SUPPLEMENTARY MATERIAL FOR "MANIFOLD FITTING BY RIDGE ESTIMATION: A LOCAL KERNEL DENSITY ESTIMATION APPROACH"

BY ZHIGANG YAO, AND ZHENG ZHAI

National University of Singapore

In this supplement we present the technical proofs for the main work. Equation and theorem references made to the main document do not contain letters.

APPENDIX A: PROOFS FOR MAIN THEOREMS AND LEMMAS

In this section, we prove all key theorems and lemmas in the order they appear.

A.1. Ridge Derivative Lemma.

LEMMA A.1. *For any R_1, R_2, and any point $x_1 \in R_1$, the pairwise distance from x_1 to R_2 yields the order of:*

$$\min_{x_2 \in R_2} \|x_1 - x_2\|_2 = O(\|H_1(x_1) - H_2(x_1)\|_F + \|g_1(x_1) - g_2(x_1)\|_2)$$

where $H_1(x_1), g_1(x_1)$ are the Hessian and gradient of some estimated density function $p_1(x_1)$ evaluated at x_1; $H_2(x_1)$ and $g_2(x_1)$ are the Hessian and gradient of the density function of $p_2(x)$ evaluated at x_1, respectively.

The following proof is a revised simple version of a similar proof in [1]. For completeness, we also include it in our paper.

PROOF. For two ridges R_1, R_2, we have two density functions $p_1(x)$ and $p_2(x)$ such that the points on each ridge satisfy the solution of $\Pi_{H_1}(x)g_1(x) = 0$ and $\Pi_{H_2}(x)g_2(x) = 0$ respectively. For any starting point $x_a \in R_1$, we can build a unit speed curve $\gamma(s)$ derived from the gradient and Hessian of $p_2(x)$ as

$$\gamma_2(0) = x_a \in R_1, \quad \gamma_2(t_0) = x_b \in R_2, \quad \gamma_2'(s) = \frac{\Pi_{H_2}(\gamma(s))g_2(\gamma(s))}{\|\Pi_{H_2}(\gamma(s))g_2(\gamma(s))\|_2}.$$

Note that the curve $\gamma(t)$ connect x_a with R_2 by x_b. Define the univariate function $\xi(s)$ as

$$\xi_2(s) = p_2(\gamma_2(t_0)) - p_2(\gamma_2(s)), \quad 0 < s < t_0.$$

Through a simple computation, we know

$$\xi_2'(s) = -\langle g_2(\gamma_2(s)), \gamma_2'(s)\rangle = -\|\Pi_{H_2}(\gamma_2(s))g_2(\gamma_2(s))\|_2, \quad \xi_2'(t_0) = 0.$$

The distance from x_a to R_2 can be bounded by the curve length of $\gamma_2(t)$ which is t_0

$$d(x_a, R_2) = \|x_a - P_{R_2}(x_a)\|_2 \le \|x_a - x_b\|_2 = \|\gamma_2(t_0) - \gamma_2(0)\|_2 \le t_0.$$

Finally, the problem becomes to bound t_0. Suppose $\sup_u \xi_2''(u) > \frac{1}{c}$, by the mean-value theorem, we have

$$t_0 = \frac{\xi_2'(t_0) - \xi_2'(0)}{\xi_2''(u)} = \frac{\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2}{\xi_2''(u)} \le c\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2.$$

Next, we show that $\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2$ is of the same order with an approximation error of $H_2(x)$ and $g_2(x)$:

$$\begin{aligned}
\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2 =&\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_1(\gamma_2(0))\|_2 \\
\le&\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_2(\gamma_2(0))\|_2 + ... \\
&+ \|\Pi_{H_1}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_1(\gamma_2(0))\|_2 \quad . \\
\le&\|\Pi_{H_2}(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))\|_F\|g_2(\gamma_2(0))\|_2 + ... \\
&+ \|\Pi_{H_1}(\gamma_2(0))\|_F\|g_2(\gamma_2(0)) - g_1(\gamma_2(0))\|_2,
\end{aligned}$$

where the last inequality is obtained by applying the Cauchy-Schwartz inequality on each row of A, to get the result $\|Ax\|_2 \le \|A\|_F\|x\|_2$. According to the Davis-Kahan theorem, $\|\Pi_{H_2}(\gamma_2(t)) - \Pi_{H_1}(\gamma_2(t))\|_F \le \beta\|H_2(\gamma_2(t)) - H_1(\gamma_2(t))\|_F$. The conclusion is proved! □

A.2. Derivatives' Bias Bound.

THEOREM A.2. *The bias of the first order and second order of the $\hat{p}_h(x)$ is*

$$|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)| = \frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int \|u\|_2^2K(u)du + o(h^2),$$

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)| = \frac{h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int \|u\|_2^2K(u)du + o(h^2).$$

where Δ is the Laplace-Beltrami operator.

PROOF. Suppose the kernel function vanishes at infinity for each dimension, i.e., it satisfies $\lim_{u_s\to\infty}K(u) = 0$ for each dimension. Then, using the integration-by-parts formula, we obtain the expectation of first-order derivatives:

$$\begin{aligned}
\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) =& \frac{1}{h^D}\int_{y\in\mathbb{R}^D}\partial_{x_s}K(\frac{x-y}{h})p(y)dy \\
=& \frac{1}{h^{D+1}}\int \partial_{z_s}K(z)|_{z=\frac{x-y}{h}}p(y)dy = h^{-1}\int_{u\in\mathbb{R}^D}\partial_{u_s}K(u)p(x-hu)du \\
=& h^{-1}\int_{u\in\mathbb{R}^D}K(u)\partial_{u_s}p(x-hu)du = \int_{u\in\mathbb{R}^D}K(u)\partial_{z_s}p(z)|_{z=x-hu}du.
\end{aligned} \tag{1}$$

For the multivariate function $\partial_{x_s}p(x)$, we have the Taylor expansion up to order 2 as

$$\partial_{z_s}p(z)|_{z=x-hu} = \partial_{x_s}p(x) - hu^T\nabla\partial_{x_s}p(x) + \frac{1}{2}h^2u^TH(\partial_{x_s}p(x))u + o(h^2). \tag{2}$$

Since $u^T\nabla\partial_{x_s}p(x)K(u)$ is an odd function with respect to each variable u_s, we have the integration $\int u^T\nabla\partial_{x_s}p(x)K(u)du = 0$ in a symmetric region.

For the term $u^TH(\partial_{x_s}p(x))u$, we know it is related with the Laplace Beltrami operator of $\Delta(\partial_{x_s}p(x))$ by

$$\begin{aligned}
\int u^TH(\partial_{x_s}p(x))uK(u)du =&\langle\int uu^TK(u)du, H(\partial_{x_s}p(x))\rangle \\
=&\frac{\int\|u\|_2^2K(u)du}{D}\langle I, H(\partial_{x_s}p(x))\rangle \\
=&\frac{\int\|u\|_2^2K(u)du}{D}\Delta(\partial_{x_s}p(x))
\end{aligned}$$

Merging the above results, we know the bias

$$|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)| = \frac{|\Delta(\partial_{x_s}p(x))|}{2D}h^2\int \|u\|_2^2 K(u)du + o(h^2), \tag{3}$$

where $\Delta(\partial_{x_s}p(x))$ is the Laplace-Beltrami operator of $\partial_{x_s}p(x)$, which is also the summation of the diagonal elements of the Hessian matrix $H(\partial_{x_s}p(x))$. Similarly, repeating the same procedure as (1)(2), we have the second-order bias as

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)| = \frac{|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}h^2\int \|u\|_2^2 K(u)du + o(h^2). \tag{4}$$

The same with (3), $\Delta(\partial_{x_s}\partial_{x_t}p(x))$ is the Laplace-Beltrami operator of $\partial_{x_s}\partial_{x_t}p(x)$ which is also the summation of the eigenvalues of the matrix $M_{s,t}$ whose i,j-th element is $\frac{\partial^4}{\partial x_s \partial x_t \partial x_i \partial x_j}p(x)$. □

A.3. Derivatives' Variance Bound.

THEOREM A.3. *The variance of the first and second order derivatives for $\hat{p}_h(x)$ has a bound as*

$$\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))| = \sqrt{\frac{\phi_s(x)}{nh^{D+2}}} + O(\frac{1}{n^{1/2}h^{(D+1)/2}}),$$

$$\mathrm{E}|\partial_{x_s}\partial_{x_t}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x))| = \sqrt{\frac{\phi_{s,t}(x)}{nh^{D+4}}} + O(\frac{1}{n^{1/2}h^{(D+3)/2}}).$$

PROOF. Because of the i.i.d. assumption and the characters of the variance, the first-order derivative yields

$$\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) = \mathrm{Var}(\frac{1}{nh^D}\sum_k \partial_{x_s}(K(\frac{x-y_k}{h})))$$

$$=\frac{1}{nh^{2D}}\mathrm{Var}(\partial_{x_s}K(\frac{x-y}{h})))) = \frac{1}{nh^{2D+2}}\mathrm{Var}(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}}).$$

Next, we derive the variance by using the equality of variance and expectation $\mathrm{Var}(a) = \mathrm{E}(a^2) - \mathrm{E}^2(a)$. In addition, let $M(\frac{x-y}{h}) = \partial_{u_s}K(u)|_{u=\frac{x-y}{h}}$, which will lead to

$$\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) = \frac{1}{nh^{2D+2}}(\mathrm{E}_y(M^2(\frac{x-y}{h})) - \mathrm{E}_y^2(M(\frac{x-y}{h}))). \tag{5}$$

Noting the bias result from (1) and (3), we have

$$\begin{aligned} &\mathrm{E}_y(M(\frac{x-y}{h})) \\ =&h^{D+1}(\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) \le h^{D+1}(\partial_{x_s}p(x) + \frac{\Delta(\partial_{x_s}p(x))}{2D}h^2\int \|u\|^2 K(u)du + o(h^2)). \end{aligned} \tag{6}$$

Taking the square of (6) on both sides, we obtain

$$\mathrm{E}_y^2(M(\frac{x-y}{h})) = h^{2D+2}((\partial_{x_s}p(x))^2 + O(h^2)).$$

Taking the expectation of $M^2(\frac{x-y}{h})$, and changing the variable $u = \frac{x-y}{h}$, we obtain

$$\mathrm{E}_y(M^2(\frac{x-y}{h})) = \frac{1}{h^D}\int M^2(u)p(x-uh)du = \frac{1}{h^D}(p(x)\int M^2(u)du + O(h)). \tag{7}$$

Using (6),(7) we have

$$
\begin{aligned}
&\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) \\
&= \frac{1}{nh^{2D+2}}(\mathrm{E}_y M^2(\frac{x-y}{h})) - \mathrm{E}_y^2(M(\frac{x-y}{h})) = \frac{1}{nh^{D+2}}(p(x)\int M^2(u)du + O(h)).
\end{aligned} \tag{8}
$$

Because the square-root function is concave, we use Jensen's inequality to determine that

$$\sqrt{\mathrm{Var}(\partial_{x_s}\hat{p}_h(x))} = \sqrt{\mathrm{E}(\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x)))^2} \geq \mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))|. \tag{9}$$

Combining (8) and (9) yields

$$\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))| \leq \sqrt{\frac{p(x)\int M^2(u)du}{nh^{D+2}}} + O(\frac{1}{n^{1/2}h^{(D+1)/2}}). \tag{10}$$

Repeating the procedures (5) - (9), we obtain

$$\mathrm{E}|\partial_{x_s}\partial_{x_t}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x))| \leq \sqrt{\frac{p(x)\int N^2(u)du}{nh^{D+4}}} + O(\frac{1}{n^{1/2}h^{(D+3)/2}}), \tag{11}$$

where $N(\frac{x-y}{h})$ is defined as $N(\frac{x-y}{h}) = \partial_{u_s}\partial_{u_t}K(u)|_{u=\frac{x-y}{h}}$ in a similar way. (10) and (11) have different orders with respect to h, which could lead to an optimal-parameter dilemma, as shown in the next section. □

A.4. Derivatives' Bias for l-SCRE.

LEMMA A.4. *For the derivatives of $\hat{p}_{r,h}(x)$, we have the bias relationship for first and second order derivatives as*

$$
\begin{aligned}
|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)| &\leq B_s(x|r,h,p), \\
|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t}p(x)| &\leq B_{s,t}(x|r,h,p),
\end{aligned}
$$

furthermore, if

$$r \geq \max\{h, \sqrt{\frac{2|\partial_{x_s}p(x)|}{|\Delta(\partial_{x_s}p(x))|}}, \sqrt{\frac{2|\partial_{x_s}\partial_{x_t}p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}}\}, \tag{12}$$

the bound of the pairwise derivatives' bias for $\hat{p}_{r,h}(x)$ will be bounded by that of $\hat{p}_h(x)$, in other words,

$$
\begin{aligned}
|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)| &\leq |\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|, \\
|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t}p(x)| &\leq |\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)|.
\end{aligned}
$$

PROOF. Recall that, in the bias for kernel density estimation, we also have the expression of expectation and the Taylor expansion:

$$
\begin{aligned}
&\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) = \int_{u\in\mathbb{R}^D} K_r(u)\partial_{z_s}p(z)|_{z=x-hu}du, \\
&\partial_{z_s}p(z)|_{z=x-hu} = \partial_{x_s}p(x) - hu^T\nabla\partial_{x_s}p(x) + \frac{1}{2}h^2u^TH(\partial_{x_s}p(x))(x)u + o(h^2).
\end{aligned}
$$

Thus, we have

$$\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) = \partial_{x_s}p(x)\int_{u\in\mathbb{R}^D} K_r(u)du + \frac{\Delta(\partial_{x_s}p(x))}{2D}h^2\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du + o(h^2). \tag{13}$$

Note that $\int_{u\in\mathbb{R}^D} K_r(u)du = \int_{\|u\|\le r/h} K(u)du$ and

$$\int_{\|u\|\le r/h} K(u)du + \int_{\|u\|>r/h} K(u)du = 1.$$

Subtracting $\partial_{x_s} p(x)$ in (13) from both sides, we have

$$\begin{aligned}&\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s} p(x)\\ =&-\partial_{x_s} p(x)\int_{\|u\|\ge r/h} K(u)du + \frac{\Delta(\partial_{x_s} p(x))}{2D} h^2 \int_{\|u\|\le r/h} \|u\|_2^2 K(u)du.\end{aligned} \tag{14}$$

Using the absolute value inequality, we have

$$\begin{aligned}&|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s} p(x)|\\ \le&|\partial_{x_s} p(x)|\int_{\|u\|\ge r/h} K(u)du + \frac{|\Delta(\partial_{x_s} p(x))|}{2D} h^2 \int_{\|u\|\le r/h} \|u\|_2^2 K(u)du.\end{aligned} \tag{15}$$

Recalling that, the original term for the upper bound of bias in (3) is

$$\frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int \|u\|_2^2 K(u)du,$$

By comparing (15) with (3), we reduce the original term for the upper bound of bias to the locally restrict version as

$$\frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int_{\|u\|\le r/h} \|u\|_2^2 K(u)du,$$

except for introducing an extra term $|\partial_{x_s} p(x)| \int_{\|u\|>r/h} K(u)du$. Next, we compare the summation of the two terms

$$|\partial_{x_s} p(x)| \int_{\|u\|>r/h} K(u)du + \frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int_{\|u\|\le r/h} \|u\|_2^2 K(u)du, \tag{16}$$

with the single term

$$\frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int \|u\|_2^2 K(u)du. \tag{17}$$

It can be easily observed that, to make (16) less than (17) , we only need to make sure the following inequality is satisfied:

$$\frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int_{\|u\|>r/h} \|u\|_2^2 K(u)du > |\partial_{x_s} p(x)| \int_{\|u\|>r/h} K(u)du. \tag{18}$$

The condition in (18) is equivalent to

$$\frac{\int_{\|u\|>r/h} \|u\|_2^2 K(u)du}{\int_{\|u\|>r/h} K(u)du} > \frac{2|\partial_{x_s} p(x)|}{h^2 |\Delta(\partial_{x_s} p(x))|}. \tag{19}$$

Note that, when $r > h$ which implies $\|u\| > 1$, the left side of (19) has a lower bound as

$$\frac{\int_{\|u\|>r/h} \|u\|_2^2 K(u)du}{\int_{\|u\|>r/h} K(u)du} \ge r^2/h^2. \tag{20}$$

Note that, when $r > h$, the condition

$$r^2/h^2 > \frac{2|\partial_{x_s} p(x)|}{h^2 |\Delta(\partial_{x_s} p(x))|}, \tag{21}$$

implies (19). (21) indicates that if we choose a proper $r > \max\{h, \frac{2|\partial_{x_s} p(x)|}{|\Delta(\partial_{x_s} p(x))|}\}$, the sufficient condition for (18) will be met automaticly, which means

$$|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s} p(x)| \le |\mathrm{E}(\partial_{x_s}\hat{p}_r(x)) - \partial_{x_s} p(x)| \tag{22}$$

Similarly, if we choose a proper such that $r > \max\{h, \frac{2|\partial_{x_s}\partial_{x_t} p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t} p(x))|}\}$, we will get

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t} p(x)| \le |\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_r(x)) - \partial_{x_s}\partial_{x_t} p(x)| \tag{23}$$

□

If choosing r such that $r > \max\{h, \frac{2|\partial_{x_s}\partial_{x_t} p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t} p(x))|}, \frac{2|\partial_{x_s} p(x)|}{|\Delta(\partial_{x_s} p(x))|}\}\}$, we will have the conditions in (22) and (23) satisfied, simultaneously.

A.5. Derivatives' Variance for l-SCRE.

THEOREM A.5. *The variance of derivative of $\hat{p}_{r,h}(x)$ is controlled by*

$$\mathrm{Var}(\partial_{x_s}\hat{p}_{r,h}(x)) \le \frac{1}{nh^{D+2}}(p(x)\int(\partial_{u_s}K(u))^2du + O(h)).$$

PROOF. Because of $\mathrm{Var}(u) = \mathrm{E}(u-\mathrm{E}u)^2 = \mathrm{E}(u^2) - (\mathrm{E}(u))^2$, by neglecting the low order term $(\mathrm{E}(u))^2$, we have

$$\mathrm{Var}(\partial_{x_s}\hat{p}_{r,h}(x)) \le \mathrm{E}((\partial_{x_s}\hat{p}_{r,h}(x))^2). \tag{24}$$

Also noting that $\hat{p}_{r,h}(x) = \frac{1}{nh^D}\sum_k K_r(\frac{x-x_k}{h})$ and taking the expectation with respect to the random variable x_k, we have

$$\mathrm{E}((\partial_{x_s}\hat{p}_{r,h}(x))^2) = \frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K_r(\frac{x-y}{h}))^2). \tag{25}$$

Using $|\frac{\partial}{\partial x_s}K_r(\frac{x-x_i}{h})| \le |\frac{\partial}{\partial x_s}K(\frac{x-x_i}{h})|, \forall \|x-x_i\| \le r$, we have

$$\frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K_r(\frac{x-y}{h}))^2) \le \frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K(\frac{x-y}{h}))^2). \tag{26}$$

Because of the chain rule of derivatives, we have

$$\frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K(\frac{x-y}{h}))^2) = \frac{1}{nh^{2D+2}}\int(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}})^2 p(y)dy. \tag{27}$$

Using the rule for changing the integrating variable from y to u, we have

$$\frac{1}{nh^{2D+2}}\int(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}})^2 p(y)dy = \frac{1}{nh^{D+2}}\int(\partial_{u_s}K(u))^2 p(x-uh)du. \tag{28}$$

In the same way as before, by Taylor expansion $p(x-uh) = p(x) + O(h)$, we have

$$\frac{1}{nh^{D+2}}\int(\partial_{u_s}K(u))^2 p(x-uh)du = \frac{1}{nh^{D+2}}(p(x)\int(\partial_{u_s}K(u))^2du + O(h)). \tag{29}$$

Combining the inequalities in (24)-(29), we can obtain the result. □

A.6. Minimum Relation.

LEMMA A.6. *For two functions* $\nu(h) = a_0h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}$ *and* $\nu_\ell(h) = \ell a_0h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}$ *with* $m = 2, 4, \ell \in (0, 1)$. *Then, the optimal minimums of them have a relationship:* $\min_h \nu_\ell(h) = \ell^{\frac{D+2}{D+6}} \min_h \nu(h)$

PROOF. For a function $\nu(h) = a_0h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}, m = 2, 4$, the global optimal minimum is achieved at $h^* = (\frac{a_1^2}{na_0^2})^{\frac{1}{D+m+4}}$, with the function value being

$$\nu(h^*) = 2(\frac{a_1^2 a_0^{\frac{D+m}{2}}}{n})^{\frac{2}{D+m+4}} = 2a_0^{\frac{D+m}{D+m+4}} a_1^{\frac{1}{D+m+4}} n^{-\frac{2}{D+m+4}}.$$

Consider another function by replacing a_0 in $\nu(h)$ with ℓa_0, where $\ell \in (0, 1)$

$$\nu_\ell(h) = \ell a_0h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}. \tag{30}$$

The modified function $\nu_\ell(h)$ will lead to a new minimum optimum point as

$$h^{**} = \arg\min \nu_\ell(h) = (\frac{a_1^2}{n\ell^2 a_0^2})^{\frac{1}{D+m+4}}.$$

Substituting it into (30), by a simple calculation, we obtain $\nu_\ell(h^{**}) = \ell^{\frac{D+m}{D+m+4}}\nu(h^*)$. Since $\frac{D+4}{D+8} > \frac{D+2}{D+6}$ and ℓ^x is a decreasing function for $\ell \in (0, 1)$, we have $\max\{\ell^{\frac{D+4}{D+8}}, \ell^{\frac{D+2}{D+6}}\} = \ell^{\frac{D+2}{D+6}}$. □

A.7. Confidence Region.

THEOREM A.7. *For any* $\alpha \in (0, 1)$, *there exist* $a_n(\alpha), b_n(\alpha)$ *such that, when* $n \to \infty$, *we have*

$$P(\mathcal{M} \subset \hat{C}_{r,h}(a_n(\alpha), b_n(\alpha))) \geq 1 - \alpha.$$

PROOF. Since the estimation of eigenvectors of the Hessian has a slower rate of convergence than the estimation of gradient, we can approximate $V^T(\hat{H}(x))\hat{g}(x) - V^T(H(x))g(x)$ by a linear combination of $\hat{H}(x)$ and $H(x)$ as:

$$\sup_{x\in\mathcal{M}} \|V_{\hat{H}}^T(x)\hat{g}(x) - V_H^T(x)g(x) - M\text{vech}(\hat{H}(x) - H(x))\hat{g}(x)\| = O_p(\sqrt{\frac{\log n}{nh^{D+4}}}).$$

Thus, we only need to ensure, with high probability,

$$\sup_{x\in\mathcal{M}} \|\hat{Q}(x)M\text{vech}(\hat{H}(x) - H(x))\hat{g}(x)\| \leq a_n. \tag{31}$$

By bringing a parameter z in the $D - d - 1$ dimensional sphere ($\|z\| = 1$), the norm in (31) equals to

$$\sup_{x\in\mathcal{M}, z\in\mathbb{S}^{D-d-1}} z^T\hat{Q}(x)M\text{vech}(\hat{H}(x) - H(x))\hat{g}(x) \leq a_n. \tag{32}$$

A sufficient condition for (32) is

$$\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}} z^T\hat{Q}(x)M\mathrm{vech}(\hat{H}(x)-\mathrm{E}(\hat{H}(x)))\hat{g}(x)+... \tag{33}$$
$$+\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}} z^T\hat{Q}(x)M\mathrm{vech}(\mathrm{E}(\hat{H}(x))-H(x))\hat{g}(x)\le a_n.$$

The second term is deterministic. Next, we show the limit distribution for the first term of (33) is normal. Let

$$g_{x,z}(X)=\frac{1}{\sqrt{h^D}}z^T\hat{Q}(x)M\mathrm{vech}(\nabla\nabla K_h(X-x))\nabla K_h(X-x).$$

Define an empirical process $\{\mathbb{G}_n(g_{x,z}), x\in\mathcal{M}, z\in\mathbb{S}^{D-d-1}\}$ as

$$\mathbb{G}_n(g_{x,z})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{x,z}(X_i)-\mathrm{E}g_{x,z}(X_1)).$$

By the central limit theorem, the limit distribution of $\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}}\mathbb{G}_n(g_{x,z})$ is the normal distribution $N(0,\sigma)$ with n approaching infinity, i.e.,

$$\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}}\mathbb{G}_n(g_{x,z})\to N(0,\sigma),$$

where σ is the variance of $\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}}\mathbb{G}_n(g_{x,z})$. Thus, we can choose

$$a_n=\sigma\sqrt{2}\mathrm{erf}^{-1}(1-2\alpha)+\sup_{x\in\mathcal{M},z\in\mathbb{S}^{D-d-1}} z^T\hat{Q}(x)M\mathrm{vech}(\mathrm{E}(\hat{H}(x))-H(x))$$

such that the condition in (32) satisfied with the probability at least $1-\alpha$. $\square$

APPENDIX B

B.1. Eigenspace Differences betweeen $C_r(x)$ and $J_r(x)$.

THEOREM B.1. *If $\|c_r(x)-x\|_2^2<\lambda_d(C_r(x)))$, the eigenspaces corresponding to the top d eigenvalues of $C_r(x)$ and $J_r(x)$ coincide, i.e., the distance* [1]

$$D(\mathcal{V}_d(C_r(x)),\mathcal{V}_d(J_r(x)))=0,$$

Otherwise, if $\|c_r(x)-x\|_2^2\ge\lambda_d(C_r(x)))$, then $D(\mathcal{V}_d(C_r(x)),\mathcal{V}_d(J_r(x)))=1$.

PROOF. If we denote the eigenvalue decomposition of $C_r(x)$ as

$$C_r(x)=[V_d,V_{D-d}]\Lambda[V_d,V_{D-d}]^T,$$

the principal space is spanned by the vectors consisting each of the columns of V_d. Here, V_d relys on x through $c_r(x)$, as a result, we denote V_d as $V_d(c_r(x))$. Similarly, we have the space which is orthogonal to the principal space, which we denote as $V_{D-d}(c_r(x))$.

Based on the assumption, we can represent $c_r(x)-x$ and $\{x_i-c_r(x), i=1:n\}$ by the coordinates in their corresponding space as:

$$c_r(x)-x=V_{D-d}(c_r(x))\alpha(x);\quad x_i-c_r(x)=V_d(c_r(x))\alpha(x,x_i). \tag{34}$$

[1] The distance between two subspace $\mathcal{V}$ and $\mathcal{U}$ is defined as the operator norm of the error of two projection matrices, i.e., $D(\mathcal{V},\mathcal{U})=\|P_{\mathcal{V}}-P_{\mathcal{U}}\|_2$

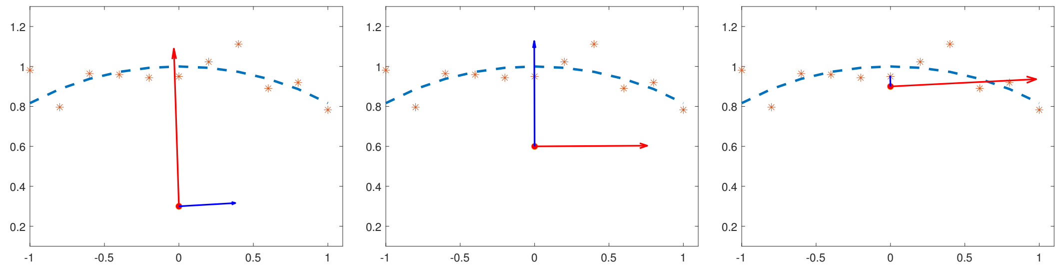

FIG 1. *The process of $J(x)$'s eigenspace's variation with x approaching the manifold*

Substitute (34) into $J_r(x)$ and let

$$A(x) = V_{D-d}(c_r(x))\alpha(x)\alpha(x)^T V_{D-d}(c_r(x))^T,$$

$$B(x) = \sum_i w(x, x_i)V_d(c_r(x))\alpha(x, x_i)\alpha(x, x_i)^T V_d(c_r(x))^T;$$

We have

$$J_r(x) = A(x) + B(x).$$

In this case, we know $\mathrm{rank}(A(x)) = 1, \mathrm{rank}(B(x)) = d$. For the rank-one matrix, we can get the eigenvalue $\lambda(A(x))$ by normalizing $A(x)$. Note that

$$A(x) = V_{D-d}(c_r(x))\frac{\alpha(x)}{\|\alpha(x)\|}\|\alpha(x)\|_2^2\frac{\alpha^T(x)}{\|\alpha(x)\|}V_{D-d}^T(c_r(x))$$

Thus, the eigenvalue of $A(x)$ is $\|c_r(x) - x\|_2^2$ and $V_{D-d}(c_r(x))\frac{\alpha(x)}{\|\alpha(x)\|}$ is a unitary vector in the space of $V_{D-d}(c_r(x))$. For simplicity, we denote

$$v(x) = V_{D-d}(c_r(x))\frac{\alpha(x)}{\|\alpha(x)\|}, \quad \Xi(x) = \sum_i w(x, x_i)\alpha(x, x_i)\alpha(x, x_i)^T,$$

which will be used in our following discussion. Similarly, the matrix $B(x)$ shares the same eigenvalues with $\Xi(x)$, because the unitary transformation keep the singular values unchanged. Denote the eigenvalue decomposition of $\Xi(x)$ as $\Xi(x) = \Theta(x)\Lambda(x)\Theta(x)^T$. Then, we have

$$B(x) = V_d(c(x))\Theta(x)\Lambda(x)\Theta(x)^T V_d(c(x)).$$

Note that $\Upsilon(x) = V_d(c(x))\Theta(x)$ is an orthonormal matrix satisfying $\Upsilon^T(x)\Upsilon(x) = I_d$. This can be divided into three cases based on the relationship between $\lambda(A(x))$ and $\lambda_{\min}(\Xi(x)), \lambda_{\max}(\Xi(x))$.

Depending on the relation between the eigenvalue of $\lambda(A(x))$ and the eigenvalues of $\Xi(x)$, there are three different cases, as described in the next few paragraphs. Because the scale of the eigenvalue of $\lambda(A(x))$ can vary greatly, we cannot recover V_d by just selecting the top d eigenvectors of $J(x)$.

Case i: $\lambda(A(x)) > \lambda_{\max}(\Xi(x))$. This case corresponds to the leftmost diagram in Figure 1, where x is far away from the data points. The covariance matrix $\sum_i w_h(x_i, x)(x_i - x)(x_i - x)^T$ will have large eigenvalues in the subspace of $V_{D-d}(C_r(x))$. Then, the eigenvector $V_{D-d}(C_r(x))\alpha(x)$ is the principal eigenvector, and we can distinguish it from the top eigenvector through eigenvalue decomposition. Then, the eigen-decomposition of $J_r(x)$ is

$$J_r(x) = [v(x), \Upsilon(x)]\begin{bmatrix}\lambda(A(x)) & \mathbf{0}\\ \mathbf{0} & \Lambda(x)\end{bmatrix}[v(x), \Upsilon(x)]^T. \tag{35}$$

From (35), we can recover the space spanned by $V_{D-d}(c_r(x))\alpha(x)$ by choosing the eigenvectors corresponding to the largest eigenvector. The space corresponding to $V_d(c_r(x))$ can be recovered by choosing the 2nd to $(d+1)$-th eigenvectors of $J_r(x)$.

Case ii: $\lambda_{\min}(\Xi(x)) \le \lambda(A(x)) \le \lambda_{\max}(\Xi(x))$. This case corresponds to the middle figure in Figure 1, where x is in the middle range of distance from the data points. In this case, the eigenvalue corresponding to $V_{D-d}(C_r(x))\alpha(x)$ is disguised by the eigenvalues of $B(x)$. Here, we cannot distinguish the eigenspace of $V_{D-d}(C_r(x))\alpha(x)$ by simply choosing the eigenvector corresponding to the largest or the smallest eigenvalue. The eigen-decomposition of $J(x)$ yields the following form:

$$J_r(x) = [\Upsilon_1(x), v(x), \Upsilon_2(x)] \begin{bmatrix} \Lambda_1(x) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda(A(x)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_2(x) \end{bmatrix} [\Upsilon_1(x), v(x), \Upsilon_2(x)]^T, \tag{36}$$

where each diagonal element of $\Lambda_1(x)$ is greater than $\lambda(A(x))$, and each diagonal element of $\Lambda_2(x)$ is less than $\lambda(A(x))$.

For *Case i* and *Case ii*, the d-dimensional projection $P(J_r(x))$ corresponding to the eigenspaces of $J_r(x)$ is different with that of $C_r(x)$. The error of the two projections is

$$P(J_r(x)) - P(C_r(x)) = vv^T - u_d u_d^T,$$

where u_d is the eigenvector corresponding to the d-th largest eigenvalue of $B(x)$. Clearly, we have the operator norm

$$\|P(J_r(x)) - P(C_r(x))\|_2 = 1.$$

Case iii: $\lambda(A(x)) < \lambda_{\min}(\Xi(x))$. This case corresponds to the rightmost diagram in Figure 1, where x is in a small range of distance from the data points. The covariance matrix $\sum_i w_h(x_i, x)(x_i - x)(x_i - x)^T$ will have large eigenvalues corresponding to the eigenvectors parallel with the tangent space at $c_r(x)$. Then, the variance along $V_{D-d}(x)\alpha(x)$ will become relatively small, causing the eigen-decomposition form of $J(x)$ to yield the following form:

$$J_r(x) = [\Upsilon(x), v(x)] \begin{bmatrix} \Lambda(x) & \mathbf{0} \\ \mathbf{0} & \lambda(A(x)) \end{bmatrix} [\Upsilon(x), v(x)]^T. \tag{37}$$

In this case, to recover $V_d(C_r(x))$, we can simply choose the eigenvectors corresponding to the top d eigenvalues of $\sum_i w_h(x_i, x)(x_i - x)(x_i - x)^T$. As a result, we can replace $C_r(x)$ by $J_r(x)$ to compute the space $V_d(C_r(x))$.

For *Case iii*, the d-dimensional projection $P(J_r(x))$ corresponding to the eigenspaces of $J_r(x)$ is the same with that of $C_r(x)$. Thus, the error of the two projections is

$$P(J_r(x)) - P(C_r(x)) = 0,$$

and the operator norm $\|P(J_r(x)) - P(C_r(x))\|_2 = 0$.

□

APPENDIX C

C.1. Rank-one Modification Enlarges Projection.

LEMMA C.1. *For any symmetric matrix B, let $A = B + \lambda uu^T, \forall\lambda \ge 0$. We then have $\|\Pi_A u\|_2 \ge \|\Pi_B u\|_2$, where Π_A, Π_B are the projections onto the space spanned by the d top principal eigenvectors of A and B, respectively.*

PROOF. Because of the variational inequality of eigenvectors, the top d eigenvectors can be written as the solution of the maximum optimal problem

$$U_A = \arg\max_{U^TU=I_d} \operatorname{trace}(U^TAU), \quad U_B = \arg\max_{U^TU=I_d} \operatorname{trace}(U^TBU).$$

Denote $\Pi_A = U_A{U_A}^T, \Pi_B = U_B{U_B}^T$. Using variational results about eigenvalues on A and B, we have

$$\langle \Pi_B, B\rangle \geq \langle \Pi_A, B\rangle, \quad \langle \Pi_A, A\rangle \geq \langle \Pi_B, A\rangle. \tag{38}$$

For the definition of the inner product $\langle\cdot,\cdot\rangle$ of two matrices with the same shape, please refer to the footnote.[2] Because of $\langle \Pi_B, B\rangle \geq \langle \Pi_A, B\rangle$, we have

$$\langle \Pi_B, B\rangle + \langle \Pi_A, \lambda uu^T\rangle \leq \langle \Pi_A, B+\lambda uu^T\rangle. \tag{39}$$

Recalling the definition of A, the right side of (39) equals to

$$\langle \Pi_A, B+\lambda uu^T\rangle = \langle \Pi_A, A\rangle. \tag{40}$$

Using variational results about eigenvalues on A, we have:

$$\langle \Pi_A, A\rangle \geq \langle \Pi_B, B+\lambda uu^T\rangle = \langle \Pi_B, B\rangle + \langle \Pi_B, \lambda uu^T\rangle. \tag{41}$$

Combining (39), (40), (41) and eliminating the constant $\langle \Pi_B, B\rangle$, we have $\langle \Pi_A, uu^T\rangle \geq \langle \Pi_B, uu^T\rangle$. Since

$$\langle \Pi_A, uu^T\rangle = u^T\Pi_A u = u^T\Pi_A\Pi_A u = \|\Pi_A u\|_2^2,$$

As a consequence, we have $\|\Pi_A u\|_2 \geq \|\Pi_B u\|_2$. □

C.2. Inclusion Lemma.

LEMMA C.2. *For any monotonously increasing and concave function $f(y)$, we have $R(f(p)) \subset R(p)$ and $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2$, where $p(x)$ is a twice-differentiable function.*

PROOF. For any two projections $\Pi_{H_p}(x), \Pi_{H_{f(p)}}(x)$ and their orthogonal complement projection $\Pi^{\perp}_{H_p}(x), \Pi^{\perp}_{H_{f(p)}}(x)$, we have the following two equalities:

$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2^2 + \|\Pi_{H_p}(x)\nabla p(x)\|_2^2 = \|\nabla p(x)\|_2^2,$$

$$\|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2^2 + \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2^2 = \|\nabla p(x)\|_2^2.$$

Note that $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2$ is equivalent to

$$\|\Pi_{H_p}(x)\nabla p(x)\|_2 \geq \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2.$$

To prove $\|\Pi_{H_p}(x)\nabla p(x)\|_2 \geq \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2$ is equivalent to prove

$$\nabla p(x)^T\Pi_{H_p}(x)\nabla p(x) \geq \nabla p(x)^T\Pi_{H_{f(p)}}(x)\nabla p(x), \tag{42}$$

which is clear, as the d principal components of $H_p(x)$ are enlarged by adding a rank-one modification in the direction of $\nabla p(x)\nabla p(x)^T$ from $H_f(x)$; this is proved in Lemma B.2.

If $x \in R_{f(p(x))}$, we have $\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x) = 0$, in other words,

$$\nabla p(x) \in \operatorname{span}\{u^1_{H_f}(x), ..., u^d_{H_f}(x)\}.$$

[2] The inner product of two matrices with the same shape is defined as: $\langle M, N\rangle = \sum_{ij} M_{ij}N_{ij}$.

Note that $H_p(x)$ is a rank-one modification with $H_f(x)$ by

$$H_p(x) = \frac{1}{f'(p(x))} H_{f(p)}(x) - \frac{f''(p(x))}{f'(p(x))} \nabla p(x) \nabla^T p(x). \tag{43}$$

Because $f(y)$ is a monotonously increasing and concave function, we know

$$-f''(p(x))/f'(p(x)) > 0.$$

Because of $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2$, $\|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2 = 0$ indicates that we also have $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 = 0$, i.e. $x \in R(p(x))$, which implies $R(f(p(x))) \subset R(p(x))$. □

C.3. Transformed Inequality.

THEOREM C.3. *For the ridge $R(f(p))$ defined by the transformed nonlinear increasing and concave function f, we have:*

$$\mathrm{Haus}(R(f(p)), \mathcal{M}_{R(f(p))}) \leq \mathrm{Haus}(R(p), \mathcal{M}_{R(p)}),$$

where $R(p)$ and $R(f(p))$ are the d-dimensional ridges corresponding to p and $f(p)$, $\mathcal{M}_{R(p)}$ and $\mathcal{M}_{R(f(p))}$ are the projections of $R(p)$ and $R(f(p))$ onto $\mathcal{M}$, respectively.

PROOF. Since the projection from R to $\mathcal{M}_R$ is surjective, for any $y^* \in \mathcal{M}_R$, such as $\inf_{x\in R} \|x - y^*\| = \sup_{y \in \mathcal{M}_R} \inf_{x \in R} \|x - y\|_2$, there is $x_{y^*} \in R$ such as $y = P_{\mathcal{M}_R}(x_{y^*})$.

$$\begin{aligned} &\sup_{y\in\mathcal{M}_R} \inf_{x\in R} \|x-y\|_2 \\ &= \inf_{x \in R} \|x - y^*\|_2 \leq \|x_{y^*} - y^*\|_2 = \inf_{z \in \mathcal{M}_R} \|x_{y^*} - z\|_2 \leq \sup_{x\in R} \inf_{z\in\mathcal{M}_R} \|x - z\|_2. \end{aligned}$$

Since $\mathrm{Haus}(R, \mathcal{M}_R) = \max\{\sup_{x\in R} \inf_{y\in\mathcal{M}_R} \|x-y\|_2, \sup_{x\in\mathcal{M}_R} \inf_{y\in R} \|x-y\|_2\}$, we can conclude that

$$\mathrm{Haus}(R, \mathcal{M}_R) = \sup_{x\in R} \inf_{y\in\mathcal{M}_R} \|x - y\|_2.$$

Also, noting that $R = R/R_f \cup R_f$, we know that

$$\sup_{x\in R} \inf_{y\in\mathcal{M}_R} \|x-y\|_2 = \max\{\sup_{x\in R/R_f} \inf_{y\in\mathcal{M}_R} \|x-y\|_2, \sup_{x\in R_f} \inf_{y\in\mathcal{M}_R} \|x-y\|_2\}. \tag{44}$$

Because of (44), we can easily obtain

$$\sup_{x\in R_f} \inf_{y\in\mathcal{M}_R} \|x-y\|_2 \leq \mathrm{Haus}(R, \mathcal{M}_R). \tag{45}$$

Because of $R_f \subset \mathcal{M}_R$,

$$\sup_{x\in R_f} \inf_{y\in\mathcal{M}_{R_f}} \|x-y\|_2 = \sup_{x\in R_f} \inf_{y\in\mathcal{M}_R} \|x-y\|_2, \tag{46}$$

Since the projection from R_f to $\mathcal{M}_{R_f}$ is surjective, we also have

$$\mathrm{Haus}(R_f, \mathcal{M}_{R_f}) = \sup_{x\in R_f} \inf_{y\in\mathcal{M}_{R_f}} \|x-y\|_2, \tag{47}$$

Merging (45),(46),(47), we have: $\mathrm{Haus}(R_f, \mathcal{M}_{R_f}) \leq \mathrm{Haus}(R, \mathcal{M}_R)$. □

REFERENCES

[1] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I., WASSERMAN, L. et al. (2014). Nonparametric ridge estimation. *The Annals of Statistics* **42** 1511–1545.