits and 700 as good. Good credits are regarded as censored as it is not known if they are going to default; thus the amount of censoring is 70%. The rest variables (covariates) in the data set, which altogether describe the customers profile, include the age of the credit recipient ('Age', continuous, in years), its credit amount ('Amount' continuous, in Deutsche Marks (DM)), the level of balance in their checking account (qualitative, measured in four levels representing the range of DM in which the account balance falls in), the duration of their employment (qualitative, measured in five ranges representing periods of years) and the level of their balance in their savings account (qualitative as the range of DM in which the balance falls in).

Table 1: Approximate MISE's of $\hat{S}_{j,id}, \hat{S}_{j,p}, j = 1, 2, 3$ for the three conditional survival function models.

Cens.	n	$S_1(y \mathbf{x} = (0.3, 0.6))$			$S_2(y x_1 = 0.3)$			$S_3(y x_2 = 0.6)$		
		$\hat{S}_{1,id}$	$\hat{S}_{1,p}$	% diff.	$\hat{S}_{2,id}$	$\hat{S}_{2,p}$	% diff.	$\hat{S}_{3,id}$	$\hat{S}_{3,p}$	% diff.
15%	50	0.341	0.616	44.64%	0.206	0.375	45.07%	0.358	0.533	32.83%
	100	0.262	0.357	26.61%	0.153	0.208	26.44%	0.266	0.326	18.4%
	150	0.226	0.295	23.39%	0.121	0.149	18.79%	0.211	0.251	15.94%
	250	0.196	0.21	6.67%	0.09	0.106	15.09%	0.157	0.168	6.55%
30%	50	0.434	0.703	38.26%	0.219	0.377	41.91%	0.381	0.616	38.15%
	100	0.355	0.422	15.88%	0.213	0.291	26.8%	0.37	0.485	23.71%
	150	0.324	0.361	10.25%	0.193	0.226	14.6%	0.336	0.384	12.50%
	250	0.29	0.315	7.94%	0.178	0.189	5.82%	0.311	0.325	4.31%
50%	50	0.506	0.758	33.25%	0.289	0.417	30.7%	0.538	0.773	30.40%
	100	0.468	0.583	19.73%	0.266	0.345	22.9%	0.515	0.713	27.77%
	150	0.361	0.404	10.64%	0.243	0.305	20.33%	0.457	0.516	11.43%
	250	0.395	0.417	5.28%	0.228	0.247	7.69%	0.431	0.457	5.69%

Estimator $\hat{F}_L(y|\mathbf{x})$ is employed to model the time to default, i.e. to estimate the probability that the duration of a credit will extend beyond a specific point in time. This information provides very useful guidance to bank managers when deciding whether to approve a loan to a prospective applicant based on his/her profiles. This is because there are two types of risks associated with the bank's decision: if the applicant is a good credit risk, i.e. is likely to repay the loan, disapproving the loan results in a loss of business to the bank. On the other hand, if the applicant is a bad credit risk, i.e. is not likely to repay the loan, approval of the loan results in a financial loss to the bank. According to the data set description, it is worse to classify a customer as good when they are bad, than it is to classify a customer as bad when they are good.

In order to facilitate the illustration of the method, we consider a model with two covariates: Amount and Age. That is, the matrix $\mathbf{X}$ has two columns (the vectors Amount and Age) and 1000 rows, while the Y sample is the Duration variable and the δ variable is Creditability. Note that the Pearson, Kendall and Spearman correlation coefficients between Amount and Age are 0.0327, 0.0173 and 0.0262 respectively. Further a simple scatterplot suggests that no other (nonlinear) pattern can be used to describe the relationship between the two variables. These two facts suggest that the assumption of independence among the two covariates is fulfilled. With these data we approximate the probability that a certain duration will exceed a specified time point conditional on the amount and the age of the loan recipient by

$$B = P(\text{Duration} \geq y | \text{Amount} = x_1, \text{Age} = x_2) \simeq 1 - \hat{F}_L(\text{Duration} \leq y | \text{Amount} = x_1, \text{Age} = x_2).$$

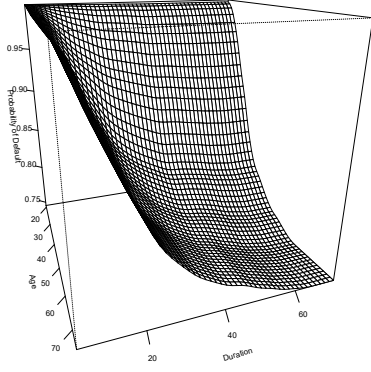
In all instances below, estimator $\hat{F}_L$ is implemented via (7) using bandwidths $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$ calculated as described in Section 3. See also Appendix A.6 for closed form expressions of the functionals involved in the present computations, as these result by the general expressions of Section 3.

By fixing for each model one covariate at a time at a specific value, we are able to demonstrate the resulting estimator as 3-dimensional plots. Four such instances are demonstrated in Fig.1. These are

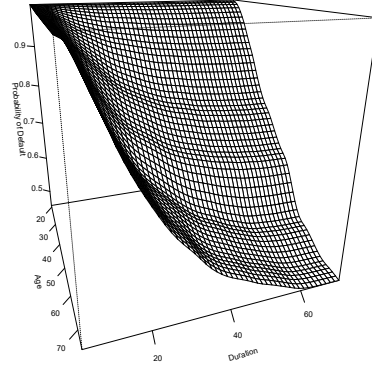
$$B_1 = P(\text{Duration} \geq y | \text{Amount} = 3588, \text{Age} = x_2), B_2 = P(\text{Duration} \geq y | \text{Amount} = 11006, \text{Age} = x_2), \\ B_3 = P(\text{Duration} \geq y | \text{Amount} = x_1, \text{Age} = 29), B_4 = P(\text{Duration} \geq y | \text{Amount} = x_1, \text{Age} = 52).$$

All four models suggest that the probability to default diminishes with time but at a different rate. For credit amounts equal to 3588DM the probability to default ranges from 1 (at the start of the loan) to 0.75 (at 65 months after the loan was granted), depending on the age: the decline on the default probability is much sharper for older loan recipients than it is for younger ones. A parallel scenario is true for credit amounts of 11006DM, (i.e. for model B_2) with the difference that the probability to default ranges from 1 to 0.45, for the same time period. Fixing age, we can visualize the probability to default in terms of credit amount. I.e. for loan recipients of 29 years of age, the probability to default ranges across the 65 months period from 1 to 0.75 while for loan recipients of 52 years of age the same probability ranges from 1 to 0.4 as the amount increases.

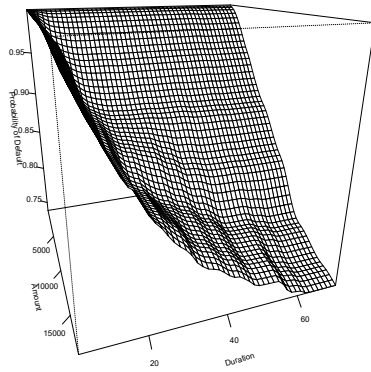
The example illustrates that the proposed method offers a powerful methodological tool which enables customized – according to the characteristics of each individual – results, with concrete, practical implications for empirical work in a variety of scientific fields, including medicine, reliability, e.t.c. The example illustrates that even with two covariates, the method helps uncover the full potential of each customized prediction in



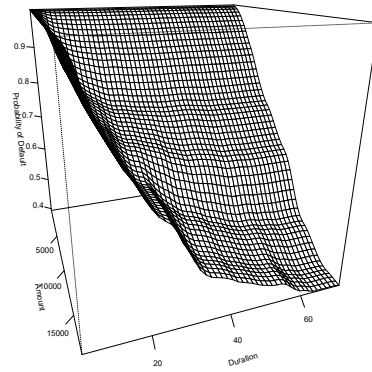
$$B_1 = P(\text{Duration} \geq y | \text{Amount} = 3588, \text{Age}).$$



$$B_2 = P(\text{Duration} \geq y | \text{Amount} = 11006, \text{Age}).$$



$$B_3 = P(\text{Duration} \geq y | \text{Amount}, \text{Age} = 29).$$



$$B_4 = P(\text{Duration} \geq y | \text{Amount}, \text{Age} = 52).$$

Fig. 1: Estimated probabilities of default given ‘Age’ for ‘Amount’ fixed to 3588 and 11006 and given ‘Amount’ for ‘Age’ fixed to 29 and 52.

the sense of the rate of change in the probability of an event to occur as the covariate values change. At the same time it highlights the fact that an optimal variable selection approach is a natural step forward for conditional estimation in multiple dimensions.

5 Conclusions and future work.

This research investigated the local linear estimation of the conditional survival function for random designs in the presence of an arbitrary number of continuous covariates in the right censored data setting. The theoretical properties of the estimator suggest consistent and reliable estimation of the underlying curve throughout the region of estimation. The MISE simulations indicate that this robust behavior is also valid for finite samples. This performance is achieved due to the proposed data bandwidth selector the consistency of which (in terms of convergence to the ideal bandwidth) is established analytically together with its uniqueness and its MISE optimality.

The methodological advances explored herein can be extended towards multiple directions. A natural step forward which is already underway is a variable selection procedure based on minimization of the estimator’s MISE. In parallel, establishment of the proposed estimator’s asymptotic distribution will enable the development of confidence intervals and facilitate employment of the estimator by practitioners. Another extension, is the incorporation in the functional form of the estimator of additional variable types, e.g. discrete random variables. Multiple other topics arise by considering the adjustment and analytical study of this technique to different censoring schemes.

6 Technical details.

6.1 Assumptions.

First, set $\mathbf{K}_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} \mathbf{K}(\mathbf{H}^{-1}\mathbf{u})$. Also note that the c.d.f. of Y satisfies $1 - F = (1 - F_T)(1 - G)$. Conditioning on covariate information, and since T and U are independent, it is common to assume that $T|\mathbf{X}$ and $U|\mathbf{X}$ are independent from each other. As a consequence it holds that $1 - F(\cdot|\mathbf{X}) = [1 - F_T(\cdot|\mathbf{X})][1 - G(\cdot|\mathbf{X})]$.

A1. The kernel $\mathbf{K} : T \rightarrow \mathbb{R}^+$ is bounded and such that

$$\int \mathbf{u} \mathbf{u}^\top \mathbf{K}(\mathbf{u}) d\mathbf{u} = \mu_2(\mathbf{K}) \mathbf{I}_d,$$

where $\mu_2(\mathbf{K}) \neq 0$ is a scalar and $\mathbf{I}_d$ is the $d \times d$ identity matrix. In addition, all odd-order moments of $\mathbf{K}$ vanish, that is,

$$\int u_1^{l_1} \dots u_d^{l_d} \mathbf{K}(\mathbf{u}) d\mathbf{u} = 0,$$

for all nonnegative integers $l_1, \dots, l_d$ such that their sum is odd. Note that this last condition is satisfied by spherically symmetric kernels and product kernels based on symmetric univariate kernels. Furthermore, $\mathbf{K}$ satisfies

$$R(\mathbf{K}) \equiv \int \mathbf{K}^2 < +\infty, \int |\mathbf{u}^2 \mathbf{K}| < +\infty, \int \mathbf{K} = 1 \text{ and } \int \mathbf{u} \mathbf{K} = 0.$$

A2. The kernel w is symmetric about the origin and satisfies

$$\int w^2 < +\infty, \int |u^2 w| < +\infty, \text{ and } \int uw = 0.$$

A3. As $n \rightarrow +\infty$, $b \rightarrow 0$ and $nb \rightarrow \infty$.

A4. The point $\mathbf{x} = (x_1, \dots, x_d)^\top \in S = \text{supp}(f_{\mathbf{X}})$ which is continuous at $\mathbf{x}$. For $(y, \mathbf{x}) \in [0, T_y] \times S$, $F_T(y|\mathbf{x})$ is twice continuously partially differentiable with respect to both variables in a neighbourhood of $(y, \mathbf{x})$ and $\partial^2 F_T(y|\mathbf{x})/\partial y^2$ exists and is bounded. Further, $\mathcal{D}_{F_T}(y|\mathbf{x})$, $\mathcal{H}_{F_T}(y|\mathbf{x})$, $\dot{f}_{\mathbf{X}}(\mathbf{x})$ and $\mathcal{H}_{f_{\mathbf{X}}}(\mathbf{x})$ exist, are continuous and bounded in $(y, \mathbf{x}) \in [0, T_y] \times S$.

A5. The sequence of bandwidth matrices $\mathbf{H}$ is such that $n^{-1}|\mathbf{H}|$ and each entry of $\mathbf{H}$ tends to zero as $n \rightarrow \infty$ with $\mathbf{H}$ remaining symmetric and positive definite. Also, there is a fixed constant L such that the condition number of $\mathbf{H}$ (i.e., the ratio of its largest to its smallest eigenvalue) is at most L for all n .

A6. For the smoothing parameter b it holds that $b \rightarrow 0$ as $n \rightarrow \infty$. Further b and $|\mathbf{H}|$ converge to zero in such a way so that $nb|\mathbf{H}| \rightarrow \infty$ and $b|\mathbf{H}| \log n \rightarrow 0$.

Set

$$\begin{aligned} \text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} &= \sum_{i=1}^d h_i \frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial x_i}, \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\} = \sum_{i=1}^d h_i \frac{\partial F_T(y|\mathbf{x})}{\partial x_i}, \\ \text{tr} \left\{ \mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H} \right\} &= \sum_{i=1}^d h_i^2 \frac{\partial^2 F_T(y|\mathbf{x})}{\partial x_i^2}, \text{tr} \left\{ \mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H} \right\} = \sum_{i=1}^d h_i^2 \frac{\partial^2 f_{\mathbf{X}}(\mathbf{x})}{\partial x_i^2}. \end{aligned} \quad (21)$$

6.2 Auxiliary lemmas.

Lemma 1 For fixed $i \in \{1, \dots, n\}$, $0 < \beta < 1/2$, as $n \rightarrow \infty$ and $b \rightarrow 0$,

$$\mathbb{E} \left[\frac{W \{ (Y_i - y)b^{-1} \} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] = F_T(y|\mathbf{x}) \{ 1 + O(b^2) + O_p(n^{-\beta}) \}.$$

Proof By (3), the rate of uniform strong convergence of $\hat{G}(y)$ to $G(y)$ over $[0, T_Y]$ is $n^{-\beta}$, $0 < \beta < 1/2$. Using this in the last step below

$$\frac{1}{1 - \hat{G}(y)} = \frac{1}{1 - G(y)} \frac{1}{1 + \frac{G(y) - \hat{G}(y)}{1 - G(y)}} = \frac{1}{1 - G(y)} \sum_{k=0}^{+\infty} (-1)^k \left(\frac{G(y) - \hat{G}(y)}{1 - G(y)} \right)^k = \frac{1}{1 - G(y)} \{ 1 + o(n^{-\beta}) \}. \quad (22)$$

Now, for fixed i , $\{Y_i | \mathbf{X}_i = \mathbf{x} : \delta_i = 1\}$ is a random variable with density $f_T(y|\mathbf{x})(1 - G(y))/p(\mathbf{x})$ where

$$p(\mathbf{x}) = \int_0^\infty f_T(y|\mathbf{x})(1 - G(y)) dy,$$

and δ_i is a Bernoulli random variable with mean $E(\delta_i) = p(y)$. Conditioning on $\delta_i = 1$ and using (3),

$$\begin{aligned} E \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] &= E \left(E \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \delta_i = 1 \right] \middle| \mathbf{X}_i = \mathbf{x} \right) \\ &= E \left(E \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - G(Y_i)} \middle| \delta_i = 1 \right] \middle| \mathbf{X}_i = \mathbf{x} \right) \{1 + o(n^{-\beta})\}. \end{aligned} \quad (23)$$

Using the change of variable $u - y = bs$ in the third step below and subsequently a Taylor expansion

$$\begin{aligned} E \left(E \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - G(Y_i)} \middle| \delta_i = 1 \right] \middle| \mathbf{X}_i = \mathbf{x} \right) &= \frac{1}{p(y)} (E\delta_i) \int \frac{f_T(u|\mathbf{x})(1 - G(u))}{1 - G(u)} W \left(\frac{u - y}{b} \right) du = \\ &= \int F_T(u|\mathbf{x}) w_b(u - y) du = \int w(s) F_T(y + bs|\mathbf{x}) ds = \\ &= \int w(s) \left\{ F_T(y|\mathbf{x}) + bs \frac{\partial F_T(y|\mathbf{x})}{\partial y} + \frac{(bs)^2}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} + o(b^2) \right\} ds = F_T(y|\mathbf{x}) + O(b^2). \end{aligned} \quad (24)$$

Using (24) back to (23) finishes the proof.

Lemma 2 For fixed $i \in \{1, \dots, n\}$, $0 < \beta < 1/2$, as $n \rightarrow \infty$ and $b \rightarrow 0$,

$$\begin{aligned} \text{Var} \left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] &= \\ &= \{P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1} - F_T^2(y|\mathbf{x})\} \{1 + O(b^2) + o(n^{-\beta})\}. \end{aligned}$$

Proof First, assuming fixed i , write

$$\begin{aligned} \text{Var} \left[\frac{\delta_i W \{(Y_i - y)b^{-1}\}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] &= \\ &= E \left(\left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \right]^2 \middle| \mathbf{X}_i = \mathbf{x} \right) - \left[E \left\{ \frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right\} \right]^2. \end{aligned} \quad (25)$$

Now, using (22) in the second step below

$$\begin{aligned} E \left(\left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \right]^2 \middle| \mathbf{X}_i = \mathbf{x} \right) &= E \left(E \left[\frac{W^2 \{(Y_i - y)b^{-1}\} \delta_i^2}{(1 - G(Y_i))^2} \middle| \delta_i = 1 \right] \middle| \mathbf{X}_i = \mathbf{x} \right) (1 + o(n^{-\beta})) \\ &= E \left\{ \int \frac{f_T(u|\mathbf{x})(1 - G(u))}{(1 - G(u))^2} W^2 \left(\frac{y - u}{b} \right) du \middle| \mathbf{X}_i = \mathbf{x} \right\} (1 + o(n^{-\beta})) \\ &= \iint f_{\mathbf{X}}(\mathbf{x}) \frac{f_T(u|\mathbf{x})}{1 - G(u)} W^2 \left(\frac{y - u}{b} \right) du d\mathbf{x} (1 + o(n^{-\beta})) = \int \frac{f_T(u|\mathbf{x})}{1 - G(u)} W^2 \left(\frac{y - u}{b} \right) du (1 + o(n^{-\beta})). \end{aligned} \quad (26)$$

Using the change of variable $y - u = sb$, integration by parts and a Taylor expansion yields

$$\int \frac{f_T(u|\mathbf{x})}{1 - G(u)} W^2 \left(\frac{y - u}{b} \right) du = \int_0^u \frac{f_T(y|\mathbf{x})}{1 - H(y)} dy - 2b \frac{f_T(u|\mathbf{x})}{1 - H(u)} \int uw(u)W(u) du. \quad (27)$$

Substituting (27) back to (26) yields

$$\begin{aligned} E \left(\left[\frac{W \{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)} \right]^2 \middle| \mathbf{X}_i = \mathbf{x} \right) &= \left\{ \int_0^u \frac{f_T(y|\mathbf{x})}{1 - H(y)} dy - 2b \frac{f_T(u|\mathbf{x})}{1 - H(u)} \int sw(\mu)W(\mu) ds \right\} (1 + o(n^{-\beta})) \\ &= \{P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1}\} (1 + o(n^{-\beta})). \end{aligned} \quad (28)$$

Further, by Lemma 1

$$\left(\mathbb{E} \left[\frac{W \{ (Y_i - y)b^{-1} \} \delta_i}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] \right)^2 = F_T^2(y|\mathbf{x}) \{1 + O(b^2) + o(n^{-\beta})\}. \quad (29)$$

Use (28) and (29) back to (25) to get

$$\text{Var} \left[\frac{\delta_i W \{ (Y_i - y)b^{-1} \}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i = \mathbf{x} \right] = \{P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1}\} (1 + o(n^{-\beta})) - \{F_T(y|\mathbf{x}) + O(b^2)\}^2 (1 + o(n^{-\beta})),$$

from which Lemma 2 immediately follows.

Lemma 3 Assume that $|\mathbf{H}| \rightarrow 0$ and $n|\mathbf{H}| \rightarrow \infty$. Then, for $\mathbf{x}$ in the boundary of S , i.e. for $\mathbf{x} \in S_{b,1}$ or $\mathbf{x} \in S_{b,r}$

$$S_{n,j}(\mathbf{x}) = |\mathbf{H}|^{j+1} f_{\mathbf{X}}(\mathbf{x}) S_{j,c} \{1 + o_p(1)\}, j = 0, 1, 2,$$

while in the interior, i.e. for $\mathbf{x} \in S_{\text{int}}$,

$$S_{n,0}(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}) + o_p(1), S_{n,1}(\mathbf{x}) = \mu_2(\mathbf{K}) \mathbf{H} \mathcal{D}_f(\mathbf{x}) + o_p \left(\sum_{i=1}^d h_i \right), S_{n,2}(\mathbf{x}) = \mu_2(\mathbf{K}) f_{\mathbf{X}}(\mathbf{x}) \mathbf{H} + o_p \left(\sum_{i=1}^d h_i \right).$$

Proof The proof for the first part of lemma is given for $j = 2$ and for $\mathbf{x} \in S_{b,1}$. The cases $j = 0, 1$ are proved in an entirely similar manner. First, write

$$S_{n,2}(\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} (\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top = \mathbb{E} \{ S_{n,2}(\mathbf{x}) \} + \text{Var} \{ S_{n,2}(\mathbf{x}) \}. \quad (30)$$

Now, using the definition of $S_{2,c}$ in the last step below,

$$\begin{aligned} \mathbb{E} \{ S_{n,2}(\mathbf{x}) \} &= \mathbb{E} \left[\mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} (\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top \right] = \int_{\mathcal{D}} \mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{y} - \mathbf{x}) \} (\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})^\top f_{\mathbf{X}}(\mathbf{y}) d\mathbf{y} \\ &\stackrel{\mathbf{H}^{-1}(\mathbf{y} - \mathbf{x}) = \mathbf{u}}{=} \int_{\mathcal{D}} \mathbf{K}(\mathbf{u}) (|\mathbf{H}|\mathbf{u})(|\mathbf{H}|\mathbf{u})^\top |\mathbf{H}| f_{\mathbf{X}}(\mathbf{x} + |\mathbf{H}|\mathbf{u}) d\mathbf{u} \\ &= |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) \int_{\mathcal{D}} \mathbf{u} \mathbf{u}^\top \mathbf{K}(\mathbf{u}) d\mathbf{u} \{1 + o_p(|\mathbf{H}|)\} = |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) S_{2,c} \{1 + o_p(|\mathbf{H}|)\}. \end{aligned} \quad (31)$$

Similarly,

$$\begin{aligned} \text{Var} \{ S_{n,2}(\mathbf{x}) \} &= \\ n^{-1} |\mathbf{H}|^{-1} \mathbb{E} \left[\mathbf{K} \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \}^2 (\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top ((\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top)^\top \right] &= O_p(n^{-1} |\mathbf{H}|^{-1}). \end{aligned} \quad (32)$$

Substituting (31) and (32) back to (30) completes the proof. The case where $\mathbf{x} \in S_{\text{int}}$ is obtained by letting $c > 1$ in the above proof.

Now let $M_+(d)$ denote the set of all $d \times d$ positive definite matrices and let $\mathbf{M}$ denote any matrix $(M_{ij})_{i,j=1}^d \in M_+(d)$. Set $h_i^2 = v_i$ so that $\mathbf{v} = (v_1, \dots, v_d)^\top \in \mathbb{R}_+^d \equiv (0, \infty)^d$, thus $\mathbf{v}^\top = (\text{vech}^\top \mathbf{H}^2)$. Then (12) can be written as

$$\text{AMISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} = \mathbf{v}^\top m_1 \mathbf{v} + \mathbf{v}^\top \mathbf{1} b^2 m_2 + b^4 m_3 - \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{v_1 \dots v_d}} \equiv \mathbf{Q}(\mathbf{v}, \mathbf{M}, b).$$

The following lemma is used in the proof of Theorem 2.

Lemma 4 For any fixed $\mathbf{M} \in M_+(d)$ and $h_i^2 = v_i, i = 1, \dots, d$ the function $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ is a strictly convex function of $\mathbf{v} \in \mathbb{R}_+^d$ and therefore it can be minimized by at most one vector $\mathbf{v} \in \mathbb{R}_+^d$.

Proof Recall that $\mathbf{1}$ denotes the d -variate unit vector. By direct calculation

$$\begin{aligned} \frac{\partial \mathbf{Q}(\mathbf{v}, \mathbf{M}, b)}{\partial \mathbf{v}} &= 2 \left(\sum_{i=1}^d v_i \right) m_1 \left(\sum_{i=1}^d v_i \right)' + \left(\sum_{i=1}^d v_i \right)' b^2 m_2 - \frac{bm_4 + m_5}{n} \left((v_1 \dots v_d)^{-1/2} \right)' \\ &= \mathbf{v}^\top m_1 \mathbf{1} + db^2 m_2 + \frac{bm_4 + m_5}{2n} \sum_{i=1}^d \frac{1}{v_i^{3/2} \left(\prod_{j=1, j \neq i}^d v_j \right)^{1/2}}. \end{aligned}$$

The Hessian of $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ with respect to $\mathbf{v}$ is

$$\frac{\partial \mathbf{Q}(\mathbf{v}, \mathbf{M}, b)}{\partial \mathbf{v} \partial \mathbf{v}^\top} = m_1 + \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{\prod_{i=1}^d v_i}} \begin{pmatrix} 3v_1^{-2} & v_1^{-1}v_2^{-1} & \dots & v_1^{-1}v_d^{-1} \\ v_2^{-1}v_1^{-1} & 3v_2^{-2} & \dots & v_2^{-1}v_d^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ v_d^{-1}v_1^{-1} & v_d^{-1}v_2^{-1} & \dots & 3v_d^{-2} \end{pmatrix} \equiv m_1 + \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{\prod_{i=1}^d v_i}} \mathbf{V}.$$

All terms in the above equation are non-negative definite; the second term is positive definite and therefore the Hessian of $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ with respect to $\mathbf{v}$ is positive definite. Obviously $\mathbf{v}$ is an elementwise positive vector and $b > 0$ is a scalar. To prove that $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ is convex in $\mathbf{v}$ and b it suffices to show that the Hessian of $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$, say $\mathcal{H}_{\mathbf{Q}}$, is positive semidefinite i.e. $\mathbf{u}^\top \mathcal{H}_{\mathbf{Q}} \mathbf{u} \geq 0$ for any vector $\mathbf{u} \in \mathbb{R}_+^{d+1}$. Set

$$\mathbf{Q}_{\mathbf{v}} = \frac{\partial \mathbf{Q}(\mathbf{v}, \mathbf{M}, b)}{\partial \mathbf{v}}, \mathbf{Q}_b = \frac{\partial \mathbf{Q}(\mathbf{v}, \mathbf{M}, b)}{\partial b}.$$

The gradient of $\mathbf{Q}$ is given by

$$\mathcal{D}_{\mathbf{Q}} = \begin{pmatrix} \mathbf{Q}_{\mathbf{v}} \\ \mathbf{Q}_b \end{pmatrix} = \begin{pmatrix} \mathbf{v}^\top m_1 \mathbf{1} + db^2 m_2 - \frac{bm_4 + m_5}{2n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} \\ 2b\mathbf{v}^\top \mathbf{1} m_2 + 4b^3 m_3 - \frac{m_4}{n} \frac{1}{\sqrt{v_1 \dots v_d}} \end{pmatrix},$$

and the Hessian is given by

$$\mathcal{H}_{\mathbf{Q}} = \begin{pmatrix} \mathbf{Q}_{\mathbf{v}\mathbf{v}} & \mathbf{Q}_{\mathbf{v}b} \\ \mathbf{Q}_{b\mathbf{v}} & \mathbf{Q}_{bb} \end{pmatrix} = \begin{pmatrix} m_1 - \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{\prod_{i=1}^d v_i}} \mathbf{V} & 2dbm_2 - \frac{m_4}{2n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} \\ 2b\mathbf{1} m_2 - \frac{m_4}{n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} & 2\mathbf{v}^\top \mathbf{1} m_2 + 12b^2 m_3 \end{pmatrix}.$$

Let $\mathbf{u}^\top = (\mathbf{v}^\top, b)$. Then,

$$\begin{aligned} \mathbf{u}^\top \mathcal{H}_{\mathbf{Q}} \mathbf{u} &= (\mathbf{v}^\top \ b) \begin{pmatrix} m_1 - \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{\prod_{i=1}^d v_i}} \mathbf{V} & 2dbm_2 - \frac{m_4}{2n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} \\ 2b\mathbf{1} m_2 - \frac{m_4}{n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} & 2\mathbf{v}^\top \mathbf{1} m_2 + 12b^2 m_3 \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ b \end{pmatrix} \\ &= \mathbf{v}^\top \mathbf{v} \left\{ m_1 - \frac{bm_4 + m_5}{n} \frac{1}{\sqrt{\prod_{i=1}^d v_i}} \mathbf{V} \right\} + b\mathbf{v} \left\{ 2b\mathbf{1} m_2 - \frac{m_4}{n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} \right\} \\ &\quad + \mathbf{v}^\top \mathbf{v} b \left\{ 2dbm_2 - \frac{m_4}{2n} \sum_{i=1}^d \frac{1}{v_i^{3/2} (\prod_{j=1, j \neq i}^d v_j)^{1/2}} \right\} + b^2 \{ 2\mathbf{v}^\top \mathbf{1} m_2 + 12b^2 m_3 \} \\ &\stackrel{n \rightarrow \infty}{\rightarrow} \mathbf{v}^\top \mathbf{v} m_1 + 2b^2 \mathbf{v}^\top \mathbf{1} m_2 + \mathbf{v}^\top \mathbf{v} 2db^2 m_2 + b^2 \{ 2\mathbf{v}^\top \mathbf{1} m_2 + 12b^2 m_3 \} \geq 0, \end{aligned}$$

since $\mathbf{v}^\top \mathbf{v} > 0$, $\mathbf{v}$ is element-wise positive, $b > 0$, $m_1, m_3 > 0$ as integrals of a square terms. Also $m_2 > 0$ because the terms in the product that define its integrand have the same sign. As a result, for any fixed $\mathbf{M} \in M_+(d)$ and $h_i^2 = v_i, i = 1, \dots, d$, $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ is a strictly convex function of $\mathbf{v}$ and therefore it is minimized by a unique vector $\mathbf{v}$. Also $\mathbf{M} \in M_+(d)$ implies

$$\mathbf{Q}(\mathbf{v}, \mathbf{M}, b) \geq \left(\sum_{i=1}^d v_i \right)^2 m_1 + \left(\sum_{i=1}^d v_i \right) b^2 m_2 + b^4 m_3 \geq c(\mathbf{M}) \left(\sum_{i=1}^d v_i^2 \right)^2 = c(\mathbf{M}) \|\mathbf{v}\|^2,$$

for all $\mathbf{v} \in \mathbb{R}_+^d$ and Euclidian norm $\|\mathbf{v}\| = \left(\sum_{i=1}^d v_i^2 \right)^{1/2}$, so that

$$\lim_{\|\mathbf{v}\| \rightarrow \infty} \mathbf{Q}(\mathbf{v}, \mathbf{M}, b) \geq \lim_{\|\mathbf{v}\| \rightarrow \infty} c(\mathbf{M}) \|\mathbf{v}\|^2 = \infty.$$

Also it always holds that

$$\begin{aligned} \lim_{\|\mathbf{v}\| \rightarrow \infty} \mathbf{Q}(\mathbf{v}, \mathbf{M}, b) &\geq \\ \lim_{\|\mathbf{v}\| \rightarrow \infty} \left\{ \mathbf{v}^\top m_1 \mathbf{1} db^2 m_2 + \frac{bm_4 + m_5}{2n} \left(\frac{1}{v_1^{3/2} (v_2 \dots v_d)^{1/2}} + \dots + \frac{1}{(v_1 \dots v_{d-1})^{1/2} v_d^{3/2}} \right) \right\} &= \infty. \end{aligned}$$

Thus the infimum of $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ is reached at a (unique as shown above) vector $\mathbf{v}$ whose norm is bounded away from ∞ and 0. Additionally, $\mathbf{Q}(\mathbf{v}_{\text{opt}}, \mathbf{M}, b) < 0$ for any $\mathbf{M} \in M_+(d)$. This ensures the existence of a unique optimal bandwidth vector which minimizes (12).

Lemma 5 Let $\mathbf{X}$ be an invertible $d \times d$ matrix. Then,

$$\frac{\partial |\mathbf{X}|^{-1}}{\partial \text{vech} \mathbf{X}} = -|\mathbf{X}|^{-1} D_d^\top \text{vec} \mathbf{X}^{-1}, \quad (33)$$

$$\frac{\partial^2 |\mathbf{X}|^{-1}}{(\partial \text{vech} \mathbf{X})(\partial \text{vech} \mathbf{X})^\top} = 2|\mathbf{X}|^{-1} [(d \text{vec} \mathbf{X})^\top (\mathbf{X}^{-1} \otimes \mathbf{I}_d)^\top (\text{vec} \mathbf{I}_d)(\text{vec} \mathbf{I}_d)^\top + 2\mathbf{I}_d] (\mathbf{I}_d \otimes \mathbf{X}^{-1})^\top D_d. \quad (34)$$

Proof Starting with (33) we have,

$$\begin{aligned} d|\mathbf{X}|^{-1} &= -|\mathbf{X}|^{-2} d|\mathbf{X}| \text{ since } du^\alpha = \alpha u^{\alpha-1} du = -|\mathbf{X}|^{-2} |\mathbf{X}| \text{tr}(\mathbf{X}^{-1} d\mathbf{X}) \\ &= -|\mathbf{X}|^{-1} (\text{vec} \mathbf{X}^{-1})^\top \text{vec}(d\mathbf{X}) = -|\mathbf{X}|^{-1} D_d^\top (\text{vech} \mathbf{X}^{-1})^\top d \text{vec} \mathbf{X} = -|\mathbf{X}|^{-1} D_d^\top \text{vec} \mathbf{X}^{-1}. \end{aligned}$$

Regarding (34), first write

$$d^2 |\mathbf{X}|^{-1} = d(d|\mathbf{X}|^{-1}) = d(-|\mathbf{X}|^{-2} d|\mathbf{X}|) = -d|\mathbf{X}|^{-2} d|\mathbf{X}| - |\mathbf{X}|^{-2} d^2 |\mathbf{X}| = 2|\mathbf{X}|^{-3} (d|\mathbf{X}|)^2 - |\mathbf{X}|^{-2} d^2 |\mathbf{X}|. \quad (35)$$

Using $d|\mathbf{X}| = |\mathbf{X}| \text{tr}(\mathbf{X}^{-1} d\mathbf{X})$ and $d^2 |\mathbf{X}| = |\mathbf{X}| [\text{tr}^2(\mathbf{X}^{-1} d\mathbf{X}) - \text{tr} \{ \mathbf{X}^{-1} (d\mathbf{X}) \mathbf{X}^{-1} (d\mathbf{X}) \}]$ back to (35) yields

$$\begin{aligned} d^2 |\mathbf{X}|^{-1} &= 2|\mathbf{X}|^{-3} |\mathbf{X}|^2 \text{tr}^2(\mathbf{X}^{-1} d\mathbf{X}) - |\mathbf{X}|^{-2} |\mathbf{X}| [\text{tr}^2(\mathbf{X}^{-1} d\mathbf{X}) - \text{tr} \{ \mathbf{X}^{-1} (d\mathbf{X}) \mathbf{X}^{-1} (d\mathbf{X}) \}] \\ &= 2|\mathbf{X}|^{-1} [(d \text{vec} \mathbf{X})^\top (\mathbf{X}^{-1} \otimes \mathbf{I}_d)^\top (\text{vec} \mathbf{I}_d)(\text{vec} \mathbf{I}_d)^\top + 2\mathbf{I}_d] (\mathbf{I}_d \otimes \mathbf{X}^{-1})^\top D_d, \end{aligned}$$

as required.

6.3 Proof of Theorem 1.

For any d -variate real function g and a sequence of d -dimensional vectors $\mathbf{a}_n$ with all components tending to zero, let $\dot{g}(\mathbf{x})$ denote the $d \times 1$ vector of first order partial derivatives of g and $\ddot{g}(\mathbf{x})$ the Hessian matrix of g , i.e. the $d \times d$ matrix with the (i, j) entry equal to

$$\frac{\partial^2}{\partial x_i \partial x_j} g(\mathbf{x}).$$

The proof uses the multivariate version of Taylor's theorem, according to which, assuming that all entries of $\ddot{g}(\mathbf{x})$ are continuous in a neighborhood of $\mathbf{x}$,

$$g(\mathbf{x} + \mathbf{a}_n) = g(\mathbf{x}) + \mathbf{a}_n^\top \dot{g}(\mathbf{x}) + \frac{1}{2} \mathbf{a}_n^\top \ddot{g}(\mathbf{x}) \mathbf{a}_n + o(\mathbf{a}_n^\top \mathbf{a}_n). \quad (36)$$

Also, by (22) $\hat{F}_L(y|\mathbf{x})$ can be equivalently written as

$$\hat{F}_L(y|\mathbf{x}) = \tilde{F}_L(y|\mathbf{x}) \{1 + o(n^{-\beta})\},$$

with

$$\tilde{F}_L(y|\mathbf{x}) = \frac{1}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^n \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \frac{W \{ (Y_i - y)b^{-1} \} \delta_i}{1 - G(Y_i)}.$$

Thus, asymptotically it is equivalent to prove the theorem using $\tilde{F}_L(y|\mathbf{x})$ instead of $\hat{F}_L(y|\mathbf{x})$. The proof assumes that $\mathbf{x} \in \varepsilon_{b,l}$ i.e. $\mathbf{x}$ is a left boundary point, as treatment of the right boundary is similar. The case of $\mathbf{x} \in \varepsilon_{int}$ is obtained by taking $\mathbf{1} \leq \mathbf{c} \leq \mathbf{M}^\top \mathbf{H}^{-1} - \mathbf{1} - \mathbf{J}^\top \mathbf{H}^{-1}$ and assumption A1. Starting with the bias expression, using the fact that the $(\mathbf{X}_i, Y_i, \delta_i)$ are i.i.d., fixing i in the third step below and using Lemma 1 in the fourth step, we have

$$\begin{aligned} \mathbb{E} \{ \tilde{F}_L(y|\mathbf{x}) \} &= \mathbb{E} \left\{ \frac{1}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^n \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \frac{W \{ (Y_i - y)b^{-1} \} \delta_i}{1 - G(Y_i)} \right\} \\ &= \frac{1}{|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \mathbb{E} \left(\mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \mathbb{E} \left[\mathbb{E} \left\{ \frac{W \{ (Y_i - y)b^{-1} \} \delta_i}{1 - G(Y_i)} \middle| \mathbf{X}_i \right\} \right] \right) \\ &= \frac{1}{|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \mathbb{E} \left[\mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \mathbb{E} \left\{ F_T(y|\mathbf{X}_i) + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{X}_i)}{\partial y^2} + o(b^2) \right\} \right] \\ &= \frac{1}{|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{u} - \mathbf{x}) \} f_{\mathbf{X}}(\mathbf{u}) \left\{ F_T(y|\mathbf{u}) + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{u})}{\partial y^2} + o(b^2) \right\} d\mathbf{u}. \end{aligned}$$

Using the change of variable $\mathbf{H}^{-1}(\mathbf{u} - \mathbf{x}) = \mathbf{z}$ so that $d\mathbf{u} = |\mathbf{H}| d\mathbf{z}$ gives

$$\mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \right\} = \frac{|\mathbf{H}|}{|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int \left[\mathbf{K}_c(\mathbf{z}) f_{\mathbf{X}}(\mathbf{x} + \mathbf{H}\mathbf{z}) \left\{ F_T(y|\mathbf{x} + \mathbf{H}\mathbf{z}) + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x} + \mathbf{H}\mathbf{z})}{\partial y^2} + o(b^2) \right\} \right] d\mathbf{z}. \quad (37)$$

Applying (36) with $g(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{a}_n = \mathbf{H}\mathbf{z}$ gives,

$$f_{\mathbf{X}}(\mathbf{x} + \mathbf{H}\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) - (\mathbf{H}\mathbf{z})^\top \dot{f}_{\mathbf{X}}(\mathbf{x}) + \frac{1}{2}(\mathbf{H}\mathbf{z})^\top \ddot{f}_{\mathbf{X}}(\mathbf{x})(\mathbf{H}\mathbf{z}) + o\{\text{tr}(\mathbf{H}^2)\}. \quad (38)$$

Further, (36) with $g(\mathbf{x}) = F_T(y|\mathbf{x})$, $\mathbf{a}_n = \mathbf{H}\mathbf{z}$ gives,

$$F_T(y|\mathbf{x} + \mathbf{H}\mathbf{z}) = F_T(y|\mathbf{x}) - (\mathbf{H}\mathbf{z})^\top \dot{F}_T(y|\mathbf{x}) + \frac{1}{2}(\mathbf{H}\mathbf{z})^\top \mathcal{H}_{F_T}(y|\mathbf{x})(\mathbf{H}\mathbf{z}) + o\{\text{tr}(\mathbf{H}^2)\}. \quad (39)$$

Also we note that

$$\begin{aligned} & \left[f_{\mathbf{X}}(\mathbf{x}) - (\mathbf{H}\mathbf{z})^\top \dot{f}_{\mathbf{X}}(\mathbf{x}) + \frac{1}{2}(\mathbf{H}\mathbf{z})^\top \ddot{f}_{\mathbf{X}}(\mathbf{x})(\mathbf{H}\mathbf{z}) + o\{\text{tr}(\mathbf{H}^2)\} \right] \times \\ & \quad \left[F_T(y|\mathbf{x} + \mathbf{H}\mathbf{z}) + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x} + \mathbf{H}\mathbf{z})}{\partial y^2} + o(b^2) \right] \\ & = f_{\mathbf{X}}(\mathbf{x}) F_T(y|\mathbf{x}) + (\mathbf{H}\mathbf{z})^\top \dot{f}_{\mathbf{X}}(\mathbf{x})(\mathbf{H}\mathbf{z})^\top \dot{F}_T(y|\mathbf{x}) + \frac{1}{2}(\mathbf{H}\mathbf{z})^\top \ddot{f}_{\mathbf{X}}(\mathbf{x})(\mathbf{H}\mathbf{z}) F_T(y|\mathbf{x}) + \\ & \quad f_{\mathbf{X}}(\mathbf{x}) \frac{1}{2}(\mathbf{H}\mathbf{z})^\top \ddot{F}_T(y|\mathbf{x})(\mathbf{H}\mathbf{z}) + f_{\mathbf{X}}(\mathbf{x}) \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} + o(\mathbf{H}^\top \mathbf{H}) + o(b^2). \end{aligned} \quad (40)$$

Use (38) and (40) back to (37), rearranging and using (56) yields

$$\begin{aligned} \mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \right\} & = F_T(y|\mathbf{x}) + \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \left\{ \mathbf{H}^\top \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \left\{ \mathbf{H}^\top \dot{F}_T(y|\mathbf{x}) \right\} \mu_2(\mathbf{K}_c) \\ & \quad + \frac{1}{2f_{\mathbf{X}}(\mathbf{x})} \left\{ \mathbf{H}^\top \ddot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \left\{ \mathbf{H} F_T(y|\mathbf{x}) \right\} \mu_2(\mathbf{K}_c) \\ & \quad + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \mathbf{H}^\top \ddot{F}_T(y|\mathbf{x}) \mathbf{H} \mu_2(\mathbf{K}_c) + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} + o(\mathbf{H}^\top \mathbf{H}) + o(b^2). \end{aligned}$$

Now, $\mathbf{H}$ is a diagonal matrix and thus $\mathbf{H}^\top = \mathbf{H}$ and for any $d \times 1$ vector $\mathbf{A}$, $\mathbf{H}^\top \mathbf{A} = \text{tr}(\mathbf{H}^\top \mathbf{A})$. As a result

$$\begin{aligned} \mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \right\} & = F_T(y|\mathbf{x}) + \mu_2(\mathbf{K}_c) \left[\frac{\text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\}}{f_{\mathbf{X}}(\mathbf{x})} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{tr} \left\{ (\mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H}) \right\} \right. \\ & \quad \left. + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{tr} \left\{ (\mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H}) \right\} \right] + \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} + o(\mathbf{H}^\top \mathbf{H}) + o(b^2), \end{aligned} \quad (41)$$

as required. The variance is calculated by combining Lemma 2 and the law of total variance

$$\text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \right\} = \text{Var} \left[\mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] + \mathbb{E} \left[\text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right]. \quad (42)$$

First, from Lemma 2 and for fixed i in the second and third steps below we have,

$$\begin{aligned} \text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} & = \text{Var} \left\{ \frac{1}{nf_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^n \mathbf{K}_c \left\{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \right\} \frac{\delta_i W \left\{ (Y_i - y)b^{-1} \right\}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i \right\} \\ & = \frac{n}{n^2 f_{\mathbf{X}}^2(\mathbf{x})} \mathbf{K}_c \left\{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \right\}^2 \text{Var} \left\{ \frac{\delta_i W \left\{ (Y_i - y)b^{-1} \right\}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i \right\} = \\ & \quad \frac{1}{nf_{\mathbf{X}}^2(\mathbf{x})} \mathbf{K}_c \left\{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \right\}^2 \left\{ P(y|\mathbf{X}_i) - 2bp(y|\mathbf{X}_i)A_{1,1} - F_T^2(y|\mathbf{X}_i) \right\} \left\{ 1 + O(b^2) + o(n^{-\beta}) \right\}. \end{aligned}$$

Then

$$\mathbb{E} \left[\text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] =$$

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{nf_{\mathbf{X}}^2(\mathbf{x})} \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \}^2 \{ P(y|\mathbf{X}_i) - 2bp(y|\mathbf{X}_i)A_{1,1} - F_T^2(y|\mathbf{X}_i) \} \right] \{ 1 + O(b^2) + o(n^{-\beta}) \} = \\ & \frac{1}{nf_{\mathbf{X}}^2(\mathbf{x})} \int \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{z} - \mathbf{x}) \}^2 f_{\mathbf{X}}(\mathbf{z}) \{ P(y|\mathbf{z}) - 2bp(y|\mathbf{z})A_{1,1} - F_T^2(y|\mathbf{z}) \} d\mathbf{z} \{ 1 + O(b^2) + o(n^{-\beta}) \}. \end{aligned} \quad (43)$$

Applying the change of variable $\mathbf{z} - \mathbf{x} = \mathbf{H}\mathbf{u}$ so that $d\mathbf{z} = \mathbf{H} d\mathbf{u}$ and substituting back to (43) gives

$$\begin{aligned} & \mathbb{E} \left[\text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] = n^{-1} |\mathbf{H}|^{-1} f_{\mathbf{X}}^{-2}(\mathbf{x}) \times \\ & \int \mathbf{K}_c(\mathbf{u})^2 f_{\mathbf{X}}(\mathbf{x} + \mathbf{H}\mathbf{u}) \{ P(y|\mathbf{x} + \mathbf{H}\mathbf{u}) - 2bp(y|\mathbf{x} + \mathbf{H}\mathbf{u})A_{1,1} - F_T^2(y|\mathbf{x} + \mathbf{H}\mathbf{u}) \} d\mathbf{u} \{ 1 + O(b^2) + o(n^{-\beta}) \}. \end{aligned} \quad (44)$$

Subsequently applying (36) on $f_{\mathbf{X}}(\mathbf{x} + \mathbf{H}\mathbf{u})$, and similarly to (39) expand $P(y|\mathbf{x} + \mathbf{H}\mathbf{u})$, $p(y|\mathbf{x} + \mathbf{H}\mathbf{u})$ and $F_T^2(y|\mathbf{x} + \mathbf{H}\mathbf{u})$ in Taylor series around $\mathbf{x}$, and substituting back to (44) gives

$$\mathbb{E} \left[\text{Var} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] = \frac{R(\mathbf{K}_c)}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \{ P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1} - F_T^2(y|\mathbf{x}) \} \{ 1 + O(b^2) + O(\mathbf{H}^\top \mathbf{H}) + o(n^{-\beta}) \}. \quad (45)$$

Also, by Lemma 1 and fixing i in the second and third steps below

$$\begin{aligned} \mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} &= \mathbb{E} \left\{ \frac{1}{nf_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^n \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \frac{\delta_i W \{ (Y_i - y)b^{-1} \}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i \right\} \\ &= \frac{n}{nf_{\mathbf{X}}(\mathbf{x})} \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} \mathbb{E} \left\{ \frac{\delta_i W \{ (Y_i - y)b^{-1} \}}{1 - \hat{G}(Y_i)} \middle| \mathbf{X}_i \right\} \\ &= \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} F_T(y|\mathbf{X}_i) \{ 1 + O(b^2) + O_p(n^{-\beta}) \}. \end{aligned}$$

Thus

$$\begin{aligned} \text{Var} \left[\mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] &= \text{Var} \left[\frac{1}{f_{\mathbf{X}}(\mathbf{x})} \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} F_T(y|\mathbf{X}_i) \right] \{ 1 + O(b^2) + O_p(n^{-\beta}) \} = \\ & \frac{1}{f_{\mathbf{X}}^2(\mathbf{x})} \left[\int \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{z} - \mathbf{x}) \}^2 F_T(y|\mathbf{z})^2 f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} - \left(\int \mathbf{K}_c \{ \mathbf{H}^{-1}(\mathbf{z} - \mathbf{x}) \} F_T(y|\mathbf{z}) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} \right)^2 \right] \\ & \quad \times \{ 1 + O(b^2) + O_p(n^{-\beta}) \}. \end{aligned}$$

Using once more the change of variable $\mathbf{z} - \mathbf{x} = \mathbf{H}\mathbf{u}$ and subsequently expanding in Taylor series around $\mathbf{x}$ both $F_T(y|\mathbf{x} + \mathbf{H}\mathbf{u})^2$ and $F_T(y|\mathbf{x} + \mathbf{H}\mathbf{u})$, substituting back to the above expression and after straightforward algebra gives

$$\text{Var} \left[\mathbb{E} \left\{ \tilde{F}_L(y|\mathbf{x}) \middle| \mathbf{X} \right\} \right] = O(\mathbf{H}^\top \mathbf{H}). \quad (46)$$

Using (45) and (46) back in (42) completes the proof of the theorem.

6.4 Proof of Theorem 2.

The first step of the proof is to write (12) in matrix form. To see this, first recall the matrix properties

$$\text{tr}(\mathbf{A}^\top \mathbf{B}) = (\text{vec}^\top \mathbf{A})(\text{vec} \mathbf{B}), \text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}), \text{vec} \mathbf{A} = \mathbf{D}_d \text{vech} \mathbf{A}.$$

Then we have

$$\begin{aligned} \text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} &= \text{vec}^\top \mathbf{H} \{ \text{vec} \dot{f}_{\mathbf{X}}(\mathbf{x}) \} = \text{vech}^\top \mathbf{H} \mathbf{D}_d^\top \{ \text{vec} \dot{f}_{\mathbf{X}}(\mathbf{x}) \}, \\ \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\} &= \text{vec}^\top \mathbf{H} \{ \text{vec} \dot{F}_T(y|\mathbf{x}) \} = \text{vech}^\top \mathbf{H} \mathbf{D}_d^\top \{ \text{vec} \dot{F}_T(y|\mathbf{x}) \}, \\ \text{tr} \left\{ \mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H} \right\} &= \text{vec}^\top \mathbf{H}^2 \{ \text{vec} \ddot{F}_T(y|\mathbf{x}) \} = \text{vech}^\top \mathbf{H}^2 \mathbf{D}_d^\top \{ \text{vec} \ddot{F}_T(y|\mathbf{x}) \}, \\ \text{tr} \left\{ \mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H} \right\} &= \text{vec}^\top \mathbf{H}^2 \{ \text{vec} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \} = \text{vech}^\top \mathbf{H}^2 \mathbf{D}_d^\top \{ \text{vec} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \}. \end{aligned}$$

Thus,

$$\begin{aligned}
& \frac{\text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\}}{f_{\mathbf{X}}(\mathbf{x})} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{tr} \left\{ (\mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H}) \right\} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{tr} \left\{ (\mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H}) \right\} = \\
& \frac{\text{vech}^\top \mathbf{H} \mathbf{D}_d^\top \{ \text{vec} \dot{f}_{\mathbf{X}}(\mathbf{x}) \} \text{vech}^\top \mathbf{H} \mathbf{D}_d^\top \{ \text{vec} \dot{F}_T(y|\mathbf{x}) \}}{f_{\mathbf{X}}(\mathbf{x})} \\
& + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{vech}^\top \mathbf{H}^2 \mathbf{D}_d^\top \{ \text{vec} \ddot{F}_T(y|\mathbf{x}) \} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{vech}^\top \mathbf{H}^2 \mathbf{D}_d^\top \{ \text{vec} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \} \\
& = \text{vech}^\top \mathbf{H}^2 \left(\frac{\mathbf{D}_d^\top \{ \text{vec} \dot{f}_{\mathbf{X}}(\mathbf{x}) \} \mathbf{D}_d^\top \{ \text{vec} \dot{F}_T(y|\mathbf{x}) \}}{f_{\mathbf{X}}(\mathbf{x})} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \mathbf{D}_d^\top \{ \text{vec} \ddot{F}_T(y|\mathbf{x}) \} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \mathbf{D}_d^\top \{ \text{vec} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \} \right) \\
& \equiv \text{vech}^\top \mathbf{H}^2 \psi(y, \mathbf{x}),
\end{aligned}$$

where

$$\psi(y, \mathbf{x}) = \frac{\mathbf{D}_d^\top \{ \text{vec} \dot{f}_{\mathbf{X}}(\mathbf{x}) \} \mathbf{D}_d^\top \{ \text{vec} \dot{F}_T(y|\mathbf{x}) \}}{f_{\mathbf{X}}(\mathbf{x})} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \mathbf{D}_d^\top \{ \text{vec} \ddot{F}_T(y|\mathbf{x}) \} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \mathbf{D}_d^\top \{ \text{vec} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \}.$$

Hence

$$\begin{aligned}
& \text{AMISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} = \\
& \mu_2^2(\mathbf{K}) (\text{vech}^\top \mathbf{H}^2)^2 \int_{\varepsilon_{y,o}} \psi^2(y, \mathbf{x}) dy d\mathbf{x} + 2\mu_2(\mathbf{K}) \frac{b^2 \mu_2(w)}{2} \text{vech}^\top \mathbf{H}^2 \int_{\varepsilon_{y,o}} \psi(y, \mathbf{x}) \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} dy d\mathbf{x} \\
& + \frac{b^4 \mu_2^2(w)}{4} \int_{\varepsilon_{y,o}} \left\{ \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\}^2 dy d\mathbf{x} \\
& + \frac{R(\mathbf{K})}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int_{\varepsilon_{y,o}} \{P(y|\mathbf{x}) - F_T^2(y|\mathbf{x})\} dy d\mathbf{x} - \frac{2bR(\mathbf{K})A_{1,1}}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int_{\varepsilon_{y,o}} p(y|\mathbf{x}) dy d\mathbf{x}.
\end{aligned}$$

Now, rewrite m_1, m_2 as the $d(d+1)/2 \times d(d+1)/2$ matrices

$$m_1 = \mu_2^2(\mathbf{K}) \int_{\varepsilon_{y,o}} \psi^2(y, \mathbf{x}) dy d\mathbf{x}, m_2 = \mu_2(\mathbf{K}) \mu_2(w) \int_{\varepsilon_{y,o}} \psi(y, \mathbf{x}) \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} dy d\mathbf{x},$$

and m_3, m_4, m_5 are scalars defined as before. We will prove that the vector $(\mathbf{v}_{\text{opt}}, b_{\text{opt}})$, where

$$\mathbf{v}_{\text{opt}} = \left(n^{-1/(4+d)} \left\{ \frac{m_{5,1}}{m_{1,1}} \right\}^{1/(4+d)}, \dots, n^{-1/(4+d)} \left\{ \frac{m_{5,d}}{m_{1,d}} \right\}^{1/(4+d)} \right)^\top, b_{\text{opt}} = \left\{ \frac{m_4}{2(m_2 + 2m_3)} \right\}^{2/(5+d)} n^{-2/(5+d)},$$

is the unique vector that minimizes the $\text{AMISE}\{\hat{F}_L(h_1, \dots, h_d, b)\} = \mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$. First it must be shown that for any matrix $\mathbf{M} = (M_{ij})_{i,j=1}^d \in M_+(d)$, the set of all $d \times d$ positive definite matrices it holds that

$$\frac{\partial}{\partial \mathbf{v}} \mathbf{Q}(\mathbf{v}_{\text{opt}}, \mathbf{M}, b_{\text{opt}}) = 0 \quad \text{and} \quad \frac{\partial}{\partial b} \mathbf{Q}(\mathbf{v}_{\text{opt}}, \mathbf{M}, b_{\text{opt}}) = 0.$$

To see this, note that

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{v}} \mathbf{Q}(\mathbf{v}_{\text{opt}}, \mathbf{M}, b_{\text{opt}}) &= \mathbf{v}_{\text{opt}}^\top m_1 \mathbf{1} + db^2 m_2 + \frac{bm_4 + m_5}{2n} \sum_{i=1}^d \frac{1}{v_{i,\text{opt}}^{3/2} \left(\prod_{j=1, j \neq i}^d v_{j,\text{opt}} \right)^{1/2}} = \\
& n^{-1/(4+d)} m_1 \sum_{i=1}^d \left(\frac{m_{5,i}}{m_{1,i}} \right)^{1/(4+d)} + d \left\{ \frac{m_4}{2(m_2 + 2m_3)} \right\}^{4/(5+d)} n^{-4/(5+d)} \\
& + \frac{bm_4 + m_5}{2n} \sum_{i=1}^d \frac{1}{v_{i,\text{opt}}^{3/2} \left(\prod_{j=1, j \neq i}^d v_{j,\text{opt}} \right)^{1/2}} \xrightarrow{n \rightarrow \infty} 0,
\end{aligned}$$

since by assumption A4, m_1, m_2, m_4, m_5 are bounded. Similarly

$$\begin{aligned}
\frac{\partial}{\partial b} \mathbf{Q}(\mathbf{v}_{\text{opt}}, \mathbf{M}, b_{\text{opt}}) &= 2b \mathbf{v}_{\text{opt}}^\top \mathbf{1} m_2 + 4b^3 m_3 - \frac{m_4}{n} \frac{1}{\sqrt{v_{1,\text{opt}} \dots v_{d,\text{opt}}}} \\
&= 2n^{-1/(4+d)} m_1 \sum_{i=1}^d \left(\frac{m_{5,i}}{m_{1,i}} \right)^{1/(4+d)} \left\{ \frac{m_4}{2(m_2 + 2m_3)} \right\}^{2/(5+d)} n^{-2/(5+d)} m_2
\end{aligned}$$

$$+ 4 \left\{ \frac{m_4}{2(m_2 + 2m_3)} \right\}^{3/(5+d)} n^{-6/(5+d)} m_3 - \frac{m_4}{n} \frac{1}{\sqrt{v_{1,\text{opt}} \cdots v_{d,\text{opt}}}} \xrightarrow{n \rightarrow \infty} 0.$$

The proof is completed by application of Lemma 4 by which for any $\mathbf{M} \in M_+(d)$ the vector $(\mathbf{v}_{\text{opt}}, b_{\text{opt}})$, where $\mathbf{v}_{\text{opt}} \equiv \mathbf{v}(\mathbf{M}) \in \mathbb{R}_+^d$ which minimizes $\mathbf{Q}(\mathbf{v}, \mathbf{M}, b)$ is unique.

6.5 Proof of Theorem 3.

The first step in proving (19) is to use (12) and write the AMISE $\{\hat{F}_L(\mathbf{H}, b)\} \equiv \mathbf{Q}(\mathbf{H}, b)$ as

$$\mathbf{Q}(\mathbf{H}, b) = (\text{vech}^\top \mathbf{H}^2) m_1 (\text{vech}^\top \mathbf{H}^2) + (\text{vech}^\top \mathbf{H}^2) b^2 m_2 + b^4 m_3 - \frac{b m_4}{n |\mathbf{H}|} + \frac{m_5}{n |\mathbf{H}|}.$$

Set

$$\nabla \mathbf{Q}(\mathbf{H}, b) = \frac{\partial \mathbf{Q}(\mathbf{H}, b)}{\partial \text{vech}(\mathbf{H}^2)}, \quad \nabla^2 \mathbf{Q}(\mathbf{H}, b) = \frac{\partial^2 \mathbf{Q}(\mathbf{H}, b)}{\{\partial \text{vech}(\mathbf{H}^2)\} \{\partial \text{vech}(\mathbf{H}^2)\}^\top}.$$

Then, (33) and (34) respectively give

$$\nabla \mathbf{Q}(\mathbf{H}, b) = -n^{-1} |\mathbf{H}|^{-1} (m_5 - b m_4) D_d^\top \text{vec} \mathbf{H}^{-1} + b^2 m_2 + m_1 \text{vech}^\top \mathbf{H}^2, \quad (47)$$

$$\begin{aligned} \nabla^2 \mathbf{Q}(\mathbf{H}, b) &= 2n^{-1} |\mathbf{H}|^{-1} (m_5 - b m_4) [(d \text{vec} \mathbf{H})^\top (\mathbf{H}^{-1} \otimes \mathbf{I}_d)^\top (\text{vec} \mathbf{I}_d) (\text{vec} \mathbf{I}_d)^\top + 2 \mathbf{I}_d] \\ &\quad \times (\mathbf{I}_d \otimes \mathbf{H}^{-1})^\top D_d. \end{aligned} \quad (48)$$

Now, fixing b_{opt} , a Taylor expansion of $\nabla \mathbf{Q}(\mathbf{H}, b)$ around $\mathbf{H}$ gives

$$\nabla \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}}) = \nabla \mathbf{Q}(\hat{\mathbf{H}}_{\text{opt}}, b_{\text{opt}}) + \nabla^2 \mathbf{Q}(\hat{\mathbf{H}}_{\text{opt}}, b_{\text{opt}}) \text{vech}(\mathbf{H}_{\text{opt}}^2 - \hat{\mathbf{H}}_{\text{opt}}^2). \quad (49)$$

But $\nabla \mathbf{Q}(\hat{\mathbf{H}}_{\text{opt}}, b_{\text{opt}}) = 0$. Substituting back to (49) and solving for $\text{vech}(\mathbf{H}_{\text{opt}}^2 - \hat{\mathbf{H}}_{\text{opt}}^2)$ gives

$$\text{vech}(\mathbf{H}_{\text{opt}}^2 - \hat{\mathbf{H}}_{\text{opt}}^2) \simeq -\{\nabla^2 \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}})\}^{-1} \nabla \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}}). \quad (50)$$

By (47) for $\mathbf{H} = \mathbf{H}_{\text{opt}}$, $\nabla \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}}) \simeq 2(\hat{m}_1 - m_1) \text{vech}(\mathbf{H}_{\text{opt}}^2) + (\hat{m}_2 - m_2) b^2$. Use this back to (50) yields

$$\text{vech}(\hat{\mathbf{H}}_{\text{opt}}^2 - \mathbf{H}_{\text{opt}}^2) = -\{\nabla^2 \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}})\}^{-1} \{2(m_1 - \hat{m}_1) \text{vech}(\mathbf{H}_{\text{opt}}^2) + (m_2 - \hat{m}_2) b_{\text{opt}}^2\}. \quad (51)$$

Let $\mathbf{C}$ be a constant positive definite matrix and set $\mathbf{H}_{\text{opt}} = \mathbf{C} n^{-2/(d+4)}$. From (48), $\nabla^2 \mathbf{Q}(\mathbf{H}_{\text{opt}}, b_{\text{opt}}) \rightarrow \mathbf{\Omega}$ as $n \rightarrow \infty$, where $\mathbf{\Omega}$ is a constant non-singular matrix. Thus

$$\text{vech}(\hat{\mathbf{H}}_{\text{opt}}^2 - \mathbf{H}_{\text{opt}}^2) \simeq \mathbf{\Omega} (m_1 - \hat{m}_1) \text{vech}(\mathbf{H}_{\text{opt}}^2) + \mathbf{\Omega} (m_2 - \hat{m}_2) b_{\text{opt}}^2. \quad (52)$$

The above equation indicates explicitly that the asymptotic performance of $\hat{\mathbf{H}}_{\text{opt}}^2$ is driven by the precision of estimation of m_1 by $\hat{m}_1$ and m_2 by $\hat{m}_2$ as measured by $\text{MSE}(\hat{m}_1)$ and $\text{MSE}(\hat{m}_2)$. Of course this precision depends also on the pilot bandwidth $\mathbf{G}$ of $\hat{m}_1, \hat{m}_2$. Working as in Theorem 2, Chacón and Duong (2010), it is straightforward to prove that, assuming that the pilot bandwidth matrix $\mathbf{G}_{\text{AMSE}}$ is of order $n^{-2/(r+d+2)}$ and for $d \geq 2$, we have

$$\text{MSE}(\hat{m}_1) = O(n^{-\min\{r+d+2, 4\}/(r+d+2)}), \quad \text{MSE}(\hat{m}_2) = O(n^{-\min\{r+d+2, 4\}/(r+d+2)}).$$

Additionally, from (15) $b_{\text{opt}} = O(n^{-2/(5+d)})$. Using this and the above equation back to (52) gives

$$\begin{aligned} \text{vech}(\hat{\mathbf{H}}_{\text{opt}}^2 - \mathbf{H}_{\text{opt}}^2) &\simeq \mathbf{\Omega} n^{-\min\{r+d+2, 4\}/(r+d+2)} \text{vech}(\mathbf{H}_{\text{opt}}^2) + \mathbf{\Omega} n^{-\min\{r+d+2, 4\}/(r+d+2)} n^{-2/(5+d)} \\ &= O_p \left(\mathbf{I}_d^* n^{-\min\{r+d+2, 4\}/(r+d+2)} \right). \end{aligned}$$

The proof of the theorem is completed by taking $r = 4$.

References

- Allen LN, Rose LC, (2006) Financial survival analysis of defaulted debtors. *J Oper Res Soc* 57:630–636.
- Amico M, Van Keilegom I (2018) Cure models in survival analysis. *Annu Rev Stat Appl* 5:311–342.
- Beran R (1981) Nonparametric regression with randomly censored survival data. Technical Report, University of California.
- Beran J, Djaïdja A (2007) Credit risk modeling based on survival analysis with immunes. *Stat Methodol*, 4:251–276.
- Bowman A, Hall P, Prvan T (1998) Bandwidth selection for the smoothing of distribution functions. *Biometrika* 85:799–808.
- Buch-Kromann, T Nielsen JP (2012) Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated or censored data. *Ann Instit Statist Math* 64:167–192.
- Cao R, Vilar JM, Devia A (2009) Modelling consumer credit risk via survival analysis. *SORT* 33:3–30.
- Chacón JE, Duong T, (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *TEST* 19:375–398.
- Chacón JE, Duong T (2018) *Multivariate Kernel Smoothing and Its Applications*, CRC press, London.
- Chen K, Lo S-H, (1997) On the rate of uniform convergence of the product-limit estimator: strong and weak laws. *Ann Statist* 25:1050–1087.
- Cheng M-Y (1997) A bandwidth selector for local linear density estimators. *Ann Statist* 25:1001–1013.
- Cheng M-Y, Fan J, Marron JS (1997) On automatic boundary corrections. *Ann Statist* 25:1691–1708.
- Cheng M-Y, Peng L (2002) Regression modeling for nonparametric estimation of distribution and quantile functions. *Stat Sin* 12:1043–1060.
- Dirick L, Claeskens G, Baesens B, (2003) Time to default in credit scoring using survival analysis: a benchmark study. *J Oper Res Soc* 68:652–665.
- Duong T, and Hazelton ML, (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *J Nonparametr Stat* 15:17–30.
- Duong T, Hazelton ML, (2005) Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J Multivar Analys* 93:417–433.
- Duong T, Hazelton ML, (2005) Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand J Statist* 32:485–506.
- Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83:189–206.
- Fan J, Yim TH (2004) A crossvalidation method for estimating conditional densities. *Biometrika* 91:819–834.
- Geerdens C, Acar EF, Janssen P (2018) Conditional copula models for right-censored clustered event time data. *Biostatistics* 19: 247–262.
- Hall P, Marron JS (1987) On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann Statist* 15:163–181.
- Gulati S, Padgett WJ, (1996) Families of smooth confidence bands for the survival function under the general random censorship model. *Lifetime Data Anal* 2:349–362.
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Amer Statist Assoc* 53:457–481.
- Karunamuni R, Yang S (1991) Weak and strong uniform consistency rates of kernel density estimates for randomly censored data. *Canad J Statist* 19:349–359.
- Kim C, Oh M, Yang SJ, Choi H (2010) A local linear estimation of conditional hazard function in censored data. *J Korean Stat Soc* 39:347–355.
- Li Q, Racine JS (2008) Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *J Bus Econ Stat* 26:423–434.
- López-Cheda A (2018) Nonparametric inference in mixture cure models. PhD thesis. University of Coruña.
- López-Cheda A, Cao R, Jácome MA, (2017) Nonparametric latency estimation for mixture cure models. *TEST* 26:353–376.
- López-Cheda A, Cao R, Jácome MA, Van Keilegom I, (2017) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput Stat Data Anal* 105:144–165.
- Meira-Machado L, Sestelo M, Goncalves A (2015) Nonparametric estimation of the survival function for ordered multivariate failure time data: A comparative study. *Biomet J* 58: 623–634.
- Müller H-G, Stadtmüller U (1999) Multivariate boundary kernels and a continuous least squares principle. *J R Statist Soc B* 61:439–458.
- Marron JS, Padgett WJ (1987) Asymptotically optimal bandwidth selection for kernel density estimators from randomly right censored samples. *Ann Statist* 15:1520–1535.
- Ruppert D, Wand MP (1994) Multivariate locally weighted least squares regression. *Ann Statist* 22:1346–1370.
- Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Ser B* 53:683–690.
- Spierdijk L (2008) Nonparametric conditional hazard rate estimation: a local linear approach. *Comput Stat Data Anal* 52:2419–2434.
- Strzalkowska-Kominiak E, Cao R (2013) Maximum likelihood estimation for conditional distribution single-index models under censoring. *J Multivar Analys* 114:74–98.
- Peláez R, Cao R, Vilar Fernández JM (2021) Probability of default estimation in credit risk using a nonparametric approach. *TEST* 30:383–405.
- Peláez R, Cao R, Vilar Fernández JM (2021) Nonparametric estimation of probability of default with double smoothing. *SORT* 30:93–120.
- Peláez R, Cao R, Vilar Fernández JM (2022) Nonparametric estimation of the conditional survival function with double smoothing. *J Nonparam Statist* 34:1063–1090.
- Peláez R, Cao R, Vilar Fernández JM (2022) Bootstrap bandwidth selection and confidence regions for double smoothed default probability estimation. *Mathematics* 10:1523.
- Peláez R, Van Keilegom I, Cao R, Vilar Fernández JM (2024). Probability of default estimation in credit risk using mixture cure models. *Comput Stat Data Anal*, 189: 107853.
- Yang L, Tschernig R (1999) Multivariate bandwidth selection for local linear regression. *J R Statist Soc B*. 61:793–815.
- Wand MP, Jones MC (1994) Multivariate plug-in bandwidth selection. *Comput Statist* 9:97–116.

Supplementary material: Nonparametric conditional survival function estimation and plug-in bandwidth selection with multiple covariates.

by

Dimitrios Bagkavos, Montserrat Guillen, Jens P. Nielsen.

A Auxiliary calculations.

A.1 Equivalent asymptotic representation of the estimator and automatic boundary adjustment.

Let $\mathbf{K}^*$ denote the so-called *effective* kernel, defined by

$$\mathbf{K}^*(\mathbf{u}; \mathbf{x}) \equiv \mathbf{K}^*(\mathbf{u}) = \mathbf{e}_1^\top (\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x)^{-1} (1 \ (\mathbf{u} - \mathbf{x})^\top)^\top \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{u} - \mathbf{x})\},$$

Note that $\hat{F}_L(y|\mathbf{x})$ can be written as

$$\hat{F}_L(y|\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbf{K}^*(\mathbf{X}_i; \mathbf{x}) \frac{W\{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)}, \quad (53)$$

and observe that for any $\mathbf{x} \in S$,

$$\sum_{i=1}^n \mathbf{K}^*(\mathbf{X}_i; \mathbf{x}) = 1 \text{ and } \sum_{i=1}^n \mathbf{K}^*(\mathbf{X}_i; \mathbf{x})(\mathbf{X}_i - \mathbf{x}) = 0,$$

and as a result, $\mathbf{K}^*$ enables $\hat{F}_L(y|\mathbf{x})$ to achieve – up to normalizing constants – automatic bias correction in the boundary. To see this, let $\mathbf{M} = (M_1, \dots, M_d)^\top$, $\mathbf{1} = (1, \dots, 1)^\top$, $\mathbf{0} = (0, \dots, 0)^\top$ and recall that $\mathbf{J} = (J_1, \dots, J_d)^\top$, $\mathbf{c} = (c_1, \dots, c_d)^\top$. In the sequel all inequalities involving vectors are assumed to hold elementwise. Then

$$\mathbf{x} = \begin{cases} \mathbf{J} + \mathbf{H}\mathbf{c} \in S_{b,l} & \text{if } \mathbf{0} < \mathbf{c} < \mathbf{1}, \\ \mathbf{J} + \mathbf{H}\mathbf{c} \in S_{\text{int}} & \text{if } \mathbf{1} \leq \mathbf{c} \leq \mathbf{M}^\top \mathbf{H}^{-1} - \mathbf{1} - \mathbf{J}^\top \mathbf{H}^{-1}, \\ \mathbf{J} + \mathbf{H}\mathbf{c} \in S_{b,r} & \text{if } \mathbf{M}^\top \mathbf{H}^{-1} - \mathbf{1} - \mathbf{J}^\top \mathbf{H}^{-1} \leq \mathbf{c} \leq \mathbf{M}^\top \mathbf{H}^{-1} - \mathbf{J}^\top \mathbf{H}^{-1}. \end{cases} \quad (54)$$

Define the scalars $S_{n,0}(\mathbf{x})$, $S_{n,2}(\mathbf{x})$ and the $d \times 1$ vector $S_{n,1}(\mathbf{x})$ by

$$\begin{aligned} S_{n,0}(\mathbf{x}) &= n^{-1} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\}, \quad S_{n,1}(\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} (\mathbf{X}_i - \mathbf{x}), \\ S_{n,2}(\mathbf{x}) &= n^{-1} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} (\mathbf{X}_i - \mathbf{x})^\top (\mathbf{X}_i - \mathbf{x}), \end{aligned}$$

By straightforward matrix algebra,

$$\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x = n \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} & \frac{1}{n} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} (\mathbf{X}_i - \mathbf{x})^\top \\ \frac{1}{n} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} (\mathbf{X}_i - \mathbf{x}) & \frac{1}{n} \sum_{i=1}^n \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} (\mathbf{X}_i - \mathbf{x})^\top (\mathbf{X}_i - \mathbf{x}) \end{pmatrix} = n \begin{pmatrix} S_{n,0}(\mathbf{x}) & S_{n,1}(\mathbf{x})^\top \\ S_{n,1}(\mathbf{x}) & S_{n,2}(\mathbf{x}) \end{pmatrix}.$$

By Lemma 3, $S_{n,i}(\mathbf{x}) = |\mathbf{H}|^{i+1} f_{\mathbf{X}}(\mathbf{x}) S_{i,c} \{1 + o_p(1)\}$, $i = 0, 1, 2$ and hence

$$\begin{aligned} \begin{pmatrix} S_{n,0}(\mathbf{x}) & S_{n,1}(\mathbf{x})^\top \\ S_{n,1}(\mathbf{x}) & S_{n,2}(\mathbf{x}) \end{pmatrix}^{-1} &= \begin{pmatrix} |\mathbf{H}| f_{\mathbf{X}}(\mathbf{x}) S_{0,c} \{1 + o_p(1)\} & |\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} \\ |\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} & |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) S_{2,c} \{1 + o_p(1)\} \end{pmatrix}^{-1} \\ &= \frac{1}{|\mathbf{H}|^4 f_{\mathbf{X}}^2(\mathbf{x})} \frac{1 + o_p(1)}{S_{0,c} S_{2,c} - S_{1,c}^2} \begin{pmatrix} |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) S_{2,c} \{1 + o_p(1)\} & -|\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} \\ -|\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} & |\mathbf{H}| f_{\mathbf{X}}(\mathbf{x}) S_{0,c} \{1 + o_p(1)\} \end{pmatrix}. \end{aligned} \quad (55)$$

Substituting (55) to the definition of $\mathbf{K}^*(\mathbf{X}_i; \mathbf{x})$ and substituting the result to (53) yields

$$\begin{aligned} \hat{F}_L(y|\mathbf{x}) &= \frac{n^{-1} \mathbf{e}_1^\top}{|\mathbf{H}|^4 f_{\mathbf{X}}^2(\mathbf{x})} \frac{1 + o_p(1)}{S_{0,c} S_{2,c} - S_{1,c}^2} \begin{pmatrix} |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) S_{2,c} \{1 + o_p(1)\} & -|\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} \\ -|\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} \{1 + o_p(1)\} & |\mathbf{H}| f_{\mathbf{X}}(\mathbf{x}) S_{0,c} \{1 + o_p(1)\} \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & \mathbf{X}_1 - \mathbf{x} \\ \vdots & \vdots \\ 1 & \mathbf{X}_n - \mathbf{x} \end{pmatrix}^\top \begin{pmatrix} \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_1 - \mathbf{x})\} & 0 \\ & \ddots \\ 0 & \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_n - \mathbf{x})\} \end{pmatrix} \begin{pmatrix} \frac{W\{(Y_1 - y)b^{-1}\} \delta_1}{1 - \hat{G}(Y_1)} \\ \vdots \\ \frac{W\{(Y_n - y)b^{-1}\} \delta_n}{1 - \hat{G}(Y_n)} \end{pmatrix} \\ &= \frac{n^{-1}}{|\mathbf{H}|^4 f_{\mathbf{X}}^2(\mathbf{x})} \frac{1 + o_p(1)}{S_{0,c} S_{2,c} - S_{1,c}^2} \sum_{i=1}^n \{ |\mathbf{H}|^3 f_{\mathbf{X}}(\mathbf{x}) S_{2,c} - |\mathbf{H}|^2 f_{\mathbf{X}}(\mathbf{x}) S_{1,c} (\mathbf{X}_i - \mathbf{x}) \} \mathbf{K} \{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} \frac{W\{(Y_i - y)b^{-1}\} \delta_i}{1 - \hat{G}(Y_i)}, \end{aligned}$$

which after some trivial algebra yields (8). That is, a useful and asymptotically equivalent representation of $\mathbf{K}^*$ is obtained by the relationship

$$\mathbf{K}^*(\mathbf{u}; \mathbf{x}) \equiv \mathbf{K}^*(\mathbf{u}) = \frac{1}{|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \frac{S_{2,c} - S_{1,c}(\mathbf{u} - \mathbf{x})|\mathbf{H}|^{-1}}{S_{2,c}S_{0,c} - S_{1,c}^2} \mathbf{K}\{\mathbf{H}^{-1}(\mathbf{u} - \mathbf{x})\} \{1 + o_p(1)\},$$

which implies that $\mathbf{K}^*(\mathbf{u}; \mathbf{x}) = |\mathbf{H}|^{-1}f_{\mathbf{X}}(\mathbf{x})^{-1}\mathbf{K}_{(c)}(\mathbf{u})\{1 + o_p(1)\}$. The weighting scheme imposed on the equivalent kernel results in $\mathbf{K}_{(c)}$ satisfying the moment conditions

$$\int_{\mathcal{D}} \mathbf{K}_{(c)}(\mathbf{u}) d\mathbf{u} = \mathbf{1}, \int_{\mathcal{D}} \mathbf{u} \mathbf{K}_{(c)}(\mathbf{u}) d\mathbf{u} = \mathbf{0}, \int_{\mathcal{D}} \mathbf{u} \mathbf{u}^{\top} \mathbf{K}_{(c)}(\mathbf{u}) d\mathbf{u} \neq \mathbf{0}, \quad (56)$$

throughout the support of the curve. Note that by using the standard approximation (see Section 3.3 in Gulati and Padgett (1996))

$$\frac{W\{(Y_i - y)b^{-1}\}\delta_i}{1 - \hat{G}(Y_i)} = w_b(Y_i - y)\hat{F}_T(Y_i)\{1 + o_p(1)\}, i = 1, \dots, n,$$

where $w_b(\cdot) = b^{-1}w(\cdot/b)$, $\hat{F}_L(y|\mathbf{x})$ can be equivalently written also as

$$\hat{F}_L(y|\mathbf{x}) = n^{-1}|\mathbf{H}|^{-1}f_{\mathbf{X}}^{-1}(\mathbf{x}) \sum_{i=1}^n \mathbf{K}_{(c)}\{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} w_b(Y_i - y)\hat{F}_T(Y_i)\{1 + o_p(1)\}, \quad (57)$$

which is useful for studying the estimator's statistical properties. Note that specifically for $d = 1$

$$\begin{aligned} \hat{F}_L(y|\mathbf{x}) &= n^{-1}e_1^{\top} \begin{pmatrix} S_{n,0}(\mathbf{x}) & S_{n,1}(\mathbf{x}) \\ S_{n,1}(\mathbf{x}) & S_{n,2}(\mathbf{x}) \end{pmatrix}^{-1} \begin{pmatrix} 1 & \mathbf{X}_1 - \mathbf{x} \\ \vdots & \vdots \\ 1 & \mathbf{X}_n - \mathbf{x} \end{pmatrix}^{\top} \begin{pmatrix} \mathbf{K}\{\mathbf{H}^{-1}(\mathbf{X}_1 - \mathbf{x})\} & 0 \\ & \ddots \\ 0 & \mathbf{K}\{\mathbf{H}^{-1}(\mathbf{X}_n - \mathbf{x})\} \end{pmatrix} \begin{pmatrix} \frac{W\{(Y_1 - y)b^{-1}\}\delta_1}{1 - \hat{G}(Y_1)} \\ \vdots \\ \frac{W\{(Y_n - y)b^{-1}\}\delta_n}{1 - \hat{G}(Y_n)} \end{pmatrix} \\ &= n^{-1} \sum_{i=1}^n \frac{S_{n,2}(\mathbf{x}) - S_{n,1}^{\top}(\mathbf{x})(\mathbf{X}_i - \mathbf{x})}{S_{n,2}(\mathbf{x})S_{n,0}(\mathbf{x}) - S_{n,1}^{\top}(\mathbf{x})S_{n,1}(\mathbf{x})} \mathbf{K}\{\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\} \frac{W\{(Y_i - y)b^{-1}\}\delta_i}{1 - \hat{G}(Y_i)}, \end{aligned}$$

which is the estimator discussed in Kim et al. (2010).

It is worth noting here that an important consequence of taking $S = \text{supp}(f_{\mathbf{X}})$ is that S is compact, closed and bounded with smooth boundary, satisfying $\text{cl}(S_{\text{int}}) = S_{\text{int}}$, where $\text{cl}(\cdot)$ is the closure of the set $\cdot$ and

$$S_{\text{int}} = [J_1 + h_1, M_1 - h_1] \times \dots \times [J_d + h_d, M_d - h_d].$$

Then, since for any p.d.f. $f_{\mathbf{X}}(\mathbf{x}) \geq 0$ on S it follows that a discontinuity will occur across the boundary of S , denoted by ∂S , while $f_{\mathbf{X}}$ is smooth on S .

Since $T = [-1, 1]^d$ is a compact and simply connected set which satisfies $\text{cl}(T_{\text{int}}) = T$, then the *effective kernel support* or *smoothing window* of $\mathbf{K}$ is $T_n(\mathbf{x}) \cap S$, where $T_n(\mathbf{x}) = \mathbf{x} - \mathbf{H}T$. The assumptions on S and the properties of T ensure that $T_n(\mathbf{x}) \cap S$ is measurable and has a positive measure for all $\mathbf{x} \in S$. Now, let $\mathbf{x} \in S$. Transposing and rescaling the support T of the kernel, leads to the set

$$T_{\mathbf{x}} = \mathbf{H}^{-1}(\mathbf{x} - T_n(\mathbf{x}) \cap S) = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z} = \mathbf{H}^{-1}(\mathbf{x} - \mathbf{y}), \mathbf{y} \in T_n(\mathbf{x}) \cap S\}.$$

Since only data falling in $T_n(\mathbf{x}) \cap S$ contribute to $\hat{F}_L(y|\mathbf{x})$, for $\mathbf{x} \in S_{\text{int}}$, it holds that $T_n(\mathbf{x}) \cap S = T_n(\mathbf{x}) \subset S$ which implies $T_{\mathbf{x}} = T$. Consequently, when $\mathbf{x} \in S_{\text{int}}$, the moment conditions of assumption A1 hold and the estimate behaves as expected. Note that the same is true asymptotically for all $\mathbf{x} \in S$ as by assumption A5, $T_{\mathbf{x}} = T$ for $n \rightarrow \infty$. However, from a finite sample standpoint, when $\mathbf{x}$ is in the boundary of S , i.e. for $\mathbf{x} \in S_{b,l}$ or $\mathbf{x} \in S_{b,r}$ where

$$S_{b,l} = [J_1, J_1 + h_1] \times \dots \times [J_d, J_d + h_d], \quad S_{b,r} = (M_1 - h_1, M_1] \times \dots \times (M_d - h_d, M_d],$$

then $T_n(\mathbf{x}) \cap S \neq T_n(\mathbf{x})$ or equivalently $T_n(\mathbf{x}) \cap S \not\subset S$ and hence part of the smoothing window is devoid of data. For example, in the case where $\mathbf{x} \in S_{b,l}$, the effective support of $\mathbf{K}$ is actually $T_{\mathbf{x}} = [-1, c_1] \times \dots \times [-1, c_d] \equiv \mathcal{D}$ with $0 < c_i < 1$ for all $i \in \{1, \dots, d\}$ and consequently the moment conditions for $\mathbf{K}$ in assumption A1 no longer hold. For the same reason, the effective kernel support of $\mathbf{K}$ is also asymmetric in the right boundary as $T_{\mathbf{x}} = [-c_1, 1] \times \dots \times [-c_d, 1]$ for $\mathbf{x} \in S_{b,r}$. Thus employing $\mathbf{K}$ without any modification for kernel estimation of $F_T(y|\mathbf{x})$ is expected to result in inflated bias at any $\mathbf{x} \in S_b = S_{b,l} \cup S_{b,r}$ in the usual asymptotic analysis. Focusing on the left boundary, as treatment of the right boundary is similar, the automatic boundary adjustment of $\hat{F}_L(y|\mathbf{x})$ is discussed next.

A.2 Bandwidth selection for a single covariate.

When $d = 1$, $\mathbf{H} \equiv h$ is a scalar and $\mathbf{K} = k$. Also $m_1 - m_5$ in (58) simplify (in an obvious manner) substantially; see A.5 for their analytic expressions. Substituting to (12) gives

$$\text{AMISE}\{\hat{F}_L(h, b)\} = h^4 m_1 + h^2 b^2 m_2 + b^4 m_3 - \frac{bm_4}{nh} + \frac{m_5}{nh}. \quad (58)$$

The first-order conditions required for minimization of (58) are

$$\frac{\partial \text{AMISE}\{\hat{F}_L(h, b)\}}{\partial h} = 4h^3 m_1 + 2hb^2 m_2 + \frac{bm_4}{nh^2} - \frac{m_5}{nh^2} \stackrel{\text{set}}{=} 0, \quad (59)$$

$$\frac{\partial \text{AMISE} \left\{ \hat{F}_L(h, b) \right\}}{\partial b} = 2h^2 b m_2 + 4b^3 m_3 - \frac{m_4}{nh} \stackrel{\text{set}}{=} 0. \quad (60)$$

Following Li and Racine (2008), three different cases which cover all possible relationships between b and h are distinguished. The first case is $b < h$, i.e. b and h are such that $b = o(h)$, which by definition implies that $\lim_{n \rightarrow \infty} b/h = 0$. Additionally note that $bm_4/(nh^2)$ is negligible asymptotically compared to $m_5/(nh^2)$. By (59), (60) and these two facts, the optimal values h and b are

$$h = \left(\frac{m_5}{2n(2m_1 + m_2)} \right)^{1/5}, \quad b = n^{-2/5} \left\{ \frac{2(2m_1 + m_2)}{m_5} \right\}^{3/5} \frac{m_4}{2m_2 + 4m_3}.$$

Thus the first case suggests $h \sim n^{-1/5}$, $b \sim n^{-2/5}$. This is rather expected and corroborates with the result of Li and Racine (2008) which also suggest in the y direction a bandwidth of an order smaller compared to the bandwidth in the x direction. The similarity in the results is a consequence of the fact that censoring leaves the MISE rate of convergence unaffected. Thus the optimal bandwidth will have the same order as in the uncensored case, but with different constants due to the effect of censoring on the estimate's variance.

Now, consider the case that h and b are such that $h = o(b)$, i.e. $\lim_{n \rightarrow \infty} h/b = 0$. By this relationship and (60),

$$h = \frac{m_4}{2nb^3(m_2 + 2m_3)}. \quad (61)$$

Combining (61) with (59), (60) and after some straightforward algebra

$$b = n^{-3/7} \left\{ \frac{m_4^3(2m_1 + m_2)}{4(m_2 + 2m_3)^3 m_5} \right\}^{1/7}.$$

Substituting back to (61) yields $h \sim n^{-1/7}$. This implies $b = o(h)$ which obviously contradicts the initial hypothesis $h = o(b)$. Thus the second option $h = o(b)$ is rejected. Finally we consider the case that h and b are of the same order, i.e. $\lim_{n \rightarrow \infty} h/b = c > 0$, which gives

$$4h^3 m_1 + 2ch^3 m_2 - \frac{m_5}{nh^2} = 0, \quad 2ch^3 m_2 + 4ch^3 m_3 - \frac{m_4}{nh} = 0.$$

The above system implies two different bandwidth orders for h ($n^{-1/5}$ and $n^{-1/4}$) which is infeasible in practice and therefore this option is also rejected. Summarizing, the only feasible option for b and h is $b = o(h)$ which from the above analysis yields the optimal values

$$h_{\text{opt}} = n^{-1/5} \left(\frac{m_5}{2(2m_1 + m_2)} \right)^{1/5}, \quad b_{\text{opt}} = n^{-2/5} \left\{ \frac{2(2m_1 + m_2)}{m_5} \right\}^{3/5} \frac{m_4}{2m_2 + 4m_3}.$$

Substituting $h_{\text{opt}} \equiv vn^{-1/5}$ and $b_{\text{opt}} \equiv sn^{-2/5}$ into (58) gives

$$\text{AMISE} \left\{ \hat{F}_L(h_{\text{opt}}, b_{\text{opt}}) \right\} = n^{-4/5} v^4 m_1 + n^{-2/5} n^{-4/5} v^2 s^2 m_2 + n^{-8/5} s^4 m_3 - \frac{sn^{-2/5} m_4}{nn^{-1/5} v} + \frac{m_5}{nn^{-1/5}} = O(n^{-4/5}).$$

That is, the optimal h_{opt} and b_{opt} values yield an AMISE of order $n^{-4/5}$. In comparison, the univariate kernel c.d.f. estimator uses optimal bandwidth which is of order $n^{-1/3}$ and optimal AMISE of order n^{-1} , see Bowman et al. (1998). As expected, the convergence properties are better in the univariate case, since we are effectively reducing the number of points used in the estimates when we condition on the value of X . In the present setting, it is intuitive to use a larger bandwidth (of order $n^{-2/5}$) so as to compensate the sparseness of data points due to the smoothing in both the y and the covariates direction.

A.3 Bandwidth selection for the multivariate kernel density estimator.

The AMISE optimal bandwidth selector of Duong and Hazelton (2003) is given by

$$\hat{\mathbf{H}}_l = \arg \min_{\mathbf{H}_l \in} \left\{ n^{-1} R(\mathbf{K}) |\mathbf{H}_l|^{-1} + \mu_2^2(\mathbf{K}) (\text{vech}^\top \mathbf{H}_l^2) \hat{\Psi}_4 (\text{vech}^\top \mathbf{H}_l^2) \right\},$$

where $\hat{\Psi}_4$ is a $d(d+1)/2 \times d(d+1)/2$ matrix, having as elements the estimates of the functionals

$$\psi_{r_1 + \dots + r_d} = \int \left\{ \frac{\partial^{r_1 + \dots + r_d}}{\partial^{r_1} x_1 \partial^{r_2} x_2 \dots \partial^{r_d} x_d} f(\mathbf{x}) \right\} f(\mathbf{x}) d\mathbf{x},$$

where $r_1 + \dots + r_d = 4$. A sensible estimate, say $\hat{\psi}_{r_1 + \dots + r_d}(\mathbf{G})$ of $\psi_{r_1 + \dots + r_d}$ is

$$\begin{aligned} \hat{\psi}_{r_1 + \dots + r_d}(\mathbf{G}) &= n^{-1} \int \left\{ \frac{\partial^{r_1 + \dots + r_d}}{\partial^{r_1} x_1 \partial^{r_2} x_2 \dots \partial^{r_d} x_d} \sum_{i=1}^n \mathbf{K}_{\mathbf{G}}^*(\mathbf{X}_i - \mathbf{x}) \right\} f(\mathbf{x}) d\mathbf{x} \\ &= n^{-2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{K}_{\mathbf{G}}^{*(r_1 + \dots + r_d)}(\mathbf{X}_i - \mathbf{X}_j), \end{aligned} \quad (62)$$

and $\mathbf{G}$ is a pilot bandwidth matrix, usually taken to be of the form $\mathbf{G} = g\mathbf{I}_d$. The data-driven counterpart of g , say $\hat{g}$ is the minimizer of the sum of AMSE criterion

$$\hat{g} = \arg \min_g \sum_{r_1 + \dots + r_d = 4} \text{AMSE} \left\{ \hat{\psi}_{r_1 + \dots + r_d}(g) \right\}.$$

A.4 Derivation of (12).

From (10) and (11) we have

$$\begin{aligned} \text{AMISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} &= \mu_2^2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left[\frac{\text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\}}{f_{\mathbf{X}}(\mathbf{x})} + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{tr} \left\{ (\mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H}) + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{tr} \left\{ (\mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H}) \right\} \right]^2 dy d\mathbf{x} \\ &\quad + \int_{\varepsilon_{y,o}} \left\{ \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\}^2 dy d\mathbf{x} + 2\mu_2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left[\frac{\text{tr} \left\{ \mathbf{H} \dot{f}_{\mathbf{X}}(\mathbf{x}) \right\} \text{tr} \left\{ \mathbf{H} \dot{F}_T(y|\mathbf{x}) \right\}}{f_{\mathbf{X}}(\mathbf{x})} \right. \\ &\quad \left. + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \text{tr} \left\{ (\mathbf{H} \ddot{F}_T(y|\mathbf{x}) \mathbf{H}) + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \text{tr} \left\{ (\mathbf{H} \ddot{f}_{\mathbf{X}}(\mathbf{x}) \mathbf{H}) \right\} \right] \left\{ \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\} dy d\mathbf{x} \\ &\quad + \frac{R(\mathbf{K})}{n|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x})} \int_{\varepsilon_{y,o}} [P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1} - F_T^2(y|\mathbf{x})] dy d\mathbf{x}. \end{aligned}$$

Substitute (21) in the above AMISE formula and rearranging gives

$$\begin{aligned} \text{AMISE} \left\{ \hat{F}_L(\mathbf{H}, b) \right\} &= \mu_2^2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left[\frac{1}{f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \sum_{j=1}^d h_i h_j \frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial x_i} \frac{\partial F_T(y|\mathbf{x})}{\partial x_j} \right. \\ &\quad \left. + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \sum_{i=1}^d h_i^2 \frac{\partial^2 F_T(y|\mathbf{x})}{\partial x_i^2} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d h_i^2 \frac{\partial^2 f_{\mathbf{X}}(\mathbf{x})}{\partial x_i^2} \right]^2 dy d\mathbf{x} \\ &\quad + 2\mu_2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left[\frac{1}{f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \sum_{j=1}^d h_i h_j \frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial x_i} \frac{\partial F_T(y|\mathbf{x})}{\partial x_j} \right. \\ &\quad \left. + \frac{f_{\mathbf{X}}(\mathbf{x})}{2} \sum_{i=1}^d h_i^2 \frac{\partial^2 F_T(y|\mathbf{x})}{\partial x_i^2} + \frac{F_T(y|\mathbf{x})}{2f_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d h_i^2 \frac{\partial^2 f_{\mathbf{X}}(\mathbf{x})}{\partial x_i^2} \right] \left\{ \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\} dy d\mathbf{x} \\ &\quad + \int_{\varepsilon_{y,o}} \left\{ \frac{b^2 \mu_2(w)}{2} \frac{\partial^2 F_T(y|\mathbf{x})}{\partial y^2} \right\}^2 dy d\mathbf{x} + \frac{R(\mathbf{K})f_{\mathbf{X}}^{-1}(\mathbf{x})}{n \left(\prod_{i=1}^d h_i \right)} \int_{\varepsilon_{y,o}} [P(y|\mathbf{x}) - 2bp(y|\mathbf{x})A_{1,1} - F_T^2(y|\mathbf{x})] dy d\mathbf{x}. \end{aligned} \quad (63)$$

Now, using

$$(\text{vech}^\top \mathbf{H}^2)^2 = (\text{vech}^\top \mathbf{H}^2)(\text{vech}^\top \mathbf{H}^2) = \text{tr} [\mathbf{H}^2 \mathbf{H}^2] = \left(\sum_{i=1}^d h_i^2 \right)^2,$$

substituting to (63) and rearranging yields (12).

A.5 Closed form expressions for $m_1 - m_5$ for $d = 1$ and for $\hat{m}_1 - \hat{m}_5$ for $d > 1$.

For a single covariate the closed form expressions for $m_1 - m_5$ in (58) simplify to

$$\begin{aligned} m_1 &= \mu_2^2(k) \int_{\varepsilon_{y,o}} \left[\frac{f'_X(x)}{f_X(x)} \frac{\partial F_T(y|x)}{\partial x} + \frac{f_X(x)}{2} \frac{\partial^2 F_T(y|x)}{\partial x^2} + \frac{F_T(y|x)}{2f_X(x)} f''_X(x) \right]^2 dy, m_2 = m_1 m_3, \\ m_3 &= \frac{\mu_2^2(w)}{4} \int_{\varepsilon_{y,o}} \left\{ \frac{\partial^2 F_T(y|x)}{\partial y^2} \right\}^2 dy, m_4 = 2R(k)A_{1,1}f_X(x)^{-1} \int_{\varepsilon_{y,o}} p(y|x) dy d\mathbf{x} = 2R(k)A_{1,1}f_X(x)^{-1} \int_{\varepsilon_{y,o}} \frac{f_T(y|x)}{1-G(y)} dy, \\ m_5 &= \frac{R(k)}{f_X(x)} \int_{\varepsilon_{y,o}} \{P(y|x) - F_T^2(y|x)\}^2 dy. \end{aligned}$$

For more than one covariates (i.e. for $d > 1$), the closed form expressions of the consistent data-driven estimates of $m_1 - m_5$, say $\hat{m}_1 - \hat{m}_5$ are

$$\begin{aligned} \hat{m}_1 &= \mu_2^2(\mathbf{K}) \int_{\varepsilon_{y,o}} \left[\frac{1}{\hat{f}_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial \hat{f}_{\mathbf{X}}(\mathbf{x})}{\partial x_i} \frac{\partial \hat{F}_T(y|\mathbf{x})}{\partial x_j} + \frac{\hat{f}_{\mathbf{X}}(\mathbf{x})}{2} \sum_{i=1}^d \frac{\partial^2 \hat{F}_T(y|\mathbf{x})}{\partial x_i^2} + \frac{\hat{F}_T(y|\mathbf{x})}{2\hat{f}_{\mathbf{X}}(\mathbf{x})} \sum_{i=1}^d \frac{\partial^2 \hat{f}_{\mathbf{X}}(\mathbf{x})}{\partial x_i^2} \right]^2 dy, \hat{m}_2 = \hat{m}_1 \hat{m}_3 \\ \hat{m}_3 &= \frac{\mu_2^2(w)}{4} \int_{\varepsilon_{y,o}} \left\{ \frac{\partial^2 \hat{F}_T(y|\mathbf{x})}{\partial y^2} \right\}^2 dy, \hat{m}_4 = 2R(\mathbf{K})A_{1,1}\hat{f}_{\mathbf{X}}^{-1}(\mathbf{x}) \int_{\varepsilon_{y,o}} \hat{p}(y|\mathbf{x}) dy d\mathbf{x} = 2R(\mathbf{K})A_{1,1}\hat{f}_{\mathbf{X}}^{-1}(\mathbf{x}) \int_{\varepsilon_{y,o}} \frac{\hat{f}_T(y|\mathbf{x})}{1-\hat{G}(y)} dy, \\ \hat{m}_5 &= \frac{R(\mathbf{K})}{\hat{f}_{\mathbf{X}}(\mathbf{x})} \int_{\varepsilon_{y,o}} \left\{ \hat{P}(y|\mathbf{x}) - \hat{F}_T^2(y|\mathbf{x}) \right\}^2 dy. \end{aligned}$$

A.6 Implementation details.

In all implementations throughout this section the number of covariates is fixed to $d = 2$. Additionally, the bivariate normal density is used in place of the kernel function in the covariates dimension. That is, for fixed i , for $\mathbf{X}_i = (X_{1i}, X_{2i})$, $\mathbf{x} = (x_1, x_2)$

and $\mathbf{K}(\cdot) = \phi(\cdot)\phi(\cdot)$,

$$\mathbf{K}^* \{ \mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}) \} = \frac{S_{n,2}(x_1, x_2) - S_{n,1}^\top(x_1, x_2) \begin{pmatrix} X_{i1} - x_1 \\ X_{i2} - x_2 \end{pmatrix}}{S_{n,2}(x_1, x_2)S_{n,0}(x_1, x_2) - S_{n,1}^\top(x_1, x_2)S_{n,1}(x_1, x_2)} \phi\left(\frac{X_{i1} - x_1}{h_1}\right) \phi\left(\frac{X_{i2} - x_2}{h_2}\right).$$

Then the classical local linear bivariate density estimate is

$$\hat{f}_{X_1, X_2}(x_1, x_2) = (n\hat{h}_1\hat{h}_2)^{-1} \sum_{i=1}^n \mathbf{K}^* \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \},$$

where $\hat{\mathbf{H}}_l = \text{diag}(\hat{h}_1, \hat{h}_2)$ and $\hat{h}_1, \hat{h}_2$ is the plug-in AMISE optimal bandwidth of Duong and Hazelton (2003), given by

$$(\hat{h}_1, \hat{h}_2) = \arg \min_{h_1, h_2} \text{AMISE}\{\hat{f}_{X_1, X_2}(x_1, x_2)\} = \arg \min_{h_1, h_2} \left\{ n^{-1} R(\mathbf{K})(h_1 h_2)^{-1} + \mu_2^2(\mathbf{K})(\text{vech}^\top \mathbf{H}_l^2) \hat{\Psi}_4 (\text{vech}^\top \mathbf{H}_l^2) \right\},$$

where, with $r_1, r_2 = 0, 1, 2, 3, 4$ and such that $r_1 + r_2 = 4$,

$$\hat{\Psi}_4 = \begin{pmatrix} \hat{\psi}_{40} & 2\hat{\psi}_{31} & \hat{\psi}_{22} \\ 2\hat{\psi}_{31} & 4\hat{\psi}_{22} & 2\hat{\psi}_{13} \\ \hat{\psi}_{22} & 2\hat{\psi}_{13} & \hat{\psi}_{40} \end{pmatrix}.$$

In accordance to (62), $\hat{\psi}_{r_1 r_2}$ is implemented by

$$\hat{\psi}_{r_1 r_2}(\hat{\mathbf{G}}) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{K}_{\hat{\mathbf{G}}}^{*(r_1+r_2)}(\mathbf{X}_i - \mathbf{X}_j),$$

where $\hat{\mathbf{G}} = \hat{g}\mathbf{I}_2$ with $\hat{g}$ being the minimizer of the sum of AMSE criterion

$$\hat{g} = \arg \min_g \sum_{r_1+r_2=4} \text{AMSE} \{ \hat{\psi}_{r_1+r_2}(g) \}.$$

Similarly, the classical univariate local linear estimate of the density of the X_i covariate is

$$\hat{f}_{X_i}(x_i) = (n\hat{h}_i)^{-1} \sum_{j=1}^n \frac{S_{n,2}(x) - S_{n,1}(x)(X_j - x)}{S_{n,2}(x)S_{n,0}(x) - S_{n,1}(x)S_{n,1}(x)} \phi\left(\frac{X_j - x}{\hat{h}_i}\right), \quad i = 1, \dots, d,$$

and is implemented with AMISE optimal plug-in bandwidth $\hat{h}_i$, provided via the R function `bw.SJ`. Note that $\hat{f}_{X_1, X_2}(x_1, x_2), \hat{f}_{X_i}(x)$, including the bandwidths $\hat{h}_1, \hat{h}_2$ and all associated functions is readily available in the `Hpi.diag` function of the R package `ks`. From (17) for $d = 2$ the closed form expression for $\hat{f}_T(y|x_1, x_2)$ is

$$\hat{f}_T(y|x_1, x_2) = \frac{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \} \{1 - \hat{G}(Y_i)\}^{-1} w_b'(Y_i - y) \delta_i}{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \}}.$$

Then, the analytic expression for $\hat{F}_T(y|\mathbf{x})$ is

$$\hat{F}_T(y|x_1, x_2) = \frac{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \} \{1 - \hat{G}(Y_i)\}^{-1} W\{(Y_i - y)\tilde{b}^{-1}\} \delta_i}{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \}}.$$

Also

$$\frac{\partial^2 \hat{F}_T(y|x_1, x_2)}{\partial y^2} = \tilde{b}^{-1} \frac{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \} \{1 - \hat{G}(Y_i)\}^{-1} w_b''(Y_i - y) \delta_i}{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \}}.$$

Further, $\dot{\hat{F}}_T(y|\mathbf{x})$ and $\hat{\mathcal{H}}_{F_T}(y|\mathbf{x})$ (defined in (18)) are obtained by the functions `grad`, `jacobian` and `hessian` in the R package `numDeriv`. Finally, $\hat{p}(y|x_1, x_2) = \hat{f}_T(y|x_1, x_2)(1 - \hat{G}(y))^{-1}$, and

$$\hat{P}(x_1, x_2) = \frac{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \} \{1 - \hat{G}(Y_i)\}^{-2} [W\{(Y_i - T_Y)\tilde{b}^{-1}\} - W(Y_i\tilde{b}^{-1})] \delta_i}{\sum_{i=1}^n \mathbf{K} \{ \hat{\mathbf{H}}_l^{-1}(\mathbf{X}_i - \mathbf{x}) \}}.$$

Straightforward calculations yield $\mu_2(\mathbf{K}) = 1$ and $R(\mathbf{K}) = (4\pi)^{-1}$. Further, the Epanechnikov kernel is used in the y dimension, thus $\mu_2(w) = 1/5$. Then, specifically for $d = 2$,

$$\hat{\mathbf{H}}_{\text{opt}} = \begin{pmatrix} \hat{h}_{\text{opt},1} & 0 \\ 0 & \hat{h}_{\text{opt},2} \end{pmatrix}, \quad \hat{h}_{\text{opt},i} = \left(\frac{\hat{m}_{5,i}}{2n(2\hat{m}_{1,i} + \hat{m}_{2,i})} \right)^{1/6}, \quad i = 1, 2,$$

where $\hat{m}_{2,i} = \hat{m}_{1,i}\hat{m}_{3,i}$ and

$$\hat{m}_{1,i} = \int_{\varepsilon_{y,o}} \left[\frac{1}{\hat{f}_{X_i}(x_i)} \frac{\partial \hat{f}_{X_i}(x_i)}{\partial x_i} \frac{\partial \hat{F}_T(y|x_i)}{\partial x_i} + \frac{\hat{f}_{X_i}(x_i)}{2} \frac{\partial^2 \hat{F}_T(y|x_i)}{\partial x_i^2} + \frac{\hat{F}_T(y|x_i)}{2\hat{f}_{X_i}(x_i)} \frac{\partial^2 \hat{f}_{X_i}(x_i)}{\partial x_i^2} \right]^2 dy,$$

Table 2: Approximate MISE's of $\hat{S}_{j,CV}, \hat{S}_{j,p}, j = 1, 2, 3$ using Cross Validation optimal bandwidth.

Cens.	n	$S_1(y \mathbf{x} = (0.3, 0.6))$			$S_2(y x_1 = 0.3)$			$S_3(y x_2 = 0.6)$		
		$\hat{S}_{1,CV}$	$\hat{S}_{1,p}$	% diff.	$\hat{S}_{2,CV}$	$\hat{S}_{2,p}$	% diff.	$\hat{S}_{3,CV}$	$\hat{S}_{3,p}$	% diff.
15%	50	0.783	0.616	21.30%	0.486	0.375	22.83%	0.668	0.533	20.23%
	100	0.439	0.357	18.69%	0.253	0.208	17.79%	0.396	0.326	17.74%
	150	0.342	0.295	13.74%	0.171	0.149	12.94%	0.289	0.251	13.16%
	250	0.236	0.21	11.01%	0.119	0.106	10.89%	0.187	0.168	10.25%
30%	50	0.874	0.703	19.64%	0.478	0.377	21.12%	0.763	0.616	19.37%
	100	0.515	0.422	18.11%	0.353	0.291	17.64%	0.578	0.485	16.13%
	150	0.424	0.361	14.89%	0.249	0.226	9.63%	0.436	0.384	12.04%
	250	0.361	0.315	12.72%	0.211	0.189	10.42%	0.365	0.325	11.21%
50%	50	0.936	0.758	19.11%	0.534	0.417	21.95%	0.971	0.773	20.42%
	100	0.718	0.583	18.85%	0.414	0.345	16.72%	0.845	0.713	16.13%
	150	0.461	0.404	12.37%	0.352	0.305	13.43%	0.607	0.516	14.86%
	250	0.465	0.417	10.32%	0.278	0.247	11.22%	0.514	0.457	11.13%

$$\hat{m}_{5,i} = \frac{1}{4\pi\hat{f}_{X_i}(x_i)} \int_{\varepsilon_{y,o}} \left\{ \hat{P}(y|x_i) - \hat{F}_T^2(y|x_i) \right\}^2 dy,$$

for $i = 1, 2$, where

$$\hat{P}(y|x_i) = \int_0^{T_Y} \frac{\hat{f}_T(y|x_i)}{1 - \hat{G}(y)} dy, \hat{F}_T(y|x_i) = \int_0^y \hat{f}_T(z|x_i) dz.$$

Also

$$\hat{b}_{\text{opt}} = \left(\frac{\hat{m}_4}{2n(\hat{m}_2 + 2\hat{m}_3)} \right)^{2/7},$$

with

$$\begin{aligned} \hat{m}_2 &= \frac{1}{5} \int_{\varepsilon_{y,o}} \left[\frac{1}{\hat{f}_{X_1, X_2}(x_1, x_2)} \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial \hat{f}_{X_1, X_2}(x_1, x_2)}{\partial x_i} \frac{\partial \hat{F}_T(y|x_1, x_2)}{\partial x_j} + \frac{\hat{f}_{X_1, X_2}(x_1, x_2)}{2} \sum_{i=1}^2 \frac{\partial^2 \hat{F}_T(y|x_1, x_2)}{\partial x_i^2} \right. \\ &\quad \left. + \frac{\hat{F}_T(y|x_1, x_2)}{2\hat{f}_{X_1, X_2}(x_1, x_2)} \sum_{i=1}^2 \frac{\partial^2 \hat{f}_{X_1, X_2}(x_1, x_2)}{\partial x_i^2} \right] \frac{\partial^2 \hat{F}_T(y|x_1, x_2)}{\partial y^2} dy, \\ \hat{m}_3 &= \frac{1}{100} \int_{\varepsilon_{y,o}} \left\{ \frac{\partial^2 \hat{F}_T(y|x_1, x_2)}{\partial y^2} \right\}^2 dy, \hat{m}_4 = \frac{1}{2\pi} A_{1,1} \hat{f}_{X_1, X_2}^{-1}(x_1, x_2) \int_{\varepsilon_{y,o}} \frac{\hat{f}_T(y|x_1, x_2)}{1 - \hat{G}(y)} dy. \end{aligned}$$

B Additional numerical examples.

In addition to the comparison of the ideal and the data driven plug-in bandwidths presented in Table 1 of Section 4, Table 2 of the present section contains the results of an indicative comparison on the performance of estimator $\hat{F}_L$ when using the proposed (data driven) plug-in bandwidths $\hat{\mathbf{H}}_{\text{opt}}, \hat{b}_{\text{opt}}$ and when using Cross Validation (CV) based bandwidths. The CV criterion utilized is

$$CV(\mathbf{H}, b) = \sum_{i=1}^n \sum_{j=1}^n \left\{ \frac{W \{ (Y_j - y)b^{-1} \} \delta_j}{1 - \hat{G}(Y_j)} - (1 - \hat{F}_{L,-i}(Y_j|\mathbf{x})) \right\}^2,$$

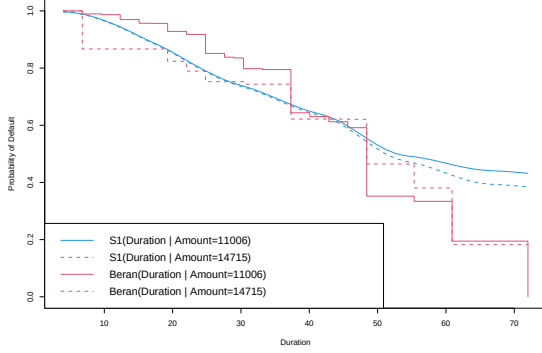
where $\hat{F}_{L,-i}$ is the leave-one-out version of $\hat{F}_L$ calculated by (7) after removing the i th row of the vector $\mathbf{e}_1^\top$, the matrix $\mathbf{X}_x$ and the vector $\mathbf{Y}$. The resulting Cross-Validation bandwidths, denoted by $\mathbf{H}_{CV}$ and b_{CV} are

$$(\mathbf{H}_{CV}, b_{CV}) = \arg \min_{\mathbf{H}, b} CV(\mathbf{H}, b).$$

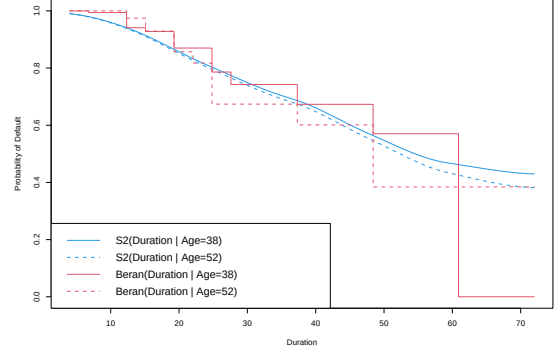
The MISE simulation is performed under exactly the same settings as the simulation study of Section 4. $\hat{S}_{i,CV}, i = 1, 2, 3$ denotes estimator $1 - \hat{F}_L$ implemented with the CV bandwidth while the MISE figures for $\hat{S}_{i,p}, i = 1, 2, 3$ are identical those in Table 1. The results indicate that the plug-in approach offers a precision gain over CV bandwidths which depending on the sample size and amount of censoring ranges from 10% - 20%.

Next, the practical benefits of $1 - \hat{F}_L$ over the routinely employed in practice and in the literature Beran's (Beran 1981; Cao et al. 2009) survival function estimator are illustrated in the analysis of the German Credit data set. For a single covariate vector X , Beran's estimator is defined by

$$\begin{aligned} \hat{S}_B(y|h) &= \prod_{i=1}^n \left(1 - \frac{\mathbf{1}_{\{Y_i \leq t, \delta_i=1\}} w_{i,n}(x)}{1 - \sum_{j=1}^n \mathbf{1}_{\{Y_j \leq Y_i\}} w_{j,n}(x)} \right), \\ w_{i,n}(x) &= \frac{K \{(x - X_i)/h\}}{\sum_{j=1}^n K(x - X_j)/h}, i = 1, \dots, n. \end{aligned}$$



Estimated probability of $\text{Duration} \geq y$ given
Amount = 11006, 14715.



Estimated probability of the Duration to exceed
a certain time point given Age = 38, 52.

Fig. 2: Estimators $1 - \hat{F}_L(y|x)$ and $\hat{S}_B(y|h)$ for the German Credit data set.

Estimator $\hat{S}_B(y|h)$ is implemented in the R package **npcore**, see López-Cheda et al. (2017a,b) and uses Cross-Validation optimal bandwidth, calculated as the largest local minimizer of the leave-one-out cross-validated criterion

$$CV(h) = \sum_{i=1}^n \sum_{j=1}^n \Delta_{i,j} \left\{ \mathbf{1}_{\{Y_i \leq Y_j\}} - (1 - \hat{S}_{B,-i}(Y_j|h)) \right\}^2.$$

In the $CV(h)$ above $\hat{S}_{B,-i}(Y_j|h)$ is the leave-one-out version of $\hat{S}_B(y|h)$ and $\Delta_{i,j} = 1$ if the pair (Y_i, Y_j) is ‘useful’ in the sense of Geerdens et al. (2018) and 0 otherwise. Both $1 - \hat{F}_L$ and $\hat{S}_B(y|h)$ are used to approximate the probability that a loan duration will exceed a certain time point, given specific ‘Amount’ and ‘Age’ levels, i.e., approximate $P(\text{Duration} \geq y | \text{Amount} = x)$ for $x = 11006, 14715$ and $P(\text{Duration} \geq y | \text{Age} = x)$ for $x = 38, 52$. The results are displayed in Fig.2; Both figures suggest that the proposed methodology drastically decreases estimation variance and consequently offers increased accuracy in the estimation of the required probabilities. Further, it addresses the limitation of discontinuity that is intrinsic in $\hat{S}_B(y|h)$: for example according to $\hat{S}_B(y|h)$ the probability that the duration of a loan will exceed 49 months (for ‘Amount’ equal to 11006) can be either 0.38 or 0.62; this is a significant difference which can greatly affect the bank manager’s decision. The smoothing on the time coordinate direction of $\hat{F}_L$ effectively solves this problem and hence helps practitioners to avoid ambiguous decisions. However, the proposed estimator is computationally more demanding compared to $\hat{S}_B(y|h)$, especially in its bandwidth selection part. As a result, assuming a similar or greater sample size as the German credit data set, a limitation that is likely to be encountered in practice is that $\hat{F}_L$ will be computationally expensive for more than 4-5 covariates. Further, practitioners might require to incorporate discrete, ordinal or correlated (continuous) covariates. The extension of the proposed estimator to these settings is already under way.

C Computational speed.

As the proposed methodology is somewhat computer intensive, in this section we present an indicative analysis on the CPU time required for the implementation of estimator $\hat{F}_L(y|x)$ and the plug-in bandwidths $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$ as a function of the sample size and the dimension d of the covariate vector. Similarly to Section 4, $\hat{F}_L(y|x)$ is always implemented via (7). For each covariate (as well as for the time dimension) we consider 50 equispaced grid points, covering the whole range of the covariate’s support. The scalar bandwidth $\hat{b}_{\text{opt}}$ and the bandwidth matrix $\hat{\mathbf{H}}_{\text{opt}}$ are implemented as described in Section A.6.

Since computational time is greatly affected by the workstation’s processor, we consider two scenarios (with two sub-scenarios in each). In the first scenario the benchmark study is implemented with a rather slow processor, an Intel Core i5–9400 CPU at 2.90GHz. The second scenario replicates the exact same process on a much faster Intel Core i7–14700F processor at 3GHz. In both situations the amount of memory installed in the corresponding workstations is 16GB of RAM, they both run Windows 11 professional and the R version is 4.2 64bit. All times reported are in seconds and were obtained by timing the calculation procedure with the `tic()` and `toc()` commands of the R package `tic`. Further, for each CPU type, we consider two sub-scenarios: one is implementation of both $\hat{F}_L(y|x)$ and the bandwidths $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$. The second sub-scenario considers only the implementation of $\hat{F}_L(y|x)$, with the bandwidths $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$ calculated separately and using their values as user input in the process. This is sought so as to also benchmark the computational time of the proposed bandwidth rule itself. Further, all CPU times are for calculation of one instance of the estimator (and bandwidths). The data generation time is not included in the CPU times reported.

The results for scenario 1 are reported on Tables 3 and 4. As expected the computational time increases drastically as the number of covariates increase. Further, as it is evident from a comparison between the two tables, on average, approximately 75% of the computational time is consumed by the bandwidth selection process.

The corresponding results for scenario 2 are reported on Tables 5 and 6. As in scenario 1, computational time increases drastically as the number of covariates increase and on average, approximately 75% of the computational time is due to bandwidth selection. However, the effect of the faster CPU is quite visible here as the overall CPU time dropped by more than a half (approximately 60% for both scenarios combined) compared to the CPU time of the i5 processor.

Table 3: CPU times(sec) for $\hat{F}_L(y|\mathbf{x})$, implemented with $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$ on an Intel i5-9400 processor.

n / # cov.	1 covariate	2 covariates	3 covariates
$n = 50$	0.548	26.221	1311.001
$n = 100$	0.553	28.058	1404.126
$n = 150$	0.568	30.528	1526.364
$n = 250$	0.741	32.859	1642.913

Table 4: CPU times(sec) for $\hat{F}_L(y|\mathbf{x})$, implemented with fixed bandwidths on an Intel i5-9400 processor.

n / # cov.	1 covariate	2 covariates	3 covariates
$n = 50$	0.164	6.818	235.981
$n = 100$	0.141	6.659	224.380
$n = 150$	0.156	7.467	411.814
$n = 250$	0.206	8.134	333.840

Table 5: CPU times(sec) for $\hat{F}_L(y|\mathbf{x})$, implemented with $\hat{\mathbf{H}}_{\text{opt}}$ and $\hat{b}_{\text{opt}}$ on an Intel i7-14700F processor.

n / # cov.	1 covariate	2 covariates	3 covariates
$n = 50$	0.231	11.411	565.172
$n = 100$	0.244	13.008	634.805
$n = 150$	0.245	13.713	674.042
$n = 250$	0.323	14.99	734.136

Table 6: CPU times(sec) for $\hat{F}_L(y|\mathbf{x})$, implemented with fixed bandwidths on an Intel i7-14700F processor.

n / # cov.	1 covariate	2 covariates	3 covariates
$n = 50$	0.069	2.967	101.731
$n = 100$	0.062	3.087	101.442
$n = 150$	0.067	3.354	181.857
$n = 250$	0.093	3.711	149.176