# NONPARAMETRIC KERNEL REGRESSION FOR MULTINOMIAL DATA

HIDENORI OKUMURA[a] and KANTA NAITO[b]

[a] *Department of Information Science and Business Management, Chugoku Junior College, Okayama 701-0197, Japan [E-mail: okumura@cjc.ac.jp];*
[b] *Department of Mathematics, Shimane University, Matsue 690-8504, Japan*

**Abstract**

This paper presents a kernel smoothing method for multinomial regression. A class of estimators of the regression functions is constructed by minimizing a localized power-divergence measure. These estimators include the bandwidth and a single parameter originating in the power-divergence measure as smoothing parameters. An asymptotic theory for the estimators is developed and the bias-adjusted estimators are obtained. A data-based algorithm for selecting the smoothing parameters is also proposed. Simulation results and an application to a real data set reveal that the proposed algorithm works efficiently.

KEYWORDS: Nonparametric regression; Multinomial data; Kernel smoothing; Power-divergence measure.

## 1 Introduction

This paper is concerned with the smoothing problem for multinomial data. Suppose that at each covariate $x$, the joint distribution of the random vector $\mathbf{Y}(x) = (Y_1(x), \ldots, Y_r(x))^T$ is the multinomial distribution $\mathrm{MN}(p_1(x), \ldots, p_r(x); N(x))$, where for any $x$, $\sum_{j=1}^r p_j(x) = 1$ and $N(x)$ is a positive constant with $N(x) = \sum_{j=1}^r Y_j(x)$. The distribution of $\mathbf{Y}(x)$ given $x$ is expressed as

$$\Pr(\mathbf{Y}(x) = \mathbf{y}(x)) = N(x)! \prod_{j=1}^{r} \frac{p_j(x)^{y_j(x)}}{y_j(x)!},$$

where $\mathbf{y}(x) = (y_1(x), \ldots, y_r(x))^T$ is an observation of $\mathbf{Y}(x)$. If $r = 2$, then the above distribution represents a binomial regression model. In this paper, our problem is to estimate the regression functions $p_j(x)(j = 1, \ldots, r)$ that yield the probabilities of each $r$ category at covariate $x$.

We observe $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ir})^T \sim \mathrm{MN}(p_1(x_i), \ldots, p_r(x_i); N(x_i))$ independently at covariates $x_i(i = 1, \ldots, K)$, where $Y_{ij} = Y_j(x_i)(j = 1, \ldots, r)$. For simplicity, assume that all $x_i$ are equispaced. Without lack of generality, it can be assumed that $x_i = (i-1)/K(i = 1, \ldots, K)$. However, this assumption can be relaxed as discussed in Müller and Schmitt [11]. Put $\bar{Y}_{ij} = Y_{ij}/N_i(i = 1, \ldots, K)$, where $N_i = N(x_i) = \sum_{j=1}^r Y_{ij}$.

Multinomial regression is used in various fields, and recently an important and promising application of multinomial regression has been the multiple classification problem. This problem aims to determine $\arg\max_j p_j(x)$; see Albert and Chib [3], and Hastie et al. ([8], Chapter 4).

It should be noted that the setting in this paper is different from that of the so-called sparse multinomial data discussed in Simonoff [18], Aerts at al. [1] and Augustyns and Wand [2]. In a sparse multinomial setting, the aim is to smooth the estimated probabilities of all categories for the situation where the number of categories increases with the sample size. Hence, it can be understood that $K = 1$ and $r \to \infty$ in our setting corresponds to a sparse multinomial setting. For such a situation, it is known that exploiting a smoothing method provides more accurate estimates than usual parametric estimates, especially for categories with a low probability. An elegant summary of tackling this issue by kernel smoothing is given in Simonoff [18].

Although our setting of the problem is different, we too utilize the kernel smoothing approach. For the same setting, Tutz [19] discussed the use of Nadaraya-Watson type estimators. The local likelihood approach was used to check the goodness of fit for the parametric model in Tutz and Kauermann [20] and Tutz [21]. In this study, we claim that the Nadaraya-Watson estimator and its variant have an advantage over the local likelihood approach. The advantage is that Nadaraya-Watson type estimators always exist; it is always expressed in an explicit form. However, the local likelihood approach does not always yield an estimator since the optimization steps sometimes cannot find a solution. This difficulty in the local likelihood approach was also pointed out in Okumura and Naito [15] in a binomial setting. Therefore, in this paper we propose the use of a more efficient estimator that is a variant of Nadaraya-Watson estimators of $p_j(x)(j = 1, \ldots, r)$.

To construct an estimator, it is important to choose a reasonable criterion to yield an estimator. Hence, it may be useful to refer to a goodness-of-fit test for multinomial data because our data is multinomial. The power-divergence measure discussed in detail in Cressie and Read [7] is famous as a measure of goodness of fit for multinomial data. The power-divergence measure is defined as

$$I_\lambda(\mathbf{p} : \mathbf{q}) = \frac{1}{\lambda(1-\lambda)} \sum_{j=1}^{r} q_j \left\{ \left( \frac{q_j}{p_j} \right)^\lambda - 1 \right\} \tag{1}$$

for $\lambda$ in $\Re$, where $\mathbf{p} = (p_1, ..., p_r)$ and $\mathbf{q} = (q_1, ..., q_r)$ are the probability distributions on $r$ categories; $I_0(\mathbf{p} : \mathbf{q}) \equiv \lim_{\lambda \to 0} I_\lambda(\mathbf{p} : \mathbf{q})$ and $I_{-1}(\mathbf{p} : \mathbf{q}) \equiv \lim_{\lambda \to -1} I_\lambda(\mathbf{p} : \mathbf{q})$. For the observed frequency vector $\mathbf{X}$ and the expected frequency vector $\mathbf{E}$, the family $\{2I_\lambda(\mathbf{X} : \mathbf{E})\}$ includes widely known goodness-of-fit statistics; see Cressie and Read [7]. Furthermore, if we put $\lambda = (\alpha - 1)/2$, then $I_\lambda$ is nothing else than the $\alpha$-divergence discussed in Amari and Nagaoka [4]. In this manner, the power-divergence includes many efficient statistics, and hence, a unified argument can be made using this divergence, as explained in Cressie and Read [7]. In this paper, we construct a family of estimators of $p_j(x)(j = 1, \ldots, r)$ including the Nadaraya-Watson estimator as a special case by minimizing a *localized* version of the power-divergence measure.

This paper is organized as follows. In Section 2, a family of kernel estimators is derived by means of a localized version of the power-divergence measure. The role of a new smoothing parameter $\lambda$, which is essentially included in (1), is also depicted. The theoretical performance of estimators is investigated in Section 3. Furthemore, a bias-adjusted estimator is naturally

obtained in Section 3. In Section 4, a method is developed for a data-based choice of smoothing parameters. We claim that $\lambda$ as well as bandwidth should be selected based on the data. A data-based choice of $\lambda$ was not discussed in Cressie and Read [7]; therefore, this is a relatively new argument in this research area. Simulation results are reported in Section 5, using which we confirm a good performance of the algorithm developed in Section 4. Applications to a real data set are discussed in Section 6, and final comments and notes are enumerated in Section 7. Outlines of the proofs for theoretical results are included in the Appendix.

# 2 Kernel Estimator

## 2.1 Criterion

In order to construct estimators, we focus on the power-divergence measure with the parameter $\lambda$ provided in Cressie and Read [7]. Essentially, we consider a localized version of the power divergence given by (1). For any fixed covariate $x$, the criterion is defined as

$$L_\lambda(\beta,\gamma) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{K}\phi_h(x_i-x)\left[\sum_{j=1}^{r}\bar{Y}_{ij}\left\{\left(\frac{\bar{Y}_{ij}}{\beta_j}\right)^{\lambda}-1\right\}+\gamma(1-\sum_{j=1}^{r}\beta_j)\right], \tag{2}$$

where $\beta=(\beta_1,\ldots,\beta_r)^T$ and $\gamma$ is a Lagrange multiplier. Further $\phi_h(\cdot)=\phi(\cdot/h)/h$, where $\phi$ is a kernel function with its support at $[-1,1]$. A few special cases are defined using continuity: $L_0(\beta,\gamma)\equiv\lim_{\lambda\to 0}L_\lambda(\beta,\gamma)$ and $L_{-1}(\beta,\gamma)\equiv\lim_{\lambda\to -1}L_\lambda(\beta,\gamma)$. For each $x$, let

$$(\hat{\beta},\hat{\gamma}) = \arg\min_{\beta,\gamma} L_\lambda(\beta,\gamma).$$

Then, the estimator of $p_j(x)$ is given as

$$\hat{p}_{j,\lambda}(x;h)=\hat{\beta}_j=\frac{\{\sum_{i=1}^{K}\phi_h(x_i-x)\bar{Y}_{ij}^{\lambda+1}\}^{\frac{1}{\lambda+1}}}{\sum_{\ell=1}^{r}\{\sum_{i=1}^{K}\phi_h(x_i-x)\bar{Y}_{i\ell}^{\lambda+1}\}^{\frac{1}{\lambda+1}}}.$$

Note that $0\le\hat{p}_{j,\lambda}(x;h)\le 1$ and $\sum_{j=1}^{r}\hat{p}_{j,\lambda}(x;h)=1$ for any $x$. If $\lambda=0$, the estimator is the Nadaraya-Watson estimator:

$$\hat{p}_{j,0}(x;h)=\frac{\sum_{i=1}^{K}\phi_h(x_i-x)\bar{Y}_{ij}}{\sum_{i=1}^{K}\phi_h(x_i-x)}. \tag{3}$$

## 2.2 Family of Estimators

Let $\hat{\mathbf{p}}_\lambda(\cdot;h)=(\hat{p}_{1,\lambda}(\cdot;h),\ldots,\hat{p}_{r,\lambda}(\cdot;h))^T$. Then we see that $\{\hat{\mathbf{p}}_\lambda(\cdot;h);\lambda\in\Re,h>0\}$ forms a family of estimators of $\mathbf{p}(\cdot)=(p_1(\cdot),\ldots,p_r(\cdot))^T$. Here, we attempt to explain the role of $\lambda$. Mathematically, introducing $\lambda$ is the same as increasing the dimension of the smoothing parameter included in the estimators by one. We can easily understand the state of a low-dimensional subspace from a higher dimensional space. By introducing $\lambda$, it appears to be possible, in some sence, to find a more efficient estimator than the Nadaraya-Watson estimator

(3), or at least, to evaluate the goodness of the Nadaraya-Watson estimator. Tackling the problem by increasing the dimension generally seems to be common in mathematical science. Recent works in statistical sciences using such an approach include Basu et al. [5], Jones et al. [9], and Naito [12, 13]. It will be shown in the subsequent discussion that introducing $\lambda$ by means of the localized power divergence is essential for the theoretical aspect of this paper.

In practical situations, $\lambda$ as well as $h$ should be selected based on the data. This data-based choice of $\lambda$ was not addressed in the parametric setting in Cressie and Read [7]. They recommended the use of $\lambda = 2/3$ based on several factors, but it was not a data-based choice. This paper addresses this problem in Section 4.

# 3 Theoretical Performance

## 3.1 Asymptotics for Estimators

Under the following regularity conditions, we can obtain the asymptotic bias, variance and normality of $\hat{p}_{j,\lambda}(x;h)$. The notations $\mu_k(f) = \int_{-1}^{1} x^k f(z)dz$ and $R(f) = \int_{-1}^{1} f(z)^2 dz$ for a function $f$ defined on $[-1,1]$ are used throughout.

**ASSUMPTION 1** $h \to 0$ and $N_i \to \infty$ as $K \to \infty$ for $i = 1, \ldots, K$ in a manner such that $Kh^{3+\varepsilon} = O(1)$ and $N_1 h^{2-\varepsilon} = O(1)$ for some $0 < \varepsilon < 1$, and $N_i/N_1 = 1 + o(h^2)$.

**ASSUMPTION 2** The support of the kernel $\phi(x)$ is $[-1,1]$ and $\phi(x)$ has continuous and bounded derivatives of order $n$ for any $x$ in $[-1,1]$ with $\phi^{(k)}(-1) = \phi^{(k)}(1) = 0$ and

$$(-1)^k \int_{-1}^{1} x^\ell \phi^{(k)}(x)dx = \begin{cases} 0, & \ell < k \text{ or } \ell = k+1, \\ \ell!, & \ell = k, \\ c_{\ell,k}, & \text{otherwise}, \end{cases}$$

where $0 \le k \le n$ and all $c_{\ell,k}$ are some positive constants.

**ASSUMPTION 3** The curves $p_j(x), j = 1, \ldots, r$ have continuous and bounded derivatives of order $n+2$ for any $x$ in [0,1] and satisfy $0 < p_j(x) < 1, j = 1, \ldots, r$ and $\sum_{j=1}^{r} p_j(x) = 1$ for any $x$ in $[0,1]$.

**Theorem 1** *Under assumption 1-3 with $n \ge 0$, we have as $K \to \infty$,*

$$\begin{aligned} \text{Bias}[\hat{p}_{j,\lambda}(x;h)] &= \frac{\lambda}{2N_1} b_{1j}(x) + h^2 b_{2j}(x) + O\left(\frac{1}{Kh} + \frac{h^2}{N_1}\right), \\ V[\hat{p}_{j,\lambda}(x;h)] &= \frac{p_j(x)(1-p_j(x))R(\phi)}{N_1 Kh} + O\left(\frac{1}{N_1^2 Kh} + \frac{h}{N_1 K}\right). \end{aligned}$$

*where*

$$b_{1j}(x) = 1 - r p_j(x),$$

$$
\begin{aligned}
b_{2j}(x) &= \frac{\mu_2(\phi)}{2}\left\{\lambda\eta_j(x) + p_j^{(2)}(x)\right\}, \\
\eta_j(x) &= p_j(x)\left\{\left(\frac{p_j^{(1)}(x)}{p_j(x)}\right)^2 - \sum_{\ell=1}^{r} p_\ell(x)\left(\frac{p_\ell^{(1)}(x)}{p_\ell(x)}\right)^2\right\}.
\end{aligned}
$$

We note that each $\eta_j(x)$ is $p_j(x)$ times centred $(p_j^{(1)}(x)/p_j(x))^2$ with its centre (weighted mean) at $\sum_{\ell=1}^r p_\ell(x)(p_\ell^{(1)}(x)/p_\ell(x))^2$. Since $p_j^{(1)}(x)/p_j(x) = (d/dx)\log p_j(x)$, $\eta_j(x)$ can be viewed as a functional of a squared *score* function (note that $p_j(x)$ itself is not a density but a smooth function). This $\eta_j$ has a key role in developing the algorithm for the data-based choice of $(\lambda, h)$, as described in Section 4. The next theorem reveals the pointwise asymptotic normality of $\hat{\mathbf{p}}_\lambda(x;h)$:

**Theorem 2** *Under assumptions 1-3 with $n \geq 0$, if $\sqrt{N_1 K h^5}$ converges to a constant $\rho$, we have for any $r$-vector $\alpha = (\alpha_1, ..., \alpha_r)^T$,*

$$
\sqrt{N_1 K h}\alpha^T\{\hat{\mathbf{p}}_\lambda(x;h) - \mathbf{p}(x) - \lambda(2N_1)^{-1}\mathbf{b}_1(x)\} \to_d N(\rho\alpha^T\mathbf{b}_2(x), R(\phi)\alpha^T\Sigma(x)\alpha),
$$

*where $\mathbf{b}_1(x) = (b_{11}(x), \ldots, b_{1r}(x))^T$, $\mathbf{b}_2(x) = (b_{21}(x), \ldots, b_{2r}(x))^T$ and*

$$
\Sigma(x) = \mathrm{diag}(\mathbf{p}(x)) - \mathbf{p}(x)\mathbf{p}(x)^T.
$$

### 3.2 Asymptotics for Bias-Adjusted Estimators

Define

$$
\tilde{p}_{j,\lambda}(x;h) = \hat{p}_{j,\lambda}(x;h) - \frac{\lambda}{2N_1}(1 - r\hat{p}_{j,\lambda}(x;h)).
$$

Then, this is a bias-adjusted estimator of $p_j(x)$, which does not include the term $O(N_1^{-1})$ that appeared in the bias of $\hat{p}_j(x;h)$ in Theorem 1. However, the asymptotic variances of $\hat{p}_{j,\lambda}(x;h)$ and $\tilde{p}_{j,\lambda}(x;h)$ are the same. Furthermore, the total sum condition $\sum_{j=1}^r \tilde{p}_{j,\lambda}(x;h) = 1$ remains to hold. We summarize these facts:

**Corollary 1** *Under assumptions 1-3 with $n \geq 0$, we have as $K \to \infty$,*

$$
\begin{aligned}
\mathrm{Bias}[\tilde{p}_{j,\lambda}(x;h)] &= h^2 b_{2j}(x) + O\left(\frac{1}{N_1^2} + \frac{1}{Kh}\right), &(4)\\
V[\tilde{p}_{j,\lambda}(x;h)] &= \frac{p_j(x)(1-p_j(x))R(\phi)}{N_1 K h} + O\left(\frac{1}{N_1^2 K h} + \frac{h}{N_1 K}\right). &(5)
\end{aligned}
$$

Note that the leading bias term of $\tilde{p}_{j,\lambda}(x;h)$ is *linear* in $\lambda$. In Section 4, it will be made apparent that this property is quite essential to identify the best estimator in $\{\tilde{\mathbf{p}}_\lambda(\cdot;h); \lambda \in \Re, h > 0\}$,

where $\tilde{\mathbf{p}}_\lambda(\cdot;h) = (\tilde{p}_{1,\lambda}(\cdot;h), \ldots, \tilde{p}_{r,\lambda}(\cdot;h))^T$. The pointwise asymptotic normality also holds for $\tilde{\mathbf{p}}_\lambda(\cdot;h)$.

**Corollary 2** *Under assumptions 1-3 with $n \geq 0$, if $\sqrt{N_1 K h^5}$ converges to a constant $\rho$, we have for any $r$-vector $\alpha = (\alpha_1, ..., \alpha_r)^T$,*

$$\sqrt{N_1 K h}\alpha^T\{\tilde{\mathbf{p}}_\lambda(x;h) - \mathbf{p}(x)\} \to_d N(\rho\alpha^T\mathbf{b}_2(x), R(\phi)\alpha^T\Sigma(x)\alpha).$$

# 4 Choice of Smoothing Parameters

## 4.1 Optimal Parameters

As a criterion to evaluate estimators, we use the MISE (mean integrated squared error) of an estimator of $p_j(\cdot)$. The MISE is defined as the integral of the MSE (mean squared error) of the estimator over the interval $[\delta_1, 1-\delta_2](\subset [h, 1-h])$, where $\delta_1$ and $\delta_2$ are positive constants. Using (4) and (5), the approximate MISE (AMISE) of $\tilde{p}_{j,\lambda}(\cdot;h)$ is given as

$$\text{AMISE}[\tilde{p}_{j,\lambda}(\cdot;h)] = \frac{h^4\mu_2(\phi)^2}{4}(B_{1j}\lambda^2 + 2B_{2j}\lambda + B_{3j}) + \frac{V_j R(\phi)}{N_1 K h}, \tag{6}$$

where for $j = 1, \ldots, r$,

$$B_{1j} = \int_{\delta_1}^{1-\delta_2} \eta_j(x)^2 dx, \quad B_{2j} = \int_{\delta_1}^{1-\delta_2} \eta_j(x) p_j^{(2)}(x) dx, \quad B_{3j} = \int_{\delta_1}^{1-\delta_2} p_j^{(2)}(x)^2 dx, \tag{7}$$

and

$$V_j = \int_{\delta_1}^{1-\delta_2} p_j(x)(1 - p_j(x)) dx.$$

$B_{1j}$,$B_{2j}$,$B_{3j}$ and $V_j$ are functionals of $p_j$ depending on $\delta_1$ and $\delta_2$. Put $B_t = \sum_{j=1}^r B_{tj}, t = 1, 2, 3$ and $V = \sum_{j=1}^r V_j$. The facts that the bias is linear in $\lambda$ and the variance does not primarily depend on $\lambda$, as pointed out in Section 3, imply that the AMISE is a quadratic function of $\lambda$ as shown in (6). A global measure of accuracy of $\tilde{\mathbf{p}}_\lambda(\cdot;h)$ is naturally defined as

$$\sum_{j=1}^r \text{AMISE}[\tilde{p}_{j,\lambda}(\cdot;h)], \tag{8}$$

which is also a quadratic function in $\lambda$, and hence, the optimal $\lambda$ can be easily derived as

$$\lambda_{\text{opt}} = -\frac{B_2}{B_1}.$$

Furthermore, the optimal bandwidth $h$ should also be defined as the minimizer of the global measure (8). A suboptimal $h$ depending on a fixed $\lambda$ can be defined as

$$h_{\text{opt}}(\lambda) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5} \left(\frac{V}{\Theta(\lambda)}\right)^{1/5} (N_1 K)^{-1/5},$$

where $\Theta(\lambda) = B_1\lambda^2 + 2B_2\lambda + B_3$. Hence the optimal $h$ is defined as

$$h_{\rm opt} = h_{\rm opt}(\lambda_{\rm opt}) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5} \left(\frac{V}{\Theta}\right)^{1/5} (N_1 K)^{-1/5},$$

where $\Theta = \Theta(\lambda_{\rm opt}) = B_3 - B_1^{-1}B_2^2$.

When we choose only the bandwidth $h$ by a data-based method, it implies that an estimate of $h_{\rm opt}(\lambda)$ is constructed for a fixed $\lambda$. On the other hand, obtaining an estimate of $(\lambda_{\rm opt}, h_{\rm opt})$ makes our procedure completely data-based.

Note that a large value of $B_1$ means that if the value of (8) for $\lambda = 0$ is not the smallest, there will exist many $\tilde{\mathbf{p}}_\lambda(\cdot; h)$ that have smaller values of (8). On the other hand, estimators in $\{\tilde{\mathbf{p}}_\lambda(\cdot; h); \lambda \in \Re, h > 0\}$ are almost equivalent in terms of the AMISE provided that the value of $B_1$ is small. This $B_1$, therefore, has a special role in evaluating the validity for considering the family $\{\tilde{\mathbf{p}}_\lambda(\cdot; h); \lambda \in \Re, h > 0\}$. Further, it is a function of the squared integral of $\eta_j(x) (j = 1, ..., r)$, which shows the importance of $\eta_j(x)$, as mentioned in Section 3.

### 4.2 Rule-of-Thumb method

The easiest and most reliable data-based method for the choice of smoothing parameters is the so-called ROT (rule of thumb) method, which exploits a certain parametric model as a target function. Here, we utilize a multinomial logit polynomial model given as

$$p_j(x; \theta) = \frac{\exp(\theta_j^T \mathbf{x})}{\sum_{\ell=1}^r \exp(\theta_\ell^T \mathbf{x})},$$

where $\theta = (\theta_1, ..., \theta_r)$, $\theta_j = (\theta_{j0}, \theta_{j1}, \ldots, \theta_{jm})^T$ and $\mathbf{x} = (1, x, \ldots, x^m)^T$. In order to uniquely obtain the maximum likelihood estimator (MLE) of $\theta$, we put $\theta_1 = \mathbf{0}$. Then we have

$$p_1(x; \theta) = \frac{1}{1 + \sum_{\ell=2}^r \exp(\theta_\ell^T \mathbf{x})} \quad \text{and} \quad p_j(x; \theta) = \frac{\exp(\theta_j^T \mathbf{x})}{1 + \sum_{\ell=2}^r \exp(\theta_\ell^T \mathbf{x})}, j = 2, ..., r.$$

The log-likelihood function excluding the constant term can be written as

$$LL(\theta) = \sum_{i=1}^K \sum_{j=2}^r \mathbf{x}_i^T \theta_j Y_{ij} - \sum_{i=1}^K N_i \log\left(1 + \sum_{j=2}^r \exp(\mathbf{x}_i^T \theta_j)\right),$$

where $\mathbf{x}_i = (1, x_i, \ldots, x_i^m)^T$, $i = 1, ..., K$. We can then obtain the MLE $\hat{\theta}$ that maximizes $LL(\theta)$ on $\theta = (\mathbf{0}, \theta_2, ..., \theta_r)$, from which we have the parametric estimators $p_j(x; \hat{\theta})$ for $j = 1, ..., r$. Note that if $m = 1$, $B_1 = B_2 = B_3$. Hence, $\lambda_{\rm opt} = -1$ can be derived; however, $h_{\rm opt}$ does not exist. In the sequel, we denote the estimators of $\lambda_{\rm opt}$ and $h_{\rm opt}$ based on the ROT method as $\hat{\lambda}_{\rm ROT}$ and $\hat{h}_{\rm ROT}$, respectively.

### 4.3 Plug-in Method

The PI (plug-in) method for the optimal parameters $(\lambda_{\rm opt}, h_{\rm opt})$ is developed by the same procedure as discussed in Ruppert et al. [17]. To construct consistent estimators of $\lambda_{\rm opt}$ and $h_{\rm opt}$,

we exploit a convenient estimator of $p_j(x)$ defined as

$$\bar{p}_j(x;g) = \hat{p}_{j,0}(x;g) = \frac{\sum_{i=1}^K \phi_g(x_i - x)\bar{Y}_{ij}}{\sum_{i=1}^K \phi_g(x_i - x)}.$$

Let us define

$$\bar{\eta}_j(x;g) = \bar{p}_j(x;g)\left\{\left(\frac{\bar{p}_j^{(1)}(x;g)}{\bar{p}_j(x;g)}\right)^2 - \sum_{\ell=1}^r \left(\frac{\bar{p}_\ell^{(1)}(x;g)}{\bar{p}_\ell(x;g)}\right)^2\right\},$$

which is a direct estimate of $\eta_j(x)$, using $\bar{p}_j(x;g)$. For practical purposes, we define $\bar{\eta}_j(x;g) = 0$ if $\bar{p}_j(x;g) = 0$. By substituting $\bar{p}_j^{(2)}(x;g)$ and $\bar{\eta}_j(x;g)$ into (7), the estimators $\bar{B}_{tj}(g)$ are obtained, and we put $\bar{B}_t(g) = \sum_{j=1}^r \bar{B}_{tj}(g), t = 1,2,3$. Then we have a consistent estimator of $\lambda_{\rm opt}$ with the bandwidth $g_1$ defined as

$$\bar{\lambda}_{\rm opt}(g_1) = -\frac{\bar{B}_2(g_1)}{\bar{B}_1(g_1)}.$$

To select the optimal $g_1$ that minimizes the MSE of $\bar{\lambda}_{\rm opt}(g_1)$, we use the following assumption:

**ASSUMPTION 4** $g_1 \to 0$ and $N_i \to \infty$ as $K \to \infty$ for $i = 1, \ldots, K$ in a manner such that $Kg_1^{4+\varepsilon} = O(1)$ and $N_1 g_1^{1-\varepsilon} = O(1)$ for some $1/2 < \varepsilon < 1$, and $N_i/N_1 = 1 + o(g_1^2)$.

**Theorem 3** *Under assumptions 2-4 with $n \geq 2$, we have as $K \to \infty$,*

$$\mathrm{MSE}[\bar{\lambda}_{\rm opt}(g_1)] = \left(g_1^2\Delta_{11} + \frac{\Delta_{12}}{N_1 K g_1^3}\right)^2 + \frac{\Delta_{13}}{N_1 K^3 g_1^{10}} + O\left(\frac{1}{N_1^2 K^3 g_1^{10}}\right), \tag{9}$$

*where $\Delta_{1t}, t = 1,2,3$ are given in Appendix.*

The first term on the right hand side of (9) is the squared leading bias and the second term is the leading term of the asymptotic variance. Under the assumptions in Theorem 3, the second term converges faster than the first term. We obtain the optimal bandwidth $g_1^\dagger$ that minimizes the first term:

$$g_1^\dagger = C_1\left(\frac{\Delta_{12}}{\Delta_{11}}\right)^{1/5}(N_1 K)^{-1/5},$$

where

$$C_1 = \begin{cases} -1 & , \quad \Delta_{11}\Delta_{12} < 0, \\ (3/2)^{1/5} & , \quad \Delta_{11}\Delta_{12} > 0. \end{cases}$$

This choice of $g_1$ yields $\bar{\lambda}_{\rm opt}(g_1^\dagger)/\lambda_{\rm opt}(g_1^\dagger) - 1 = O_P((N_1K)^{-2/5})$ as $K \to \infty$.

Next, we focus on the data-based choice of $h$. It is evident that an estimator of $h_{\rm opt}(\lambda)$ with the bandwidth $g_2$ can be obtained as

$$\bar{h}_{\rm opt}(g_2,\lambda) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5}\left(\frac{\bar{V}}{\bar{\Theta}(g_2,\lambda)}\right)^{1/5}(N_1K)^{-1/5},$$

where $\bar{V} = \sum_{j=1}^r \bar{V}_j, \bar{V}_j = K^{*-1}\sum_i{}^* N_i(N_i-1)^{-1}(\bar{Y}_{ij} - \bar{Y}_{ij}^2)$ in which $K^*$ is the number of $x_i$ falling into $[\delta_1, 1-\delta_2]$ and $\sum_i{}^*$ is the summation for those $x_i$, and $\bar{\Theta}(g_2,\lambda) = \bar{B}_1(g_2)\lambda^2 +$

$2\bar{B}_2(g_2)\lambda + \bar{B}_3(g_2)$. Put $\bar{\Theta}(g_2) = \bar{\Theta}(g_2, -\bar{B}_1(g_2)^{-1}\bar{B}_2(g_2)) = \bar{B}_3(g_2) - \bar{B}_1(g_2)^{-1}\bar{B}_2(g_2)^2$. Then, an estimator of $h_{\rm opt}$ with the bandwidth $g_2$ should be

$$\bar{h}_{\rm opt}(g_2) = \bar{h}_{\rm opt}(g_2, -\bar{B}_1(g_2)^{-1}\bar{B}_2(g_2)) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5}\left(\frac{\bar{V}}{\bar{\Theta}(g_2)}\right)^{1/5}(N_1K)^{-1/5}.$$

The following assumption is required to obtain the optimal $g_2$ based on the MSE of $\bar{h}_{\rm opt}(g_2)$.

**ASSUMPTION 5** $g_2 \to 0$ and $N_i \to \infty$ as $K \to \infty$ for $i = 1, \ldots, K$ in a manner such that $Kg_2^{6+\varepsilon} = O(1)$ and $N_1 g_2^{1-\varepsilon} = O(1)$ for some $0 < \varepsilon < 1$, and $N_i/N_1 = 1 + o(h^2)$.

**Theorem 4** *Under assumptions 2, 3 and 5, we have as* $K \to \infty$,

$$\mathrm{MSE}[\bar{\Theta}(g_2)] = \left(g_2^2\Delta_{21} + \frac{\Delta_{22}}{N_1Kg_2^5}\right)^2 + \frac{\Delta_{23}}{N_1K^2g_2^9} + O\left(\frac{1}{N_1^2K^2g_2^9}\right), \tag{10}$$

*where* $\Delta_{2t}, t = 1, 2, 3$ *are given in Appendix.*

We shall provide the same explanation as in the paragraph following Theorem 3. The first term on the right hand side of (10) is the squared leading bias and the second term is the leading term of the asymptotic variance. Under the assumptions in Theorem 4, the second term converges faster than the first term. We obtain the optimal bandwidth $g_2^\dagger$ that minimizes the first term:

$$g_2^\dagger = C_2\left(\frac{\Delta_{22}}{\Delta_{21}}\right)^{1/7}(N_1K)^{-1/7},$$

where

$$C_2 = \begin{cases} -1 & , \quad \Delta_{21}\Delta_{22} < 0, \\ (5/2)^{1/7} & , \quad \Delta_{21}\Delta_{22} > 0. \end{cases}$$

This choice of $g_2$ yields $\bar{h}_{\rm opt}(g_2^\dagger)/h_{\rm opt}(g_2^\dagger) - 1 = O_P((N_1K)^{-2/7})$ as $K \to \infty$ since $\bar{V} - V = O_P((N_1K)^{-1/2})$.

Instead of the optimal bandwidths $g_1^\dagger$ and $g_2^\dagger$, we use the ROT estimators $\hat{g}_1$ and $\hat{g}_2$ in the same manner as described in subsection 4.1. Finally, we obtain the data-driven parameter $(\bar{\lambda}_{\rm opt}(\hat{g}_1), \bar{h}_{\rm opt}(\hat{g}_2))$, which is expressed as $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ in the sequel.

### 4.4 Summary of the Algorithm

In order to increase the level of sophistication, we summarize the algorithm of the PI method for selecting $\tilde{\lambda}_{\rm PI}$ and $\tilde{h}_{\rm PI}$ and the ROT method for selecting $\hat{\lambda}_{\rm ROT}$ and $\hat{h}_{\rm ROT}$, as follows:
Algorithm 1: The PI power-divergence selector $\tilde{\lambda}_{\rm PI}$

1. Obtain the estimator of $g_1^\dagger$, which is denoted as $\hat{g}_1$, by the ROT method.
2. Calculate $\bar{B}_t(\hat{g}_1), t = 1, 2$.

3. The power-divergence selector is

$$\tilde{\lambda}_{\mathrm{PI}} = \bar{\lambda}_{\mathrm{opt}}(\hat{g}_1) = -\frac{\bar{B}_2(\hat{g}_1)}{\bar{B}_1(\hat{g}_1)}.$$

Algorithm 2: The PI bandwidth $\tilde{h}_{\mathrm{PI}}$

1. Obtain the estimator of $g_2^{\dagger}$, which is denoted as $\hat{g}_2$, by the ROT method.
2. Calculate $\bar{B}_t(\hat{g}_2), t = 1, 2, 3$ and $\bar{V}$.
3. The bandwidth is

$$\tilde{h}_{\mathrm{PI}} = \bar{h}_{\mathrm{opt}}(\hat{g}_2) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5} \left(\frac{\bar{V}}{\bar{\Theta}(\hat{g}_2)}\right)^{1/5} (N_1 K)^{-1/5},$$

where $\bar{\Theta}(\hat{g}_2) = \bar{B}_3(\hat{g}_2) - \bar{B}_1(\hat{g}_2)^{-1}\bar{B}_2(\hat{g}_2)^2$.

Algorithm 3: The ROT parameters $\hat{\lambda}_{\mathrm{ROT}}$ and $\hat{h}_{\mathrm{ROT}}$

1. Obtain the estimator of $B_t (t = 1, 2, 3)$ and $V$ by the ROT method, which are denoted as $\hat{B}_t (t = 1, 2, 3)$ and $\hat{V}$, respectively.
2. The ROT parameters are

$$\hat{\lambda}_{\mathrm{ROT}} = -\frac{\hat{B}_2}{\hat{B}_1} \text{ and } \hat{h}_{\mathrm{ROT}} = \hat{h}_{\mathrm{ROT}}(\hat{\lambda}_{\mathrm{ROT}}),$$

where

$$\hat{h}_{\mathrm{ROT}}(\lambda) = \left(\frac{R(\phi)}{\mu_2(\phi)^2}\right)^{1/5} \left(\frac{\hat{V}}{\hat{B}_1\lambda^2 + 2\hat{B}_2\lambda + \hat{B}_3}\right)^{1/5} (N_1 K)^{-1/5}.$$

# 5 Simulation Study

The performance of the proposed methods was evaluated using estimates of the sums of MISEs of the estimators $\tilde{p}_{\lambda,j}(\cdot, h)$. In addition, the classical method corresponding to $\lambda = 0$ was compared with the proposed methods. We calculated estimates of the sums of MISEs of estimators with the following six pairs of $(\lambda, h)$: $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$,$(0, \bar{h}_0)$,$(\hat{\lambda}_{\mathrm{ROT}}, \hat{h}_{\mathrm{ROT}})$, $(0, \hat{h}_0)$, $(\lambda_{\mathrm{opt}}, h_{\mathrm{opt}})$ and $(0, h_0)$, where $h_0 = h_{\mathrm{opt}}(0)$, $\bar{h}_0 = \bar{h}_{\mathrm{opt}}(\hat{g}_2, 0)$ and $\hat{h}_0 = \hat{h}_{\mathrm{ROT}}(0)$; see Algorithms 1-3 in the previous section. Two models for $r = 3$ were adopted as the true model. Let

$$p_j(x) = \frac{f_j(x)}{\sum_{\ell=1}^{3} f_\ell(x)},$$

for $j = 1, 2, 3$. Model 1 is a multinomial logit model defined as

$$f_1(x) = 1, \quad f_2(x) = \exp(2.5 - 6x + 2x^2), \quad f_3(x) = \exp(-2.5 + 6x - 2x^2),$$

| | $K = 100$ | | | | | |
|---|---|---|---|---|---|---|
| $N_1$ | $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ | $(0, \bar{h}_0)$ | $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$ | $(0, \hat{h}_0)$ | $(\lambda_{\rm opt}, h_{\rm opt})$ | $(0, h_0)$ |
| 100 | 1194 | 251 | 17848 | 249 | 10679(.214) | 247(.135) |
| 200 | 111 | 135 | 832 | 134 | 366(.186) | 133(.118) |
| 300 | 76 | 100 | 92 | 99 | 91(.172) | 99(.109) |
| 400 | 60 | 82 | 59 | 82 | 58(.162) | 81(.103) |
| | $K = 120$ | | | | | |
| $N_1$ | $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ | $(0, \bar{h}_0)$ | $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$ | $(0, \hat{h}_0)$ | $(\lambda_{\rm opt}, h_{\rm opt})$ | $(0, h_0)$ |
| 100 | 228 | 217 | 15270 | 215 | 9624(.206) | 214(.130) |
| 200 | 92 | 117 | 293 | 116 | 225(.180) | 116(.114) |
| 300 | 65 | 86 | 73 | 86 | 72(.166) | 86(.105) |
| 400 | 51 | 69 | 50 | 68 | 49(.156) | 68(.099) |

Table 1: The MISEs ($\times 10^6$) for six pairs of $(\lambda, h)$ for Model 1 with $K = 100, 120$ and $N_1 = 100, 200, 300, 400$: $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$,$(0, \bar{h}_0)$, $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$, $(0, \hat{h}_0)$, $(\lambda_{\rm opt}, h_{\rm opt})$ and $(0, h_0)$. The values of the optimal bandwidth $h_{\rm opt}$ for each $(K, N_1)$ are given in parentheses and $\lambda_{\rm opt} = -1.039$.

| | $K = 100$ | | | | | |
|---|---|---|---|---|---|---|
| $N_1$ | $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ | $(0, \bar{h}_0)$ | $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$ | $(0, \hat{h}_0)$ | $(\lambda_{\rm opt}, h_{\rm opt})$ | $(0, h_0)$ |
| 100 | 405 | 333 | 91007 | 370 | 329(.117) | 320(.100) |
| 200 | 173 | 178 | 7488 | 202 | 156(.102) | 172(.087) |
| 300 | 123 | 132 | 1044 | 149 | 114(.098) | 129(.080) |
| 400 | 100 | 107 | 306 | 123 | 92(.089) | 104(.075) |
| | $K = 120$ | | | | | |
| $N_1$ | $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ | $(0, \bar{h}_0)$ | $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$ | $(0, \hat{h}_0)$ | $(\lambda_{\rm opt}, h_{\rm opt})$ | $(0, h_0)$ |
| 100 | 428 | 277 | 90979 | 309 | 270(.113) | 266(.096) |
| 200 | 151 | 158 | 6229 | 179 | 137(.098) | 153(.084) |
| 300 | 109 | 117 | 745 | 133 | 101(.091) | 113(.077) |
| 400 | 84 | 91 | 255 | 105 | 78(.086) | 89(.073) |

Table 2: The MISEs ($\times 10^6$) for six pairs of $(\lambda, h)$ for Model 2 with $K = 100, 120$ and $N_1 = 100, 200, 300, 400$: $(\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$, $(0, \bar{h}_0)$,$(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$, $(0, \hat{h}_0)$, $(\lambda_{\rm opt}, h_{\rm opt})$ and $(0, h_0)$. The values of the optimal bandwidth $h_{\rm opt}$ for each $(K, N_1)$ are given in parentheses, and $\lambda_{\rm opt} = -1.355$.

and Model 2, which is more complicated than Model 1, is defined as $f_1(x) = 1$,

$$f_2(x) = (0.5\sin(4x) + 1)\exp(2.5 - 6x + 2x^2), \quad f_3(x) = (0.25\sin(8x) + 1)\exp(-2.5 + 6x - 2x^2).$$

In this simulation, $N_i (i = 1, ..., K)$ were set to be equal and $\delta_1 = \delta_2 = 0.1$. We utilized $m = 2$ as a parameter in the multinomial logit polynomial model used for the ROT method. The setting of $m = 2$ is evidently advantageous for the ROT method in Model 1 because the assumed parametric model in this method is nothing but the true target. For the kernels of $\tilde{p}_{j,\lambda}(\cdot; h)$ and $\bar{p}_j(\cdot; g)$, the Epanechnikov kernel $(3/4)(1 - x^2)I_{(-1,1)}(x)$ and the triweight kernel $(35/32)(1-x^2)^3 I_{(-1,1)}(x)$ were employed, respectively. We calculated the MISE estimates using 500 Monte Carlo runs with $K = 100, 120$ and $N_1 = 100, 200, 300, 400$.

Table 1 shows the result of the simulation for Model 1. The estimator with $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ performs well unless $N_1 = 100$. The outer points in $[\delta_1, 1 - \delta_2]$ might affect the performance of the estimator with $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ for the case when $N_1 = 100$. The fact that the estimator with $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ is superior to that with $(\lambda_{\mathrm{opt}}, h_{\mathrm{opt}})$ for $N_1 = 200, 300$ reveals the usefulness of the data-based choice of smoothing parameters. Such superiorities are also recognized for estimators with $\lambda = 0$ fixed, which shows in particular the importance of the choice of $\lambda$. Further, it is worth noting that the estimator with $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ performs better than that with $(\hat{\lambda}_{\mathrm{ROT}}, \hat{h}_{\mathrm{ROT}})$ despite the previously mentioned advantage of Model 1 for the ROT method.

Table 2 shows the result for Model 2. In this table, the tendency is the same as that in Table 1. The estimator with $(\hat{\lambda}_{\mathrm{ROT}}, \hat{h}_{\mathrm{ROT}})$ does not perform well since this is a misspecified case. On the other hand, the estimator with $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ exhibits a stable performance.

Figure 1 displays the density estimates for the relative error of $\tilde{\lambda}_{\mathrm{PI}}$ to $\lambda_{\mathrm{opt}}$ defined as $\tilde{\lambda}_{\mathrm{PI}}/\lambda_{\mathrm{opt}} - 1$ for Model 2 with $K = 100$ and $N_1 = 100, 200, 300, 400$. Figure 2 shows the density estimates for the relative error of $\tilde{h}_{\mathrm{PI}}$ to $h_{\mathrm{opt}}$ for Model 2 with $K = 100$ and $N_1 = 100, 200, 300, 400$. Both the figures show the convergence property of the PI method as $N_1$ increases. Similar results were obtained for Model 1.

The density estimates of the relative errors of $\tilde{h}_{\mathrm{PI}}$ and $\hat{h}_{\mathrm{ROT}}$ to $h_{\mathrm{opt}}$ are comparatively provided in Figure 3. For a large sample, it would be more appropriate to select the PI method instead of the ROT method.

## 6 Applications

We applied our approach to a real data set obtained from the UCI Repository of machine learning databases (Blake and Merz [6]).

In the abalone database, we focus on the relationship between the abalone's age and the type of sex: male ($j = 1$), female ($j = 2$) and infant ($j = 3$). To extract this relationship, the problem was formulated as a multinomial regression with $r = 3$ and the shell ring as the covariate. It is known that the shell ring plus 1.5 gives the age of the abalone. The interval of the ring, $[3, 23]$, was first transformed linearly to the interval $[0, 1]$, after which it was transformed inversely to the original scale. We see that $K = 21$ and $N_i$ are not equal to each other. Taking this fact into

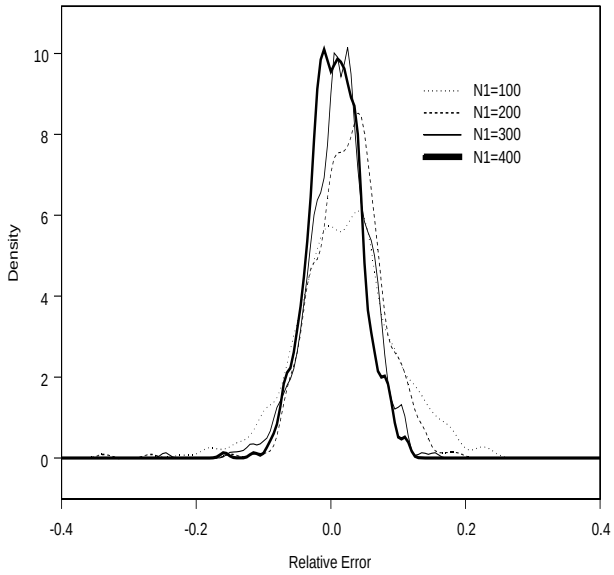


Figure 1: Density estimates of the relative error of $\tilde{\lambda}_{\mathrm{PI}}$ to $\lambda_{\mathrm{opt}}$: $\tilde{\lambda}_{\mathrm{PI}}/\lambda_{\mathrm{opt}} - 1$ for Model 2 with $K = 100$ and $N_1 = 100, 200, 300, 400$.

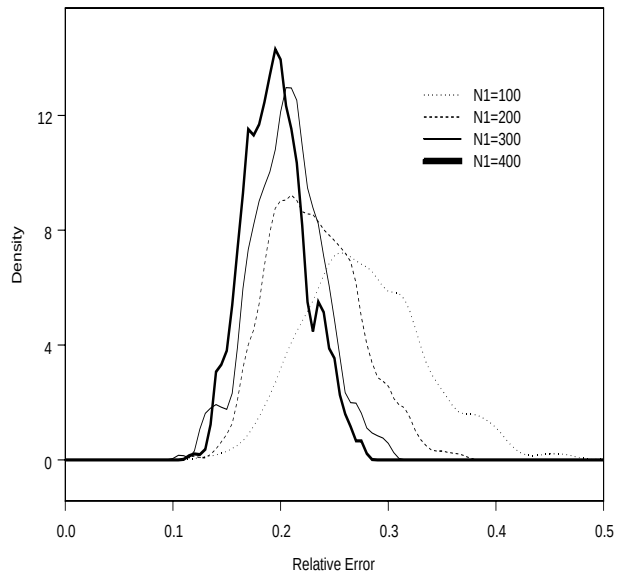


Figure 2: Density estimates of the relative error of $\tilde{h}_{\mathrm{PI}}$ to $h_{\mathrm{opt}}$: $\tilde{h}_{\mathrm{PI}}/h_{\mathrm{opt}} - 1$ for Model 2 with $K = 100$ and $N_1 = 100, 200, 300, 400$.

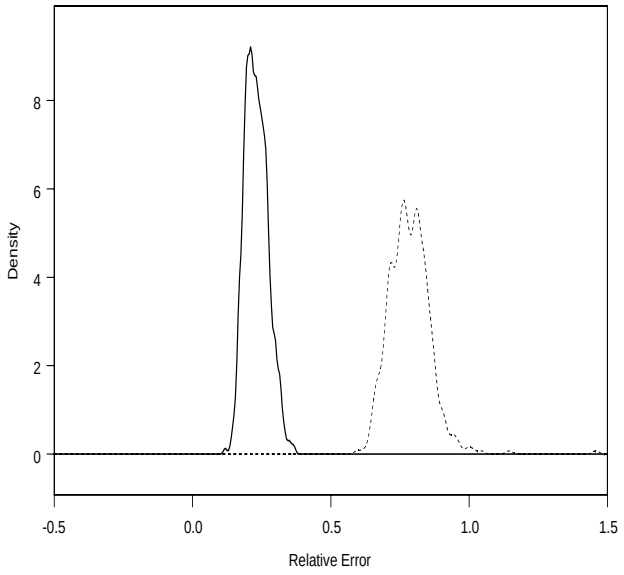


Figure 3: Density estimates of the relative errors of $\tilde{h}_{\mathrm{PI}}$ and $\hat{h}_{\mathrm{ROT}}$ to $h_{\mathrm{opt}}$: $\tilde{h}_{\mathrm{PI}}/h_{\mathrm{opt}} - 1$ (solid line) and $\hat{h}_{\mathrm{ROT}}/h_{\mathrm{opt}} - 1$ (dashed line ) for Model 2 with $K = 100$ and $N_1 = 200$.

consideration, the following estimator, discussed in Okumura and Naito [14], was used instead of $\tilde{p}_{\lambda,j}(x;h)$:

$$\dot{p}_{j,\lambda}(x;h) = \hat{p}_{j,\lambda}(x;h) - \frac{\lambda}{2\bar{N}(x;h)}(1 - r\hat{p}_{j,\lambda}(x;h)),$$

where $\bar{N}(x;h) = (\sum_{i=1}^{K} N_i\phi_h(x_i - x))/(\sum_{i=1}^{K}\phi_h(x_i - x))$. $\dot{p}_{j,\lambda}(x;h)$ is known to have the same asymptotic property as $\tilde{p}_{j,\lambda}(x;h)$. Hence, for any $x$, an approximate $100(1-\alpha)\%$ confidence interval for $p_j(x)$ can be constructed as follows:

$$\begin{aligned}&\mathrm{CI}_{100(1-\alpha),j}(x;\lambda,h)\\ &= [\dot{p}_{\lambda,j}(x;h) - z_{\alpha/2}\dot{\sigma}_j(x;\lambda,h),\ \dot{p}_{\lambda,j}(x;h) + z_{\alpha/2}\dot{\sigma}_j(x;\lambda,h)],\end{aligned}$$

where $\dot{\sigma}_j(x;\lambda,h) = \sqrt{\dot{p}_{\lambda,j}(x;h)(1-\dot{p}_{\lambda,j}(x;h))R(\phi)/(\bar{N}(x)Kh)}$ and $z_{\alpha/2}$ is the upper $100(\alpha/2)\%$ point of the standard normal distribution. We adopted $N_1 = 218$ (the integer part of the average of $N_i$), $\delta_1 = \delta_2 = 0.05$ and $m = 3$ in the ROT method. The kernels used in the simulation in Section 5 were employed for this analysis too, which gives $R(\phi) = 0.6$. From this data set, we obtained $(\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}}) = (-0.602, 0.069)$ and $(\hat{\lambda}_{\mathrm{ROT}}, \hat{h}_{\mathrm{ROT}}) = (-0.529, 0.100)$.

Figure 4 shows the result for $\dot{p}_{j,\lambda}(x;h)$ with $(\lambda,h) = (\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}}), (\hat{\lambda}_{\mathrm{ROT}}, \hat{h}_{\mathrm{ROT}})$ and the parametric estimates $p_j(x;\hat{\theta})$ based on the MLE. Figure 4 (a), (b) and (c) exhibit the relationship between the ring and the type of sex along with the approximate 95% confidence intervals $\mathrm{CI}_{95,j}(x;\tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$.

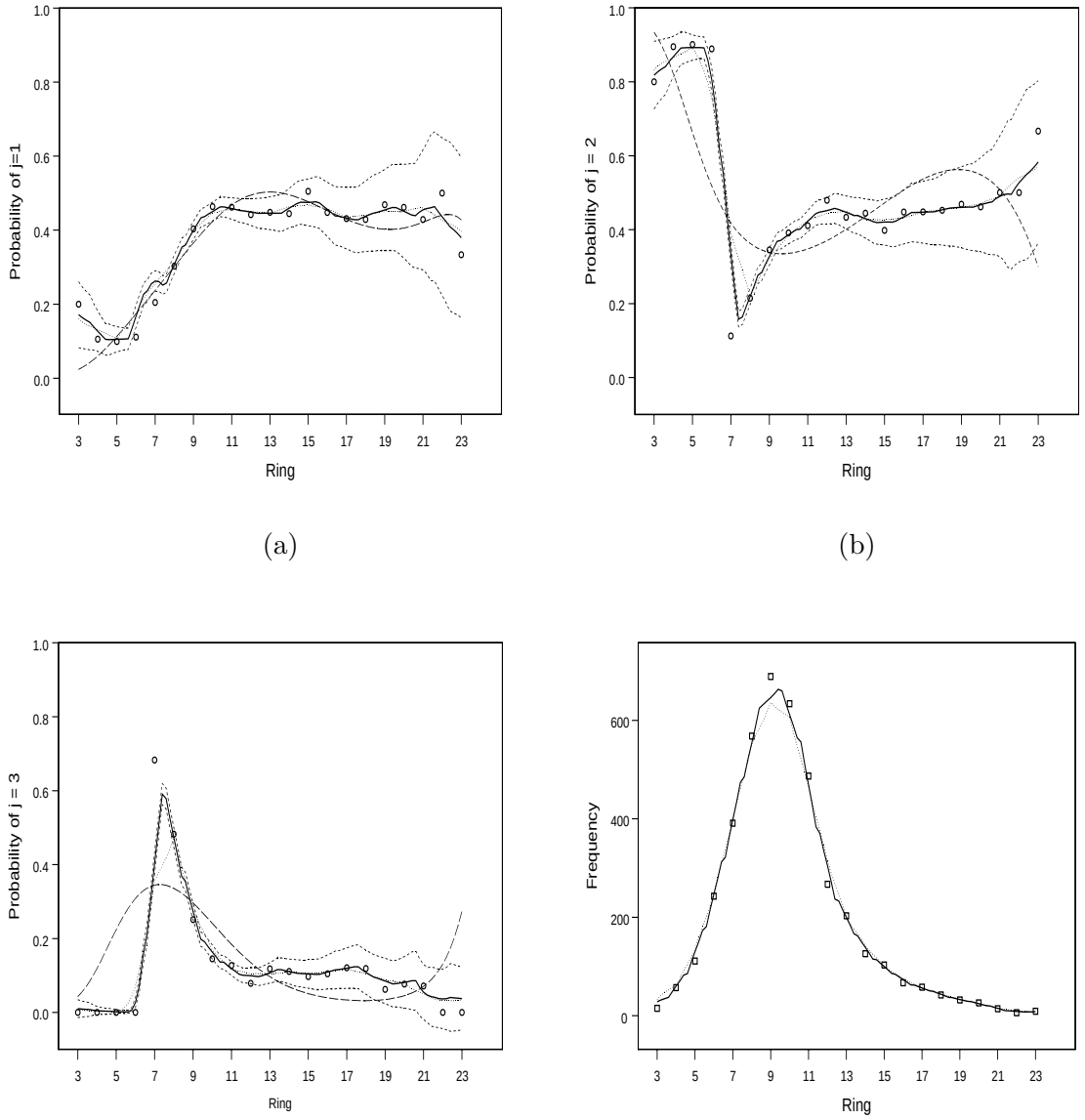


Figure 4: (a), (b) and (c): The estimators $\dot{p}_{j,\lambda}(x;h)$ for $(\lambda,h) = (\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ (solid line) and $(\hat{\lambda}_{\rm ROT}, \hat{h}_{\rm ROT})$ (dotted line), the MLE $p_j(x;\hat{\theta})$ (dashed line) and an approximate 95% confidence interval ${\rm CI}_{95,j}(x;\tilde{\lambda}_{\rm PI}, \tilde{h}_{\rm PI})$ (short dashed line) for $j = 1, 2$ and 3, respectively. (d): The frequency $N_i$ at each covariate and $\bar{N}(x;h)$ for $h = \tilde{h}_{\rm PI}$ (solid line) and $\hat{h}_{\rm ROT}$ (dotted line) for the abalone data.

From these figures, we infer that the parametric estimates based on MLEs show oversmoothing and deterioration in the accuracy at boundaries. On the other hand, the proposed PI estimates could capture a rapid change in data, and the ROT estimates produced curves that were almost similar to those of the PI estimates. It is noteworthy that the ROT estimates show a slight oversmoothing as compared to the PI estimates at a value of the ring between 7 and 9. This reveals that the PI estimates have less bias than the ROT estimates. This behavior appears to be common in nonparametric smoothing; see Simonoff [18] and Wand and Jones [23]. Figure 4 (d) exhibits $N_i$ and $\bar{N}(x)$, which provide us with an insight into the length of the approximate confidence $\mathrm{CI}_{95,j}(x; \tilde{\lambda}_{\mathrm{PI}}, \tilde{h}_{\mathrm{PI}})$ at each covariate $x$.

# 7 Discussion

We have proposed a new approach to construct the kernel estimators by means of the localized power-divergence, as shown in (2). The proposed class of estimators includes the Nadaraya-Watson estimator as a special case, and we have shown the existence of estimators that are better than the Nadaraya-Watson estimator in terms of the AMISE. Further, we again note that our proposed estimators always exist; however, local likelihood estimators do not.

A method of data-based selection for the parameters $(\lambda, h)$ has also been developed. Hence, our estimators are completely data-based. The efficiency of the proposed PI method for selecting $(\lambda, h)$ has been demonstrated through a simulation study for a large sample. In particular, the results reveal the effectiveness of choosing not only the bandwidth $h$ but also $\lambda$ based on the data set. In practical situations, $\hat{g}_1$ and $\hat{g}_2$ would take values larger than $\delta_1$ and $\delta_2$, which are included in the integral interval. For such a case, the ROT method is recommended on the basis of our experiments.

We selected the bandwidths $g_1$ and $g_2$ by minimizing the squared asymptotic bias. However, in a finite case, we believe that the effect of the variance term should be considered. Okumura and Naito [16] discussed this in the setting of a binomial regression, and in fact, a substantial improvement was reported in their paper as a result of considering the effect of variance. We expect that the same improvement would be obtained in the setting of a multinomial regression. Furthermore, the boundary effect is often very important. It is known that the estimators are immune to the boundary effect if a certain type of boundary kernel is used. Since our proposed estimators require a nonnegative kernel, the boundary kernel given in Karunamuni and Alberts [10] could be utilized.

In application, the random design for covariates with $N_i = 1, i = 1, ..., K$ will be occurred. We will attempt to devise approaches for these in our future resarch.

# Appendix

Proofs of theorems in this paper are presented. The lemmas required in the proofs are only enumerated in the end of Appendix.

*Proof of Theorem 1* For any $K$-dimensional vector $\mathbf{t}_{Kj} = (t_{1j}, \ldots, t_{Kj})^T$, we put

$$f_K(\mathbf{t}_{Kj}) = \left( \frac{1}{K} \sum_{i=1}^{K} w_i t_{ij}^{\lambda+1} \right)^{\frac{1}{\lambda+1}},$$

where $w_i = \phi_h(x_i - x)$. Then we obtain

$$\frac{\partial}{\partial t_{kj}} f_K(\mathbf{t}_{Kj}) = \frac{1}{K} t_{kj}^{\lambda} w_k f_K(\mathbf{t}_{Kj})^{-\lambda},$$

$$\frac{\partial^2}{\partial t_{kj}^2} f_K(\mathbf{t}_{Kj}) = -\frac{1}{K^2} \lambda t_{kj}^{2\lambda} w_k^2 f_K(\mathbf{t}_{Kj})^{-1-2\lambda} + \frac{1}{K} \lambda t_{kj}^{-1+\lambda} w_k f_K(\mathbf{t}_{Kj})^{-\lambda},$$

and

$$\frac{\partial^2}{\partial t_{kj} \partial t_{\ell j}} f_K(\mathbf{t}_{Kj}) = -\frac{1}{K^2} \lambda t_{kj}^{\lambda} t_{\ell j}^{\lambda} w_k w_\ell f_K(\mathbf{t}_{Kj})^{-1-2\lambda}.$$

It is easy to see that the proposed estimator $\hat{p}_{j,\lambda}(x; h)$ can be rewritten as

$$\hat{p}_{j,\lambda}(x; h) = \frac{f_K(\bar{\mathbf{Y}}_{Kj})}{\sum_{\ell=1}^{r} f_K(\bar{\mathbf{Y}}_{K\ell})},$$

where $\bar{\mathbf{Y}}_{Kj} = (\bar{Y}_{1j}, \ldots, \bar{Y}_{Kj})^T$. We put $\mathbf{p}_{Kj} \equiv E[\bar{\mathbf{Y}}_{Kj}] = (p_{1j}, \cdots, p_{Kj})^T$. Taylor expansion of $f_K(\bar{\mathbf{Y}}_{Kj})$ around $\mathbf{p}_{Kj}$ is given as

$$\begin{aligned} f_K(\bar{\mathbf{Y}}_{Kj}) &= f_K(\mathbf{p}_{Kj}) + f_K(\mathbf{p}_{Kj})^{-\lambda} \frac{1}{K} \sum_{i=1}^{K} w_i p_{ij}^{\lambda} (\bar{Y}_{ij} - p_{ij}) \\ &\quad + \frac{\lambda}{2} \left\{ f_K(\mathbf{p}_{Kj})^{-\lambda} \frac{1}{K} \sum_{i=1}^{K} p_{ij}^{-1+\lambda} w_i (\bar{Y}_{ij} - p_{ij})^2 \right. \\ &\quad \left. - f_K(\mathbf{p}_{Kj})^{-1-2\lambda} \frac{1}{K^2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} w_k w_\ell p_{kj}^{\lambda} p_{\ell_j}^{\lambda} (\bar{Y}_{kj} - p_{kj})(\bar{Y}_{\ell j} - p_{\ell j}) \right\} + \cdots . \quad \text{(A.1)} \end{aligned}$$

Hence we obtain

$$\begin{aligned} E[f_K(\bar{\mathbf{Y}}_{Kj})] &= f_K(\mathbf{p}_{Kj}) + \frac{\lambda}{2} \left\{ f_K(\mathbf{p}_{Kj})^{-\lambda} \frac{1}{K} \sum_{i=1}^{K} p_{ij}^{-1+\lambda} w_i \frac{p_{ij}(1-p_{ij})}{N_i} \right. \\ &\quad \left. - f_K(\mathbf{p}_{Kj})^{-1-2\lambda} \frac{1}{K^2} \sum_{k=1}^{K} w_k^2 \frac{p_{kj}(1-p_{kj})}{N_k} \right\} + \cdots . \end{aligned}$$

Direct calculations also give

$$f_K(\mathbf{p}_{Kj}) = p_j(x) + h^2 \frac{\mu_2(\phi)}{2} \left\{ \lambda \frac{p_j^{(1)}(x)^2}{p_j(x)} + p_j^{(2)}(x) \right\} + O\left( \frac{1}{Kh} + h^4 \right),$$

and therefore

$$\begin{aligned} E[f_K(\bar{\mathbf{Y}}_{Kj})] &= p_j(x) + \frac{\lambda}{2N_1}(1-p_j(x)) \\ &\quad + h^2\frac{\mu_2(\phi)}{2}\left\{\lambda\frac{p_j^{(1)}(x)^2}{p_j(x)} + p_j^{(2)}(x)\right\} + O\left(\frac{1}{Kh}+\frac{h^2}{N_1}\right), \end{aligned} \tag{A.2}$$

and

$$E\left[\sum_{j=1}^{r} f_K(\bar{\mathbf{Y}}_{Kj})\right] = 1 + \frac{\lambda h^2\mu_2(\phi)}{2}\sum_{j=1}^{r}\frac{p_j^{(1)}(x)^2}{p_j(x)} + O\left(\frac{1}{Kh}+\frac{h^2}{N_1}\right).$$

Moreover,

$$\begin{aligned} \hat{p}_{j,\lambda}(x;h) &= \frac{f_K(\bar{\mathbf{Y}}_{Kj})}{\sum_{\ell=1}^{r} f_K(\bar{\mathbf{Y}}_{K\ell})} \\ &= \frac{f_K(\bar{\mathbf{Y}}_{Kj})}{\sum_{\ell=1}^{r} E[f_K(\bar{\mathbf{Y}}_{Kj})] + \sum_{\ell=1}^{r}\{f_K(\bar{\mathbf{Y}}_{K\ell}) - E[f_K(\mathbf{Y}_{K\ell})]\}} \\ &= \frac{f_K(\bar{\mathbf{Y}}_{Kj})}{\sum_{\ell=1}^{r} E[f_K(\mathbf{Y}_{K\ell})]} \\ &\quad - \frac{f_K(\bar{\mathbf{Y}}_{Kj})}{\{\sum_{\ell=1}^{r} E[f_K(\mathbf{Y}_{K\ell})]\}^2}\left[\sum_{\ell=1}^{r}\{f_K(\bar{\mathbf{Y}}_{K\ell}) - E[f_K(\mathbf{Y}_{K\ell})]\}\right] + \cdots, \end{aligned} \tag{A.3}$$

it follows from direct calculations for (A.1) and (A.3) using Lemma 1 that

$$\begin{aligned} E[\hat{p}_{j,\lambda}(x;h)] &= \frac{E[f_K(\bar{\mathbf{Y}}_{Kj})]}{\sum_{\ell=1}^{r} E[f_K(\mathbf{Y}_{K\ell})]} + O\left(\frac{h}{N_1K}\right) \\ &= p_j(x) + \frac{\lambda}{2N_1}(1-rp_j(x)) \\ &\quad + \frac{h^2\mu_2(\phi)}{2}\left\{\lambda\left(\frac{p_j^{(1)}(x)^2}{p_j(x)} - p_j(x)\sum_{\ell=1}^{r}\frac{p_\ell^{(1)}(x)^2}{p_\ell(x)}\right) + p_j^{(2)}(x)\right\} + O\left(\frac{1}{N_1K}+\frac{h^2}{N_1}\right), \end{aligned}$$

which gives the expression of the bias in Theorem 1. On the other hand, by directly calculating the variance for (A.1) and (A.3) through Lemma 1, we have

$$\begin{aligned} V[\hat{p}_{j,\lambda}(x;h)] &= \frac{V[f_K(\bar{\mathbf{Y}}_{Kj})]}{(\sum_{\ell=1}^{r} E[f_K(\mathbf{Y}_{K\ell})])^2} + O\left(\frac{1}{N_1^2Kh}+\frac{h}{N_1K}\right) \\ &= \frac{p_j(x)(1-p_j(x))R(\phi)}{N_1Kh} + O\left(\frac{1}{N_1^2Kh}+\frac{h}{N_1K}\right), \end{aligned}$$

which gives the variance expression.

*Proof of Theorem 2* From (A.1), (A.2) and Theorem 1, we obtain

$$\begin{aligned} \hat{p}_{j,\lambda}(x;h) - E[\hat{p}_{j,\lambda}(x;h)] &= \frac{f_K(\mathbf{p}_{Kj})^{-\lambda}}{K}\sum_{i=1}^{K} p_j(x_i)^\lambda\phi_h(x_i-x)(\bar{Y}_{ij}-p_j(x_i)) + o_P\left(\frac{1}{\sqrt{N_1Kh}}\right) \\ &= \frac{1}{K}\sum_{i=1}^{K}\phi_h(x_i-x)(\bar{Y}_{ij}-p_j(x_i)) + o_P\left(\frac{1}{\sqrt{N_1Kh}}\right). \end{aligned}$$

From Lemma 4, it follows for any $r$-vector that $\alpha$, $\sqrt{N_1Kh}\alpha^T\{\hat{\mathbf{p}}_\lambda(x;h) - E[\hat{\mathbf{p}}_\lambda(x;h)]\}$ converges in distribution to $N(0, R(\phi)\alpha^T\Sigma(x)\alpha)$. Since $\sqrt{N_1Kh}\alpha^T\{E[\hat{\mathbf{p}}_\lambda(x;h)] - \mathbf{p}(x) - \lambda(2N_1)^{-1}\mathbf{b}_1(x)\}$ converges to a constant $\rho\alpha^T\mathbf{b}_2(x)$, the proof has been completed by Slutsky's theorem.

*Proof of Theorem 3* The Taylor expansion of $\bar{\lambda}_{\rm opt}(g_1)$ around $E[\bar{B}_1(g_1)]$ is given as

$$\bar{\lambda}_{\rm opt}(g_1) = -\frac{\bar{B}_2(g_1)}{\bar{B}_1(g_1)} = -\frac{\bar{B}_2(g_1)}{E[\bar{B}_1(g_1)]} + \frac{\bar{B}_2(g_1)(\bar{B}_1(g_1) - E[\bar{B}_1(g_1)])}{E[\bar{B}_1(g_1)]^2} - \cdots.$$

Since

$$\begin{aligned} E[\bar{B}_{1j}(g_1)] &= E\left[\int_{\delta_1}^{1-\delta_2} \bar{\eta}_j(x;g_1)^2 dx\right] \\ &= E\left[\int_{\delta_1}^{1-\delta_2} \frac{\bar{p}_j^{(1)}(x;g_1)^4}{\bar{p}_j(x;g_1)} dx\right] - 2E\left[\int_{\delta_1}^{1-\delta_2} \bar{p}_j^{(1)}(x;g_1)^2 \sum_{\ell=1}^r \frac{\bar{p}_\ell^{(1)}(x;g_1)^2}{\bar{p}_\ell(x;g_1)} dx\right] \\ &\quad + E\left[\int_{\delta_1}^{1-\delta_2} \bar{p}_j(x;g_1)^2 \left(\sum_{\ell=1}^r \frac{\bar{p}_\ell^{(1)}(x;g_1)^2}{\bar{p}_\ell(x;g_1)}\right)^2 dx\right], \end{aligned}$$

through direct calculations using Lemma 1, it follows that

$$E[\bar{B}_1(g_1)] = B_{1j} + g_1^2 H_{11j} + \frac{H_{12j}}{N_1 K g_1^3} + O\left(g_1^4 + \frac{1}{N_1 K g_1}\right),$$

where

$$\begin{aligned} H_{11j} &= 2\int_{\delta_1}^{1-\delta_2} \eta_j(x) d_j(x) dx, \\ H_{12j} &= \frac{R(\phi^{(1)})}{N_1 K g_1^3} \int_{\delta_1}^{1-\delta_2} \left[6\frac{(1-2p_j(x))p_j^{(1)}(x)^2}{p_j(x)} - 2\{p_j(x) - r p_j(x)^2\} \sum_{\ell=1}^r \frac{p_\ell^{(1)}(x)^2}{p_\ell(x)} + 2r p_\ell^{(1)}(x)^2\right] dx \\ Q_j(x) &= d_j(x) - p_j(x) \sum_{\ell=1}^r d_\ell(x), \end{aligned}$$

in which

$$d_j(x) = -\frac{\mu_3(\phi^{(1)})}{3} \frac{p_j^{(1)}(x) p_j^{(3)}(x)}{p_j(x)} - \frac{\mu_2(\phi)}{2} \frac{p_j^{(1)}(x)^2 p_j^{(2)}(x)}{p_j(x)^2}.$$

Hence

$$E[\bar{B}_1(g_1)] = \sum_{j=1}^r \bar{B}_{1j}(g_1) = B_1 + g_1^2 H_{11} + \frac{H_{12}}{N_1 K g_1^3} + O\left(g_1^4 + \frac{1}{N_1 K g_1}\right),$$

where

$$\begin{aligned} H_{11} &= \sum_{j=1}^r H_{11j} = 2\sum_{j=1}^r \int_{\delta_1}^{1-\delta_2} \eta_j(x) Q_j(x) dx, \\ H_{12} &= \sum_{j=1}^r H_{12j} = R(\phi^{(1)}) \int_{\delta_1}^{1-\delta_2} \left\{(4 + r\sum_{j=1}^r p_j(x)^2) \sum_{j=1}^r \frac{p_j^{(1)}(x)^2}{p_j(x)} + 2(r-6)\sum_{j=1}^r p_j^{(1)}(x)^2\right\} dx. \end{aligned}$$

On the other hand, since

$$\begin{aligned} E[\bar{B}_{2j}(g_1)] &= E\left[\int_{\delta_1}^{1-\delta_2} \bar{\eta}_j(x;g_1)\bar{p}_j^{(2)}(x;g_1)dx\right] \\ &= E\left[\int_{\delta_1}^{1-\delta_2} \frac{\bar{p}_j^{(1)}(x;g_1)^2\bar{p}_j^{(2)}(x;g_1)}{\bar{p}_j(x;g_1)}dx\right] - E\left[\int_{\delta_1}^{1-\delta_2} \bar{p}_j(x;g_1)\bar{p}_j^{(2)}(x;g_1)\sum_{\ell=1}^{r}\frac{\bar{p}_\ell^{(1)}(x;g_1)^2}{\bar{p}_\ell(x;g_1)}dx\right], \end{aligned}$$

through direct calculations using Lemma 1, it follows that

$$E[\bar{B}_{2j}(g_1)] = B_{2j} + g_1^2 H_{21j} + \frac{H_{22j}}{N_1 K g_1^3} + O\left(g_1^4 + \frac{1}{N_1 K g_1}\right),$$

where

$$\begin{aligned} H_{21j} &= \int_{\delta_1}^{1-\delta_2} \{d_j(x)p_j^{(2)}(x) + \frac{\mu_4(\phi^{(2)})}{24}\eta_j(x)p_j^{(4)}(x)\}dx, \\ H_{22j} &= \int_{\delta_1}^{1-\delta_2} \left[R(\phi^{(1)})\{p_j^{(2)}(x) - rp_j(x)p_j^{(2)}(x)\} - \mu(\phi^{(2)})p_j(x)(1-p_j(x))\sum_{\ell=1}^{r}\frac{p_\ell^{(1)}(x)^2}{p_\ell(x)}\right]dx, \end{aligned}$$

in which $\mu(f) = \int_{-1}^{1} f(t)\phi(t)dt$. Hence we obtain

$$E[\bar{B}_2(g_1)] = \sum_{j=1}^{r}\bar{B}_{2j}(g_1) = B_2 + g_1^2 H_{21} + \frac{H_{22}}{N_1 K g_1^3} + O\left(g_1^4 + \frac{1}{N_1 K g_1}\right),$$

where

$$\begin{aligned} H_{21} &= \sum_{j=1}^{r}\int_{\delta_1}^{1-\delta_2}\left(d_j(x)p_j^{(2)}(x) + \frac{\mu_4(\phi^{(2)})}{24}\eta_j(x)p_j^{(4)}(x)\right)dx, \\ H_{22} &= -\int_{\delta_1}^{1-\delta_2}\left\{R(\phi^{(1)})\sum_{j=1}^{r}p_j(x)p_j^{(2)}(x) + \mu(\phi^{(2)})(1-\sum_{j=1}^{r}p_j(x)^2)\sum_{j=1}^{r}\frac{p_j^{(1)}(x)^2}{p_j(x)}\right\}dx. \end{aligned}$$

Moreover, through straightforward calculations, the $E[\bar{\lambda}_{\text{opt}}(g_1)]$ is written as

$$\begin{aligned} E\left[-\frac{\bar{B}_2(g_1)}{\bar{B}_1(g_1)}\right] &= -\frac{E[\bar{B}_2(g_1)]}{E[\bar{B}_1(g_1)]} + \frac{E[\bar{B}_2(g_1)(\bar{B}_1(g_1) - E[\bar{B}_1(g_1)])]}{E[\bar{B}_1(g_1)]^2} - \cdots \\ &= -\frac{E[\bar{B}_2(g_1)]}{E[\bar{B}_1(g_1)]} + O\left(\frac{1}{N_1 K^2 g_1^5}\right). \end{aligned}$$

Hence we have

$$E[\bar{\lambda}_{\text{opt}}(g_1)] = -\frac{B_2}{B_1} + g_1^2\Delta_{11} + \frac{\Delta_{12}}{N_1 K g_1^3} + O\left(g_1^4 + \frac{1}{N_1 K g_1}\right),$$

where

$$\Delta_{1t} = -\frac{1}{B_1}\left(H_{t2} - \frac{B_2}{B_1}H_{t1}\right), \quad t = 1, 2.$$

On the other hand, we can also obtain from calculations using Lemma 1, 6 and 7 that

$$\begin{aligned} V[\bar{\lambda}_{\mathrm{opt}}(g_1)] &= \frac{1}{E[\bar{B}_1(g_1)]^2} V[\bar{B}_2(g_1)] + O\left(\frac{1}{N_1 K^2 g_1^5}\right) \\ &= \frac{1}{B_1^2} V\left[\sum_{j=1}^{r} \int_{\delta 1}^{1-\delta 2} p_j(x)^{-1} \bar{p}_j^{(1)}(x;g_1)^2 \bar{p}_j^{(2)}(x;g_1) dx\right] + O\left(\frac{1}{N_1 K^2 g_1^5}\right) \\ &= \frac{\Delta_{13}}{N_1 K^3 g_1^{10}} + O\left(\frac{1}{N_1^2 K^3 g_1^{10}}\right), \end{aligned}$$

where

$$\Delta_{13} = \frac{1}{B_1^2}\left\{6R(\phi^{(1)} * \phi^{(1)})\mu_0(\phi^{(2)} * \phi^{(2)}) + 12R(\phi^{(1)} * \phi^{(2)})\mu_0(\phi^{(1)} * \phi^{(1)})\right\} \int_{\delta_1}^{1-\delta_2} \mathbf{p}(x)^T \Sigma(x) \mathbf{p}(x) dx,$$

which completes the proof.

*Proof of Theorem 4* The Taylor expansion of $\bar{\Theta}(g_2)$ around $E[\bar{B}_1(g_2)]$ is given as

$$\begin{aligned} \bar{\Theta}(g_2) &= \bar{B}_3(g_2) - \frac{\bar{B}_2(g_2)^2}{\bar{B}_1(g_2)} \\ &= \bar{B}_3(g_2) - \frac{\bar{B}_2(g_2)^2}{E[\bar{B}_1(g_2)]} + \frac{\bar{B}_2(g_2)^2}{E[\bar{B}_1(g_2)]^2}(\bar{B}_1(g_2) - E[\bar{B}_1(g_2)]) - \cdots. \end{aligned}$$

the expectation of $\bar{\Theta}(g_2)$ can be expressed as

$$E[\bar{\Theta}(g_2)] = E[\bar{B}_3(g_2)] - E[\bar{B}_2(g_2)]^2 E[\bar{B}_1(g_2)]^{-1} + O\left(\frac{1}{N_1 K g_2^3}\right).$$

Through direct calculatons using Lemma 1, we obtain

$$E[\bar{B}_{3j}(g_2)] = B_{3j} + g_2^2 H_{31j} + \frac{H_{32j}}{N_1 K g_2^5} + O\left(g_2^4 + \frac{1}{N_1 K g_2^3}\right),$$

where

$$H_{31j} = \frac{\mu_4(\phi^{(2)})}{12} \int_{\delta_1}^{1-\delta_2} p_j^{(2)}(x) p_j^{(4)}(x) dx \ , \quad H_{32j} = R(\phi^{(2)}) \int_{\delta_1}^{1-\delta_2} p_j(x)(1 - p_j(x)) dx.$$

Hence

$$E[\bar{B}_3(g_2)] = B_{3j} + g_2^2 H_{31} + \frac{H_{32}}{N_1 K g_2^5} + O\left(g_2^4 + \frac{1}{N_1 K g_2^3}\right),$$

where

$$H_{31} = \sum_{j=1}^{r} H_{31j} = \frac{\mu_4(\phi^{(2)})}{12} \sum_{j=1}^{r} \int_{\delta_1}^{1-\delta_2} p_j^{(2)}(x) p_j^{(4)}(x) dx, \quad H_{32} = \sum_{j=1}^{r} H_{32j} = \int_{\delta_1}^{1-\delta_2} \{1 - \sum_{j=1}^{r} p_j(x)^2\} dx.$$

Since

$$E[\bar{B}_2(g_2)]^2 = B_2^2 + 2g_2^2 B_2 H_{21} + O\left(g_2^4 + \frac{1}{N_1 K g_2^3}\right),$$

these are combined to give

$$E[\bar{\Theta}(g_2)] = g_2^2\Delta_{21} + \frac{\Delta_{22}}{N_1 K g_2^5} + O\left(g_2^4 + \frac{1}{N_1 K g_2^3}\right),$$

where $\Delta_{21} = \{B_2/B_1\}^2 H_{11} - 2\{B_2/B_1\}H_{21} + H_{31}$ and $\Delta_{22} = H_{32}$. From calculations using Lemma 1, 6 and 8, evaluations of the variance is progressed as

$$\begin{aligned} V[\bar{\Theta}(g_2)] &= V\left[\sum_{j=1}^{r} \bar{B}_{3j}(g_2)dx\right] + O\left(\frac{1}{N_1 K^2 g_2^8}\right) \\ &= V\left[\sum_{j=1}^{r} \int_{\delta_1}^{1-\delta_2} \bar{p}_j^{(2)}(x; g_2)^2 dx\right] + O\left(\frac{1}{N_1 K^2 g_2^8}\right) \\ &= \frac{\Delta_{23}}{N_1 K^2 g_2^9} + O\left(\frac{1}{N_1^2 K^2 g_2^9}\right), \end{aligned}$$

where

$$\Delta_{33} = 4R(\phi^{(2)} * \phi^{(2)}) \int_{\delta_1}^{1-\delta_2} \mathbf{p}(x)^T \Sigma(x) \mathbf{p}(x) dx.$$

The proof has been completed.

**Lemma 1** *If the random vector* $\mathbf{Y} = (Y_1, ..., Y_r)$ *follows the multinomial distribution* $\mathrm{MN}(p_1, ..., p_r; N)$, *then*

$$E\left[\prod_{k=1}^{d} \bar{Y}_{j_k}^{m_k}\right] = \prod_{k=1}^{d} p_{j_k}^{m_k} \left\{1 + \frac{1}{2N}\left(\sum_{k=1}^{d} m_{j_k}(m_{j_k} - 1)\frac{1 - p_{j_k}}{p_{j_k}} - 2\sum_{i<k} m_{j_i} m_{j_k}\right)\right\} + O\left(\frac{1}{N^2}\right),$$

*where* $\bar{Y}_j = Y_j/N, j = 1, ..., r$.

*Proof of Lemma 1* The quantity $\bar{Y}_{j_k}^{m_k}$ is written as

$$\bar{Y}_{j_k}^{m_k} = p_{j_k}^{m_k} + \binom{m_k}{m_k - 1} p_{j_k}^{m_k - 1}(\bar{Y}_{j_k} - p_{j_k}) + \binom{m_k}{m_k - 2} p_{j_k}^{m_k - 2}(\bar{Y}_{j_k} - p_{j_k})^2 + \cdots + (\bar{Y}_{j_k} - p_{j_k})^{m_k}.$$

Since $E[\bar{Y}_{j_k} - p_{j_k}] = 0$, $E[(\bar{Y}_{j_k} - p_{j_k})^2] = p_{j_k}(1 - p_{j_k})/N$, $E[(\bar{Y}_{j_i} - p_{j_i})(\bar{Y}_{j_k} - p_{j_k})] = -p_{j_i} p_{j_k}/N$,

$$\begin{aligned} E[(\bar{Y}_{j_k} - p_{j_k})^3] &= \frac{p_{j_k}(1 - p_{j_k})(1 - 2p_{j_k})}{N^2}, \\ E[(\bar{Y}_{j_k} - p_{j_k})(\bar{Y}_{j_\ell} - p_{j_\ell})^2] &= -\frac{p_{j_k} p_{j_\ell}(1 - 2p_{j_\ell})}{N^2}, \\ E[(\bar{Y}_{j_k} - p_{j_k})(\bar{Y}_{j_\ell} - p_{j_m})(\bar{Y}_{j_\ell} - p_{j_m})] &= \frac{2p_{j_k} p_{j_\ell} p_{j_m}}{N^2}, \\ E[(\bar{Y}_{j_k} - p_{j_k})^4] &= \frac{p_{j_k}(1 - p_{j_k})\{3(n-2)p_{j_k}(1 - p_{j_k}) + 1\}}{N^3}, \end{aligned}$$

and furthermore, for $j_1, ..., j_k \in \{1, .., r\} (k \geq 4)$

$$\begin{aligned} E\left[\prod_{i=1}^{k}(\bar{Y}_{j_i} - p_{j_i})\right] &\leq E\left[|\prod_{i=1}^{4}(\bar{Y}_{j_i} - p_{j_i})|\right] \\ &\leq \prod_{i=1}^{4} E\left[(\bar{Y}_{j_i} - p_{j_i})^4\right]^{1/4} = O\left(\frac{1}{N^2}\right), \end{aligned}$$

the lemma follows.

**Lemma 2** *Let $\{\mathbf{X}_{ni}, 1 \leq i \leq k_n, n \geq 1\}$ be a double array of independent r-dimensional random vectors with $E[\sum_{i=1}^{k_n} \alpha^T \mathbf{X}_{ni}] = \mathbf{0}$ and $V[\sum_{i=1}^{k_n} \alpha^T \mathbf{X}_{ni}] \to \alpha^T \Sigma \alpha$ for any r-vector $\alpha$. If for any $n$ and $i$, there is a finite constant $M_{ni}$ such that $|\alpha^T \mathbf{X}_{ni}| \leq M_{ni}$* a.e., *and* $\max_{1 \leq i \leq k_n} M_{k_n} \to 0$, *then*

$$\sum_{i=1}^{k_n} \alpha^T \mathbf{X}_{ni} \to_d N(0, \alpha^T \Sigma \alpha).$$

Put $\breve{\mathbf{p}}(x;h) = (\breve{p}_1(x;h), ..., \breve{p}_j(x;h))^T$, where $\breve{p}_j(x;h) = K^{-1} \sum_{i=1}^{K} \phi_g(x_i - x) \bar{Y}_{ij}, j = 1, .., r$. Then $\breve{p}_j(x;h) - \bar{p}_j(x;h) = o_P((\sqrt{N_1 K h})^{-1}), j = 1, .., r$.

**Lemma 3** *Under assumptions 1-3, we have for any r-vector $\alpha = (\alpha_1, ..., \alpha_r)^T$*

$$V[\alpha^T \breve{\mathbf{p}}(x;h)] = \frac{R(\phi) \alpha^T \Sigma(x) \alpha}{N_1 K h} + O\left(\frac{h}{N_1 K} + \frac{1}{N_1 K^2 h^2}\right).$$

*Proof of Lemma 3* It follows from direct calculations that

$$V[\breve{p}_j(x;h)] = \frac{p_j(x)(1 - p_j(x)) R(\phi)}{N_1 K h} + O\left(\frac{h}{N_1 K} + \frac{1}{N_1 K^2 h^2}\right),$$

$$Cov[\breve{p}_i(x;h), \breve{p}_i(x;h)] = -\frac{p_i(x) p_j(x) R(\phi)}{N_1 K h} + O\left(\frac{h}{N_1 K} + \frac{1}{N_1 K^2 h^2}\right)$$

for any $i, j (i \neq j)$.

**Lemma 4** *Under assumptions 1-3, we have for any r-vector $\alpha = (\alpha_1, ..., \alpha_r)^T$*

$$\sqrt{N_1 K h^5} \alpha^T \{\breve{\mathbf{p}}(x;h) - E[\breve{\mathbf{p}}(x;h)]\} \to_d N(0, R(\phi) \alpha^T \Sigma(x) \alpha).$$

*Proof of Lemma 4* Put $S_K = \sqrt{N_1 K h^5} \alpha^T \{\breve{\mathbf{p}}(x;h) - E[\breve{\mathbf{p}}(x;h)]\} = \sum_{i=1}^{K} \sum_{j=1}^{r} a_j X_{Kij}$, where

$$X_{Kij} = \frac{\sqrt{N_1 K h^5}}{K} \phi_h(x_i - x)(\bar{Y}_{ij} - E[\bar{Y}_{ij}]).$$

From $|\bar{Y}_{ij} - E[\bar{Y}_{ij}]| \leq 1$ a.s. and Lemma 3, we immediately obtain the result.

For any $k$ random variables $X_1, ..., X_k$, let $I$ denote the identity operator, that is, $If = f$, and define operators $Q_j$, $j = 1, ..., k$ by

$$Q_j f = E[f(X_{t_1}, ..., X_{t_k}) | X_{t_\alpha}, \alpha \in N_{-j}]$$

where $N_{-j} = \{1, ..., k\} - \{j\}$, $f$ is any Borel function on $R^n$ with $E[|f(X_1, ..., X_k)|] < \infty$ and $(t_1, ..., t_k)$ is any permutaion of $\{1, ..., k\}$.

**Lemma 5** *It holds that for any k random variables $X_1, ..., X_k$, let $I$ denote the identity operator, that is, $If = f$, and define operators $Q_j$, $j = 1, ..., k$ by*

$$\begin{aligned} f(X_{t_1}, ..., X_{t_k}) &= E[f] + \sum_{j=1}^{k} f_1(X_{t_j}) \\ &\quad + \sum_{1 \le j_1 < j_2 \le k} f_2(X_{t_{j_1}}, X_{t_{j_2}}) + \cdots \\ &\quad + \sum_{1 \le j_1 < \cdots < j_{k-1} \le k} f_{k-1}(X_{t_{j_1}}, ..., X_{t_{j_{k-1}}}) \\ &\quad + f_k(X_{t_1}, ..., X_{t_k}), \end{aligned}$$

*where* $f_i(X_{t_{j_1}}, ..., X_{t_{j_i}}) = \left[\prod_{j \in \{j_1, .. j_i\}} (I - Q_{t_j}) \prod_{j \in N - \{j_1, ..., j_i\}} Q_{t_j}\right] f$.

*Proof of Lemma 5* Noting that

$$[\prod_{i=1}^{k} Q_i] f = E[f | X_\alpha, \alpha \ne 1, ..., k] = E[f],$$

it follows from

$$\begin{aligned} f(X_{t_1}, ..., X_{t_k}) &= \prod_{j=1}^{k} [(I - Q_j) + Q_j] f \\ &= \sum_{i=1}^{k} \sum_{1 \le j_1 < \cdots < j_i \le k} \left[ \prod_{j \in \{j_1, .. j_i\}} (I - Q_{t_j}) \prod_{j \in N - \{j_1, ..., j_i\}} Q_{t_j} \right] f. \end{aligned}$$

**Lemma 6** *Let $X_1, ..., X_K$ denote $K$ independent random variables with $\mu_i = E[X_i], i = 1, .., K$, and $L(t_1, ..., t_k)$ denote a function on $A_k = \{(t_1, ..., t_k) \in N^k : t_i \ne t_j, 1 \le i, j \le k\}, 1 \le k \le K$, where $N = \{1, ..., K\}$. Then*

$$\begin{aligned} &\sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) X_{t_1} \cdots X_{t_k} \\ &= \sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) \mu_{t_1} \cdots \mu_{t_k} \\ &\quad + \sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) \sum_{j=1}^{k} (X_{t_j} - \mu_{t_j}) \prod_{i \ne j} \mu_{t_i} \\ &\quad + \sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) \sum_{1 \le j_1 < j_2 \le k} \prod_{j \in \{j_1, j_2\}} (X_{t_j} - \mu_{t_j}) \prod_{j \in N - \{j_1, j_2\}} \mu_j \\ &\quad + \cdots \\ &\quad + \sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) \sum_{1 \le j_1 < \cdots < j_{k-1} \le k} \prod_{j \in \{j_1, ..., j_{k-1}\}} (X_{t_j} - \mu_{t_j}) \mu_{t_{j_k}} \\ &\quad + \sum_{(t_1, ..., t_k) \in A_k} L(t_1, ..., t_k) (X_{t_1} - \mu_{t_1}) \cdots (X_{t_1} - \mu_{t_k}), \end{aligned}$$

*Moreover for any* $(j_1,...,j_k) \in A_k$, *it holds that* $E[\prod_{j\in\{j_1,...,j_i\}}(X_{t_j}-\mu_{t_j})\prod_{j\in N-\{j_1,...,j_i\}}\mu_{t_j}]=0$.

*Proof of Lemma 6* By puting $f(X_{t_1},...,X_{t_k}) = X_{t_1}\cdots X_{t_k}$ in Lemma 5, the decomposition can be obtained. The latter equations are clear.

Put

$$\breve{\mathbf{D}}_\gamma(g) = (\breve{D}_{1\gamma}(g),...,\breve{D}_{r\gamma}(g))^T, \gamma = 1,2,$$

where $\breve{D}_{1j}(g) = \int_{\delta_2}^{1-\delta_2} p_j(x)^{-1}\breve{p}_j^{(2)}(x;g)\breve{p}_j^{(1)}(x;g)^2dx$ and $\breve{D}_{2j}(g) = \int_{\delta_2}^{1-\delta_2}\breve{p}_j^{(2)}(x;g)^2dx$.

**Lemma 7** *Under assumptions 2-4, we have for any r-vector* $\alpha = (\alpha_1,...,\alpha_r)^T$

$$V[\alpha^T\breve{\mathbf{D}}_1(g_1)] = \frac{\alpha^T\Sigma_1\alpha}{N_1K^3g_1^{10}} + O\left(\frac{1}{N_1^2K^3g_1^{10}}\right),$$

*where* $\Sigma_1 = 6R(\phi^{(1)}*\phi^{(1)})^2R(\phi^{(2)}*\phi^{(2)})\int_{\delta_1}^{1-\delta_2}(\mathbf{P}(x)^2)^T\Sigma(x)\mathbf{P}(x)^2dx$.

*Proof of Lemma 7* . We can write $\breve{D}_{1j}(g_1) = \sum_{(t_1,t_2,t_3)\in N^3} L_K(t_1,t_2,t_3)\bar{Y}_{i_{t_1}j}\bar{Y}_{i_{t_2}j}\bar{Y}_{i_{t_3}j}$, where $L_K(t_1,t_2,t_3) = (K^3g_1^7)^{-1}\int_{\delta_1}^{1-\delta_2}\phi_{g_1}^{(2)}(x_{t_1}-x)\phi_{g_1}^{(1)}(x_{t_2}-x)\phi_{g_1}^{(1)}(x_{t_3}-x)p_j(x)^{-1}dx$. By using the decomposition of Lemma 6, we obtain the following expression

$$\breve{D}_{1j}(g_1) = \breve{D}_{1j}^*(g_1) + T,$$

where

$$\begin{aligned}\breve{D}_{1j}^*(g_1) &= \sum_{(t_1,t_2,t_3)\in A_3} L_K(t_1,t_2,t_3)p_{t_1j}p_{t_2j}p_{t_3j}\\ &= F_{1j}+F_{2j}+F_{3j},\end{aligned}$$

in which

$$\begin{aligned}F_{1j} &= \sum_{(t_1,t_2,t_3)\in A_3} L_K(t_1,t_2,t_3)\sum_{\alpha=1}^{3}(\bar{Y}_{t_\alpha j}-p_{t_\alpha j})\prod_{d\neq\alpha}p_{t_dj}\\ F_{2j} &= \sum_{(t_1,t_2,t_3)\in A_3} L_K(t_1,t_2,t_3)\sum_{1\le\alpha_1<\alpha_2\le 3}\prod_{\alpha\in\{\alpha_1,\alpha_2\}}(\bar{Y}_{t_\alpha j}-p_{t_\alpha j})\prod_{\alpha\in N-\{\alpha_1,\alpha_2\}}p_{t_\alpha j}\\ F_{3j} &= \sum_{(t_1,t_2,t_3)\in A_3} L_K(t_1,t_2,t_3)(\bar{Y}_{t_1j}-p_{t_1j})(\bar{Y}_{t_2j}-p_{t_2j})(\bar{Y}_{t_3j}-p_{t_3j}),\end{aligned}$$

and

$$\begin{aligned}T &= \breve{D}_{1j}(g_1)-\breve{D}_{1j}^*(g_1)\\ &= \sum_{(t_1,t_2,t_3)\in N^3-A_3} L_K(t_1,t_2,t_3)E[\bar{Y}_{t_1}\bar{Y}_{t_2}\bar{Y}_{t_3}]\\ &= T_1+T_2+T_3-2T_4,\end{aligned}$$

in which

$$T_1 = \sum_{i_1,i_2} L_K(i_1,i_1,i_2)(\bar{Y}_{i_1j}^2\bar{Y}_{i_2j} - E[\bar{Y}_{i_1j}^2\bar{Y}_{i_2j}]), \qquad T_2 = \sum_{i_1,i_2} L_K(i_1,i_2,i_1)(\bar{Y}_{i_1j}^2\bar{Y}_{i_2j} - E[\bar{Y}_{i_1j}^2\bar{Y}_{i_2j}])$$

$$T_3 = \sum_{i_1,i_2} L_K(i_2,i_1,t_1)(\bar{Y}_{i_1j}^2\bar{Y}_{i_2j} - E[\bar{Y}_{i_1j}^2\bar{Y}_{i_2j}]), \qquad T_4 = \sum_{i=1}^{K} L_K(i,i,i)(\bar{Y}_{ij}^3 - E[\bar{Y}_{ij}^3]).$$

Through a similar manner given in Ruppert, Sheather and Wand [17], we obtain

$$\begin{aligned} V[F_1] &= \frac{c}{N_1K^3g_1^{10}}\int_{\delta_1}^{1-\delta_2} p_j(x)^3(1-p_j(x))dx + O\left(\frac{1}{N_1K}\right), \\ V[F_2] &= O\left(\frac{1}{N_1^2K^3g_1^{10}}\right), \\ V[F_3] &= O\left(\frac{1}{N_1^3K^3g_1^{10}}\right) \end{aligned}$$

and $V[T] = O((N_1K^4g_1^{11})^{-1})$, where

$$c = 6R(\phi^{(1)} * \phi^{(1)})\mu_0(\phi^{(2)} * \phi^{(2)}) + 12R(\phi^{(1)} * \phi^{(2)})\mu_0(\phi^{(1)} * \phi^{(1)}).$$

Noting that $Cov[F_i, F_j] = 0 (i \neq j)$, we obtain

$$V[\breve{D}_{1j}(g_1)] = \frac{c}{N_1K^3g_1^{10}}\int_{\delta_1}^{1-\delta_2} p_j(x)^3(1-p_j(x))dx + O\left(\frac{1}{N_1^2K^3g_1^{10}}\right).$$

In the same manner, we obtain

$$Cov[\breve{D}_{1i}(g_1), \breve{D}_{1j}(g_1)] = -\frac{c}{N_1K^3g_1^{10}}\int_{\delta_1}^{1-\delta_2} p_i(x)^2p_j(x)^2dx + O\left(\frac{1}{N_1^2K^3g_1^{10}}\right).$$

**Lemma 8** *Under assumptions 2, 3 and 5, we have for any r-vector* $\alpha = (\alpha_1, ..., \alpha_r)^T$

$$V[\alpha^T\breve{\mathbf{D}}_2(g_2)] = \frac{\alpha^T\Sigma_2\alpha}{N_1K^2g_2^9} + O\left(\frac{1}{N_1^2K^2g_2^9}\right),$$

*where* $\Sigma_2 = 4R(\phi^{(2)} * \phi^{(2)})\int_{\delta_1}^{1-\delta_2} \mathbf{P}(x)^T\Sigma(x)\mathbf{P}(x)dx$.

*Proof of Lemma 8* We can obtain the result by calculating through the simular manner in the proof of Lemma 7

**Acknowledgement:** A part of this work was carried out by the second author while enjoying the hospitality of the Department of Statistics at the University of New South Wales.

## References


[1] Aerts, M., Augustyns, I. and Janssen, P. (1997) Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*. **8**, 127-147.

[2] Augustyns, I. and Wand, M. P. (1998). Bandwidth selection for local polynomial smoothing of multinomial data. *Computational Statistics*, **13**, 447-461.

[3] Albert, J. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669-679.

[4] Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry*. Translations of Mathematical Monographs, **191**, American Mathematical Society / Oxford Univ. Press.

[5] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549-559.

[6] Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[7] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal. Statistical Society, Series B **46**, 440-464.

[8] Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning*, Springer.

[9] Jones, M. C., Hjort, N. L., Harris and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88**, 865-873.

[10] Karunamuni, R. J. and Alberts, T. (2003). A Locally Adaptive Generalized Reflection Method of Boundary Correction in Kernel Density Estimation. submitted.

[11] Müller, H. and Schmitt, T. (1988). Kernel and Probit Estimates in Quantal Bioassay. *Journal of the American Statistical Association*, **83**, 750-759.

[12] Naito, K. (2001). On a certain class of nonparametric density estimators with reduced bias. *Statistics and Probability Letters*, **51**, 71-78.

[13] Naito, K. (2004). Semiparametric density estimation by local $L_2$-fitting. *Annals of Statistics*, **32**, 1162-1191.

[14] Okumura, H. and Naito, K. (2003). Kernel smoothing in quantal bioassay. *Japanese Journal of Applied Statistics*. **32**, 127-144.

[15] Okumura, H. and Naito, K. (2004). Weighted kernel estimators in nonparametric binomial regression. *Journal of Nonparametric Statistics*, **16**, 39-62.

[16] Okumura, H. and Naito, K. (2004). Bandwidth selection for nonparametric binomial regression. submitted.

[17] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257-1270.

[18] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

[19] Tutz, G. (1990). Smoothed Categorical Regression Based on Direct Kernel Estimates. *Journal of Statistical Computing and Simulation*, **36**, 139-156.

[20] Tutz, G. (1995). Smoothing for categorical data: Discrete kernel regression and local likelihood approaches. In: H. H. Bock, W. Polasek (Eds.), *Data Analysis and Information Systems*, 261-271, Springer-Verlag.

[21] Tutz, G. and Kauermann, G. (1995). Varying conffidients in multivariate generalized linear models: A local likelihood approach. *Forschungsberichte des Fachbereichs Informatik*, TU Berlin, 95/4.

[22] Tutz, G. and Scholz, T. (2004). Semiparametric modelling of multicategorical data. *Journal of Statistical Computation and Simulation*, **74**, 183-200.

[23] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London.