A note on a nonparametric regression test through P-spline

Huaihou Chen[1], Yuanjia Wang[2], Runze Li[3], and Katherine Shear[2]

[1]*New York University,* [2]*Columbia University, and* [3]*Pennsylvania State University*

Supplementary Material

First we state our assumptions (see also Zhou et al. (1998) and Claeskens et al. (2009)).

Assumption 1. Let $\delta_j = \tau_{j+1} - \tau_j$ and $\delta = \max\limits_{0\le j\le K} \delta_j$, where $\tau_1, \cdots, \tau_K$ are the K knots. There exists a constant $M > 0$, such that $\delta/(\min\limits_{0\le j\le K} \delta_j) \le M$ and $\delta \sim K^{-1}$. This assumption is a weak restriction on the knot distribution, and assures that $M^{-1} < K\delta < M$, which is required for stable numerical computations.

Assumption 2. For design points $u_i \in [a, b]$, $i = 1, \cdots, n$, there exists a distribution function Q with corresponding positive continuous design density ρ such that, with Q_n the empirical distribution of $u_1, \cdots, u_n$, $\sup\limits_{u\in[a,b]} |Q_n(u) - Q(u)| = o(K^{-1})$.

Assumption 3. The number of knots $K = o(n)$.

S1 Proof of Theorem 1

We state a Lemma proved in Eubank and Spiegelman (1990).

Lemma 1 *(Eubank and Spiegelman 1990). Let $\boldsymbol{M}_n$ denote a sequence of $n \times n$ symmetric positive semidefinite matrices with eigenvalues $\tau_{1n} \le \cdots \le \tau_{nn}$. Assume that $\boldsymbol{y}_n \sim \boldsymbol{N}_n(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{I}_n)$. Then*

$$\frac{\boldsymbol{y}_n^{\mathrm{T}} \boldsymbol{M}_n \boldsymbol{y}_n - \sigma^2 trace(\boldsymbol{M}_n) - \boldsymbol{\mu}_n^{\mathrm{T}} \boldsymbol{M}_n \boldsymbol{\mu}_n}{\sigma^2 \{2 trace(\boldsymbol{M}_n^2)\}^{1/2}} \to N(0,1), \quad as \quad n \to \infty \quad if$$

(A). $\max\limits_i \tau_{ni}^2 / \sum_{i=1}^n \tau_{ni}^2 \to 0$ *and*

(B). $\boldsymbol{\mu}_n^{\mathrm{T}} \boldsymbol{M}_n^2 \boldsymbol{\mu}_n / trace(\boldsymbol{M}_n^2) \to 0.$

Next we state Lemma A3 in Claeskens et al. (2009) which is adapted from Speckman (1985).

Lemma 2 *(Claeskens et al. 2009) Under Assumption A2 and for the eigenvalues obtained in S,*

$$s_1 = \cdots = s_q = 0, \quad s_j = n^{-1}(j-q)^{2q}\widehat{c}_1 \text{ for } j = q+1, \cdots, K+p+1, \tag{S1.1}$$

where $\tilde{c}_1 = c_1(1+o(1))$ *with* c_1 *as a constant depending only on* q *and the design density and o(1) converges to 0 as* $n \to \infty$ *uniformly for* $j_{1n} \le j \le j_{2n}$ *for any sequences* $j_{1n} \to \infty$ *and* $j_{2n} = o(n^{\frac{2}{2q+1}})$.

Proof of Theorem 1. Theorem 1 follows as a direct application of Lemma 1. We verify condition (A) in Lemma 1. Using Lemma 2, we have

$$\begin{aligned}
\text{trace}(\boldsymbol{H}_n^4) &= \text{trace}[\{\boldsymbol{A}(\boldsymbol{I}_{K+p+1} + \lambda \boldsymbol{S})^{-1}\boldsymbol{A}^{\mathrm{T}}\}^4] = \sum_{j=1}^{K+p+1} \frac{1}{(1+\lambda s_j)^4} \\
&= q + \sum_{j=q+1}^{K+p+1} \frac{1}{\{1+\lambda n^{-1}\tilde{c}_1 (j-q)^{2q}\}^4} \\
&= q + (\frac{\lambda \tilde{c}_1}{n})^{-\frac{1}{2q}} \int_0^{K_q} \frac{1}{(1+u^{2q})^4} du + r_n,
\end{aligned}$$

where $r_n = O(1)$ is the residual term from the Euler-Maclaurin formula. If $K_q = o(1)$, then

$$\begin{aligned}
\text{trace}(\boldsymbol{H}_n^4) &= q + (\frac{\lambda \tilde{c}_1}{n})^{-\frac{1}{2q}} \int_0^{K_q} \frac{1}{(1+u^{2q})^4} du + r_n \\
&= q + (\frac{\lambda \tilde{c}_1}{n})^{-\frac{1}{2q}} K_q c + \tilde{r}_n \\
&= q + cK + \tilde{r}_n,
\end{aligned}$$

where c is a bounded constant $\tilde{r}_n = O(1)$. The second equality follows from integral intermediate value theorem. If $K_q = O(1)$, then

$$\int_0^{K_q} \frac{1}{(1+u^{2q})^4} du \le 1 + \int_1^{\infty} u^{-8q} du = 2,$$

therefore

$$\text{trace}(\boldsymbol{H}_n^4) = \left(\frac{\lambda}{n}\right)^{-\frac{1}{2q}} + O(1). \tag{S1.2}$$

To verify condition (A) in Lemma 1, note that when $K_q = o(1)$ or $K_q = O(1)$,

$$\frac{\max_j \{1/(1+\lambda s_j)^2\}}{\text{trace}(\boldsymbol{H}_n^4)} \quad = \quad O(K^{-1}) \to 0.$$

Since under the null hypothesis $\boldsymbol{\mu}_n = \mathbf{0}$, condition (B) is automatically satisfied. Therefore T_n is asymptotically normal under the H_0. $\square$

S2 Proof of Theorem 2

The following notation will be used. Set

$$\tilde{\sum_{ij}} = \sum_{i,i\neq j}\sum_{j}, \quad \tilde{\sum_{ijk}} = \sum_{i,i\neq j}\sum_{j,j\neq k}\sum_{k}, \quad \text{and} \quad \tilde{\sum_{ijkl}} = \sum_{i,i\neq j,j\neq l}\sum_{j,j\neq k,k\neq l}\sum_{k,i\neq l}\sum_{l,j\neq k}.$$

The following Lemma is from Chen (1994), which is an application of the results in De Jong (1987).

Lemma 3 *(Chen 1994) Let $\boldsymbol{y}_n = (y_1, \cdots, y_n)^T$ be a random vector and set $\boldsymbol{\mu}_n = (f_1, \cdots, f_n)^T$. Define $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T = \boldsymbol{y}_n - \boldsymbol{\mu}_n$, and suppose $\epsilon_1, \cdots, \epsilon_n$ are independent, identically distributed random variables with $E(\epsilon_1) = 0$, $var(\epsilon_1) = \sigma^2$ and $0 < E(\epsilon_1^4) < \infty$. Let $\boldsymbol{M}_n$ be a symmetric $n \times n$ matrix of constants and m_{lj} be its (l,j)th element with $m_{lj}^{(k)}$ denoting the (l,j)th element of $\boldsymbol{M}_n^k$, for $k = 2, 3, \cdots$. Define $\sigma^2(n) = \sum_{j=1}^{n}(m_{jj}^{(2)} - m_{jj}^2)$, $\alpha_1 = \tilde{\sum_{l,j}} m_{lj}^4$, $\alpha_2 = \tilde{\sum_{l,j,k}} m_{ij}^2 m_{lk}^2$ and $\alpha_3 = \tilde{\sum_{i,j,k,l}} m_{ij}m_{ik}m_{lj}m_{lk}$. Then,*

$$A_n = \frac{\boldsymbol{y}_n^T\boldsymbol{M}_n\boldsymbol{y}_n - \sigma^2 trace(\boldsymbol{M}_n) - \boldsymbol{\mu}_n^T\boldsymbol{M}_n\boldsymbol{\mu}_n}{\sigma^2\sqrt{2trace(\boldsymbol{M}_n^2)}} \to N(0,1), \quad as \quad n \to \infty \quad if \tag{S2.1}$$

A. $\sum_j m_{jj}^2/trace(\boldsymbol{M}_n^2) \to 0$ as $n \to \infty$,

B. $\boldsymbol{\mu}_n^T\boldsymbol{M}_n^2\boldsymbol{\mu}_n/trace(\boldsymbol{M}_n^2) \to 0$ as $n \to \infty$, and

C. $\alpha_j = o(\sigma^4(n))$ for $j = 1, 2, 3$ as $n \to \infty$.

Define $\boldsymbol{H}_{K,n} = \frac{1}{n}(\boldsymbol{N}^T\boldsymbol{N} + \lambda\boldsymbol{D}_q)$. The following lemma is adapted from the Lemma A1 in Claesken et al. (2009).

Lemma 4 *There exists a constant $c_0 > 0$ independent of K and n such that $|\{\boldsymbol{H}_{K,n}^{-1}\}_{i,j}| \leq c_0K$ for $K_q = o(1)$ and $|\{\boldsymbol{H}_{K,n}^{-1}\}_{i,j}| \leq c_0K(1 + K_q^{2q})^{-1}$ for $K_q = O(1)$.*

Proof of Theorem 2. Let $\boldsymbol{H}_n = \boldsymbol{N}(\boldsymbol{N}^T\boldsymbol{N} + \lambda\boldsymbol{D}_q)^{-1}\boldsymbol{N}^T = \frac{1}{n}\boldsymbol{N}\boldsymbol{H}_{K,n}^{-1}\boldsymbol{N}^T$. When $K_q = o(1)$, the (i,j)th element of $\boldsymbol{H}_n$ can be bounded as following:

$$|H_{ij}| = \frac{1}{n}|\sum_{k=1}^{K+p+1}\sum_{l=1}^{K+p+1} N_{ik}\{\boldsymbol{H}_{K,n}^{-1}\}_{kl}N_{jl}| \leq \frac{1}{n}c_0K\sum_{k=1}^{K+p+1}\sum_{l=1}^{K+p+1} N_{ik}N_{jl} = \frac{c_0K}{n}.$$

Let h_{ij} be the (i,j)th element of $\boldsymbol{H}_n^2$, then

$$|h_{ij}| = |\sum_{k=1}^{n} H_{ik}H_{kj}| \leq c_0^2\sum_{k=1}^{n}(\frac{K}{n})^2 = c_0^2\frac{K^2}{n}.$$

Note that

$$\frac{\sum_{i=1}^{n} h_{ii}^2}{\text{trace}\{\boldsymbol{H}_n^4\}} \sim \frac{n(\frac{K^2}{n})^2}{K} = \frac{K^3}{n} \to 0, \quad \text{as} \quad n \to \infty. \tag{S2.2}$$

This shows that condition (A) in Lemma 3 holds. It is obvious that condition (B) in Lemma 3 is true, since $\mu_n = 0$ under the null hypothesis. Thus it remains to prove condition (C). Following (S2.2) to obtain $\sigma^2(n) \sim \frac{K^4}{n}$. We have

$$\begin{aligned}
\alpha_1 &= \tilde{\sum_{i,j}} h_{ij}^4 \le c_0^8 \frac{K^8}{n^2} = c_0^8 (\frac{K^4}{n})^2 = o(\sigma^4(n)). \\
\alpha_2 &= \tilde{\sum_{i,j,k}} h_{ij}^2 h_{ik}^2 = \tilde{\sum_{i,j}} h_{ij}^2 (h_{ii}^{(2)} - h_{ii}^2 - h_{jj}^2) \sim \tilde{\sum_{i,j}} h_{ij}^2 h_{ii}^{(2)} \\
& \quad \tilde{\sum_{i,j}} h_{ij}^2 h_{ii}^{(2)} \le \tilde{\sum_{i,j}} h_{ii}^{(2)} c_0^4 (\frac{K^2}{n})^2 \sim \frac{K^5}{n} = o(\sigma^4(n)). \\
\alpha_3 &= \tilde{\sum_{i,j,k,l}} h_{ij} h_{ik} h_{lj} h_{lk} = \tilde{\sum_{k,j}} (h_{jk}^{(2)} - h_{jj} h_{jk} - h_{kj} h_{kk})^2 - \alpha_2.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\tilde{\sum_{k,j}} (h_{jk}^{(2)})^2 \le \sum_{j=1}^{n} h_{jj}^{(4)} = o(\sigma^4(n)); \\
&\tilde{\sum_{k,j}} h_{jj}^2 h_{jk}^2 \sim (\frac{K^4}{n})^2 = o(\sigma^4(n)); \\
&\tilde{\sum_{k,j}} h_{jk}^{(2)} h_{jj} h_{jk} \sim \tilde{\sum_{k,j}} h_{jk}^{(2)} (\frac{K^2}{n})^2 \sim \frac{K^5}{n} = o(\sigma^4(n)); \\
&\tilde{\sum_{k,j}} h_{jk}^2 h_{jj} h_{jk} \sim (\frac{K^4}{n})^2 = o(\sigma^4(n)).
\end{aligned}$$

Therefore, $\alpha_3 = o(\sigma^4(n))$ and condition (C) holds. A direct application of Lemma 3 completes the proof for $K_q = o(1)$ case.

Similarly when $K_q = O(1)$, we have $|h_{ij}| \sim K^2 n^{-1} K_q^{-4q}$. Note that

$$\frac{\sum_{i=1}^{n} h_{ii}^2}{\text{trace}\{\boldsymbol{H}_n^4\}} \sim \frac{n(\frac{K^2}{n})^2 K_q^{-8q}}{(\frac{\lambda}{n})^{-1/2q}} = \frac{1}{n(\frac{\lambda}{n})^{3/2q} K_q^{7q-3}} \to 0, \quad \text{as} \quad n \to \infty. \tag{S2.3}$$

This shows that condition (A) in Lemma 3. It is obvious that condition (B) in Lemma 3 is true, since $\mu_n = 0$ under the null hypothesis. To prove condition (C), following (S2.3) to get $\sigma^2(n) \sim (\frac{\lambda}{n})^{-1/2q}$. We obtain

$$\alpha_1 = \tilde{\sum_{i,j}} h_{ij}^4 \sim \frac{1}{n^2(\frac{\lambda}{n})^{4/q} K_q^{16q-8}} = o(\sigma^4(n)).$$

$$\alpha_2 = \tilde{\sum_{i,j,k}} h_{ij}^2 h_{ik}^2 = \tilde{\sum_{i,j}} h_{ij}^2 (h_{ii}^{(2)} - h_{ii}^2 - h_{jj}^2) \sim \tilde{\sum_{i,j}} h_{ij}^2 h_{ii}^{(2)} = o(\sigma^4(n)).$$

$$\alpha_3 = \tilde{\sum_{i,j,k,l}} h_{ij} h_{ik} h_{lj} h_{lk} = \tilde{\sum_{k,j}} (h_{jk}^{(2)} - h_{jj} h_{jk} - h_{kj} h_{kk})^2 - \alpha_2 = o(\sigma^4(n)).$$

A direct application of Lemma 3 finishes the proof of Theorem 2. □

S3 Proof of Theorem 3 and its remarks

Under the alternative hypothesis, we obtain

$$\begin{aligned} T_n^* &= (\boldsymbol{Y} - \boldsymbol{\mu}_n)^T \boldsymbol{H}_n^2 (\boldsymbol{Y} - \boldsymbol{\mu}_n) = T_n + \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n - 2\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{Y} \\ &= T_n - \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n - 2\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n, \end{aligned}$$

therefore

$$T_n = T_n^* + \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n + 2\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n.$$

Note that

$$\begin{aligned} \frac{T_n - \sigma^2 \text{trace}(\boldsymbol{H}_n^2)}{\sigma^2 \{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} &= \frac{T_n^* - \sigma^2 \text{trace}(\boldsymbol{H}_n^2) + \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n + 2\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n}{\sigma^2 \{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} \\ &= \frac{T_n^* - \sigma^2 \text{trace}(\boldsymbol{H}_n^2)}{\sigma^2 \{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} + \frac{\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n}{\sigma^2 \{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} + \frac{2\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n}{\sigma^2 \{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} \\ &\triangleq s_{n1} + s_{n2} + s_{n3}. \end{aligned}$$

From the proof of Theorem 1, we obtain $s_{n1} \to^d N(0,1)$. In addition, it is straightforward that

$$\text{var}(s_{n3}) = \text{var}(\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n) = E(\boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n) = \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 E(\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^T) \boldsymbol{H}_n^2 \boldsymbol{\mu}_n = \sigma^2 \boldsymbol{\mu}_n^T \boldsymbol{H}_n^4 \boldsymbol{\mu}_n.$$

Since

$$\frac{\sigma^2 \boldsymbol{\mu}_n^T \boldsymbol{H}_n^4 \boldsymbol{\mu}_n}{\{\sigma^2 \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n\}^2} \to 0,$$

by Chebyshev's inequality we obtain $s_{n3}/s_{n2} \to^P 0$. Further notice that

$$\lambda_{min}(\boldsymbol{H}_n^2) \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n \le \boldsymbol{\mu}_n^T \boldsymbol{H}_n^2 \boldsymbol{\mu}_n \le \lambda_{max}(\boldsymbol{H}_n^2) \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n,$$

where $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ denote the smallest and largest eigenvalue of the matrix $\boldsymbol{M}$, and

$$\frac{1}{n}\boldsymbol{\mu}_n^T\boldsymbol{\mu}_n = \frac{1}{n}\|\boldsymbol{\mu}_n\|^2 = \frac{1}{n}\sum_{i=1}^n f^2(u_i) = E[f^2(u_1)] + o(1) = \|f\|_u^2 + o(1).$$

From $\lambda_{\max}(\boldsymbol{H}_n^2) = 1$ and $\lambda_{\min}(\boldsymbol{H}_n^2) = \dfrac{1}{(1+K_q^{2q})^2}$, we obtain $\boldsymbol{\mu}_n^T\boldsymbol{H}_n^2\boldsymbol{\mu}_n = O(n)$. To obtain detectable rates under local alternatives, note that for $K_q = o(1)$, we have

$$\frac{\boldsymbol{\mu}_n^T\boldsymbol{H}_n^2\boldsymbol{\mu}_n}{\{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} = O\left(\frac{n}{K^{1/2}}\right),$$

and for $K_q = O(1)$ or $K_q \to \infty$, we obtain

$$\frac{\boldsymbol{\mu}_n^T\boldsymbol{H}_n^2\boldsymbol{\mu}_n}{\{2\text{trace}(\boldsymbol{H}_n^4)\}^{1/2}} = O\left(\frac{n}{(\frac{\lambda}{n})^{-\frac{1}{4q}}}\right). \tag{S3.1}$$

To examine the optimal rate of K and λ, note that

$$\frac{1}{n}\boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{H}_n^2\boldsymbol{\mu}_n = \frac{1}{n}(E\widehat{\boldsymbol{f}_n})^{\mathrm{T}}E\widehat{\boldsymbol{f}_n}.$$

Theorem 2 in Claeskens et al. (2009) and Theorem 1 in Chen and Wang (2011) gives convergence rate of $E[\widehat{f}(u_i)]$. Therefore when $K_q = o(1)$ we obtain

$$\frac{1}{n}\boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{H}_n^2\boldsymbol{\mu}_n = O\left(\frac{\lambda^2K^{2q}}{n^2}\right) + O\left(\frac{1}{K^{2(p+1)}}\right) + O(\|f\|_u^2).$$

In this case, local alternatives are detectable at the rate $h(n) = 1/\sqrt{nK^{-1/2}}$. Therefore at these detectable local alternatives denoted by f^*, we have $\|f^*\|^2 = O(h^2(n)) = 1/(nK^{-1/2})$. The optimal rate for K and λ is obtained from

$$h^2(n) = \frac{1}{K^{2(p+1)}}, \quad \text{and} \quad h^2(n) \geq \frac{\lambda^2K^{2q}}{n^2},$$

which implies

$$K = O(n^{\frac{2}{4p+5}}), \text{ and } \lambda = O(n^{\nu}) \quad \text{for} \quad \nu \leq \frac{2p-2q+3}{4p+5}.$$

Similarly, for $K_q = O(1)$ we obtain

$$\frac{1}{n}\boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{H}_n^2\boldsymbol{\mu}_n = O\left(\frac{\lambda}{n}\right) + O\left(\frac{1}{K^{2(p+1)}}\right) + O(\|f\|^2),$$

and the local alternatives are detectable at the rate $g(n) = \{n(\lambda/n)^{1/4q}\}^{-1/2}$. Therefore the optimal rate of K and λ for testing is obtained from

$$g^2(n) = \frac{\lambda}{n}, \text{ and } g^2(n) \geq \frac{1}{K^{2(p+1)}},$$

which implies

$$\lambda = O(n^{\frac{1}{4q+1}}), \text{ and } K = O(n^{\nu}) \text{ for } \nu \geq \frac{2q}{(4q+1)(p+1)}.$$

□

Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, NY 10016, U.S.A.

E-mail: huaihou.chen@nyumc.org

Department of Biostatistics, Columbia University, New York, NY 10032, U.S.A.

E-mail: yw2016@columbia.edu

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, Pennsylvania, 16802 USA.

E-mail: rzli@psu.edu

School of Social Work, Columbia University, New York, NY 10027, U.S.A.

E-mail: ks2394@columbia.edu