Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIAL FOR "MANIFOLD FITTING BY RIDGE ESTIMATION: A SUBSPACE-CONSTRAINED 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 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 in which 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 $H_1(x_1)$ and $g_1(x_1)$ are the Hessian and gradient of some estimated density function $p_1(x_1)$ evaluated at x_1, respectively; $H_2(x_1)$ 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, simplified version of a similar proof in Genovese et al. (2014). For completeness, we 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 $p_2(x)$:

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

Note that the curve $\gamma(t)$ connects x_a with R_2 by x_b. The univariate function $\xi(s)$ is defined 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:

$$\begin{aligned} & d(x_a, R_2) \\ =&\|x_a - P_{R_2}(x_a)\|_2 \\ \leq&\|x_a - x_b\|_2 \\ =&\|\gamma_2(t_0) - \gamma_2(0)\|_2 \leq t_0. \end{aligned}$$

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

$$\begin{aligned} t_0 =&\frac{\xi_2'(t_0) - \xi_2'(0)}{\xi_2''(u)} \\ =&\frac{\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2}{\xi_2''(u)} \\ \leq&c\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2. \end{aligned}$$

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

$$\begin{aligned} & \|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0))\|_2 \\ =&\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_1(\gamma_2(0))\|_2 \end{aligned} \tag{1}$$

Using the triangle inequality with respect to the $\|\|_2$ norm, we determine that (1) is dominated by

$$\begin{aligned} &\|\Pi_{H_2}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_2(\gamma_2(0))\|_2 + ... \\ &\qquad + \|\Pi_{H_1}(\gamma_2(0))g_2(\gamma_2(0)) - \Pi_{H_1}(\gamma_2(0))g_1(\gamma_2(0))\|_2 \end{aligned} \tag{2}$$

For any matrix A, using the Cauchy-Schwartz inequality on each row of A, we will get the result $\|Ax\|_2 \leq \|A\|_F\|x\|_2$. Thus, similarly, we determine that (2) is dominated by

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

According to the Davis-Kahan theorem, $\|\Pi_{H_2}(\gamma_2(t)) - \Pi_{H_1}(\gamma_2(t))\|_F \leq \beta\|H_2(\gamma_2(t)) - H_1(\gamma_2(t))\|_F$. The conclusion is therefore proved. □

A.2. Bound of the Derivatives' Bias.

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

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

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

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:

$$\begin{aligned}
&\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) \\
\overset{s.1}{=}&\frac{1}{h^D}\int_{y\in\mathbb{R}^D}\partial_{x_s}K(\frac{x-y}{h})p(y)dy \\
\overset{s.2}{=}&\frac{1}{h^{D+1}}\int\partial_{z_s}K(z)|_{z=\frac{x-y}{h}}p(y)dy \\
\overset{s.3}{=}&h^{-1}\int_{u\in\mathbb{R}^D}\partial_{u_s}K(u)p(x-hu)du \\
\overset{s.4}{=}&h^{-1}\int_{u\in\mathbb{R}^D}K(u)\partial_{u_s}p(x-hu)du \\
\overset{s.5}{=}&\int_{u\in\mathbb{R}^D}K(u)\partial_{z_s}p(z)|_{z=x-hu}du.
\end{aligned} \tag{3}$$

The equation $s.1$ is the definition of the expectation. The equation $s.2$ is obtained from the derivative of the function composition. The equation $s.3$ is obtained from the formula of the integration by changing variables. The equation $s.4$ is obtained from the formula of the integration by partition. The equation $s.5$ is similar to equation $s.2$.

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

$$\begin{aligned}
&\partial_{z_s}p(z)|_{z=x-hu} \\
=&\partial_{x_s}p(x) - hu^T\nabla\partial_{x_s}p(x) + \frac{1}{2}h^2u^TH(\partial_{x_s}p(x))u + o(h^2).
\end{aligned} \tag{4}$$

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 $u^T H(\partial_{x_s} p(x))u$, we know it is related to the Laplace-Beltrami operator of $\Delta(\partial_{x_s} p(x))$ by

$$\begin{aligned}\int u^T H(\partial_{x_s} p(x))uK(u)du =& \langle \int uu^T K(u)du, H(\partial_{x_s} p(x))\rangle \\ \overset{s.6}{=}& \frac{\int \|u\|_2^2 K(u)du}{D}\langle I, H(\partial_{x_s} p(x))\rangle \\ =& \frac{\int \|u\|_2^2 K(u)du}{D}\Delta(\partial_{x_s} p(x)),\end{aligned}$$

where the equation of $s.6$ is obtained from $\|u\|_2^2 = \sum_k u_k^2$, and

$$\int u_k u_s K(u)du = 0 \quad k \neq s,$$

because of symmetric domain of integration and the independence of each of the dimensions.

Combining the above results, we know the bias

$$|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s} p(x)| = \frac{|\Delta(\partial_{x_s} p(x))|}{2D} h^2 \int \|u\|_2^2 K(u)du + o(h^2), \tag{5}$$

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 for (3)(4), we obtain the second-order bias,

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t} p(x)| = \frac{|\Delta(\partial_{x_s}\partial_{x_t} p(x))|}{2D} h^2 \int \|u\|_2^2 K(u)du + o(h^2). \tag{6}$$

Likewise, with (5), $\Delta(\partial_{x_s}\partial_{x_t} p(x))$ 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. Bound of the Derivatives' Variance.

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

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

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

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

$$\begin{aligned}
&\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) \\
\overset{s.7}{=}&\mathrm{Var}(\frac{1}{nh^D}\sum_k \partial_{x_s}(K(\frac{x-y_k}{h}))) \\
\overset{s.8}{=}&\frac{1}{nh^{2D}}\mathrm{Var}(\partial_{x_s}K(\frac{x-y}{h}))) \\
\overset{s.9}{=}&\frac{1}{nh^{2D+2}}\mathrm{Var}(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}}).
\end{aligned}$$

As with the process used to derive the bias, equation $s.7$ is the definition of variance, equation $s.8$ is obtained from the independence of the samples of y_k, and equation $s.9$ is the result of the derivative for the composite functions.

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, letting $M(\frac{x-y}{h}) = \partial_{u_s}K(u)|_{u=\frac{x-y}{h}}$, the variance will be

$$\begin{aligned}
&\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) \\
=&\frac{1}{nh^{2D+2}}(\mathrm{E}_y(M^2(\frac{x-y}{h})) - \mathrm{E}_y^2(M(\frac{x-y}{h}))).
\end{aligned} \tag{7}$$

Noting the bias result from (3) and (5), we have

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

Taking the square of (8) 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

$$\begin{aligned}
&\mathrm{E}_y(M^2(\frac{x-y}{h})) \\
=&\frac{1}{h^D}\int M^2(u)p(x-uh)du \\
=&\frac{1}{h^D}(p(x)\int M^2(u)du + O(h)).
\end{aligned} \tag{9}$$

Combining the results in (8) and (9), we have the order of the variance:

$$\begin{aligned}
&\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) \\
=&\frac{1}{nh^{2D+2}}(\mathrm{E}_y M^2(\frac{x-y}{h})) - \mathrm{E}_y^2(M(\frac{x-y}{h})) \\
=&\frac{1}{nh^{D+2}}(p(x)\int M^2(u)du + O(h)).
\end{aligned} \tag{10}$$

Because the square-root function is concave, we use Jensen's inequality to show that $\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))|$ is dominated by the square root of the variance:

$$\begin{aligned}
&\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))| \\
\leq&\sqrt{\mathrm{E}(\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x)))^2} \\
=&\sqrt{\mathrm{Var}(\partial_{x_s}\hat{p}_h(x))}.
\end{aligned} \tag{11}$$

Combining (10) and (11) yields

$$\begin{aligned}
&\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))| \\
\leq&\sqrt{\frac{p(x)\int M^2(u)du}{nh^{D+2}}} + O(\frac{1}{n^{1/2}h^{(D+1)/2}}).
\end{aligned} \tag{12}$$

Repeating procedures (7)-(11), we obtain

$$\begin{aligned}
&\mathrm{E}|\partial_{x_s}\partial_{x_t}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x))| \\
\leq&\sqrt{\frac{p(x)\int N^2(u)du}{nh^{D+4}}} + O(\frac{1}{n^{1/2}h^{(D+3)/2}}),
\end{aligned} \tag{13}$$

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. (12) and (13) 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 the first- and second-order derivatives:*

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

Furthermore, if

$$r \geq \max\{h, \sqrt{\frac{2|\partial_{x_s}p(x)|}{|\Delta(\partial_{x_s}p(x))|}}, \sqrt{\frac{2|\partial_{x_s}\partial_{x_t}p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}}\}, \tag{14}$$

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,

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

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

$$\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) = \int_{u\in\mathbb{R}^D} K_r(u)\partial_{z_s}p(z)|_{z=x-hu}du,$$

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

Thus, we have

$$\begin{aligned}
&\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) \\
=&\partial_{x_s}p(x)\int_{u\in\mathbb{R}^D} K_r(u)du + ... \\
&+\frac{\Delta(\partial_{x_s}p(x))}{2D}h^2\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du + o(h^2).
\end{aligned} \tag{15}$$

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

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

Subtracting $\partial_{x_s}p(x)$ in (15) from both sides, we have

$$\begin{aligned}
&\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x) \\
=&-\partial_{x_s}p(x)\int_{\|u\|\geq r/h} K(u)du + ... \\
&+\frac{\Delta(\partial_{x_s}p(x))}{2D}h^2\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du.
\end{aligned} \tag{16}$$

Using the absolute value inequality, we have

$$\begin{aligned}
&|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)| \\
\leq&|\partial_{x_s}p(x)|\int_{\|u\|\geq r/h} K(u)du + ... \\
&+\frac{|\Delta(\partial_{x_s}p(x))|}{2D}h^2\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du.
\end{aligned} \tag{17}$$

Recall that the original term for the upper bound of the bias in (5) is

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

By comparing (17) with (5), we reduce the original term for the upper bound of the bias to the locally restricted version:

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

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

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

with the single term

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

It can be easily observed that, to make (18) less than (19), we only need to make sure the following inequality is satisfied:

$$\begin{aligned} &\frac{|\Delta(\partial_{x_s} p(x))|}{2} h^2 \int_{\|u\|>r/h} \|u\|_2^2 K(u) du \\ >&|\partial_{x_s} p(x)| \int_{\|u\|>r/h} K(u) du. \end{aligned} \tag{20}$$

The condition in (20) is equivalent to

$$\frac{\int_{\|u\|>r/h} \|u\|_2^2 K(u) du}{\int_{\|u\|>r/h} K(u) du} > \frac{2|\partial_{x_s} p(x)|}{h^2 |\Delta(\partial_{x_s} p(x))|}. \tag{21}$$

Note that, when $r > h$, this implies $\|u\| > 1$, and so the left side of (21) has the following lower bound:

$$\frac{\int_{\|u\|>r/h} \|u\|_2^2 K(u) du}{\int_{\|u\|>r/h} K(u) du} \ge r^2/h^2. \tag{22}$$

Also, when $r > h$, the condition

$$r^2/h^2 > \frac{2|\partial_{x_s} p(x)|}{h^2 |\Delta(\partial_{x_s} p(x))|} \tag{23}$$

implies (21). (23) 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 (20) will be met automatically, which means

$$|\mathrm{E}(\partial_{x_s} \hat{p}_{r,h}(x)) - \partial_{x_s} p(x)| \le |\mathrm{E}(\partial_{x_s} \hat{p}_r(x)) - \partial_{x_s} p(x)| \tag{24}$$

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

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t}p(x)| \leq |\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r}(x)) - \partial_{x_s}\partial_{x_t}p(x)| \tag{25}$$

□

If we choose r such that $r > \max\{h, \frac{2|\partial_{x_s}\partial_{x_t}p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}, \frac{2|\partial_{x_s}p(x)|}{|\Delta(\partial_{x_s}p(x))|}\}\}$, the conditions in (24) and (25) will be satisfied, simultaneously.

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

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

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

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

$$\mathrm{Var}(\partial_{x_s}\hat{p}_{r,h}(x)) \leq \mathrm{E}((\partial_{x_s}\hat{p}_{r,h}(x))^2). \tag{26}$$

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

$$\mathrm{E}((\partial_{x_s}\hat{p}_{r,h}(x))^2) = \frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K_r(\frac{x-y}{h}))^2). \tag{27}$$

For the x that satisfies $\|x - x_i\|_2 \leq r$, we have

$$K_r(\frac{x-x_i}{h}) = K(\frac{x-x_i}{h}),$$

which implies

$$|\frac{\partial}{\partial x_s}K_r(\frac{x-x_i}{h})| = |\frac{\partial}{\partial x_s}K(\frac{x-x_i}{h})|.$$

Otherwise, for the x that satisfies $\|x - x_i\|_2 > r$, we have

$$K_r(\frac{x-x_i}{h}) = 0,$$

which implies

$$|\frac{\partial}{\partial x_s}K_r(\frac{x-x_i}{h})| = 0.$$

Thus, when $\|x - x_i\| \neq r$, we always have the following inequality satisfied:

$$|\frac{\partial}{\partial x_s}K_r(\frac{x-x_i}{h})| \leq |\frac{\partial}{\partial x_s}K(\frac{x-x_i}{h})|. \tag{28}$$

Using (28), we have

$$\frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K_r(\frac{x-y}{h}))^2) \leq \frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K(\frac{x-y}{h}))^2). \tag{29}$$

Because of the chain rule of derivatives, we have

$$\begin{aligned}&\frac{1}{nh^{2D}}\mathrm{E}_y((\partial_{x_s}K(\frac{x-y}{h}))^2)\\ =&\frac{1}{nh^{2D+2}}\int(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}})^2p(y)dy.\end{aligned} \tag{30}$$

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

$$\begin{aligned}&\frac{1}{nh^{2D+2}}\int(\partial_{u_s}K(u)|_{u=\frac{x-y}{h}})^2p(y)dy\\ =&\frac{1}{nh^{D+2}}\int(\partial_{u_s}K(u))^2p(x-uh)du.\end{aligned} \tag{31}$$

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

$$\begin{aligned}&\frac{1}{nh^{D+2}}\int(\partial_{u_s}K(u))^2p(x-uh)du\\ =&\frac{1}{nh^{D+2}}(p(x)\int(\partial_{u_s}K(u))^2du+O(h)).\end{aligned} \tag{32}$$

Combining the inequalities in (45)-(32), we can obtain the result. □

A.6. Minimum Relation.

LEMMA A.6. *We have 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, their optimal minimums have the following 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

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

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

$$\nu_\ell(h)=\ell a_0h^2+a_1\sqrt{\frac{1}{nh^{D+m}}}. \tag{33}$$

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

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

Substituting this into (33), 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 the 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 $H(x)$:

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

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

$$\sup_{x\in\mathcal{M}} \|\hat{Q}(x)M\text{vech}(\hat{H}(x) - H(x))\hat{g}(x)\| \leq a_n. \tag{34}$$

By bringing a parameter z in the $D-d-1$ dimensional sphere ($\|z\| = 1$), we make the norm in (34) equal to

$$\sup_{x\in\mathcal{M}, z\in\mathbb{S}^{D-d-1}} z^T\hat{Q}(x)M\text{vech}(\hat{H}(x) - H(x))\hat{g}(x) \leq a_n. \tag{35}$$

A sufficient condition for (35) is

$$\begin{aligned}\sup_{x\in\mathcal{M}, z\in\mathbb{S}^{D-d-1}} & z^T\hat{Q}(x)M\text{vech}(\hat{H}(x) - \mathrm{E}(\hat{H}(x)))\hat{g}(x) + ... \\ + \sup_{x\in\mathcal{M}, z\in\mathbb{S}^{D-d-1}} & z^T\hat{Q}(x)M\text{vech}(\mathrm{E}(\hat{H}(x)) - H(x))\hat{g}(x) \leq a_n.\end{aligned} \tag{36}$$

The second term is deterministic. Next, we show that the limit distribution for the first term of (36) is normal. Let

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

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

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

According to 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 (35) is satisfied with a probability of at least $1-\alpha$. □

APPENDIX B

B.1. Eigenspace Differences between $C_r(x)$ 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*

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

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

PROOF. If the eigenvalue decomposition of $C_r(x)$ is denoted by

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

the principal space is spanned by the vectors consisting of the columns of V_d. Here, V_d relies on x through $c_r(x)$, and as a result we let V_d be expressed by $V_d(c_r(x))$. Similarly, we have the space that is orthogonal to the principal space, denoted by $V_{D-d}(c_r(x))$.

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

$$\begin{aligned} c_r(x) - x &= V_{D-d}(c_r(x))\alpha(x); \\ x_i - c_r(x) &= V_d(c_r(x))\alpha(x, x_i). \end{aligned} \tag{37}$$

Substitute (37) into $J_r(x)$ and let

$$\begin{aligned} A(x) &= V_{D-d}(c_r(x))\alpha(x)\alpha(x)^T V_{D-d}(c_r(x))^T, \\ B(x) &= \sum_i w(x, x_i) V_d(c_r(x))\alpha(x, x_i)\alpha(x, x_i)^T V_d(c_r(x))^T. \end{aligned}$$

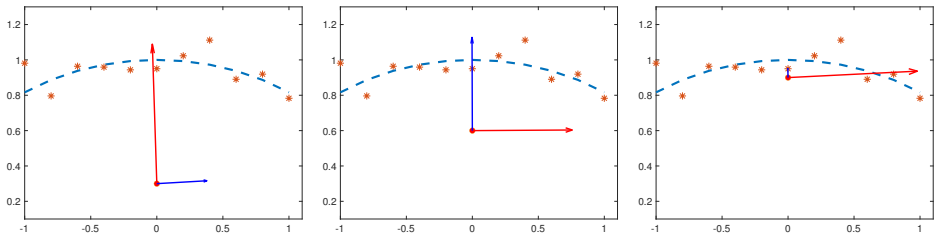

FIG 1. *The process of the variation of $J_r(x)$'s eigenspace, with x approaching the manifold*

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 let

$$v(x) = V_{D-d}(c_r(x))\frac{\alpha(x)}{\|\alpha(x)\|},$$

$$\Xi(x) = \sum_i w(x, x_i)\alpha(x, x_i)\alpha(x, x_i)^T,$$

which will be used in the following discussion. Similarly, the matrix $B(x)$ shares the same eigenvalues with $\Xi(x)$, because the unitary transformation keeps the singular values unchanged. Let the eigenvalue decomposition of $\Xi(x)$ be denoted by $\Xi(x) = \Theta(x)\Lambda(x)\Theta(x)^T$. 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 left 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

$$J_r(x) = [v(x), \Upsilon(x)] \begin{bmatrix} \lambda(A(x)) & \mathbf{0} \\ \mathbf{0} & \Lambda(x) \end{bmatrix} [v(x), \Upsilon(x)]^T. \tag{38}$$

From (38), 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 2^{nd} to $(d+1)$-th eigenvectors of $J_r(x)$.

Case ii: $\lambda_{\min}(\Xi(x)) \leq \lambda(A(x)) \leq \lambda_{\max}(\Xi(x))$. 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:

$$J_r(x) = [\Upsilon_1(x), v(x), \Upsilon_2(x)] \begin{bmatrix} \Lambda_1(x) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda(A(x)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_2(x) \end{bmatrix} [\Upsilon_1(x), v(x), \Upsilon_2(x)]^T, \tag{39}$$

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 from 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 right 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:

$$J_r(x) = [\Upsilon(x), v(x)] \begin{bmatrix} \Lambda(x) & \mathbf{0} \\ \mathbf{0} & \lambda(A(x)) \end{bmatrix} [\Upsilon(x), v(x)]^T. \tag{40}$$

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 $C_r(x)$ with $J_r(x)$ 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 as 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 is $\|P(J_r(x)) - P(C_r(x))\|_2 = 0$.

□

APPENDIX C

C.1. Rank-One Modification Enlarges Projection.

LEMMA C.1. *For any symmetric matrix B, let $A = B + \lambda uu^T, \forall \lambda \geq 0$. We have $\|\Pi_A u\|_2 \geq \|\Pi_B u\|_2$, where Π_A and Π_B are the projections onto the space spanned by the eigenvectors corresponding to the d largest eigenvalues 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

$$U_A = \arg \max_{U^T U = I_d} \operatorname{trace}(U^T A U),$$

$$U_B = \arg \max_{U^T U = I_d} \operatorname{trace}(U^T B U).$$

Let $\Pi_A = U_A {U_A}^T, \Pi_B = U_B {U_B}^T$. For any Z and W with the same shape, the trace and inner product are equal according to the following equation:

$$\operatorname{trace}(Z^T W) = \langle Z, W \rangle,$$

where the inner product of two matrices with the same shape is defined as $\langle Z, W \rangle = \sum_{ij} Z_{ij} W_{ij}$. Therefore, we have:

$$\operatorname{trace}(U_A^T A U_A) = \operatorname{trace}(U_A U_A^T A) = \langle U_A U_A^T, A \rangle = \langle \Pi_A, A \rangle.$$

Similarly,

$$\operatorname{trace}(U_B^T B U_B) = \operatorname{trace}(U_B U_B^T B) = \langle U_B U_B^T, B \rangle = \langle \Pi_B, B \rangle.$$

Using variational results for the eigenvalues on A and B, we have

$$\langle \Pi_B, B\rangle \geq \langle \Pi_A, B\rangle, \quad \langle \Pi_A, A\rangle \geq \langle \Pi_B, A\rangle. \tag{41}$$

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

$$\langle \Pi_B, B\rangle + \langle \Pi_A, \lambda uu^T\rangle \leq \langle \Pi_A, B + \lambda uu^T\rangle. \tag{42}$$

Recalling the definition of A, the right side of (42) equals

$$\langle \Pi_A, B + \lambda uu^T\rangle = \langle \Pi_A, A\rangle. \tag{43}$$

Using variational results for the eigenvalues of A, we have:

$$\langle \Pi_A, A\rangle \geq \langle \Pi_B, B + \lambda uu^T\rangle = \langle \Pi_B, B\rangle + \langle \Pi_B, \lambda uu^T\rangle. \tag{44}$$

Combining (42), (43), and (44) and eliminating the term $\langle \Pi_B, B\rangle$, we have

$$\langle \Pi_A, uu^T\rangle \geq \langle \Pi_B, uu^T\rangle. \tag{45}$$

Because

$$\langle \Pi_A, uu^T\rangle = u^T\Pi_A u = u^T\Pi_A\Pi_A u = \|\Pi_A u\|_2^2, \tag{46}$$

using (45) and (46), we will obtain $\|\Pi_A u\|_2 \geq \|\Pi_B u\|_2$. □

C.2. Rank-One Modification on Subspace.

LEMMA C.2. *For any symmetric matrix B, let $A = B + \lambda uu^T$, $\forall \lambda \geq 0$ and any nonzero vector $u \in \mathrm{span}\{u_1(B(x)), u_2(B(x)), ..., u_d(B(x))\}$; the $(d+1)$-th to D-th largest eigenvalues of A and B yield*

$$\lambda_{d+k}(A) = \lambda_{d+k}(B), \quad k = 1, ..., D - d$$

where $u_k(B(x))$ and $\lambda_k(B(x))$ are the eigenvector and eigenvalue corresponding to the k-th largest eigenvalues of $B(x)$.

PROOF. We use $\Lambda^{(1)}$ and $\Lambda^{(2)}$ to stand for the $d \times d$ and $(n-d)\times(n-d)$ diagonal matrix corresponding to the eigenvalue decomposition of B. Let the eigenvalue decomposition of B be denoted by

$$B = [U_d, U_{n-d}] \begin{bmatrix} \Lambda^{(1)} & \\ & \Lambda^{(2)} \end{bmatrix} [U_d, U_{n-d}]^T.$$

Since $u \in \mathcal{S}_d$, we can write u by combining the columns of U_d, as $u = U_d\alpha$. Then,

$$B + \lambda uu^T = [U_d, U_{n-d}] \begin{bmatrix} \Lambda^{(1)} & \\ & \Lambda^{(2)} \end{bmatrix} [U_d, U_{n-d}]^T + \lambda U_d\alpha\alpha^T U_d^T.$$

Let the eigenvalue decomposition be denoted by

$$U_d\Lambda^{(1)}U_d^T + \lambda U_d\alpha\alpha^T U_d^T = \hat{U}_d\Lambda\hat{U}_d^T,$$

where the diagonal elements in $\Lambda, \Lambda^{(1)}$ are placed in decreasing order. Using Weyl's theorem for eigenvalues Horn and Johnson (2012), we know

$$\Lambda_{ii} \geq \Lambda^{(1)}_{ii}, \quad i = 1, ..., d.$$

and

$$\Lambda_{ii} \geq \Lambda^{(1)}_{ii} \geq \Lambda^{(2)}_{jj}, \forall, i = 1, ..., d, \forall, j = 1, ..., D - d.$$

Thus, the eigenvalue decomposition of A is

$$A = [\hat{U}_d, U_{n-d}] \begin{bmatrix} \Lambda & \\ & \Lambda^{(2)} \end{bmatrix} [\hat{U}_d, U_{n-d}]^T.$$

Note that, because the columns of U_d and $\hat{U}_d$ span the same subspace and the columns of U_d are orthogonal with the columns of U_{n-d}, we know that the columns of $\hat{U}_d$ are also orthogonal with the columns of U_{n-d}.

Because of the uniqueness of the eigenvalue decomposition, we have proved that

$$\lambda_{d+1}(A) = \lambda_{d+1}(B) = \Lambda^{(2)}_{11}.$$

Furthermore, for the remaining eigenvalues, we have a similar result:

$$\lambda_{d+k}(A) = \lambda_{d+k}(B) = \Lambda^{(2)}_{kk} \quad \forall, k = 1, ..., D - d.$$

□

C.3. Inclusion Lemma.

LEMMA C.3. *For any monotonously increasing and concave function $f(y)$, i.e, $f'(x) > 0, f''(x) \leq 0$, for $x \in R(f(p))$, we have the following satisfied simultaneously:*

$$\lambda_{d+1}(H_p(x)) < 0,$$
$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2 = 0,$$

The conditions $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 = 0$ implies $\Pi^{\perp}_{H_p}(x)\nabla p(x) = \mathbf{0}$, which indicates that $x \in R(p)$. Thus, $R(f(p)) \subset R(p)$.

PROOF. Recall that $H_p(x)$ is a rank-one modification with $H_f(x)$, by

$$H_p(x) = \frac{1}{f'(p(x))}H_{f(p)}(x) - \frac{f''(p(x))}{f'(p(x))}\nabla p(x)\nabla^T p(x). \tag{47}$$

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

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

Thus, $H_p(x)$ is obtained from a nonnegative rank-one modification. Lemma C.1 implies

$$\|\Pi_{H_p}(x)\nabla p(x)\|_2 \geq \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2, \tag{48}$$

where the projection matrix Π_{H_p} and $\Pi_{H_{f_p}}$ are defined as

$$\Pi_{H_p}(x) = U_d(H_p(x))U_d^T(H_p(x)),$$
$$\Pi_{H_{f_p}}(x) = U_d(H_{f(p)}(x))U_d^T(H_{f_p}(x)).$$

Here, $U_d(H_p(x))$and $U_d^T(H_p(x))$ are the eigenvectors corresponding to the largest d eigenvalues of $H_p(x)$ and $H_{f_p}(x)$, respectively.

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

$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2^2 + \|\Pi_{H_p}(x)\nabla p(x)\|_2^2 = \|\nabla p(x)\|_2^2. \tag{49}$$

Similarly, for $\Pi^{\perp}_{H_{f(p)}}(x)$ and $\Pi_{H_{f(p)}}(x)$, because of the orthogonal properties, we have

$$\|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2^2 + \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2^2 = \|\nabla p(x)\|_2^2. \tag{50}$$

Because of (49) and (50), we know that the condition (squaring both sides in (48))

$$\|\Pi_{H_p}(x)\nabla p(x)\|_2^2 \geq \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2^2$$

implies

$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2^2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2^2. \tag{51}$$

Taking the square root of both sides in (51) will lead to

$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2. \tag{52}$$

It is easy to obtain $\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x) = 0$. For any $x \in R_{f(p(x))}$, using the definition of a ridge, we have

$$\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x) = 0, \quad \lambda_{d+1}(H_{f(p)}(x)) < 0$$

Because of $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2$, we have

$$\|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2 = 0,$$

which implies that we also have $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 = 0$. Next, we show $\lambda_{d+1}(H_p(x)) < 0$. Recall that,

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

Thus, $-f''(p(x))\nabla p(x)\nabla^T p(x)$ is a semi-positive definite modification. $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 = 0$, which implies that $\nabla p(x)$ is in the space spanned by

$$\mathcal{S}_{H_{f(p)}(x)} = \text{span}\{u_1(H_{f(p)}(x)), u_2(H_{f(p)}(x)), ..., u_d(H_{f(p)}(x))\}.$$

Because $\nabla p(x) \in \mathcal{S}_{H_{f(p)}(x)}$, the semi-positive rank-one modification

$$-f''(p(x))\nabla p(x)\nabla^T p(x)$$

on the matrix $H_{f(p)}(x)$ is equivalent to a rank-one modification on the principal d-dimensional subspace $\mathcal{S}_{H_{f(p)}(x)}$, which will not affect the orthogonal complement subspace $\mathcal{S}^{\perp}_{H_{f(p)}(x)}$, which in turn means that the eigenvalues from the $(d+1)$-th largest eigenvalue to the D-th will remain unchanged by Lemma (C.2). Thus, we have :

$$\lambda_{d+1}(H_p(x)) = \frac{1}{f'(p(x))}\lambda_{d+1}(H_{f(p)}(x)) < 0,$$

because of $f'(p(x)) > 0$ and $f''(p(x)) < 0$. In conclusion, x also satisfies the ridge condition derived by $p(x)$, i.e., $x \in R(p(x))$, which implies $R(f(p(x))) \subset R(p(x))$.

□

C.4. Transformed Inequality.

THEOREM C.4. *For the ridge $R(f(p))$ defined by the transformed nonlinear increasing and concave function f, we have*

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

where $R(p)$ and $R(f(p))$ are the d-dimensional ridges corresponding to p and $f(p)$, and $\mathcal{M}_{R(p)}$ 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 that $\inf_{x\in R}\|x - y^*\| = \sup_{y\in\mathcal{M}_R}\inf_{x\in R}\|x - y\|_2$, 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

$$\sup_{x\in R}\inf_{y\in\mathcal{M}_R}\|x-y\|_2 = \max\{\sup_{x\in R/R_f}\inf_{y\in\mathcal{M}_R}\|x-y\|_2, \sup_{x\in R_f}\inf_{y\in\mathcal{M}_R}\|x-y\|_2\}. \tag{53}$$

Because of (53), we can easily obtain

$$\sup_{x\in R_f}\inf_{y\in\mathcal{M}_R}\|x-y\|_2 \le \mathrm{Haus}(R, \mathcal{M}_R). \tag{54}$$

Because of $R_f \subset R$, we have $\mathcal{M}_{R_f} \subset \mathcal{M}_R$. Also, note that $\mathcal{M}_{R_f}$ is the projection of R_f onto $\mathcal{M}$ and $\mathcal{M}_R$ is the projection of R onto $\mathcal{M}$. We have

$$\sup_{x\in R_f}\inf_{y\in\mathcal{M}_{R_f}}\|x-y\|_2 = \sup_{x\in R_f}\inf_{y\in\mathcal{M}_R}\|x-y\|_2. \tag{55}$$

The projection from R_f to $\mathcal{M}_{R_f}$ is surjective, which implies that the Hausdorff distance equals the quasi-Hausdorff, Chen et al. (2015), i.e.,

$$\mathrm{Haus}(R_f, \mathcal{M}_{R_f}) = \sup_{x\in R_f}\inf_{y\in\mathcal{M}_{R_f}}\|x-y\|_2, \tag{56}$$

Combining (54),(55), and (56), we have $\mathrm{Haus}(R_f, \mathcal{M}_{R_f}) \le \mathrm{Haus}(R, \mathcal{M}_R)$. □

REFERENCES

CHEN, Y.-C., GENOVESE, C. R., WASSERMAN, L. et al. (2015). Asymptotic theory for density ridges. *The Annals of Statistics* **43** 1896–1928.

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

HORN, R. A. and JOHNSON, C. R. (2012). *Matrix analysis*. Cambridge university press.