Electronic Journal of Statistics
Vol. 0 (0000) 1
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Supplementary Material: On Forward Sufficient Dimension Reduction for Categorical and Ordinal Responses

Harris Quach and Bing Li

The Pennsylvania State University, University Park, USA.
e-mail: hxq5@psu.edu; bxl9@psu.edu

Contents

1. Derivations for Categorical and Ordinal-Categorical Linear Exponential Family

We derive the multivariate link functions for categorical and ordinal-categorical variables and express their distributions as linear exponential families using the multivariate links.

1.1. Multivariate Logit link and Multivariate Expit function

For notation, refer to section 3 of the paper. The density and log-likelihood of S are

$$\begin{aligned} f(S;p) &\propto \left(\prod_{j=1}^{m} p_j^{S^j}\right)\left(1-p_1-\cdots-p_m\right)^{1-S^1-\cdots-S^m}, \\ \ell(\theta;S) &= \theta(p)^{\top}S-\log(1-\mathbf{1}_m^{\top}p). \end{aligned}$$

respectively, where

$$\theta(p)=\log\left(\frac{p}{1-\mathbf{1}_m^{\top}p}\right),$$

defines the canonical link, which is known as the Multivariate Logit link. The inverse canonical link is

$$p=(I_m+e^{\theta}\mathbf{1}_m^{\top})^{-1}e^{\theta}=e^{\theta}-\frac{e^{\theta}\mathbf{1}_m^{\top}e^{\theta}}{1+\mathbf{1}_m^{\top}e^{\theta}}=\frac{e^{\theta}}{1+\mathbf{1}_m^{\top}e^{\theta}},$$

where the second equality follows from the Woodbury formula. This inverse of the multivariate logistic link is called the Multivariate Expit function.

1.2. Adjacent-Categories (Ad-Cat) Logit Link

Recall that the density of T, via the multinomial distribution of S, is

$$f(T;p)\propto\prod_{j=1}^{m_0}(p_j)^{S^j}=p_1\left(\frac{p_2}{p_1}\right)^{T^1}\cdots\left(\frac{p_{m_0}}{p_{m_0-1}}\right)^{T^{m_0-1}},$$

since $T^0=1$, $T^{m_0}=0$, and $S^j=T^{j-1}-T^j$. The log-likelihood for $T^1,\dots,T^m$ is then

$$\ell(p;T)=\sum_{j=1}^{m}T^j\log\left(\frac{p_{j+1}}{p_j}\right)+\log(p_1).$$

Letting $\theta(p)=(\theta_1(p),\dots,\theta_m(p))$, where, for $j=1,...,m$,

$$\theta_j(p)=\log\left(\frac{p_{j+1}}{p_j}\right),$$

we rewrite the log-likelihood as $\ell(p;T)=\theta(p)^{\top}T+\log(p_1)$. To derive the canonical link, we need to determine the relation between θ and the mean of T. We derive the mean of T and the covariance of T below.

For $j=1,\dots,m$, let $\gamma_j=p_1+\cdots+p_j$. Let $\gamma_0=0$, $\gamma_{m_0}=1$. Let $\tau_j=1-\gamma_j$ for $j=0,\dots,m_0$. Let $\tau=(\tau_1,\dots,\tau_m)^{\top}$. Since $T^j=\mathbb{1}\{Y>j\}$, we have

$$E(T^j)=P(Y>j)=1-\gamma_j=\tau_j.$$

Since, for $j \leq l$, $T^j T^l = T^l$, we have

$$E(T^j T^l) = E(T^l) = 1 - \gamma_l = \tau_l.$$

Likewise, for $j > l$, we have $E(T^j T^l) = E(T^j) = \tau_j$. Hence

$$\operatorname{var}(T^j) = \tau_j(1-\tau_j), \quad \operatorname{cov}(T^j, T^l) = \tau_{\max\{j,l\}}(1-\tau_{\min\{j,l\}}).$$

Summarizing the above results in matrix form, we have

$$E(T) = \tau, \quad \operatorname{var}(T) = \Gamma - \tau\tau^{\top}, \quad \text{where} \quad \Gamma = \begin{pmatrix} \tau_1 & \tau_2 & \cdots & \tau_{m-1} & \tau_m \\ \tau_2 & \tau_2 & \cdots & \tau_{m-1} & \tau_m \\ \vdots & \vdots & & \vdots & \vdots \\ \tau_{m-1} & \tau_{m-1} & \cdots & \tau_{m-1} & \tau_m \\ \tau_m & \tau_m & \cdots & \tau_m & \tau_m \end{pmatrix}.$$

Now we can express $\theta_1, \ldots, \theta_m$ as functions of τ as follows

$$\theta_j = \log\left(\frac{p_{j+1}}{p_j}\right) = \log\left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_j - \gamma_{j-1}}\right) = \log\left(\frac{\tau_j - \tau_{j+1}}{\tau_{j-1} - \tau_j}\right),$$

which is the canonical link, also known as the Ad-Cat link. Let

$$P = \begin{pmatrix} 0 & 1 \\ I_m & 0 \end{pmatrix}$$

be the permutation matrix that maps $(a^1, ..., a^m)$ to $(a^m, a^1, ..., a^{m-1})$. Then it is easy to check that the Ad-Cat link can be written in matrix notation as

$$\theta = \theta(\tau) = \log\{[\operatorname{diag}\{(P^{-1} - I)\tau\}]^{-1}(P^{-1} - I)\tau\}.$$

We next compute the inverse canonical link mapping θ to τ. Note that, for $r = 1, \ldots, m$,

$$\sum_{s=1}^{r} \theta_s = \log\left(\frac{\tau_r - \tau_{r+1}}{1 - \tau_r}\right). \tag{S1}$$

Hinted by this, we define

$$\phi_j(\theta) = \sum_{r=1}^{j} \prod_{s=1}^{r} \exp(\theta_s) = \sum_{r=1}^{j} \exp\left(\sum_{s=1}^{r} \theta_s\right), \quad j = 1, \ldots, m. \tag{S2}$$

Let $\phi(\theta) = (\phi_1(\theta), \ldots, \phi_m(\theta))^{\top}$, L the lower triangular matrix of 1's (including the diagonal). Then the above equations can be written in matrix form as $\phi(\theta) = L\exp(L\theta)$. Meanwhile, by (S1) and (S2),

$$\phi(\theta) = L\exp(L\theta) = \frac{1}{1-\tau_1}(\tau_1 - \tau_2, \ldots, \tau_1 - \tau_m, \tau_1)^{\top}.$$

Permute $\phi(\theta)$ by using P to obtain

$$P\phi(\theta) = \frac{1}{1-\tau_1}(\tau_1, \tau_1 - \tau_2, \ldots, \tau_1 - \tau_m)^\top = \frac{1}{1-e_1^\top \tau} Q\tau,$$

where Q is a difference matrix given by

$$Q = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & -1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} - I_m = -I_m + (\mathbf{1}_m + e_1)e_1^\top.$$

Solving the equation $P\phi(\theta) = (1 - e_1^\top \tau)^{-1} Q\tau$ for τ gives us

$$\tau(\theta) = [-I_m + \{\mathbf{1}_m + e_1 + P\phi(\theta)\}e_1^\top]^{-1} P\phi(\theta).$$

By the Sherman-Woodbury formula, we have

$$\begin{aligned}[-I_m + \{\mathbf{1} + e_1 + P\phi(\theta)\}e_1^\top]^{-1} &= -I_m - \frac{(-I_m)\{\mathbf{1}_m + e_1 + P\phi(\theta)\}e_1^\top(-I_m)}{1 + e_1^\top(-I_m)\{\mathbf{1}_m + e_1 + P\phi(\theta)\}} \\ &= -I_m - \frac{\{\mathbf{1}_m + e_1 + P\phi(\theta)\}e_1^\top}{1 - e_1^\top\{\mathbf{1}_m + e_1 + P\phi(\theta)\}}.\end{aligned}$$

Use the relation $1 - e_1^\top\{\mathbf{1}_m + e_1 + P\phi(\theta)\} = -\{1 + e_1^\top P\phi(\theta)\}$, to further simplify the above as

$$\tau(\theta) = \frac{-P\phi(\theta) + \mathbf{1}_m e_1^\top P\phi(\theta) + e_1 e_1^\top P\phi(\theta)}{1 + e_1^\top P\phi(\theta)} = \frac{QPL\exp(L\theta)}{1 + e_1^\top PL\exp(L\theta)}.$$

This is the inverse Ad-Cat Logit link.

Also note that

$$\log\{1 + e_1^\top PL\exp(L\theta)\} = \log\left(1 + \frac{\tau_1}{1-\tau_1}\right) = -\log(1-\tau_1) = -\log(p_1).$$

Hence, the log-likelihood of T has the following linear exponential family form

$$l(\theta; T) = \theta^\top T - b(\theta),$$

where $b(\theta) = \log\{1 + e_1^\top PL\exp(L\theta)\}$.

2. Newton-Raphson for Step 3 in Algorithm 3

In this section, we describe in detail the minimization in step 3 of Algorithm 3 in the paper. That is, the minimization of the negative local log-likelihood

$$\ell(a_1, ..., a_n, C_1, ..., C_n, \beta; Y_{1:n}, X_{1:n}) = \sum_{j=1}^n \ell_j(a_j, C_j, \beta; Y_{1:n}, X_{1:n}),$$

over β for fixed $a_1, \dots, a_n, C_1, \dots, C_n$, where

$$\begin{aligned}\ell_j(a_j, C_j, \beta; Y_{1:n}, X_{1:n}) = \sum_{i=1}^{n} W_{ij}(h)[\{a_j + C_j^\top \beta^\top (X_i - X_j)\}^\top Y_i \\ - b(a_j + C_j^\top \beta^\top (X_i - X_j))].\end{aligned}$$

Let $\beta_v = \mathrm{vec}(\beta^\top) \in \mathbb{R}^{dp}$ and $U_{ij} = (X_i - X_j)^\top \otimes C_j^\top \in \mathbb{R}^{m \times dp}$. Suppressing the fixed variables from the notation, we express the negative log-likelihood, score and information, as a function of β_v, as

$$\begin{aligned}\ell(\beta_v) &= n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{n} W_{ij}(h) \{\beta_v^\top U_{ij}^\top Y_i - b(a_j + U_{ij}\beta_v)\}, \\ S(\beta_v) &= n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{n} W_{ij}(h) U_{ij}^\top \{Y_i - \frac{\partial}{\partial \beta_v} b(a_j + U_{ij}\beta_v)\}, \\ J(\beta_v) &= -n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{n} -W_{ij}(h) U_{ij}^\top \frac{\partial b(a_j + U_{ij}\beta_v)}{\partial \beta_v \partial \beta_v^\top} U_{ij}.\end{aligned}$$

Given an initial estimate for β_v, we iterate $\hat{\beta}_v^{(r+1)} = \hat{\beta}_v^{(r)} + J^{-1}(\hat{\beta}_v^{(r)}) S(\hat{\beta}_v^{(r)})$ until convergence according to some criteria. One option for the initial value of β is to use the OPCG estimate, $\hat{\beta}_{\mathrm{opcg}}$. We denote the converged iterate result as $\hat{\beta}_v$ and set $\hat{\beta} = \mathrm{mat}(\hat{\beta}_v)^\top$.

3. Unbiasedness and Exhaustiveness of OPCG

In this section, we prove a more general version of Proposition 4.1 in the paper. Proposition 4.1 in the paper follows by taking $t(X)$ below to be the canonical parameter $\theta(X)$.

Proposition 4.1. *Suppose $X \in \Omega_X \subseteq \mathbb{R}^p$ and $Y \in \Omega_Y \subseteq \mathbb{R}^m$ satisfy $Y \perp\!\!\!\perp E(Y|X)|\beta^\top X$ for $\beta \in \mathbb{R}^{p \times d}$. Let $t : \Omega_X \to \Omega_Y$ be a differentiable function such that $t(X) = \tilde{t}(\beta^\top X)$. Then*

(a) The columns of the gradient, $\partial t(x)^\top / \partial x \in \mathbb{R}^{p \times m}$, belong to $\mathscr{S}_{E(Y|X)}$.

(b) Let $U = \beta^\top X$. If the $d \times d$ matrix

$$A = E\left\{ \frac{\partial \tilde{t}(U)^\top}{\partial u} \frac{\partial \tilde{t}(U)}{\partial u^\top} \right\}$$

is full rank, then

$$\mathrm{span}\left[E\left\{ \frac{\partial t(X)^\top}{\partial x} \frac{\partial t(X)}{\partial x^\top} \right\} \right] = \mathscr{S}_{E(Y|X)}.$$

(c) If U is supported on a convex set with a nonempty interior, then

$$\mathrm{span}\left[E\left\{ \frac{\partial t(X)^\top}{\partial x} \frac{\partial t(X)}{\partial x^\top} \right\} \right] = \mathscr{S}_{E(Y|X)}.$$

Proof. For (a), we have, by the chain rule, $\partial t(x)^\top/\partial x = \beta\partial\tilde{t}(u)^\top/\partial u$, where $\partial\tilde{t}(u)^\top/\partial u \in \mathbb{R}^{d\times m}$. This implies

$$\operatorname{span}\left\{\frac{\partial t(x)^\top}{\partial x}\right\} = \operatorname{span}\left\{\beta\frac{\partial\tilde{t}(u)^\top}{\partial u}\right\} \subseteq \operatorname{span}(\beta) = \mathscr{S}_{E(Y|X)}.$$

Therefore, the columns of $\partial t(x)^\top/\partial x$ belong to $\mathscr{S}_{E(Y|X)}$.

For (b), note that

$$\left[E\left\{\frac{\partial t(X)^\top}{\partial x}\frac{\partial t(X)}{\partial x^\top}\right\}\right] = \beta\left[E\left\{\frac{\partial\tilde{t}(U)^\top}{\partial u}\frac{\partial\tilde{t}(U)}{\partial u^\top}\right\}\right]\beta^\top = \beta A\beta^\top.$$

Since A is full rank, we have

$$\operatorname{span}\left[E\left\{\frac{\partial t(X)^\top}{\partial x}\frac{\partial t(X)}{\partial x^\top}\right\}\right] = \mathscr{S}_{E(Y|X)}.$$

For (c), let

$$\Lambda = E\left(\frac{\partial t(X)^\top}{\partial x}\frac{\partial t(X)}{\partial x^\top}\right) \in \mathbb{R}^{p\times p}.$$

We need to show $\operatorname{span}(\Lambda) = \operatorname{span}(\beta)$. By the chain rule,

$$\Lambda = \beta E\left(\frac{\partial\tilde{t}(U)^\top}{\partial u}\frac{\partial\tilde{t}(U)}{\partial u^\top}\right)\beta^\top,$$

implying $\operatorname{span}(\Lambda) \subseteq \operatorname{span}(\beta)$. It remains to show $\operatorname{span}(\Lambda)$ is not a proper subspace $\mathscr{S}_{E(Y|X)}$. Suppose $\operatorname{span}(\Lambda) \subsetneq \operatorname{span}(\beta)$. Then $\ker(\Lambda) \supsetneq \ker(\beta)$, so there exists $\alpha \neq 0 \in \mathbb{R}^p$ such that $\alpha \notin \ker(\beta)$ and $\alpha \in \ker(\Lambda)$. Since $\alpha \notin \ker(\beta)$, we get $\beta^\top\alpha \neq 0$. Since $\alpha \in \ker(\Lambda)$, the quadratic form $\alpha^\top\Lambda\alpha$ satisfies

$$0 = a^\top\Lambda\alpha = E\left(\alpha^\top\beta\frac{\partial\tilde{t}(U)^\top}{\partial u}\frac{\partial\tilde{t}(U)}{\partial u^\top}\beta^\top\alpha\right).$$

The non-negativity of the expression inside the expectation implies that it must also be 0 almost everywhere. The quadratic expression then implies

$$0 = \frac{\partial\tilde{t}(u)}{\partial u^\top}\beta^\top\alpha, \tag{S3}$$

for all $u \in \operatorname{supp}(U)$, where $\beta^\top\alpha = \gamma \neq 0 \in \mathbb{R}^d$. Then we see that $\frac{\partial\tilde{t}(u)}{\partial u^\top}\gamma = 0$ for all $u \in \operatorname{supp}(U)$.

Let $u_1, u_2 \in \operatorname{supp}(U)$ such that the segment is $u_2 - u_1$ is parallel to $\gamma \in \mathbb{R}^d$, which is possible because $\operatorname{supp}(U)$ contains an open ball. We take the derivative of $\tilde{t}(u)$ at any point along the line $u_2 - u_1$, say at the point $u_0 = (1-\varepsilon)u_1 + \varepsilon u_2$ for some $\varepsilon > 0$. Then

$$\frac{\partial\tilde{t}^\top((1-\varepsilon)u_1 + \varepsilon u_2)}{\partial\varepsilon} = (u_2 - u_1)^\top\frac{\partial\tilde{t}^\top((1-\varepsilon)u_1 + \varepsilon u_2)}{\partial u} = 0 \in \mathbb{R}^m,$$

where the last equality follows from (S3). This implies that $t(x)$ does not change in the direction of γ. This contradicts $\mathscr{S}_{E(Y|X)}$ being a minimal dimension reduction space, since $\mathscr{S}_{E(Y|X)} \ominus \operatorname{span}(\gamma)$ is then an even smaller SDR subspace for $E(Y|X)$. □

4. Preliminaries and Assumptions for theoretical developments

In this section, we provide some preliminaries and assumptions needed for the theoretical developments in the next few sections. We present the developments for general, strictly convex loss functions instead of the negative log-likelihood. This is because the negative log-likelihood for a linear exponential family can be associated with a strictly convex deviance loss criteria. In the developments that follow, we can always take strictly convex loss function as being the deviance function $\rho(\theta, y) = -\ell(\theta; y) + \ell(\theta_0; y)$, where θ_0 is the unique minimizer of the negative log-likelihood. Then our results in this section will be directly applicable to our results in the paper.

4.1. Preliminaries

For theoretical purpose it is much easier to work under a more general framework than the setting of multivariate GNM. We first set up this general problem. Let (X, Y) be a pair of random vectors taking values in $\Omega_X \times \Omega_Y$, where $\Omega_X \subseteq \mathbb{R}^p$ and $\Omega_Y \subseteq \mathbb{R}^m$. Let $\Theta \subseteq \mathbb{R}^m$ be the parameter space. Let $\rho : \Omega_Y \times \Theta \to \mathbb{R}$ be a loss function, and let $R_0(\theta, x) = E[\rho(Y, \theta)|X = x]$ be the conditional risk. Let

$$\theta(x) = \operatorname*{argmin}_{x \in \Omega_X} R_0(\theta, x).$$

Then OPCG and MADE amount to estimating the gradient function $\partial\theta^\top(x)/\partial x$. Due to the nonparametric nature of this problem, it is impossible to set an objective function with finite-dimensional argument. So, we instead set up an objective function with a tuning parameter h_n, hoping that, as $h_n \to 0$, the minimizer of the objective function converges to $\partial\theta^\top(x)/\partial x$.

To achieve this goal, we construct the following population- and sample-level objective functions. For $a \in \mathbb{R}^m$, $B \in \mathbb{R}^{p\times m}$, and $h_n > 0$, let

$$R(h_n, a, B, x) = E\{K(h_n^{-1}(X - x))\rho(Y, a + B^\top(X - x))\}.$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample from (X, Y). Our sample-level loss function is

$$\hat{R}(h_n, a, B, x) = E_n\{K(h_n^{-1}(X - x))\rho(Y, a + B^\top(X - x))\}.$$

For analytic and algorithmic convenience we reexpress the parameter (a, B) in a vectorized form. We also need to have a convenient notation to multiply B by a constant in (a, B). For these purposes we introduce the following functions: for $u \in \mathbb{R}^p$ and $h \in \mathbb{R}$, $a \in \mathbb{R}^m$ and $B \in \mathbb{R}^{p\times m}$, let

$$\nu(u) = \begin{pmatrix} 1 \\ u \end{pmatrix} \otimes I_m, \quad D(h) = \begin{pmatrix} 1 & 0 \\ 0 & hI_p \end{pmatrix} \otimes I_m, \quad \operatorname{wec}(a, B) = \begin{pmatrix} a \\ \operatorname{vec}(B^\top) \end{pmatrix}.$$

The notation wec is inspired by the fact that the letter w looks like two v's, indicating we are vectorizing two objects together: one vector and one matrix. Also note that there is a transpose in $\mathrm{vec}(B^{\top})$. In this notational system,

$$a + B^{\top}(X - x) = \nu(X - x)^{\top}\mathrm{wec}(a, B), \quad \nu(hu) = D(h)\nu(u) = \begin{pmatrix} 1 \\ hu \end{pmatrix} \otimes I_m.$$

Letting $c = \mathrm{wec}(a, B)$, we can reexpress $R(h_n, a, B, x)$ and $\hat{R}(h_n, a, B, x)$ as

$$E\{K(h_n^{-1}(X - x))\rho(Y, \nu(X - x)^{\top}c)\}, \quad E_n\{K(h_n^{-1}(X - x))\rho(Y, \nu(X - x)^{\top}c)\}.$$

With a slight abuse of notation, we still use $R(h_n, c, x)$ and $\hat{R}(h_n, c, x)$ to denote the above two functions. We will abbreviate h_n as h if doing so causes no ambiguity. Let

$$\begin{aligned} c_0(x) = \mathrm{wec}(\theta(x), \partial\theta^{\top}(x)/\partial x), \quad c(h, x) = \underset{c \in \mathbb{R}^{m(p+1)}}{\mathrm{argmin}}\, R(h, c, x), \\ \hat{c}(h, x) = \underset{c \in \mathbb{R}^{m(p+1)}}{\mathrm{argmin}}\, \hat{R}(h{,}c{,}x). \end{aligned}$$

We will use a to denote the first m components of c, and b the $(m+1)$th through $(m + mp)$th components of c. This applies to $c_0(x)$, $c(h, x)$, and $\hat{c}(h, x)$ as well.

We need a systematic notation to denote vectors, matrices, or multi-dimensional arrays of derivatives. For a vector-valued function of three variables, say $f(h, c, x)$, we use $\partial_1 f(h, c, x)$ and $\partial_2 f(h, c, x)$ to denote the derivatives with respect to the first and second argument of f. When c is itself a function of h, x, we use ∂_h to denote the derivative with respect to h. Specifically,

$$\begin{aligned} \partial_1 f(h, c, x) =& \partial f(h, c, x)/\partial h, \\ \partial_2 f(h, c, x) =& \partial f(h, c, x)/\partial c^{\top}, \\ \partial_h f(h, c(h, x), x) =& \partial_1 f(h, c(h, x), x) + \partial_2 f(h, c(h, x), x)\dot{c}(h, x), \end{aligned}$$

where $\dot{c}(h, x)$ is the derivative of $c(h, x)$ with respect to h. For a function $f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, the notation $\partial_1^2 f(x, y)$ and $\partial_2^2 f(x, y)$ denote the second-derivative matrices with respect to the first and second argument, respectively. For example, $\partial_1^2 f(x, y)$ is the matrix $\partial^2 f(x, y)/\partial x\partial x^{\top}$. For a function $g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$, the notations $\partial_1^2 g(y, \theta)$ and $\partial_2^2 g(y, \theta)$ are 3-dimensioal arrays: for example, $\partial_2^2 g(y, \theta)$ is the 3-dimensional array consisting of matrices $A_1, \ldots, A_m$, where $A_i = \partial^2 g_i(y, \theta)/\partial\theta\partial\theta^{\top}$, g_i being the ith entry of the m-dimensional vector g. Furthermore, if α is an m-dimensional vector, then

$$\alpha^{\top}\partial_2^2 g(y, \theta)\alpha = (\alpha^{\top}A_1\alpha, \ldots, \alpha^{\top}A_m\alpha)^{\top}.$$

We will show uniform consistency of the estimates $\hat{c}(x)$ over $x \in \Omega_X$,

$$\sup_{x \in \Omega_X} \|\hat{c}(h, x) - c_0(x)\| \xrightarrow{a.s} 0 \;\; as \;\; n \to \infty.$$

We will also derive the uniform convergence rate for the estimate $\hat{c}(x)$. To compute the convergence rate, we upper bound the supremum above by a stochastic/variance component and a deterministic/bias component:

$$\sup_{x\in\Omega_X} \|\hat{c}(h,x) - c_0(x)\| \leq \underbrace{\sup_{x\in\Omega_X} \|\hat{c}(h,x) - c(h,x)\|}_{\text{variance/stochastic}} + \underbrace{\sup_{x\in\Omega_X} \|c(h,x) - c_0(x)\|}_{\text{bias/deterministic}}.$$

4.2. Assumptions

We replace Assumptions 3, 4 and 5 in the paper with a set of general assumptions about a strictly convex loss function $\rho(y,\theta)$. We do so because the negative log-likelihood for a linear exponential family is a special case of a strictly convex deviance loss criteria, and the latter is easier to work with.

Assumption 3′. *The parameter θ is identifiable and the loss function $\rho(y,\theta)$ is strictly convex in θ, twice continuously differentiable in θ and has a unique minimum. If $g(y,\theta) = \partial\rho(y,\theta)/\partial\theta$, then there exist $s_1 > 2$ and $s_2 > 2$ such that $\|g(y,\theta)\| \leq M_1(y)$ with $E[M_1(Y)^{s_1}] < \infty$ and $\|\partial_2 g(y,\theta)\| \leq M_2(y)$ with $E[M_2(Y)^{s_2}] < \infty$.*

Assumption 4′. *For each $x \in \Omega_X$, $R_0(\theta,x)$ has a unique minimum $\theta(x)$ over Θ, where $\theta(x)$ is continuously differentiable. The parameter spaces $\Theta \subseteq \mathbb{R}^m$ is compact and convex.*

Assumption 5′. *Derivatives and integrals in*

$$\frac{\partial}{\partial\theta}\int\int \rho(y,\theta)f(x,y)dydx \qquad \text{and} \qquad \frac{\partial}{\partial x}\int g(y,\theta)f(y|x)dy.$$

are interchangeable.

5. Proofs of Fisher Consistency

In this section we prove the Fisher consistency and convergence rate of $c(h,x)$ to $c_0(x)$. This is the bias part of the asymptotic development: there is no randomness involved. Recall that $c(h,x)$ is the solution in c of the equation $\partial_2 R(h,c,x) = 0$. Let $S(h,c,x) = D(h)^{-1}\partial_2 R(h,c,x)$. Since $D(h)$ is nonsingular for $h \in (0,1)$, $c(h,x)$ is also the solution to the equation $S(h,c,x) = 0$. The motivation of changing $\partial_2 R(h,c,x)$ to $S(h,c,x)$ is to clear away any proportionality constant that depends on h. By computation, the specific form $S(h,c,x)$ is

$$S(h,c,x) = \int_{\Omega_X\times\Omega_Y} K(u)\nu(u)g(y,\nu(hu)^\top c)f(x+hu,y)dudy. \tag{S4}$$

The following lemma gives an expression for $\partial\theta(x)/\partial x^\top$, which corresponds to the last mp entries of $c_0(x)$, which we show in Lemma S.5.2.

Lemma S.5.1. *Under Assumptions* 2, $3'$, $4'$, $5'$ *we have*

$$\frac{\partial \theta(x)}{\partial x^\top} = -\left\{ E\left[\frac{\partial g(y,\theta(x))}{\partial \theta^\top} \middle| x \right] \right\}^{-1} E\left[g(y,\theta(x)) \frac{\partial \log f(y|x)}{\partial x^\top} \middle| x \right],$$

where $g(y,\theta) = \partial \rho(y,\theta)/\partial\theta$.

Proof. Since $R_0(\theta, x)$ has unique a minimizer at $\theta(x)$, taking derivatives of the conditional risk gives

$$0 = \partial R_0(\theta(x), x)/\partial\theta = E[\partial\rho(y,\theta(x))/\partial\theta | X = x] = E[g(y,\theta(x))|X = x].$$

Since the above holds for all x, taking the derivative with respect to x will again yield 0:

$$\begin{aligned}
0 &= \frac{\partial}{\partial x^\top} E[g(y,\theta(x))|X = x] \\
&= \frac{\partial}{\partial x^\top} \int g(y,\theta(x)) f(y|x) dy \\
&= \int \frac{\partial}{\partial x^\top} g(y,\theta(x)) f(y|x) dy \\
&= \int \frac{\partial g(y,\theta(x))}{\partial \theta^\top} \frac{\partial \theta(x)}{\partial x^\top} f(y|x) dy + \int g(y,\theta(x)) \frac{\partial f(y|x)}{\partial x^\top} \frac{1}{f(y|x)} f(y|x) dy \\
&= E\left\{ \frac{\partial g(y,\theta(x))}{\partial \theta^\top} \middle| x \right\} \frac{\partial \theta(x)}{\partial x^\top} + E\left\{ g(y,\theta(x)) \frac{\partial \log f(y|x)}{\partial x^\top} \middle| x \right\}.
\end{aligned}$$

Solving for $\partial\theta(x)/\partial x^\top$ yields the desired result. □

In the following, for a function $f(h)$ of h, we use $f(0+)$ to denote $\lim_{h\downarrow 0} f(h)$. The next lemma provides an expression for $c_0(x)$ and the derivative of the first m entries of $c(h,x)$ evaluated at $0+$.

Lemma S.5.2. *If Assumptions* $2, 3', 4', 5', 6$ *hold, then, for each* $x \in \Omega_X$,

$$a(0+, x) = \theta(x), \quad \dot{a}(0+, x) = 0, \quad b(0+, x) = \mathrm{vec}\{\partial\theta(x)/\partial x^\top\}.$$

Note that another way to write the first and third relation is $c(0+, x) = c_0(x)$.

Proof. Since $S(h, c(h,x), x) = 0$ for all $h > 0$, we have, by (S4),

$$S(0+, c(0+, x), x) = \int_{\Omega_X \times \Omega_Y} K(u)\nu(u) g(y, \nu(0)^\top c(0+, x)) f(x, y) du dy = 0. \tag{S5}$$

By the definition of ν, and Assumption 6 on $K(u)$, we have

$$\nu(0)^\top c(0+, x) = a(0+, x), \qquad \int_{\Omega_X} K(u)\nu(u) du = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes I_m. \tag{S6}$$

Also note that $\int_{\Omega_Y} g(y, a(0+, x)f(x, y)dy = f(x)E[g(Y, a(0+, x))|x]$, where $f(x)$ is the density of X evaluated at x. Hence (S5) can be rewritten as

$$\begin{pmatrix} 0 \\ f(x)E[g(Y, a(0+, x)|x)] \end{pmatrix} = 0.$$

Since $\theta(x)$ is the unique solution to $E[g(Y, a(0+, x)|x)] = 0$, we have $a(0+, x) = \theta(x)$.

Next, differentiating the equation $S(h, c(h, x), x) = 0$ with respect to h, we have

$$\partial_1 S(h, c(h, x), x) + \partial_2 S(h, c(h, x), x)\dot{c}(h, x) = 0. \tag{S7}$$

By computation, the first term on the left-hand side of (S7) is

$$\partial_1 S(h, c(h, x), x) = A(h, c(h, x), x) + B(h, c(h, x), x),$$

where

$$\begin{aligned} A(h, c(h, x), x) = &\int K(u)\nu(u)\partial_2 g\{y, \nu(hu)^\top c(h, x)\}\{(0, u^\top) \otimes I_m\}c(h, x) \\ &\times f(x + hu, y)dudy \\ B(h, c(h, x), x) = &\int K(u)\nu(u)g\{y, \nu(hu)^\top c(h, x)\}\partial_1 f(x + hu, y)^\top ududy, \end{aligned}$$

Taking the limit $h \downarrow 0$ for $A(h, c(h, x), x)$, we have

$$A(0+, c(0+, x), x) = \begin{pmatrix} 0_{m\times 1} \\ f(x)E[\partial_2 g(Y, \theta(x))|x]b(0+, x) \end{pmatrix},$$

where we have used $[(0, u^\top) \otimes I_m]c(0+, x) = b(0+, x)$, $\int K(u)uu^\top du = I_p$, and the relation (S6). By a similar computation,

$$B(0+, c(0+, x), x) = \begin{pmatrix} 0_{m\times 1} \\ \int_{\Omega_Y} \partial_1 f(x, y) \otimes g(y, \theta(x))dy \end{pmatrix},$$

where we have used the relations $\nu(0)^\top c(0+, x) = a(0+, x) = \theta(x)$, $\int K(u)udu = 0$, and $\int K(u)uu^\top du = I_p$. By

$$\partial_1 f(x, y) = \dot{f}(x)f(y|x) + f(x)[\partial f(y|x)/\partial x)],$$

and

$$\int_{\Omega_Y} \dot{f}(x)f(y|x) \otimes g(y, \theta(x))dy = \dot{f}(x) \otimes \int_{\Omega_Y} f(y|x)g(y, \theta(x))dy = 0,$$

we have

$$B(0+, c(0+, x), x) = \begin{pmatrix} 0_{m\times 1} \\ f(x) \int_{\Omega_Y} [\partial f(x, y)/\partial x] \otimes g(y, \theta(x))dy \end{pmatrix}.$$

The second term on the left-hand side of (S7) is

$$\int_{\Omega_X\times\Omega_Y} K(u)\nu(u)\partial_2 g\{y,\nu(hu)^{\top}c(h,x)\}\nu(hu)^{\top}\dot{c}(h,x)f(x+hu,y)dudy,$$

which, upon taking the limit $h\downarrow 0$, becomes

$$\begin{aligned}\partial_2 S(0+,c(0+,x),x)\dot{c}(0+,x) &= \int K(u)\nu(u)\partial_2 g\{y,\theta(x)\}\nu(0)^{\top}\dot{c}(0+,x)f(x,y)dudy\\ &= \int K(u)\nu(u)\partial_2 g\{y,\theta(x)\}\dot{a}(0+,x)f(x,y)dudy\\ &= \begin{pmatrix} E[\partial_2 g\{Y,\theta(x)\}|x]f(x)\dot{a}(0+,x)\\ 0_{mp\times m}\end{pmatrix}.\end{aligned}$$

To summarize the results for $A(0+,x(0+,x),x)$ and $B(0+,c(0+,x),x)$, from the above and the equation (S7), we have the following equations

$$\begin{aligned}&E[\partial_2 g\{Y,\theta(x)\}|x]\dot{a}(0+,x)=0,\\ &f(x)E[\partial_2 g\{Y,\theta(x)\}|x]b(0+,x)+f(x)\int_{\Omega_Y}[\partial f(x,y)/\partial x]\otimes g\{y,\theta(x)\}dy=0.\end{aligned}$$

The first equation gives us the second relation in Lemma S.5.2, and the second equation, combined with Lemma S.5.1 and the fact $uv^{\top}=v\otimes u$ for any vectors u,v, implies the second relation in Lemma S.5.2. □

We now use this lemma to prove Theorem 4.1 and Theorem 4.2 in the paper. We first restate them here for reference.

Theorem 4.1. *If Assumptions $2,3',4',5',6,7$ are satisfied, then, as $h\downarrow 0$,*

$$\sup_{x\in\Omega_X}\|a(h,x)-\theta(x)\|=O(h^2),\qquad \sup_{x\in\Omega_X}\|B(h,x)-\partial\theta(x)^{\top}/\partial x\|=O(h). \tag{S8}$$

Proof. By Taylor expansion about $c_0(x)$,

$$0=S(h,c(h,x),x)=S(h,c_0(x),x)+\partial_2 S(h,c^{\dagger},x)D(h^{-1})D(h)[c(h,x)-c_0(x)],$$

where $c^{\dagger}$ is a vector between the line joining $c(h,x)$ and $c(0+,x)$. Hence

$$\begin{aligned}&\|D(h)[c(h,x)-c_0(x)]\|\\ &\le \sup_{h\in[0,1],x\in\Omega_X,c\in A}\|\{\partial_2 S(h,c^{\dagger},x)D(h^{-1})\}^{-1}\|\sup_{x\in\Omega_X}\|S(h,c_0(x),x)\|.\end{aligned} \tag{S9}$$

Note that $\partial_2 S(h,c^{\dagger},x)D(h^{-1})$ is exactly the integral in part 2 of Assumption 7. Hence

$$\sup_{h\in[0,1],x\in\Omega_X,c\in A}\|\{\partial_2 S(h,c^{\dagger},x)D(h^{-1})\}^{-1}\|<\infty. \tag{S10}$$

By Taylor expanding $S(h, c_0(x), x)$ about $h_0 > 0$ and then letting $h_0 \downarrow 0$,

$$S(h, c_0(x), x) = S(0+, c_0(x), x) + \partial_1 S(0+, c_0(x), x)h + \tfrac{1}{2}\partial_1^2 S(h^\dagger, c_0(x), x)h^2,$$

where $h^\dagger$ is a point between 0 and h. The term $S(0+, c_0(x), x)$ is 0 because $S(h, c(h, x), x) = 0$ for all $h > 0$. Also, differentiating the equation $S(h, c(h, x), x) = 0$ with respect to h, we have

$$\partial_1 S(0+, c_0(x), x) + \partial_2 S(0+, c_0(x), x)\dot{c}_0(x) = 0.$$

As we have already seen in the proof of Lemma S.5.2, the first term on the left is a vector whose first m entries are 0, and the second term is a vector whose last mp entries are 0. Hence the above equation implies first term is 0, i.e. $\partial_1 S(0+, c_0(x), x) = 0$, and consequently,

$$S(h, c_0(x), x) = \tfrac{1}{2}\partial_1^2 S(h^\dagger, c_0(x), x)h^2.$$

Then

$$\sup_{x\in\Omega_X} \|S(h, c_0(x), x)\| \leq \tfrac{1}{2} \sup\nolimits_{h\in(0,1), c\in A, x\in\Omega_X} \|\partial_1^2 S(h, c, x)\| \, h^2. \tag{S11}$$

By straightforward computation,

$$\partial_1^2 S(h, c, x) = I_1(h, c, x) + I_2(h, c, x) + I_3(h, c, x) + I_4(h, c, x),$$

where $I_1(h, c, x)$, $I_2(h, c, x)$, $I_3(h, c, x)$ and $I_4(h, c, x)$ are as defined in part 3 of Assumption 7. By part 3 of Assumption 7, then

$$\sup_{h\in(0,1), c\in A, x\in\Omega_X} \|\partial_1^2 S(h, c, x)\| < \infty.$$

Hence, by (S11),

$$\sup_{x\in\Omega_X} \|S(h, c_0(x), x)\| = O(h^2). \tag{S12}$$

Combining (S9), (S10), and (S12), we have $\|D(h)(c(h, x) - c_0(x))\| = O(h^2)$, or equivalently

$$\|a(h, x) - \theta(x)\| = O(h^2), \quad h\|b(h, x) - \text{vec}\{\partial\theta(x)/\partial x^\top\}\| = O(h^2).$$

These are equivalent to (S8), since $B(h, x) = \text{mat}\{b(h, x)\}^\top$, $\|\cdot\|_\text{F} = \|\text{vec}(\cdot)\|$, and operator norm is upper bounded by Frobenius norm, where $\text{mat}(\cdot)$ maps a vector to a matrix by filling the columns of the matrix from left to right with the consecutive elements in the vector. This completes the proof. □

Let $B(h, x) = \text{mat}\{b(h, x)\}^\top \in \mathbb{R}^{p\times m}$. We use the uniform Fisher consistency above to show that the candidate matrices, and their corresponding eigenvectors, are also Fisher consistent. Statement (b) of the theorem below corresponds to Theorem 4.2 in the paper.

Theorem 4.2. *If Assumptions* $2, 3', 4', 5', 6, 7$ *hold, then, for* $h \in [0, 1]$*, and as* $h \downarrow 0$*,*

(a) Then $\Lambda(h) = E\{B(h, X)B(h, X)^\top\}$ *is Fisher consistent for*

$$\Lambda = E\bigg\{\frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\bigg\},$$

and $\|\Lambda(h) - \Lambda\| = O(h)$*.*

(b) Let $\eta(h), \eta$ *be matrices with columns comprised of the first* d *eigenvectors of* $\Lambda(h), \Lambda$ *respectively. Then*

$$m(\eta(h), \eta) = \|\eta(h)\eta(h)^\top - \eta\eta^\top\| = O(h).$$

Proof. For (a), we have

$$\begin{aligned}\|\Lambda(h) - \Lambda\| =& \bigg\|E\bigg\{B(h, X)B(h, X)^\top - \frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\bigg\}\bigg\| \\ \le& E\bigg\{\bigg\|B(h, X)B(h, X)^\top - \frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\bigg\|\bigg\},\end{aligned}$$

where the last inequality follows from Jensen's Inequality. Note that we can re-arrange the difference as

$$\begin{aligned}&B(h, X)B(h, X)^\top - \frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top} \\ =&\bigg\{B(h, X) - \frac{\partial\theta(X)^\top}{\partial x}\bigg\}\bigg\{B(h, X)^\top - \frac{\partial\theta(X)^\top}{\partial x}\bigg\}^\top + \bigg\{B(h, X) - \frac{\partial\theta(X)^\top}{\partial x}\bigg\}\frac{\partial\theta(X)}{\partial x^\top} \\ &+ \frac{\partial\theta(X)^\top}{\partial x}\bigg\{B(h, X) - \frac{\partial\theta(X)^\top}{\partial x}\bigg\}.\end{aligned}$$

Then Theorem 4.1, smoothness of $\theta(\cdot)$, and compactness of Ω_X give us

$$\begin{aligned}&E\bigg\{\bigg\|B(h, X)B(h, X)^\top - \frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\bigg\|\bigg\} \\ \le&\bigg[\sup_{x\in\Omega_X}\bigg\|B(h, x) - \frac{\partial\theta(x)^\top}{\partial x}\bigg\|\bigg]^2 + 2\sup_{x\in\Omega_X}\bigg\|B(h, x) - \frac{\partial\theta(x)^\top}{\partial x}\bigg\| \times \sup_{x\in\Omega_X}\bigg\|\frac{\partial\theta(x)}{\partial x^\top}\bigg\| \\ \le& O(h^2 + h) = O(h),\end{aligned}$$

which completes the proof for (a).

For (b), note that $\|\cdot\|_{\mathrm{F}} \le \sqrt{p}\|\cdot\|$, so $\|\Lambda(h) - \Lambda\|_{\mathrm{F}} \le \sqrt{p}\|\Lambda(h) - \Lambda\| = O(h)$. By Lemma S.8.3(b) of [1], we obtain the final result

$$\begin{aligned}\|\eta(h)\eta(h)^\top - \eta\eta^\top\| \le \sum_{k=1}^{d}\|\eta_k(h)\eta_k(h)^\top - \eta_k\eta_k^\top\| \le \sum_{k=1}^{d}\|\eta_k(h)\eta_k(h)^\top - \eta_k\eta_k^\top\|_{\mathrm{F}} \\ = O(h),\end{aligned}$$

where $\eta_k(h)$ and η_k are the k^{th} columns of $\eta(h)$ and η, respectively, completing the proof. □

6. Proofs for Consistency of OPCG

In this section, we develop the uniform convergence rate of $\sup_{x\in\Omega_X} \|\hat{c}(h,x) - c(h,x)\|$. Recall that $\hat{c}(h,x)$ is the solution in c of the equation $\partial_2 \hat{R}(h,c,x)$. By the same motivation for the definition of (S4), we define

$$\begin{aligned}\hat{S}(h,c,x) =& h^{-p} D(h)^{-1} \partial_2 \hat{R}(h,c,x),\\ =& h^{-p} D(h)^{-1} E_n\{K[(X-x)/h]\nu(X-x) g(Y, \nu(X-x)^{\top} c)\}.\end{aligned}$$

Then $\hat{c}(h,x)$ is also a solution in c of the equation $\hat{S}(h,c,x) = 0$. Next, we make the following assumption.

Assumption 9. *Let $Z = (X,Y)$ and*

(i) $m_n(c,x;Z) = hK[(X-x)/h]\nu[(X-x)/h] g(Y, \nu(X-x)^{\top} c)$,

(ii) $m_n(c,x;Z) = h^2 K[(X-x)/h]\nu[(X-x)/h] \partial g(Y, \nu(X-x)^{\top} c)/\partial\theta^{\top}$.

Suppose each of the above functions satisfy:

(a) For some $c_1, c_2 > 0$,

$$\|m_n(c,x,Z) - m_n(c,x',Z)\| \leq \|x-x'\|^{c_1} n^{c_2} L_1(Z),$$

with $E|L_1(Z)| < \infty$, and where $\|\cdot\|$ is the Euclidean norm,

(b) For some $c_3, c_4 > 0$,

$$\|m_n(c,x,Z) - m_n(c',x,Z)\| \leq \|c-c'\|^{c_3} n^{c_4} L_2(Z),$$

with $E|L_2(Z)| < \infty$, where $\|\cdot\|$ is the operator norm when applied to a matrix (left) and the Euclidean norm when applied to a vector (right).

The following lemma shows that the negative log-likelihood of a linear exponential family that we use in the paper, i.e. $\rho(y,\theta) = -\theta^{\top} y + b(\theta)$, satisfies Assumption 9.

Lemma S.6.1. *Suppose Assumptions $2, 3', 4', 5', 6, 7, 8$ hold and let $\rho(y,\theta) = -\theta^{\top} y + b(\theta)$. Then Assumption 9 holds.*

Proof. Since $g(y,\theta) = -y + b(\theta)$, the two functions in Assumption 9 are

(i) $m_n(c,x;Z) = hK[(X-x)/h]\nu[(X-x)/h]\{-Y + \partial b(\nu(X-x)^{\top} c)/\partial\theta\}$,

(ii) $m_n(c,x;Z) = h^2 K[(X-x)/h]\nu[(X-x)/h]\{\partial^2 b(\nu(X-x)^{\top} c)/\partial\theta\partial\theta^{\top}\}\nu[(X-x)/h]^{\top}$.

We first check that (i) satisfies (a) and (b) of Assumption 9. For (a), we can assume that h is chosen sufficiently small so that x and x' are in a convex subset of Ω_X, and apply the mean value theorem by appealing to the smoothness of g,

K and ν. This gives us

$$\begin{aligned}
&\|m_n(c,x;Z)-m_n(c,x';Z)\|\\
\leq&\|\partial m_n(c,x^{\ddagger};Z)/\partial x^{\top}\|\|x-x'\|\\
\leq&h\|D(h^{-1})\|\|\partial[K[(X-x^{\ddagger})/h]\mathrm{vec}\{\nu(X-x^{\ddagger})[-Y+b(\nu(X-x^{\ddagger})^{\top}c)]\}]/\partial x^{\top}\|\\
&\times\|x-x'\|\\
\leq&\|K[(X-x^{\ddagger})/h]\partial[\mathrm{vec}\{\nu(X-x^{\ddagger})[-Y+b(\nu(X-x^{\ddagger})^{\top}c)]\}]/\partial x^{\top}\|\|x-x'\|\\
&+\|h^{-1}\{\partial K[(X-x^{\ddagger})/h]/\partial u\}\mathrm{vec}\{\nu(X-x^{\ddagger})[-Y+b(\nu(X-x^{\ddagger})^{\top}c)]\}\|\|x-x'\|\\
\leq&n^{\alpha}C_g(Y)\|x-x'\|,
\end{aligned}$$

where $x^{\ddagger}$ lies between x and x'. The term $C_g(Y)$ in the last inequality follows from smoothness of the compactness of Ω_X and Θ_c, and the n^{α} term follows from $n\geq 1$, so we can take $c_1=1$ and $c_2=\alpha$.

For (b), we again appeal to the smoothness of g and convexity of Θ_c, so that we can apply mean value theorem to get:

$$\begin{aligned}
&\|m_n(c,x;Z)-m_n(c',x;Z)\|\\
\leq&\|\partial m_n(c^{\dagger},x;Z)/\partial c^{\top}\|\|c-c'\|\\
\leq&C_K C_X^2 h\|D(h)\|\|\partial^2 b([\nu(X-x)]^{\top}c^{\dagger})/\partial\theta\partial\theta^{\top}\|\|c-c'\|\\
\leq&n^{c_2}C_K C_X^2 C_{b,1}\|c-c'\|
\end{aligned}$$

where $c^{\dagger}$ lies between c and c', c_2 can be arbitrary, $c_1=1$. The bounds follow from our smoothness and compactness assumptions. This completes the proof for function (i).

Next, we check that (ii) satisfies (a) and (b) in Assumption 9 as well. Again, throughout this part of the proof, $\|\cdot\|$ is the operator norm when applied to a matrix and Euclidean norm when applied to a vector. We rely on the fact that the operator norm is bounded by the Frobenius norm, $\|\cdot\|_{\mathrm{F}}$, which gives us

$$\|m_n(c,x;Z)\|\leq\|m_n(c,x;Z)\|_{\mathrm{F}}=\|\mathrm{vec}[m_n(c,x;Z)]\|,$$

where $\|\cdot\|$ refers to operator norm for matrices and euclidean norm for vectors. For (a), we assume that h is chosen sufficiently small so that x and x' are in a convex subset of Ω_X, and apply the mean value theorem by appealing to the smoothness of g, K and ν. This gives us

$$\begin{aligned}
&\|m_n(c,x;Z) - m_n(c,x';Z)\| \\
\le&\|\mathrm{vec}[m_n(c,x;Z)] - \mathrm{vec}[m_n(c,x';Z)]\| \\
\le&\|\partial \mathrm{vec}[m_n(c,x^{\ddagger};Z)]/\partial x^{\top}\|\|x-x'\| \\
\le&h^2\|D(h^{-1})\|^2\|\partial\{K[(X-x^{\ddagger})/h]\mathrm{vec}[\nu(X-x^{\ddagger}) \\
&\times \partial^2 b(\nu(X-x^{\ddagger})^{\top}c)/\partial\theta\partial\theta^{\top}\nu(X-x^{\ddagger})^{\top}]]\}/\partial x^{\top}\|\|x-x'\| \\
\le&\|h^{-1}\{\partial K[(X-x^{\ddagger})/h]/\partial u\}\mathrm{vec}[\nu(X-x^{\ddagger})[\partial^2 b(\nu(X-x^{\ddagger})^{\top}c)/\partial\theta\partial\theta^{\top}]\nu(X-x^{\ddagger})^{\top}]\| \\
&\times\|x-x'\| \\
&+\|K[(X-x^{\ddagger})/h]\partial\{\mathrm{vec}[\nu(X-x^{\ddagger})[\partial^2 b(\nu(X-x^{\ddagger})^{\top}c)/\partial\theta\partial\theta^{\top}]\nu(X-x^{\ddagger})^{\top}]\}/\partial x^{\top}\| \\
&\times\|x-x'\| \\
\le&n^{\alpha}C_{g,2}\|x-x'\|,
\end{aligned}$$

where $x^{\ddagger}$ lies between x and x', and the bound, $C_{g,2}$, follow from our smoothness and compactness assumptions. Then (a) is satisfied with $c_1 = 1$ and $c_2 = \alpha$.

For (b), by the smoothness of g and the convexity of Θ_c, we can apply mean value theorem to get:

$$\begin{aligned}
\|m_n(c,x;Z) - m_n(c',x;Z)\| \le&\|\mathrm{vec}[m_n(c,x;Z)] - \mathrm{vec}[m_n(c',x;Z)]\| \\
\le&\|\partial\mathrm{vec}[m_n(c^{\dagger},x;Z)]/\partial c^{\top}\|\|c-c'\| \\
\le&h^2\|D(h^{-1})\|^2K[(X-x)/h]\|\nu(X-x)\otimes\nu(X-x)\| \\
&\times\|\partial\mathrm{vec}[\partial^2 b([\nu(X-x)]^{\top}c^{\dagger})/\partial\theta\partial\theta^{\top}]/\partial c^{\top}\|\|c-c'\| \\
\le&C_K C_X^2 C_{b,2}\|c-c'\|,
\end{aligned}$$

where $c^{\dagger}$ lies between c and c', so (b) holds with $c_1 = 1$ and c_2 being arbitrary since $n \ge 1$. This completes the proof for function (ii), and completes the proof of the lemma. □

We need the following proposition that shows our choice of bandwidth in Assumption 8 satisfies the second condition of Lemma S.8.1.

Proposition S.6.1. *Let $a_n = h^{p+k}$, where h satisfies Assumption 8 and $0 \le k \le 4$. Then, for $s > 2$, we have $a_n \downarrow 0$ and*

$$\frac{a_n^{s/(s-2)}n}{\log n} \to \infty, \quad \textit{as} \quad n \to \infty.$$

Proof. By Assumption 8, $h = cn^{-\alpha}$ for $\alpha > 0$, and so $a_n \downarrow 0$ as $n \to \infty$. To compute the limit, note that

$$\begin{aligned}
\frac{a_n^{s/(s-2)}n}{\log n} = \frac{[(cn^{-\alpha})^{p+k}]^{s/(s-2)}n}{\log n} &= \frac{c^{(p+k)s/(s-2)}n^{-\alpha(p+k)s/(s-2)}n}{\log n} \\
&\propto \frac{n^{-\alpha(p+k)s/(s-2)+1}}{\log n}.
\end{aligned}$$

For $s > 2$, $0 < \alpha \leq 1/p_0$, and $p_0 > (p+4)s/(s-2)$, we get for $0 \leq k \leq 4$,

$$1 > 1 - \alpha\frac{(p+k)s}{(s-2)} \geq 1 - \frac{(p+k)s}{p_0(s-2)} > 0,$$

and so $n^{-\alpha(p+k)s/(s-2)+1}/\log n$ is monotonically diverging in n, completing the proof. □

In particular, Proposition S.6.1 implies that h^{p+2} and h^{p+4} will satisfy condition 2 of Lemma S.8.1. The following lemma gives the uniform convergence of $\hat{S}(h,c,x)$ and $\partial_2\hat{S}(h,c,x)D(h)^{-1}$.

Lemma S.6.2. *Suppose Assumptions* $2, 3', 4', 5'$ *and* $6 \sim 9$ *hold. Then,*

(a) $\sup_{x\in\Omega_X, c\in\Theta_c} \|\hat{S}(h,c,x) - S(h,c,x)\| = O_{\rm as}(\delta_{ph})$,

(b) $\sup_{x\in\Omega_X, c\in\Theta_c} \|\partial_2\hat{S}(h,c,x)D(h)^{-1} - \partial_2 S(h,c,x)D(h)^{-1}\| = O_{\rm as}(\delta_{ph})$.

Proof. (a). We apply Lemma S.8.1 with $Z = (X,Y)$ and

$$m_n(c,x;Z) = hK[(X-x)/h]D(h)^{-1}\nu(X-x)g(Y,\nu(X-x)^\top c),$$

so that $\hat{S}(h,c,x) = h^{-(p+1)}E_n m_n(c,x;Z)$. We now check the three conditons in Lemma S.8.1.

For condition (i), we have $\|g(y,\theta)\| \leq M_1(y)$, by smoothness of g and compactness of Θ. Furthermore, $E\{M_1(Y)^{s_1}\} < \infty$ for some $s_1 > 2$ by Assumption $3'$. Therefore,

$$\begin{aligned}\|m_n(c,x;Z)\| \leq & hK[(X-x)/h]\|D(h)^{-1}\|\|\nu[(X-x)]\|\|g(Y,\nu(X-x)^\top c)\| \\ \leq & C_K C_X M_1(Y),\end{aligned}$$

where C_K and C_X bounds the kernel and $\|\nu(X-x)\|$ respectively, and $\|D(h)^{-1}\| = h^{-1}$ since $h \in (0,1)$. Since the bounds are free of θ and x, taking the supremum implies condition (i).

For condition (ii) of Lemma S.8.1, we have

$$\begin{aligned}\sigma_1^2 = & |E[m_n(c,x;Z)^\top m_n(c,x;Z)]| \\ = & |Eh^2K[(X-x)/h]^2 g(Y,\nu(X-x)^\top c)^\top \nu[(X-x)/h]^\top \nu[(X-x)/h] \\ & \times g(Y,\nu(X-x)^\top c)| \\ \leq & C_K h^{p+2}\int K(u)\left|g(y,\nu(hu)^\top c)^\top \nu(u)^\top \nu(u) g(y,\nu(hu)^\top c)\right| f(x+hu,y)dudy \\ \leq & C_K h^{p+2} \sup_{c\in\Theta_c, h\in[0,1], x\in\Omega_X} \int K(u)\left|g(y,\nu(hu)^\top c)^\top \nu(u)^\top \nu(u) g(y,\nu(hu)^\top c)\right| \\ & \times f(x+hu,y)dudy \\ \leq & O(h^{p+2}),\end{aligned}$$

where the boundedness follows from part (4) of Assumption 7. Similarly,

$$
\begin{aligned}
\sigma_2^2 =& |E[m_n(c,x;Z)m_n(c,x;Z)^\top]| \\
\le& \|Eh^2K[(X-x)/h]^2\nu[(X-x)/h]g(Y,\nu(X-x)^\top c)g(Y,\nu(X-x)^\top c)^\top \\
& \times \nu[(X-x)/h]^\top\| \\
\le& C_K h^{p+2}\int K(u)\left\|\nu(u)g(y,\nu(hu)^\top c)g(y,\nu(hu)^\top c)^\top\nu(u)^\top\right\| f(x+huy)dudy \\
\le& C_K h^{p+2}\sup_{c\in\Theta_c,h\in[0,1],x\in\Omega_X}\int K(u)\left\|\nu(u)g(y,\nu(hu)^\top c)g(y,\nu(hu)^\top c)^\top\nu(u)^\top\right\| \\
& \times f(x+hu,y)dudy \\
\le& O(h^{p+2}).
\end{aligned}
$$

Hence, $\sigma^2 = \max\{\sigma_1^2,\sigma_2^2\} = O_{\text{as}}(h^{p+2})$. The second part of condition (ii) regarding $a_n = h^{p+2}$ holds by Proposition S.6.1. Condition (iii) follows directly from Assumption 9.

Then, by Lemma S.8.1 and Corollary S.8.2, we have

$$
\sup_{x\in\Omega_X,c\in\Theta_c}\|E_n m_n(c,x;Z) - Em_n(c,x;Z)\| = h^{p+2}O_{\text{as}}(\delta_{(p+2)h}).
$$

Since $\delta_{(p+2)h} = h^{-1}\delta_{ph}$, we have $h^{p+2}\delta_{(p+2)h} = h^{p+1}\delta_{ph}$. Plugging $E_n m_n(c,x;Z) = h^{p+1}\hat{S}(h,c,x)$ and $Em_n(c,x;Z) = h^{p+1}S(h,c,x)$ into the supremum above, we have

$$
\sup_{x\in\Omega_X,c\in\Theta_c}\|\hat{S}(h,c,x) - S(h,c,x)\| = O_{\text{as}}(\delta_{ph}),
$$

which completes the proof of (a).

(b). We apply Lemma S.8.1 with $Z = (X,Y)$ and

$$
m_n(c,x;Z) = h^2K[(X-x)/h]\nu[(X-x)/h]\partial_2 g(Y,\nu(X-x)^\top c)\nu[(X-x)/h]^\top,
$$

so that $\partial_2\hat{S}(h,c,x)D(h)^{-1} = h^{-(p+2)}E_n m_n(c,x;Z)$. We now check the three conditions in Lemma S.8.1.

For condition (i), $\|\partial_2 g(Y,\theta)\| = M_2(y)$, where $E[M_2(Y)^{s_2}] < \infty$ for some $s_2 > 2$ by Assumption $3'$. Therefore,

$$
\begin{aligned}
\|m_n(c,x;Z)\| \le& h^2K[(X-x)/h]\|\nu[(X-x)]\|^2\|D(h)^{-1}\|^2\|\partial_2 g(Y,\nu(X-x)^\top c)\| \\
\le& C_K C_X^2 M_2(Y),
\end{aligned}
$$

where C_K and C_X bounds the kernel and $\|\nu(X-x)\|$ respectively, and the operator norm of $\|D(h)^{-1}\|^2$ is just h^{-2} since $h\in(0,1)$. Since the bounds are free of θ and x, taking supremum implies condition (i).

For condition (ii), since $m_n(c,x;Z)$ is a symmetric matrix, we just need to compute $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We have

$$\begin{aligned}
\sigma^2 =& \|E[m_n(c,x;Z)^\top m_n(c,x;Z)]\| \\
\leq& \|Eh^4 K[(X-x)/h]^2 \nu[(X-x)/h]\partial_2 g(Y, \nu(X-x)^\top c)\nu[(X-x)/h]^\top \\
&\times \nu[(X-x)/h]\partial_2 g(Y, \nu(X-x)^\top c)\nu[(X-x)/h]^\top\| \\
\leq& C_K h^{p+4}\int K(u)\left\|\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top \nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right\| \\
&\times f(x+huy)dudy \\
\leq& C_K h^{p+4}\sup_{c\in\Theta_c, h\in[0,1], x\in\Omega_X}\int K(u)\left\|\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right. \\
&\left.\times\, \nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right\| f(x+hu,y)dudy \\
=& O(h^{p+4}),
\end{aligned}$$

where the second last inequality follows from part (4) of Assumption 7. The second part of condition (ii) regarding $a_n = h^{p+4}$ holds by Proposition S.6.1. Condition (iii) follows directly from Assumption 9.

Then, by Lemma S.8.1 and Corollary S.8.2, have

$$\sup_{x\in\Omega_X, c\in\Theta_c}\|E_n m_n(c,x;Z) - Em_n(c,x;Z)\| = h^{p+4}O_{\text{as}}(\delta_{(p+4)h}).$$

Since $\delta_{(p+4)h} = h^{-2}\delta_{ph}$, we have $h^{p+4}\delta_{(p+4)h} = h^{p+2}\delta_{ph}$. Plugging

$$E_n m_n(c,x;Z) = h^{p+2}\partial_2\hat{S}(h,c,x)D(h)^{-1}$$

and $Em_n(c,x;Z) = h^{p+2}\partial_2 S(h,c,x)D(h)^{-1}$ into the supremum above, we have

$$\sup_{x\in\Omega_X, c\in\Theta_c}\|\partial_2\hat{S}(h,c,x)D(h)^{-1} - \partial_2 S(h,c,x)D(h)^{-1}\| = O_{\text{as}}(\delta_{ph}),$$

which completes the proof of (b). □

The next lemma and corollary gives the uniform convergence of $\{\partial_2\hat{S}(h,c,x)D(h)^{-1}\}^{-1}$.

Lemma S.6.3. *Let $\eta\in E$, where E is compact and $\|\cdot\|_{\text{F}}$ denote the Frobenius norm. Suppose a sequence of random invertible matrices $\{\hat{A}_n(\eta) : n = 1,2,\ldots\}$ and deterministic invertible matrices $\{A_n(\eta) : n = 1,2,\ldots\}$ satisfy*

$$\sup_{\eta\in E}\|\hat{A}_n(\eta) - A_n(\eta)\|_{\text{F}} = O_{\text{as}}(d_n),$$

where $d_n\to 0$ as $n\to\infty$. If $\sup_{\eta\in E}\|A_n(\eta)^{-1}\|_{\text{F}} = O(1)$, then

$$\sup_{\eta\in E}\|\hat{A}_n(\eta)^{-1} - A_n(\eta)^{-1}\|_{\text{F}} = O_{\text{as}}(d_n).$$

Proof. Since $\sup_{\eta\in E}\|\hat{A}_n(\eta)-A_n(\eta)\|_{\mathrm{F}} = O_{\mathrm{as}}(d_n)$, we have that $\hat{A}_n(\eta) = A_n(\eta)+D_n$, where $D_n = O_{\mathrm{as}}(d_n)$, which is meant entry-wise. Then,

$$\hat{A}_n(\eta)^{-1} =[A_n(\eta)+D_n]^{-1} = A_n(\eta)^{-1} + A_n(\eta)^{-1}D_n[A_n(\eta)+D_n]^{-1},$$

since, for matrices A and D with A invertible, we always have $(A+D)^{-1} = A^{-1}+A^{-1}D(A+D)^{-1}$. Because $d_n \to 0$, we have $\sup_{\eta\in E}\|[A_n(\eta)+D_n]^{-1}\|_{\mathrm{F}} = O(1)$ and $\sup_{\eta\in E}\|A_n(\eta)^{-1}\|_{\mathrm{F}} = O(1)$. This gives us

$$\begin{aligned}&\sup_{\eta\in E}\|\hat{A}_n(\eta)^{-1} - A_n(\eta)^{-1}\|_{\mathrm{F}}\\ \le&\sup_{\eta\in E}\|A_n(\eta)^{-1}\|_{\mathrm{F}} \times \|D_n\|_{\mathrm{F}} \times \sup_{\eta\in E}\|[A_n(\eta)+D_n]^{-1}\|_{\mathrm{F}}\\ =&O_{\mathrm{as}}(d_n),\end{aligned}$$

completing the proof. □

Corollary S.6.1. *Suppose Assumptions* $2, 3', 4', 5'$ *and* $6 \sim 9$ *hold. Then,*

$$\sup_{x\in\Omega_X, c\in\Theta_c}\|\{\partial_2\hat{S}(h,c,x)D(h)^{-1}\}^{-1} - \{\partial_2 S(h,c,x)D(h)^{-1}\}^{-1}\|_{\mathrm{F}} = O_{\mathrm{as}}(\delta_{ph}).$$

Proof. We need to verify the conditions in Lemma S.6.3. By Lemma S.6.2, we have

$$\sup_{x\in\Omega_X, c\in\Theta_c}\|\partial_2\hat{S}(h,c,x)D(h)^{-1} - \partial_2 S(h,c,x)D(h)^{-1}\|_{\mathrm{F}} = O_{\mathrm{as}}(\delta_{ph}),$$

where $\delta_{ph} \to 0$. We also have $\sup_{x\in\Omega_X, c\in\Theta_c}\|\{\partial_2 S(h,c,x)D(h)^{-1}\}^{-1}\|_{\mathrm{F}} = O(1)$ by part (4) of Assumption 7. Then, by Lemma S.6.3, we get

$$\sup_{x\in\Omega_X, c\in\Theta_c}\|\{\partial_2\hat{S}(h,c,x)D(h)^{-1}\}^{-1} - \{\partial_2 S(h,c,x)D(h)^{-1}\}^{-1}\|_{\mathrm{F}} = O_{\mathrm{as}}(\delta_{ph}),$$

completing the proof. □

The following theorem serves as a precursor to Theorem 5.1.

Theorem S.6.1. *Suppose Assumptions* $2, 3', 4', 5'$ *and* $6 \sim 9$ *hold. Then, as* $n\to\infty$*, we have*

$$\sup_{x\in\Omega_X}\|\hat{a}(h,x) - a(h,x)\| = O_{\mathrm{as}}(\delta_{ph}), \quad \sup_{x\in\Omega_X}\|\hat{B}(h,x) - B(h,x)\| = O_{\mathrm{as}}(h^{-1}\delta_{ph}).$$

Proof. A Taylor expansion of $\hat{S}(h,\hat{c}(h,x),x)$ in c about $c(h,x)$ gives us

$$0 = \hat{S}(h,\hat{c}(h,x),x) = \hat{S}(h,c(h,x),x) + \partial_2\hat{S}(h,c^{\dagger},x)[\hat{c}(h,x) - c(h,x)],$$

where $\|c^{\dagger} - c(h,x)\| \le \|\hat{c}(h,x) - c(h,x)\|$. Solving for $D(h)[\hat{c}(h,x) - c(h,x)]$, we have

$$D(h)[\hat{c}(h,x) - c(h,x)] = -\{\partial_2\hat{S}(h,c^{\dagger},x)D(h)^{-1}\}^{-1}\hat{S}(h,c(h,x),x).$$

Taking norm on both sides and observing that $E(\hat{S}(h, c(h,x), x)) = S(h, c(h,x), x) = 0$, we can upper bound the RHS as follows, where $\|\cdot\|$ refers to operator norm for matrices and Euclidean otherwise:

$$\begin{aligned}&\|D(h)[\hat{c}(h,x) - c(h,x)]\| \\ \le& \sup_{c\in\Theta_c} \|\{\partial_2 \hat{S}(h,c,x)D(h)^{-1}\}^{-1}\| \sup_{c\in\Theta_c} \|\hat{S}(h,c,x) - S(h,c,x)\|.\end{aligned}$$

where

$$\begin{aligned}T_1 &= \sup_{x\in\Omega_X, c\in\Theta_c} \|\{\partial_2 \hat{S}(h,c,x)D(h)^{-1}\}^{-1} - \{\partial_2 S(h,c,x)D(h)^{-1}\}^{-1}\| \\ T_2 &= \sup_{x\in\Omega_X, c\in\Theta_c} \|\hat{S}(h,c,x) - S(h,c,x)\| \\ T_3 &= \sup_{x\in\Omega_X, c\in\Theta_c} \|\{\partial_2 S(h,c,x)D(h)^{-1}\}^{-1}\|.\end{aligned}$$

By part 4 of Assumption 7,

$$\begin{aligned}T_3 \le& \sup_{h\in[0,1], c\in A, x\in\Omega_X} \left\| \left\{ \int K(u)\nu(u)\partial_2 g(y, \nu(hu)^{\top}c)\nu(u)^{\top} f(x+hu, y)dydu \right\}^{-1} \right\| \\ &< \infty.\end{aligned}$$

Lemma S.6.2 and Corollary S.6.1 imply $T_1 = O_{\mathrm{as}}(\delta_{ph})$ and $T_3 = O_{\mathrm{as}}(\delta_{ph})$. Plugging in these rates, we get

$$\sup_{x\in\Omega_X} \|D(h)[\hat{c}(h,x) - c(h,x)]\| = O_{\mathrm{as}}(\delta_{ph})O_{\mathrm{as}}(\delta_{ph}) + O(1)O_{\mathrm{as}}(\delta_{ph}) = O_{\mathrm{as}}(\delta_{ph}).$$

Hence the norm of the first m entries of $D(h)\{\hat{c}(h,x) - c(h,x)\}$ satisfy

$$\sup_{x\in\Omega_X} \|\hat{a}(h,x) - a(h,x)\| = O_{\mathrm{as}}(\delta_{ph}),$$

and the norm of off the last mp entries of $D(h)\{\hat{c}(h,x) - c(h,x)\}$ satisfy

$$\sup_{x\in\Omega_X} \|h\hat{b}(x) - b(h,x)\| = O_{\mathrm{as}}(\delta_{ph}) \implies \sup_{x\in\Omega_X} \|\hat{b}(x) - b(h,x)\| = O_{\mathrm{as}}(h^{-1}\delta_{ph}).$$

Because $\|\cdot\|_{\mathrm{F}} = \|\mathrm{vec}(\cdot)\|$ and the operator norm is upper bounded by the Frobenius norm, we have the desired result. □

Let $\hat{B}(x) = \mathrm{mat}(\hat{b}(x))^{\top} \in \mathbb{R}^{p\times m}$, where $\mathrm{mat}(\cdot)$ maps a vector to a matrix by filling the m columns of the matrix from left to right with the p consecutive elements in the vector. This operation depends on the dimension p of the columns, but we omit this dependence from the notation as it is usually obvious from the context. We can now prove Theorem 5.1 as a direct consequence of Theorem S.6.1.

Theorem 5.1. *Suppose Assumptions* $2, 3', 4', 5'$ *and* $6 \sim 8$ *hold. Then, as* $n \to \infty$*, we have*

$$\sup_{x \in \Omega_X} \|\hat{a}(h, x) - \theta(x)\| = O_{\text{as}}(h^2 + \delta_{ph}),$$

$$\sup_{x \in \Omega_X} \|\hat{B}(x) - \partial\theta(x)^\top / \partial x\|_{\text{F}} = O_{\text{as}}(h + h^{-1}\delta_{ph}).$$

Proof. Combine the result from Theorem S.6.1 with Theorem 4.1 to obtain

$$\begin{aligned}\sup_{x \in \Omega_X} \|\hat{a}(h, x) - \theta(x)\| \leq & \sup_{x \in \Omega_X} \|\hat{a}(h, x) - a(h, x)\| + \sup_{x \in \Omega_X} \|a(h, x) - \theta(x)\| \\ =& O_{\text{as}}(h^2 + \delta_{ph}),\end{aligned}$$

and

$$\begin{aligned}&\sup_{x \in \Omega_X} \|\hat{B}(h, x) - \partial\theta(x)^\top / \partial x\| \\ \leq & \sup_{x \in \Omega_X} \|\hat{B}(x) - B(h, x)\| + \sup_{x \in \Omega_X} \|B(h, x) - \partial\theta(x)/\partial x^\top\| \\ =& O_{\text{as}}(h + h^{-1}\delta_{ph}). \square\end{aligned}$$

Let $B(x)$ denote the true gradients $\partial\theta(x)^\top / \partial x$. Since $\hat{b}(x)$ is uniformly consistent for $\text{vec}(\partial\theta(x)/\partial x^\top)$, $\hat{B}(h, x)$ is uniformly consistent for $B(x)$. The candidate matrix for OPCG is given by $\hat{\Lambda}_n = n^{-1} \sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top$. The next theorem is an augmented version of Theorem 5.2 in the paper (the latter corresponds to part (b) of the theorem below; to avoid confusion).

Theorem 5.2. *Suppose Assumptions* $2, 3', 4', 5'$ *and* $6 \sim 8$ *hold, and let* $\delta_{ph} = (\log n / h^p n)^{1/2}$.

(a) Then $\hat{\Lambda}_n = n^{-1} \sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top$ *consistently estimates* $\Lambda = E\{B(x)B(x)^\top\}$*, and*

$$\|\hat{\Lambda}_n - \Lambda\| = O_{\text{as}}(h + h^{-1}\delta_{ph}),$$

(b) Let $\hat{\eta}, \eta \in \mathbb{R}^{p \times d}$ *be matrices with columns comprised of the first* d *eigenvectors of* $\hat{\Lambda}_n, \Lambda$ *respectively. Then*

$$m(\hat{\eta}, \eta) = \|\hat{\eta}\hat{\eta}^\top - \eta\eta^\top\| = O_{\text{as}}\left(h + h^{-1}\delta_{ph}\right).$$

Proof. (a). By the triangle inequality, we have

$$\begin{aligned}\|\hat{\Lambda}_n - \Lambda\| =& \left\| n^{-1} \sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top - EB(X)B(X)^\top \right\| \\ \leq & n^{-1} \sum_{j=1}^n \|\hat{B}(X_j)\hat{B}(X_j)^\top - B(X_j)B(X_j)^\top\| \\ & + \|E_n\{B(X)B(X)^\top - EB(X)B(X)^\top\}\|.\end{aligned}$$

We use the uniform consistency in Theorem S.6.1 to bound the first term above as follows:

$$\begin{aligned}
&\|\hat{B}(X_j)\hat{B}(X_j)^{\top} - B(X_j)B(X_j)^{\top}\| \\
\leq &\|\{\hat{B}(X_j) - B(X_j)\}\{\hat{B}(X_j) - B(X_j)\}^{\top}\| + \|\{\hat{B}(X_j) - B(X_j)\}B(X_j)^{\top}\| \\
&+ \|B(X_j)\{\hat{B}(X_j) - B(X_j)\}^{\top}\| \\
\leq &\left(\sup_{x\in\Omega_X} \|\hat{B}(x) - B(x)\|\right)^2 + 2\sup_{x\in\Omega_X}\|B(x)\| \sup_{x\in\Omega_X}\|\hat{B}(x) - B(x)\| \\
&= O_{\mathrm{as}}[\{(h + h^{-1}\delta_{ph})\}^2] + O_{\mathrm{as}}(h + h^{-1}\delta_{ph}) \\
&= O_{\mathrm{as}}(h + h^{-1}\delta_{ph}).
\end{aligned}$$

Using Lemma S.8.2 to bound the second term gives

$$\|E_n\{B(X)B(X)^{\top} - EB(X)B(X)^{\top}\}\| = O_{\mathrm{as}}(\sqrt{\log n/n}).$$

Hence

$$\|\hat{\Lambda}_n - \Lambda\| = O_{\mathrm{as}}(h + h^{-1}\delta_{ph} + \sqrt{\log n/n}).$$

Since $h^{-1}\delta_{ph} = \sqrt{\log n/(h^{p+2}n)}$, for $h \in (0,1)$, $h^{-1}\delta_{ph}$ is larger than $\sqrt{\log n/n}$, so we drop $\sqrt{\log n/n}$, completing the proof for (a).

(b). Since $\|\cdot\|_{\mathrm{F}} \leq \sqrt{p}\|\cdot\|$, we have $\|\Lambda(h) - \Lambda\|_{\mathrm{F}} \leq \sqrt{p}\|\Lambda(h) - \Lambda\| = O_{\mathrm{as}}(h + h^{-1}\delta_{ph})$. By Lemma S.8.3(b), due to [1], we get the final result

$$\|\hat{\eta}\hat{\eta}^{\top} - \eta\eta^{\top}\| \leq \sum_{k=1}^{d}\|\hat{\eta}_k\hat{\eta}_k^{\top} - \eta_k\eta_k^{\top}\| \leq \sum_{k=1}^{d}\|\hat{\eta}_k\hat{\eta}_k^{\top} - \eta_k\eta_k^{\top}\|_{\mathrm{F}} = O(h + h^{-1}\delta_{ph}),$$

where $\hat{\eta}_k$ and η_k are the k^{th} columns of $\hat{\eta}$ and η, respectively. This proves (b). □

7. Plots for Simulations and Applications

In this section, we provide some additional plots to supplement the results reported in the paper: specifically, the predictor augmentation plots used for estimating the dimension d of the central mean subspace $\mathscr{S}_{E(Y|X)}$, the five-fold k-means tuning plots for determining the bandwidth, and some plots of the sufficient predictors constructed from test sets. Figure S1 contains the F-ratio plots for our supervised k-means tuning procedure with different numbers of clusters per class. Figure S2 plots the prediction augmentation variation [2] against the dimension d of the central mean subspace. Figure S3 shows classes in the first two sufficient predictors by six different SDR methods. The central subspace is estimated using the training set; the plots are based on the test sets. Figures S1, S2, S3 are the visual support of the results in Table 1 of the paper.

For our applications, Figure S4 provides the predictor augmentation plots for pendigit and USPS in; Figure S5 provides the F-ratios for tuning pendigit, USPS, and ISOLET; Figure S6 shows the F-ratio for tuning the wine quality data; Figure S7 shows the sufficient predictors constructed on the test set for the wine quality data set.

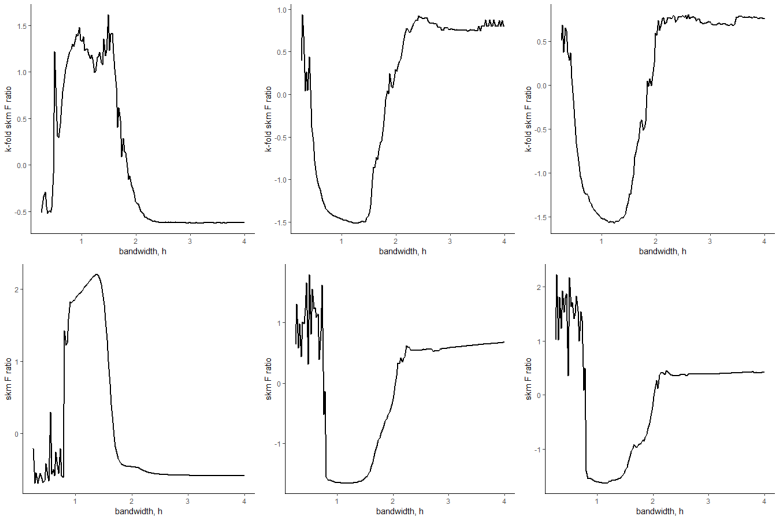


Fig S1. *F-ratios for our supervised k-means tuning procedure in simulations. The plots on top are 5-fold supervised k-means on the training set, and the plots on bottom are supervised k-means on a validation set. From left to right, the number of clusters per class is set to* $1, 2, 3$.

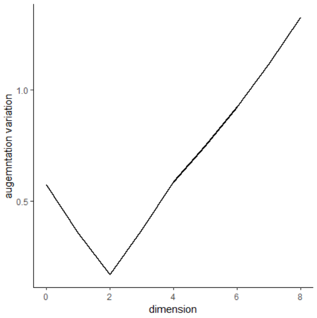


Fig S2. *The estimated dimension of* $\mathscr{S}_{E(Y|X)}$ *for our simulations. The Predictor Augmentation plot estimates* $\hat{d} = 2$.

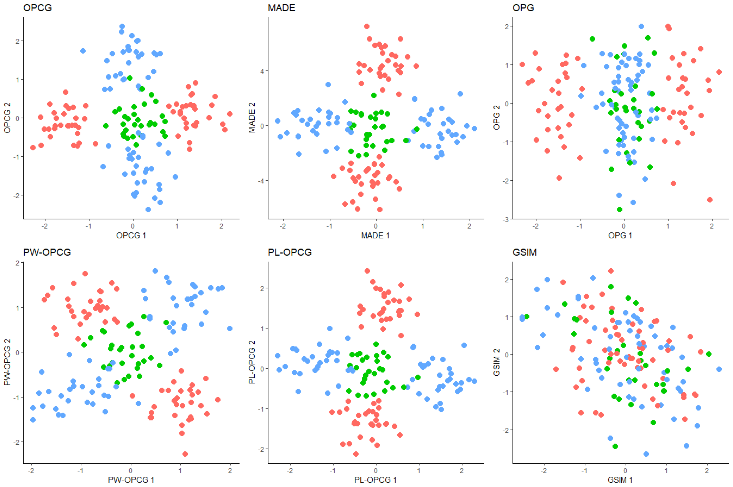
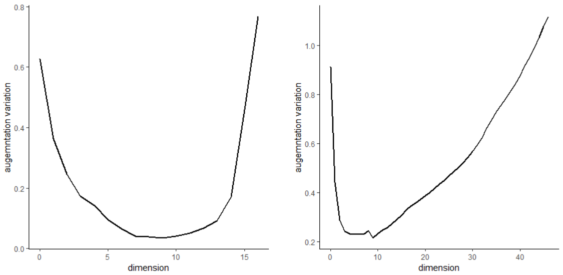


FIG S3. *Sufficient Predictors on the test set in simulations. Red is class* 1*; blue is 2; and green is 3.*



(a) pendigit (b) USPS

FIG S4. *From left to right: Predictor Augmentation plot for pendigit and USPS.*

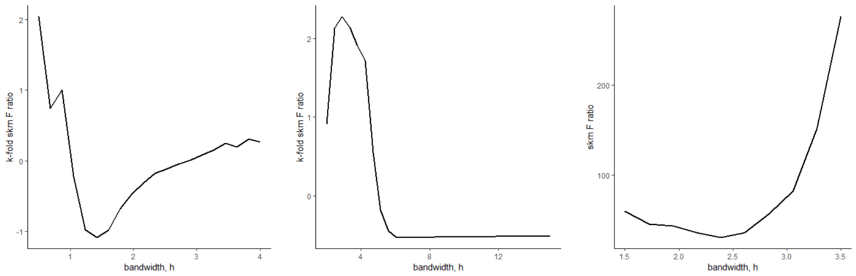
8. Background Theorems and Lemmas

In this section, we provide some preliminary theorems and lemmas used in previous sections. Some of these results are known, and we provide their proofs here for completeness.

8.1. Uniform Consistency for the mean with a parameter and index

The following lemma and proof is adapted from [5, 6].

Lemma S.8.1. *Suppose* $m_n(\eta, Z) \in \mathbb{R}^{d_1 \times d_2}$*,* $n = 1, 2 \ldots,$ *are matrix-valued functions, where* Z *is a random vector, and* η *is a parameter that ranges over a compact set* $E \subset \mathbb{R}^d$*. Suppose* $\{Z_i : i = 1, 2, \cdots\}$ *is a sequence of i.i.d. copies*



(a) pendigit (b) USPS (c) ISOLET

Fig S5. *F-ratio from supervised k-means for tuning the bandwidth.*

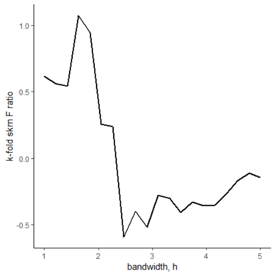


Fig S6. *Wine Quality. F-ratio from five-fold supervised k-means for tuning the bandwidth.*

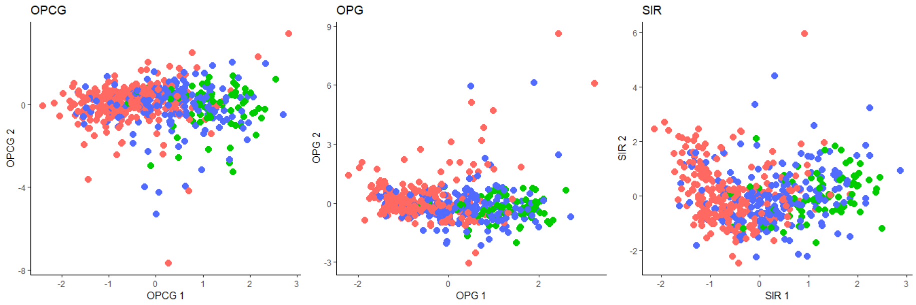


Fig S7. *Wine Quality. From Left to Right: Sufficient predictors constructed on the test set using OPCG, OPG, and SIR. Red is category* $\{3, 4, 5\}$*, blue is category* $\{6\}$*, and green is category* $\{7, 8\}$.

of Z. Suppose m_n satisfy the following conditions, where $\|\cdot\|$ denotes operator norm when referring to matrices and Euclidean norm when referring to vectors:

(i) (Uniform Boundedness) Suppose

$$\sup_{\eta \in E} \|m_n(\eta, Z)\| \leq M(Z),$$

with $E\{M^s(Z)\} < \infty$ *for some* $s > 2$*;*

(ii) (Uniformly rates of Second Moments) Let

$$\sigma_1^2 = \sup_{\eta \in E} \|E[m_n(\eta, Z)^\top m_n(\eta, Z)]\|, \quad \sigma_2^2 = \sup_{\eta \in E} \|E[m_n(\eta, Z) m_n(\eta, Z)^\top]\|.$$

We assume $\sigma^2 = \max(\sigma_1^2, \sigma_2^2) < a_n$*, where* $a_n \to 0$*, and*

$$\liminf_n \frac{a_n^{s/(s-2)} n}{\log n} = \liminf_n a_n^{2/(s-2)} \frac{a_n n}{\log n} > 0$$

(iii) (Lipschitz in η*) For all* $\eta, \eta' \in E$*,*

$$\|m_n(\eta, Z) - m_n(\eta', Z)\| \le \|\eta - \eta'\|^{c_1} n^{c_2} L(Z),$$

for some $c_1, c_2 > 0$*, and* $L(Z) \ge 0$ *with* $EL(Z) < \infty$*.*

Then,

$$\sup_{\eta \in E} \left\| E_n m_n(\eta, Z) - E m_n(\eta, Z) \right\| = O_{as}\left(\sqrt{\frac{a_n \log n}{n}} \right).$$

Proof. Since E is compact, it can be covered by N balls of radius r centered at $\eta_1, \ldots, \eta_N$. By the triangle inequality,

$$\begin{aligned}
&\sup_{\eta \in E} \left\| n^{-1} \sum_{i=1}^n [m_n(\eta, Z_i) - E m_n(\eta, Z_i)] \right\| \\
&\le \sup_{\eta \in \cup_k B_r(\eta_k)} \left\| n^{-1} \sum_{i=1}^n [m_n(\eta, Z_i) - E m_n(\eta, Z_i)] \right\| \\
&\le \max_{k=1,\ldots,N} \left\| n^{-1} \sum_{i=1}^n \Big[m_n(\eta_k, Z_i) - E m_n(\eta_k, Z_i) \Big] \right\| \\
&\quad + \max_k \sup_{\eta \in B_r(\eta_k)} \left\| n^{-1} \sum_{i=1}^n \Big[[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \Big] \right\| \\
&= \max_{k=1,\ldots,N} R_{n,k,1} + \max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \qquad \text{(S13)}
\end{aligned}$$

The strategy from this point is to use truncation and Bernstein's inequality to determine the rate of the first term in (S13), and use the Lipschitz property to control the second term in (S13).

First Term of (S13) . Define the following truncations of the random function $m_n(\eta_k, Z)$:

$$\begin{aligned}
m_n^{(O)}(\eta_k, Z) &= m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| \ge C_n\}, \\
m_n^{(I)}(\eta_k, Z) &= m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\}, \\
\xi_{k,i} &= m_n^{(I)}(\eta_k, Z_i) - E m_n^{(I)}(\eta_k, Z_i),
\end{aligned}$$

for some constant $C_n > 0$, which will be made explicit later. This gives us

$$\begin{aligned}
&\max_{k=1,\dots,N} \left\| n^{-1} \sum_{i=1}^n m_n(\eta_k, Z_i) - Em_n(\eta_k, Z_i) \right\| \\
\le &\max_{k=1,\dots,N} \left\| n^{-1} \sum_{i=1}^n m_n^{(O)}(\eta_k, Z_i) - Em_n^{(O)}(\eta_k, Z_i) \right\| + \max_{k=1,\dots,N} \left\| n^{-1} \sum_{i=1}^n \xi_{k,i} \right\| \\
\le &\max_k \left\| n^{-1} \sum_{i=1}^n m_n^{(O)}(\eta_k, Z_i) \right\| + \max_k \left\| n^{-1} \sum_{i=1}^n Em_n^{(O)}(\eta_k, Z_i) \right\| + \max_k \left\| n^{-1} \sum_{i=1}^n \xi_{k,i} \right\|
\end{aligned} \tag{S14}$$

For the last term in (S14), we have $|\xi_{k,i}| \le 2C_n$, $E\xi_{k,i} = 0$, and

$$\begin{aligned}
\|E(\xi_{k,i}^\top \xi_{k,i})\| =& \bigg\| E\bigg\{ \Big[m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\}) \Big]^\top \\
&\times \Big[m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\}) \Big] \bigg\} \bigg\| \\
\le& \bigg\| E\bigg\{ \Big[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \Big]^\top \Big[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \Big] \bigg\} \bigg\| \\
\le& \| E\{ m_n(\eta_k, Z)^\top m_n(\eta_k, Z) \} \| < \sigma_1^2 \\
\|E(\xi_{k,i} \xi_{k,i}^\top)\| =& \bigg\| E\bigg\{ \Big[m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\}) \Big] \\
&\times \Big[m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z)\mathbb{1}\{|M(Z)| < C_n\}) \Big]^\top \bigg\} \bigg\| \\
\le& \bigg\| E\bigg\{ \Big[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \Big] \Big[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \Big]^\top \bigg\} \bigg\| \\
\le& \| E\{ m_n(\eta_k, Z) m_n(\eta_k, Z)^\top \} \| < \sigma_2^2, .
\end{aligned}$$

Hence, by condition (ii),

$$\sigma_{\xi_k}^2 = \max\left\{ \left\| \sum_{i=1}^n E(\xi_{k,i}^\top \xi_{k,i}) \right\|, \left\| \sum_{i=1}^n E(\xi_{k,i} \xi_{k,i}^\top) \right\| \right\} < n\sigma^2 < na_n.$$

By Bernstein's inequality for matrices [4, 3], for any $\varepsilon_n > 0$,

$$\begin{aligned}
P\left(\left\| n^{-1} \sum_{i=1}^n \xi_{k,i} \right\| > \varepsilon_n \right) \le& P\left(\left\| \sum_{i=1}^n \xi_{k,i} \right\| > n\varepsilon_n \right) \\
\le& 2(d_1 + d_2) \exp\left\{ -\frac{n^2 \varepsilon_n^2}{2\sigma_{\xi_k}^2 + (2/3)2C_n n\varepsilon_n} \right\} \\
\le& 2(d_1 + d_2) \exp\left\{ -\frac{n \varepsilon_n^2}{2a_n + (2/3)2C_n \varepsilon_n} \right\},
\end{aligned}$$

where the last inequality is independent of k and i.

Therefore,

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(\max_k \left\| n^{-1}\sum_{i=1}^{n} \xi_{k,i}\right\| > \varepsilon_n\right) &= \sum_{n=1}^{\infty} P\Bigg(\bigcup_{k=1}^{N}\left\{n^{-1}\sum_{i=1}^{n}\|\xi_{k,i}\| > \varepsilon_n\right\}\Bigg)\\ &\le \sum_{n=1}^{\infty}\sum_{k=1}^{N} P\Bigg(n^{-1}\sum_{i=1}^{n}\|\xi_{k,i}\| > \varepsilon_n\Bigg)\\ &\le \sum_{n=1}^{\infty} N \max_k P\Bigg(n^{-1}\sum_{i=1}^{n}\|\xi_{k,i}\| > \varepsilon_n\Bigg)\\ &\le 2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{n\varepsilon_n^2}{2a_n+(4/3)C_n\varepsilon_n}\right\}.\end{aligned}$$

In particular, if we select ε_n and C_n such that $C_n\varepsilon_n \asymp a_n$, and $n\varepsilon_n^2 \asymp a_n \log n$, say $C_n\varepsilon_n = b_1 a_n$ and $n\varepsilon_n^2 = b_2 a_n \log n$ for $b_1, b_2 > 0$, then we have $\varepsilon_n \asymp (a_n \log n/n)^{1/2}$, $C_n \asymp (a_n n/\log n)^{1/2}$, and

$$\begin{aligned}2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{n\varepsilon_n^2}{2a_n+(4/3)C_n\varepsilon_n}\right\} &=2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{b_2}{2+\frac{4}{3}b_1}\log n\right\}\\ &\le 2(d_1+d_2)\sum_{n=1}^{\infty} N n^{-\frac{b_2}{2+\frac{4}{3}b_1}}.\end{aligned}$$

As we will see later, we will need N to increase with n with rate

$$N(n) \asymp n^{dc_2/c_1}(a_n \log n/n)^{-d/(2c_1)}$$

for constants $c_1, c_2 > 0$. So, if we choose $b_1, b_2 > 0$ large enough so that

$$N(n) n^{-\frac{b_2}{2+\frac{4}{3}b_1}} \prec n^{-c},$$

for some $c > 1$, then

$$\sum_{n=1}^{\infty} P\left(\max_k \left\| n^{-1}\sum_{i=1}^{m}\xi_{k,i}\right\| > b_2^{1/2}(a_n \log n/n)^{1/2}\right) < \infty.$$

By the first Borel-Cantelli Lemma,

$$P\Bigg(\left\{\omega : \max_k \left\| n^{-1}\sum_{i=1}^{n}\xi_{k,i}(\omega)\right\| > b_2^{1/2}(a_n\log n/n)^{1/2}\right\}\ \ i.o.\Bigg) = 0,$$

or equivalently,

$$\max_k \left\| n^{-1}\sum_i \xi_{k,i}\right\| = O_{as}((a_n \log n/n)^{1/2}).$$

For the second term of (S14), we have

$$\begin{aligned}\|Em_n^{(O)}(\eta_k, Z_i)\| \leq& E\|m_n^{(O)}(\eta_k, Z_i)\| \\ =& E[\|m_n(\eta_k, Z_i)\|\mathbb{1}\{|M(Z_i)| \geq C_n\}] \\ \leq& E[|M(Z_i)|\mathbb{1}\{|M(Z_i)| \geq C_n\}] \\ =& C_n^{-(s-1)}E|M(Z_i)|C_n^{s-1}\mathbb{1}\{|M(Z_i)| \geq C_n\} \\ \leq& C_n^{-(s-1)}E|M(Z_i)|^s\mathbb{1}\{|M(Z_i)| \geq C_n\} \\ \leq& C_n^{1-s}E|M(Z_i)|^s,\end{aligned}$$

where $E|M(Z_i)|^s < \infty$ for $s > 2$ by condition (ii). This implies

$$\max_k \left\| n^{-1}\sum_i Em_n^{(O)}(\eta_k, Z_i)\right\| \leq n^{-1}\sum_i C_n^{1-s}E|M(Z)|^s = C_n^{1-s}E|M(Z)|^s.$$

So we need to show $C_n^{1-s} = O(\varepsilon_n)$. But since $C_n \asymp (a_n n/\log n)^{1/2}$, we have

$$\frac{C_n^{1-s}}{\varepsilon_n} = \frac{(a_n n/\log n)^{(1-s)/2}}{(a_n \log n/n)^{1/2}} = \left[\frac{(a_n n/\log n)^{(1-s)}}{(a_n \log n/n)}\right]^{1/2}.$$

The RHS without the square root is

$$\begin{aligned}\frac{a_n^{(1-s)}n^{(1-s)}/(\log n)^{(1-s)}}{a_n \log n/n} =& \frac{a_n^{(1-s)}nn^{(1-s)}}{a_n \log n(\log n)^{(1-s)}} = a_n^{-s}\left(\frac{n}{\log n}\right)^{2-s} = \left[\frac{a_n^{-s/(2-s)}n}{\log n}\right]^{2-s} \\ =& \left[\frac{a_n^{s/(s-2)}n}{\log n}\right]^{2-s}.\end{aligned}$$

Since $s > 2$, and $\liminf_n \frac{a_n^{s/(s-2)}n}{\log n} > 0$ by condition (ii), the RHS is bounded above. Therefore, $C_n^{1-s} = O(\varepsilon_n)$, and consequently,

$$\max_k \left\| n^{-1}\sum_{i=1}^n Em_n^{(O)}(\eta_k, Z_i)\right\| = O((a_n \log n/n)^{1/2}).$$

For the first term in (S14), we have

$$\begin{aligned}\left\| n^{-1}\sum_{i=1}^n m_n^{(O)}(\eta_k, Z_i)\right\| \leq& n^{-1}\sum_{i=1}^n \|m_n(\eta_k, Z_i)\|\mathbb{1}\{|M(Z_i)| \geq C_n\} \\ \leq& C_n^{1-s}n^{-1}\sum_{i=1}^n |M(Z_i)|^s\mathbb{1}\{|M(Z_i)| \geq C_n\} \\ \leq& C_n^{1-s}n^{-1}\sum_{i=1}^n |M(Z_i)|^s.\end{aligned}$$

Taking the maximum over k gives us

$$\max_k n^{-1}\left\|\sum_i m_n^{(O)}(\eta_k, Z_i)\right\| \leq C_n^{1-s}n^{-1}\sum_{i=1}^n |M(Z_i)|^s = O_{\mathrm{as}}((a_n \log n/n)^{1/2}),$$

where the last inequality follows from the strong law of large numbers.

Second Term of (S13)**.** By condition (ii), for any $k = 1, \dots, N$ and any $\eta \in B_r(\eta_k)$, we have

$$\begin{aligned}
&\left\| n^{-1} \sum_{i=1}^{n} \left\{ [m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right\} \right\| \\
&\le n^{-1} \sum_{i=1}^{n} \left\| m_n(\eta, Z_i) - m_n(\eta_k, Z_i) \right\| + n^{-1} \sum_{i=1}^{n} E \left\| m_n(\eta, Z_i) - m_n(\eta_k, Z_i) \right\| \\
&\le n^{-1} \sum_{i=1}^{n} |L(Z_i)| \times \|\eta - \eta_k\|^{c_1} n^{c_2} + n^{-1} \sum_{i=1}^{n} E|L(Z_i)| \times \|\eta - \eta_k\|^{c_1} n^{c_2} \\
&\le r^{c_1} n^{c_2} n^{-1} \sum_{i=1}^{n} |L(Z_i)| + r^{c_1} n^{c_2} E|L(Z)|.
\end{aligned}$$

Hence

$$\max_{k=1,\dots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \le r^{c_1} n^{c_2} \left[n^{-1} \sum_{i=1}^{n} |L(Z_i)| + E|L(Z)| \right],$$

where the strong law ensures the sum converges almost surely to $E|L(Z)|$. So we just need to pick r so that $r^{c_1} n^{c_2} \asymp (a_n \log n/n)^{1/2}$, which gives $r = n^{-c_2/c_1} (a_n \log n/n)^{1/(2c_1)}$. With r thus determined, the number of balls needed to cover E satisfies $N(n) r^d \asymp V$, V being the true volume of E. Hence

$$N(n) \asymp n^{dc_2/c_1} (a_n \log n/n)^{-d/(2c_1)},$$

which we used earlier. Putting together all the results, we conclude that

$$\begin{aligned}
\sup_{\eta \in E} \left\| n^{-1} \sum_{i=1}^{n} [m_n(\eta, Z_i) - E m_n(\eta, Z_i)] \right\| \le & \max_{k=1,\dots,N} R_{n,k,1} + \max_{k} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \\
= & O_{as}((a_n \log n/n)^{1/2}),
\end{aligned}$$

completing the proof. □

The next Corollary shows the Lipschitz condition in Lemma S.8.1 can be replaced by a component-wise Lipschitz condition.

Corollary S.8.1. *Let $\eta = (\theta, x) \in E = \Theta \times \Omega_X$. Then condition (iii) in Lemma* (S.8.1) *can be replaced by the following conditions*

(i) (Lipschitz for $x \in \Omega_X$) For all $x, x' \in \Omega_X$ and $\theta \in \Theta$,

$$\|m_n(\theta, x, Z) - m_n(\theta, x', Z)\| \le \|x - x'\|^{c_1} n^{c_2} L_1(Z),$$

for some $c_1, c_2 > 0$, with $EL_1(Z) < \infty$;

(ii) (Lipschitz for $\theta \in \Theta$) For all $\theta, \theta' \in \Theta$ and $x \in \Omega_X$,

$$\|m_n(\theta, x, Z) - m_n(\theta', x, Z)\| \leq \|\theta - \theta'\|^{c_1'} n^{c_2'} L_2(Z),$$

for some $c_1', c_2' > 0$, with $EL_2(Z) < \infty$.

Proof. We just need to show that the two conditions are sufficient for bounding the second term in (S13). By the triangle inequality,

$$\begin{aligned}
R_{n,k,2} \leq & \left\| n^{-1} \sum_{i=1}^n \left\{ [m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right\} \right\| \\
& \leq n^{-1} \sum_{i=1}^n \|m_n(\eta, Z_i) - m_n(\eta_k, Z_i)\| + n^{-1} \sum_{i=1}^n E\|m_n(\eta, Z_i) - m_n(\eta_k, Z_i)\|.
\end{aligned}$$

By Lipschitz in θ and x, we get

$$\begin{aligned}
& \|m_n(\eta, Z_i) - m_n(\eta_k, Z_i)\| \\
= & \|m_n(\theta, x, Z_i) - m_n(\theta_k, x_k, Z_i)\| \\
\leq & \|m_n(\theta, x, Z_i) - m_n(\theta_k, x, Z_i)\| + \|m_n(\theta_k, x, Z_i) - m_n(\theta_k, x_k, Z_i)\| \\
& + \|m_n(\theta_k, x_k, Z_i) - m_n(\theta, x_k, Z_i)\| + \|m_n(\theta, x_k, Z_i) - m_n(\theta_k, x_k, Z_i)\| \\
\leq & \|\theta - \theta_k\|^{c_1'} n^{c_2'} L_2(Z) + \|x - x_k\|^{c_1} n^{c_2} L_1(Z) + \|\theta - \theta_k\|^{c_1'} n^{c_2'} L_2(Z) \\
& + \|x - x_k\|^{c_1} n^{c_2} L_1(Z) \\
\leq & 2r^{c_1'} n^{c_2'} L_2(Z) + 2r^{c_1} n^{c_2} L_1(Z),
\end{aligned}$$

where the last inequality follows from $\|\eta - \eta_k\| \leq r$, which implies $\|x - x_k\| \leq r$ and $\|\theta - \theta_k\| \leq r$. So

$$\begin{aligned}
\max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \leq & 2r^{c_1'} n^{c_2'} \left[n^{-1} \sum_{i=1}^n L_2(Z_i) + E|L_2(Z)| \right] \\
& + 2r^{c_1} n^{c_2} \left[n^{-1} \sum_{i=1}^n L_1(Z_i) + EL_1(Z) \right].
\end{aligned}$$

The strong law of large numbers ensures the averages converge almost surely finite constants. So we need to pick r such that $r^{c_1} n^{c_2} \asymp \varepsilon_n$ and $r^{c_1'} n^{c_2'} \asymp \varepsilon_n$. Without loss of generality, assume $r^{c_1'} n^{c_2'} \leq r^{c_1} n^{c_2}$. Then

$$\begin{aligned}
& \max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \\
\leq & 2r^{c_1} n^{c_2} \left[n^{-1} \sum_{i=1}^n |L_2(Z_i)| + E|L_2(Z)| + n^{-1} \sum_{i=1}^n |L_1(Z_i)| + E|L_1(Z)| \right].
\end{aligned}$$

We then set $r \asymp n^{-c_2/c_1} (a_n \log n/n)^{1/(2c_1)}$ and $N \asymp n^{dc_2/c_1} (a_n \log n/n)^{-d/(2c_1)}$, as we did originally in the Lemma, where V is the volume of E. This completes the proof. □

The next corollary provides a convergence rate that is uniform in one component of η, but pointwise in the other.

Corollary S.8.2. *Let $\eta = (\theta, x) \in E = \Theta \times \Omega_X$ and replace condition (iii) in Lemma S.8.1 with condition (ii) in Corollary S.8.1. Then*

$$\sup_{\theta \in \Theta} \left\| E_n m_n(\theta, x, Z) - E m_n(\theta, x, Z) \right\| = O_{as}\left(\sqrt{\frac{a_n \log n}{n}} \right).$$

Proof. Let $\tilde{m}_n(\theta, Z) = m_n(\theta, x, Z)$. Applying Lemma S.8.1 to $\tilde{m}_n$ gives us the result. □

8.2. Convergence Rates for Sums of Bounded Outer Products

Lemma S.8.2. *Suppose $X_1, ..., X_n$ are i.i.d, random vectors in $\mathbb{R}^p$. Let $g : \mathbb{R}^p \to \mathbb{R}^{d_1 \times d_2}$ be a continuous function such that $\|g(X_i)\| \le K$ almost surely, in operator norm, for all i. Let $\mu = E\{g(X)g(X)^\top\}$. Then,*

(i)

$$\left\| \frac{1}{n} \sum_{i=1}^n \{g(X_i)g(X_i)^\top - \mu\} \right\| = O_p(n^{-1/2}).$$

(ii)

$$\left\| \frac{1}{n} \sum_{i=1}^n \{g(X_i)g(X_i)^\top - \mu\} \right\| = O_{\text{as}}\left(\sqrt{\frac{\log n}{n}} \right).$$

Proof. Let $Z_i = g(X_i)g(X_i)^\top - \mu$ so that $\|Z_i\| \le 2K^2$ almost surely, $E(Z_i) = 0$, and $\text{var}(Z_i)$ has finite entries. For (i), it suffices to show that, for any fixed $\varepsilon \ge 0$, there is a sequence $c_n \asymp n^{-1/2}$ such that

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \ge c_n \right) < \varepsilon, \quad \forall n.$$

By Matrix Chebyshev's inequality [4, 3],

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \ge c_n \right) < \frac{1}{n^2 c_n^2} E \left\| \sum_i Z_i \right\|^2 \le \frac{1}{n c_n^2} E\|Z_1\|^2.$$

So, choosing $c_n = \{E(\|Z_1\|^2)\}^{1/2}(n\varepsilon)^{-1/2}$ gives

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \ge \left(\frac{E(\|Z_1\|^2)}{\varepsilon} \right)^{1/2} n^{-1/2} \right) < \varepsilon, \quad \forall n,$$

which proves (i).

For (ii), we appeal to Bernstein's Inequality for Matrices and the first Borel-Cantelli Lemma. Since $\|Z_i\| \leq 2K^2$, we have

$$\sigma^2 = \left\| \sum_{i=1}^n E(Z_i^2) \right\| \leq E\|g(X_i)g(X_i)^\top g(X_i)g(X_i)^\top\| \leq nK^4.$$

By the matrix version of Bernstein's inequality [4, 3], we have, for any $c > 0$,

$$\begin{aligned} P\left(\left\| \sum_i Z_i \right\| \geq c\sqrt{n \log n} \right) \leq & 2(p+1) \exp \left\{ -\log n \frac{c^2/2}{K^4 + \frac{1}{3}K^2 c \sqrt{\frac{\log n}{n}}} \right\} \\ \equiv & 2(p+1) n^{-d_n}, \end{aligned}$$

where $d_n = \frac{c^2}{2K^4 + \frac{2K^2 c\sqrt{\log n}}{3\sqrt{n}}}$ for $n = 1, 2, \ldots$. Note that $d_0 < d_n \uparrow d$, where $d_0 = \frac{c^2}{2K^4 + \frac{2K^2 c}{3\sqrt{e}}}$ and $d = \frac{c^2}{2K^4}$. Summing over n, we get

$$\sum_{n=1}^\infty P\left(\left\| \sum_i Z_i \right\| \geq c\sqrt{n \log n} \right) \leq 2(p+1) \sum_{n=1}^\infty n^{-d_n} \leq 2(p+1) \sum_{n=1}^\infty n^{-d_0}.$$

So we need to choose c so that the series on the right converges. This can be achieved by choosing

$$1 < \frac{c^2}{2\sigma^2 + \frac{2Kc}{3\sqrt{e}}} \implies 0 < 3\sqrt{e}c^2 - 2Kc - 6\sqrt{e}\sigma^2 \implies c > \frac{2K + \sqrt{4K^2 + 72e\sigma^2}}{6\sqrt{e}}.$$

Then the first Borel-Cantelli Lemma then gives us $\|\sum_i Z_i\| = O_{\text{as}}(\sqrt{n \log n})$, which is equivalent to (ii).

□

8.3. Bai, Miao, and Rao's Lemma

This Lemma is from [1]. In the following, let $O(c)$ denote a number whose absolute value is bounded by Mc, or a matrix whose entries have absolute values bounded by Mc, where M is any constant independent of c. Let $O(c^k)$ be similarly defined. Note that $O(c)O(c) = O(c^2)$, whether $O(c)$ is a number or a matrix.

Lemma S.8.3. *Let $A = (a_{ik})$ and $B = (b_{ik})$ be two symmetric $p \times p$ matrices with spectral decomposition*

$$A = \sum_{k=1}^p \delta_k u_k u_k^\top, \quad \delta_1 \geq \delta_2 \geq \cdots \geq \delta_p,$$

$$B = \sum_{k=1}^p \lambda_k v_k v_k^\top, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

where δ_k, λ_k and u_k, v_k are the eigenvalues and corresponding orthonormal eigenvectors of A and B, respectively. Suppose there are s distinct eigenvalues, denoted by $\tilde{\lambda}_1 > \tilde{\lambda}_2 > \cdots > \tilde{\lambda}_s$, where each $\tilde{\lambda}_h$ has multiplicity m_h. Let $n_h = m_1 + \cdots + m_h$, so that $0 < n_1 < \cdots < n_h = p$. Set $n_0 = 0$.

If $|a_{ik} - b_{ik}| < c$ for all $i, k = 1, ..., p$, then there exist constants $M_1, M_2 > 0$, independent of c, such that

(i) $|\delta_k - \lambda_k| < M_1 c, \quad k = 1, 2, ..., p$ (i.e. $\delta_k = \lambda_k + O(c)$), and

(ii) if $C^{(h)} = (c_{lk}^{(h)})$ is the $p \times p$ matrix defined by

$$C^{(h)} = \sum_{k=n_{h-1}+1}^{n_h} u_k u_k^\top - \sum_{k=n_{h-1}+1}^{n_h} v_k v_k^\top,$$

then $|c_{lk}^{(h)}| \leq M_2 c$ for all $k, l = 1, \ldots, p$ and $h = 1, \ldots, s$.

Intuitively, this lemma says that the differences between the eigenvalues and eigenvectors of two matrices are of the same order of magnitude as the differences of the entries of the two matrices.

Proof. (i). By Von-Neumann's Trace Inequality, $|\mathrm{tr}(AB)| \leq \sum_{i=1}^p \delta_i \lambda_i$, we have

$$\begin{aligned}
\sum_{i=1}^p (\delta_i - \lambda_i)^2 &= \sum_{i=1}^p \delta_i^2 + \sum_{i=1}^p \lambda_i^2 - 2\sum_{i=1}^p \delta_i \lambda_i \\
&\leq \mathrm{tr}(A^2 + B^2 - 2AB) \\
&= \mathrm{tr}\{(A-B)(A-B)\} \\
&= \sum_{i,j} (a_{ij} - b_{ij})^2 = p^2 c^2,
\end{aligned}$$

which implies $(\delta_i - \lambda_i)^2 < p^2 c^2$, proving (i) with M_1 being p.

(ii). By assertion (i),

$$A = \sum_{i=1}^p \delta_i u_i u_i^\top = \sum_{i=1}^p (\lambda_i + O(c)) u_i u_i^\top = \sum_{i=1}^p \lambda_i u_i u_i^\top + O(c) = \sum_{h=1}^s \tilde{\lambda}_h \sum_{i \in L_h} u_i u_i^\top + O(c),$$

where $L_h = \{n_{h-1} + 1, \ldots, n_h\}$. Then, from boundedness assumption $A = B + O(c)$, we get

$$\sum_{h=1}^s \tilde{\lambda}_h \sum_{i \in L_h} u_i u_i^\top + O(c) = A = B + O(c) = \sum_{h=1}^s \tilde{\lambda}_h \sum_{i \in L_h} v_i v_i^\top + O(c).$$

Let $P_h(A) \equiv \sum_{i \in L_h} u_i u_i^\top$ and $P_h(B) \equiv \sum_{i \in L_h} v_i v_i^\top$. Then the above equation can be written as $\sum_{h=1}^s \tilde{\lambda}_h P_h(A) = \sum_{h=1}^s \tilde{\lambda}_h P_h(B) + O(c)$.

We next use mathematical induction to prove (ii). If $s = 1$, then

$$\tilde{\lambda}_1 P_1(A) = \tilde{\lambda}_1 P_1(B) + O(c) \Rightarrow P_h(A) = P_h(B) + O(c),$$

so (ii) holds for $s = 1$. Assume, for induction, that (ii) holds for $s = t - 1$; that is, $P_h(A) = P_h(B) + O(c)$, for all $h = 1, ..., t - 1$. Then

$$\begin{aligned}
&\sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) + O(c) \\
&\Rightarrow \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(A) - \tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) - \tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(A) + O(c) \\
&\Rightarrow \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) - \left(\tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(B) + O(c) \right) + O(c) \\
&\Rightarrow \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) = \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(B) + O(c).
\end{aligned}$$

Hence, for any $v \in P_t(B)$, which is orthogonal to $P_1(B), ..., P_{t-1}(B)$, we have

$$\sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) v = \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(B) v + O(c) \Rightarrow \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) v = O(c).$$

Since $\tilde{\lambda}_h - \tilde{\lambda}_t > 0$ for all $h = 1, .., s$, we have $P_h(A)v = O(c)$ for all $v \in P_t(B)$, which means $\langle P_h(A), P_t(B) \rangle = O(c)$, for $h = 1, ..., t-1$, where the inner product between matrices S, T of the same dimension is defined as $\langle S, T \rangle = \mathrm{tr}(ST^\top)$.

Let U_1 be an orthonormal basis (ONB) for $\oplus_{h=1}^{t-1} P_h(A)$, U_2 an ONB for $P_t(A)$, and $U = (U_1, U_2)$. Similarly, let V_1 be an ONB for $\oplus_{h=1}^{t-1} P_h(B)$, V_2 an ONB for $P_t(B)$, and $V = (V_1, V_2)$. Then $\langle P_h(A), P_t(B) \rangle = O(c)$ implies that $\langle U_1, V_2 \rangle = O(c)$ and, by symmetry, $\langle V_1, U_2 \rangle = O(c)$. Furthermore, since $V = (V_1, V_2)$ forms a basis, we can express U_2 in terms of the bases V_1, V_2 as $U_2 = V_1 G_1 + V_2 G_2$, $G_1 \in \mathbb{R}^{n_{t-1} \times (n_t - n_{t-1})}$ and $G_2 \in \mathbb{R}^{(n_t - n_{t-1}) \times (n_t - n_{t-1})}$.

Since $U = (U_1, U_2)$ is an ONB for $\mathbb{R}^p$, we have $U_2 U_2^\top = I_p - U_1 U_1^\top$. Therefore,

$$V_2^\top U_2 U_2^\top V_2 = V_2^\top (I_p - U_1 U_1^\top) V_2 = V_2^\top V_2 + O(c)O(c) = I_{p-n_{t-1}} + O(c^2).$$

Since $\langle V_1, U_2 \rangle = O(c)$, we also have

$$O(c) = \langle V_1, U_2 \rangle = \langle V_1, V_1 G_1 + V_2 G_2 \rangle = I_{n_{t-1}} G_1 = G_1,$$

implying $G_1 = O(c)$, and so $U_2 = V_1 O(c) + V_2 G_2 + O(c) = V_2 G_2 + O(c)$. Now note that

$$\begin{aligned}
G_2 G_2^\top &= V_2^\top V_2 G_2 G_2^\top V_2^\top V_2 \\
&= V_2^\top (U_2 + O(c))(U_2^\top + O(c)) V_2 \\
&= V_2^\top U_2 U_2^\top V_2 + O(c) \\
&= I_{p-n_{t-1}} + O(c^2) + O(c) \\
&= I_{p-n_{t-1}} + O(c).
\end{aligned}$$

This gives us

$$\begin{aligned}P_t(A) = U_2U_2^\top &= V_2G_2G_2^\top V_2^\top + O(c) + O(c^2)\\ &= V_2(I_{p-n_{t-1}} + O(c))V_2^\top + O(c)\\ &= V_2V_2^\top + O(c)\\ &= P_t(B) + O(c).\end{aligned}$$

By the induction assumption, we conclude that $P_h(A) = P_h(B) + O(c)$ for $h = 1, .., t$, completing the proof. □

References

[1] Bai, D. Z., Miao, Q. B. and Rao, R. C. (1991). Estimation of directions of arrival of signals: Asymptotic results. In *Advances in spectrum analysis and array processing*, (S. Haykin, ed.) **II** 9, 327-347. Prentice Hall, Englewood Cliffs, NJ.

[2] Luo, W. and Li, B. (2020). On order determination by predictor augmentation. *Biometrika*.

[3] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*.

[4] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.

[5] Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137.

[6] Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35** 2654–2690.