SUPPLEMENTS TO SINGLE INDEX FRÉCHET REGRESSION

BY SATARUPA BHATTACHARJEE[1,a] AND HANS-GEORG MÜLLER[2,b]

[1]*Department of Statistics, Pennsylvania State University,* [a]*sfb5992@psu.edu*

[2]*Department of Statistics, University of California, Davis,* [b]*hgmueller@ucdavis.edu*

Section S.1 contains the proofs of the main results in the paper and several auxiliary lemmas. The technical assumptions required to obtain the uniform rate results for the local linear Fréchet regression estimator that is utilized to obtain an estimate of the object link function uniformly across the single-index values and the direction parameter are listed in Section S.2. A detailed remark regarding the critical assumption (A5) in Section 3 in the main article can be found in Section S.3. Additional data analysis results are in Section S.4 for human mortality data and data on a Riemannian manifold, specifically on the positive segment of a sphere as observed for mood compositional data. This section also contains additional simulations for the special case of Euclidean responses and additional numerical results for resting-state fMRI image (ADNI) data.

S.1. Proofs and auxiliary results.

In this section, we provide the proofs of the results in Section 3 of the main manuscript and state and include several auxiliary lemmas.

PROOF OF PROPOSITION 2. By assumption (A2), $m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})$ is a continuous function of $\bar{\boldsymbol{\theta}}$ for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ for almost all $\mathbf{x}$ on the compact ball $\bar{\Theta}$. This implies $m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})$ is uniformly continuous in $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ for almost all $\mathbf{x}$. That is, there exists $\delta>0$ for any $\varepsilon>0$ and $\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2\in\bar{\Theta}$ such that $\|\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2\|\leqslant\delta$, implies $d\left(m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1),m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2)\right)\leqslant\varepsilon$ for almost all $\mathbf{x}$. This implies the uniform continuity of $d^2(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}}))$ as a function of $\bar{\boldsymbol{\theta}}$, for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ for almost all $\mathbf{x},y$. To see this, let $\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_2\in\bar{\Theta}$ such that $\|\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2\|\to0$, and observe

$$
\begin{aligned}
&\left|d^2(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))-d^2(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\\
&=\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))+d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))-d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\\
&\leqslant\left(\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))\right|\right.\\
&\qquad\left.+\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\right)\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))-d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\\
&\leqslant 2D\left|d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1))-d(y,m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))\right|\\
&\leqslant 2Dd\left(m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1),m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2)\right)\to0.
\end{aligned}
$$

This holds for almost all $\mathbf{x},y$. The second inequality uses the assumption that Ω has a finite diameter D. The last inequality follows from the triangle inequality. The above technique will be used repeatedly in the subsequent proofs. By bounded convergence, $\mathbb{E}(d^2(Y,m_\oplus(\mathbf{X}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})))$ is a continuous function of $\bar{\boldsymbol{\theta}}$ for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$. Hence the map $\bar{\boldsymbol{\theta}}\mapsto H(\bar{\boldsymbol{\theta}}),\ \bar{\boldsymbol{\theta}}\in\bar{\Theta}$ is continuous.

Note that

$$
\begin{aligned}
H(\bar{\boldsymbol{\theta}})&=\mathbb{E}[d^2(Y,m_\oplus(\mathbf{X}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}}))]=\mathbb{E}[\mathbb{E}(d^2(Y,m_\oplus(\mathbf{X}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}}))|\mathbf{X})]\\
&=\mathbb{E}[\mathbb{E}(d^2(Y,m_\oplus(\mathbf{X}^\top\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})|\mathbf{X}^\top\bar{\boldsymbol{\theta}}_{\mathbf{0}})].
\end{aligned}
$$

The last equality is true since the single index model $m_\oplus$ depends on $\mathbf{X}$ only through the parameter $\bar{\boldsymbol{\theta}}_\mathbf{0} \in \bar{\Theta}$. Now, according to the single index Fréchet regression model $m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}, \bar{\boldsymbol{\theta}}_\mathbf{0})$ is the conditional Fréchet mean at $\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}$. Since the conditional Fréchet mean is assumed to be the unique minimizer of the conditional Fréchet objective function, for each $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$ with $\bar{\boldsymbol{\theta}} \neq \bar{\boldsymbol{\theta}}_\mathbf{0}$,

$$\mathbb{E}[d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}))|\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}] \geqslant \mathbb{E}[d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}, \bar{\boldsymbol{\theta}}_\mathbf{0}))|\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}],$$

where the strict inequality holds on the set $R(\bar{\boldsymbol{\theta}}) = \{\mathbf{X} \in \mathbb{R}^p : m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}) \neq m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}, \bar{\boldsymbol{\theta}}_\mathbf{0})\}$. Thus, on the set $R(\bar{\boldsymbol{\theta}})$,

$$\begin{aligned} H(\bar{\boldsymbol{\theta}}) &= \mathbb{E}[\mathbb{E}(d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}))|\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}] \\ &> \mathbb{E}[\mathbb{E}(d^2(Y, m_\oplus(\mathbf{x}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}, \bar{\boldsymbol{\theta}}_\mathbf{0})))|\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}] = \mathbb{E}[d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}, \bar{\boldsymbol{\theta}}_\mathbf{0}))] = H(\bar{\boldsymbol{\theta}}_\mathbf{0}). \end{aligned}$$

Further, under assumption (A0), this set has positive probability, i.e., $P(R(\bar{\boldsymbol{\theta}})) > 0$. Denoting the indicator function as $\mathbb{I}(A) = \begin{cases} 1 \text{ if } \mathbf{X} \in A, \\ 0 \text{ otherwise} \end{cases}$, it follows that

$$H(\bar{\boldsymbol{\theta}}) \geqslant \mathbb{E}[\mathbb{E}(d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}))|\mathbf{X}^\top \bar{\boldsymbol{\theta}}_\mathbf{0}) \mathbb{I}\left(R(\bar{\boldsymbol{\theta}})\right)] > H(\bar{\boldsymbol{\theta}}_\mathbf{0}).$$

Thus $\bar{\boldsymbol{\theta}}_\mathbf{0}$ is the unique minimizer of $H(\bar{\boldsymbol{\theta}})$, for all $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$. □

Throughout the following, $\rightsquigarrow$ denotes weak convergence as per [11], $\ell^\infty(\Omega)$ the space of bounded functions on Ω, $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^p$ and $\|\cdot\|_F$ the Frobenius norm.

Lemma 1 is adapted as stronger version of Theorem 1 of [1]. Uniformity over the single index value t was already required in [1] to achieve uniform convergence of local linear Fréchet regression. In the single index model framework, there is a new parameter vector $\bar{\boldsymbol{\theta}}$, the presence of which requires an additional uniformity requirement over $\bar{\boldsymbol{\theta}}$. The lemma can be proved following a similar argument and we provide a brief sketch of the proof at the end of this section.

PROOF OF THEOREM 1. It is shown that the map $\bar{\boldsymbol{\theta}} \mapsto H(\bar{\boldsymbol{\theta}})$ is continuous and $\bar{\boldsymbol{\theta}}_\mathbf{0}$ and $\hat{\bar{\boldsymbol{\theta}}}$ are the respective unique minimizers of $H(\bar{\boldsymbol{\theta}})$ and $V_n(\bar{\boldsymbol{\theta}})$. By Corollary 3.2.3 in [11] it is then sufficient to show the convergence of $\sup_{\bar{\boldsymbol{\theta}}} \|V_n(\bar{\boldsymbol{\theta}}) - H(\bar{\boldsymbol{\theta}})\|$, to zero in probability. To do this we first show that $V_n \rightsquigarrow H$ in $\ell^\infty(\Omega)$ and apply Theorem 1.3.6 of [11]. The weak convergence result is proved (see Theorem 1.5.4 of [11]) by checking that

(C1) $V_n(\bar{\boldsymbol{\theta}}) - H(\bar{\boldsymbol{\theta}}) = o_P(1)$ for all $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$,
(C2) V_n is asymptotically equi-continuous in probability, that is, for all $\epsilon, \eta > 0$, there exists $\delta > 0$ such that, $\limsup_{n\to\infty} P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left|V_n(\bar{\boldsymbol{\theta}}_1) - V_n(\bar{\boldsymbol{\theta}}_2)\right| > \epsilon\right] < \eta.$

We first express the difference between the sample and population objective functions as the sum of two differences by introducing the intermediate quantity $\tilde{V}_n(\cdot)$ as described in equation (2.7). Recalling $\tilde{V}_n(\bar{\boldsymbol{\theta}}) := \frac{1}{M}\sum_{l=1}^M d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}))$,

$$\|V_n(\bar{\boldsymbol{\theta}}) - H(\bar{\boldsymbol{\theta}})\| = |V_n(\bar{\boldsymbol{\theta}}) - \tilde{V}_n(\bar{\boldsymbol{\theta}})| + |\tilde{V}_n(\bar{\boldsymbol{\theta}}) - H(\bar{\boldsymbol{\theta}})|. \tag{S.1}$$

Now,

$$|\tilde{V}_n(\bar{\boldsymbol{\theta}}) - H(\bar{\boldsymbol{\theta}})| = \left|\frac{1}{M}\sum_{l=1}^M d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})) - \mathbb{E}(d^2(Y, m_\oplus(\mathbf{X}^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})))\right| = o_P(1),$$

by Weak Law of Large Numbers, since $\tilde{V}_n(\cdot)$ can be seen as an i.i.d sum. As for the first term in (S.1),

$$|V_n(\bar{\boldsymbol{\theta}}) - \tilde{V}_n(\bar{\boldsymbol{\theta}})| = \left| \frac{1}{M}\sum_{l=1}^{M} d^2(\tilde{Y}_l, \hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})) - \frac{1}{M}\sum_{l=1}^{M} d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}) \right|$$

$$\leqslant 2D \frac{1}{M}\sum_{l=1}^{M} d(\hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}), m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})). \tag{S.2}$$

The steps to obtain (S.2) are similar to those followed in the proof of Proposition 2, using the total boundedness of Ω and properties of the metric d.

It remains to show $\frac{1}{M}\sum_{l=1}^{M} d(\hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}), m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})) \xrightarrow{P} 0$. Observe that,

$$\begin{aligned} &\frac{1}{M}\sum_{l=1}^{M} d(\hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}), m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})) \\ \leqslant &\frac{1}{M}\sum_{l=1}^{M} \sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \sup_t d(\hat{m}_\oplus(t, \bar{\boldsymbol{\theta}}), m_\oplus(t, \bar{\boldsymbol{\theta}})) = \frac{1}{M}\sum_{l=1}^{M} O_P(a_n) = o_P(1), \end{aligned} \tag{S.3}$$

where a_n is the rate of uniform convergence for the local linear Fréchet regression estimate, as given in equation (3.1) in the main manuscript for Lemma 1 [1]. We use here that the O_P terms in the sum are uniform in l. Hence the result follows. Thus we have (C1). The finite distribution converges weakly since, for any $k \in \mathcal{N}$ and $\bar{\boldsymbol{\theta}}_1, \ldots, \bar{\boldsymbol{\theta}}_k \in \Theta$, we have $(V_n(\bar{\boldsymbol{\theta}}_1), \ldots, V_n(\bar{\boldsymbol{\theta}}_k)) \rightsquigarrow (H(\bar{\boldsymbol{\theta}}_1), \ldots, H(\bar{\boldsymbol{\theta}}_k))$.
It is also important to observe that, by virtue of Lemma 1,

(S.4)

$$\begin{aligned} &\sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} |V_n(\bar{\boldsymbol{\theta}}) - \tilde{V}_n(\bar{\boldsymbol{\theta}})| \\ &\leqslant \sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \frac{1}{M}\sum_{l=1}^{M} d(\hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}), m_\oplus(\tilde{\mathbf{X}}_l^\top \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})) \leqslant \frac{1}{M}\sum_{l=1}^{M} \sup_{\bar{\boldsymbol{\theta}}} \sup_t d(\hat{m}_\oplus(t, \bar{\boldsymbol{\theta}}), m_\oplus(t, \bar{\boldsymbol{\theta}})) \\ &= O_P(a_n). \end{aligned}$$

For (C2), let $\epsilon, \gamma > 0$ and $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2 \in \bar{\Theta}$.

$$\begin{aligned} &P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left| V_n(\bar{\boldsymbol{\theta}}_1) - V_n(\bar{\boldsymbol{\theta}}_2) \right| > \epsilon \right] \\ \leqslant &P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left| V_n(\bar{\boldsymbol{\theta}}_1) - \tilde{V}_n(\bar{\boldsymbol{\theta}}_1) \right| > \frac{\epsilon}{3} \right] + P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left| V_n(\bar{\boldsymbol{\theta}}_2) - \tilde{V}_n(\bar{\boldsymbol{\theta}}_2) \right| > \frac{\epsilon}{3} \right] \\ &+ P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left| \tilde{V}_n(\bar{\boldsymbol{\theta}}_1) - \tilde{V}_n(\bar{\boldsymbol{\theta}}_2) \right| > \frac{\epsilon}{3} \right]. \end{aligned} \tag{S.5}$$

Using (S.4), the first two terms of (S.5) are $O_P(a_n) = o_P(1)$, uniformly in $\bar{\boldsymbol{\theta}}_1$ and $\bar{\boldsymbol{\theta}}_2$ respectively. For the third term,

$$P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\| \leqslant \delta} \left| \tilde{V}_n(\bar{\boldsymbol{\theta}}_1) - \tilde{V}_n(\bar{\boldsymbol{\theta}}_2) \right| > \frac{\epsilon}{3} \right] \tag{S.6}$$

$$\leqslant P\left[\sup_{\|\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2\|\leqslant\delta} 2D\,\frac{1}{M}\sum_{l=1}^{M} d(m_\oplus(\tilde{\mathbf{X}}_l^\top\bar{\boldsymbol{\theta}}_1,\bar{\boldsymbol{\theta}}_1),m_\oplus(\tilde{\mathbf{X}}_l^\top\bar{\boldsymbol{\theta}}_2,\bar{\boldsymbol{\theta}}_2))>\frac{\epsilon}{3}\right].$$

By the assumption on $m_\oplus$ being Lipschitz continuous with Lipschitz constant L (see assumption (A2)),and $\mathbf{X}$ having a bounded support, (see assumptions (R1)-(R2)), choosing $\delta<\frac{\epsilon}{6DL}$: we have, (S.6) $\to 0$, as $\delta\to 0$. The asymptotic equi-continuity result for the stochastic process $V_n(\bar{\boldsymbol{\theta}})$, $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ follows. □

PROOF OF COROLLARY 1. For any $\mathbf{x}\in\mathbb{R}^p$, we observe that, by the triangle inequality of the metric,

$$\begin{aligned}(\text{S.7})\qquad & d(\hat{m}_\oplus(\mathbf{x}^\top\hat{\bar{\boldsymbol{\theta}}},\hat{\bar{\boldsymbol{\theta}}}),m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_{\mathbf{0}},\bar{\boldsymbol{\theta}}_{\mathbf{0}}))\\ \leqslant & d(\hat{m}_\oplus(\mathbf{x}^\top\hat{\bar{\boldsymbol{\theta}}},\hat{\bar{\boldsymbol{\theta}}}),m_\oplus(\mathbf{x}^\top\hat{\bar{\boldsymbol{\theta}}},\hat{\bar{\boldsymbol{\theta}}}))+d(m_\oplus(\mathbf{x}^\top\hat{\bar{\boldsymbol{\theta}}},\hat{\bar{\boldsymbol{\theta}}}),m_\oplus(\mathbf{x}^\top\bar{\boldsymbol{\theta}}_{\mathbf{0}},\bar{\boldsymbol{\theta}}_{\mathbf{0}})).\end{aligned}$$

From Lemma 1 we know that

$$\sup_{\bar{\boldsymbol{\theta}}}\sup_t d(\hat{m}_\oplus(t,\bar{\boldsymbol{\theta}}),m_\oplus(t,\bar{\boldsymbol{\theta}}))=O_P(a_n)=o_P(1),$$

where a_n is as defined in equation (3.1) in the main manuscript. Since $\hat{\bar{\boldsymbol{\theta}}}$ lies in a small neighborhood around $\bar{\boldsymbol{\theta}}_{\mathbf{0}}$ for n large enough, the first term of (S.7) converges to zero in probability. Note that by assumption (A2), $m_\oplus$ is continuous. Since, by Theorem 1 $\hat{\bar{\boldsymbol{\theta}}}\xrightarrow{P}\bar{\boldsymbol{\theta}}_{\mathbf{0}}$, using continuous mapping theorem, the second term of (S.7) also converges to zero in probability. The result follows using Slutsky's theorem. □

Before proceeding with the proof of the asymptotic normality for $\hat{\boldsymbol{\theta}}$, recall that $\bar{\boldsymbol{\theta}}=(\theta_1,\boldsymbol{\theta})^\top$, for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$. It is important to note that the full vector $\bar{\boldsymbol{\theta}}$ can be written as a function of the last $(p-1)$ elements $\boldsymbol{\theta}$ since the first element θ_1 of $\bar{\boldsymbol{\theta}}$ can be written as $\theta_1=\sqrt{1-\|\boldsymbol{\theta}\|^2}$. Thus we can view $H(\cdot)$, $\tilde{V}_n(\cdot)$, and $V_n(\cdot)$ to be effectively only functions of $\boldsymbol{\theta}\in\Theta$, respectively. Further, since the map $H:\boldsymbol{\theta}\mapsto\bar{\boldsymbol{\theta}}$ is continuous, assumption (A2) implies the $L-$Lipschitz continuity of the regression function $m_\oplus$ as a function of $\boldsymbol{\theta}\in\Theta$. Also note that $\boldsymbol{\theta}_{\mathbf{0}}$, $\tilde{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\theta}}$ are minimizers for the criteria functions $H(\cdot)$, $\tilde{V}_n(\cdot)$, and $V_n(\cdot)$ respectively. These are continuous as a function of $\boldsymbol{\theta}$, the latter two almost surely.

It is also possible to define the partial derivatives of each of the criteria functions with respect to the components of $\boldsymbol{\theta}$ in terms of limits of finite differences. The following function $f_{\mathbf{x},y}:\mathbb{R}^{p-1}\mapsto\mathbb{R}$ was defined in Section 2.3,

$$f_{\mathbf{x},y}(\boldsymbol{\theta})=f_{\mathbf{x},y}(\theta_2,\ldots,\theta_p)=d^2\left(y,m_\oplus(\mathbf{x}^\top(\theta_1,\ldots,\theta_r,\ldots,\theta_s,\ldots,\theta_p))\right),\ r,s=2,\ldots,p,$$

and the first and second ordered forward finite differences of $f_{\mathbf{x},y}$ are

$$\begin{aligned}(\text{S.8})\qquad \nabla_a(\mathbf{x},y,\theta_r)&=f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r+a,\ldots,\theta_p)-f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r,\ldots,\theta_p),\\ \nabla_a^2(\mathbf{x},y,\theta_r,\theta_s)&=f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r+a,\ldots,\theta_s+a,\ldots\theta_p)-f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r+a,\ldots,\theta_s,\ldots,\theta_p)\\ &\quad-f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r,\ldots,\theta_s+a,\ldots,\theta_p)+f_{\mathbf{x},y}(\theta_2,\ldots,\theta_r,\ldots,\theta_s,\ldots,\theta_p).\end{aligned}$$

Define the population derivatives as limits of difference quotients as follows,

$$\begin{aligned}\frac{\partial H(\boldsymbol{\theta})}{\partial\theta_r}&:=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\mathbb{E}(\nabla_\varepsilon(\mathbf{X},Y,\theta_r)),\ r=2,\ldots,p,\\ \frac{\partial^2 H(\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}&:=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon^2}\mathbb{E}(\nabla_\varepsilon^2(\mathbf{X},Y,\theta_r,\theta_s)),\ r,s=2,\ldots,p.\end{aligned}$$

Then

(S.9)

$$\boldsymbol{\Delta} H(\boldsymbol{\theta}) := \left(\frac{\partial H(\boldsymbol{\theta})}{\partial \theta_2}, \dots \frac{\partial H(\boldsymbol{\theta})}{\partial \theta_p} \right)^\top, \boldsymbol{\Delta^2} H(\boldsymbol{\theta}) := \left(\left(\frac{\partial^2 H(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \right) \right)_{r,s=2,\dots,p}, \text{ with}$$

$$\begin{cases} \frac{\partial H(\boldsymbol{\theta})}{\partial \theta_r} := \lim\limits_{\varepsilon \to 0} \frac{H(\theta_2,\dots,\theta_r+\varepsilon,\dots,\theta_p) - H(\theta_2,\dots,\theta_r,\dots,\theta_p)}{\varepsilon}, \\ \frac{\partial H(\boldsymbol{\theta})}{\partial \theta_r \theta_s} = \lim\limits_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \Big[H(\theta_2, \dots, \theta_r + \varepsilon, \dots, \theta_s + \varepsilon, \dots \theta_p) - H(\theta_2, \dots, \theta_r + \varepsilon, \dots, \theta_s, \dots, \theta_p) \\ \qquad\qquad - H(\theta_2, \dots, \theta_r, \dots, \theta_s + \varepsilon, \dots, \theta_p) + H(\theta_2, \dots, \theta_r, \dots, \theta_s, \dots, \theta_p) \Big], \end{cases}$$

for $r, s = 2, \dots, p$.

The corresponding empirical estimates are given by

$$\frac{\partial V_n(\boldsymbol{\theta})}{\partial \theta_r} := \frac{1}{hM} \sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r),\ r = 2, \dots, p,$$

$$\frac{\partial^2 V_n(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} = \frac{1}{h^2 M} \sum_{l=1}^{M} \widehat{\nabla^2}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s),\ r, s = 2, \dots, p,$$

where

$$\begin{aligned} \widehat{\nabla}_h(\mathbf{x}, y, \theta_r) &= \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r + h, \dots, \theta_p) - \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r, \dots, \theta_p), \\ \widehat{\nabla^2}_h(\mathbf{x}, y, \theta_r, \theta_s) &= \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r + h, \dots, \theta_s + h, \dots \theta_p) \\ &\quad - \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r + h, \dots, \theta_s, \dots, \theta_p) \\ &\quad - \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r, \dots, \theta_s + h, \dots, \theta_p) \\ &\quad + \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_r, \dots, \theta_s, \dots, \theta_p), \end{aligned} \tag{S.10}$$

and

$$\hat{f}_{\mathbf{x},y}(\boldsymbol{\theta}) = \hat{f}_{\mathbf{x},y}(\theta_2, \dots, \theta_p) = d^2 \left(y, \hat{m}_{\oplus}(\mathbf{x}^\top(\theta_1, \dots, \theta_r, \dots, \theta_s, \theta_p)) \right),\ r, s = 2, \dots, p.$$

Thus

(S.11)

$$\boldsymbol{\Delta} V_n(\boldsymbol{\theta}) := \left(\frac{\partial V_n(\boldsymbol{\theta})}{\partial \theta_2}, \dots \frac{\partial V_n(\boldsymbol{\theta})}{\partial \theta_p} \right)^\top,\ \boldsymbol{\Delta^2} V_n(\boldsymbol{\theta}) := \left(\left(\frac{\partial^2 V_n(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \right) \right)_{r,s=2,\dots,p}, \text{ with}$$

$$\begin{cases} \frac{\partial V_n(\boldsymbol{\theta})}{\partial \theta_r} := \frac{V_n(\theta_2,\dots,\theta_r+h,\dots,\theta_p) - V_n(\theta_2,\dots,\theta_r,\dots,\theta_p)}{h}, \\ \frac{\partial V_n(\boldsymbol{\theta})}{\partial \theta_r \theta_s} = \frac{1}{h^2} \Big[V_n(\theta_2, \dots, \theta_r + h, \dots, \theta_s + h, \dots \theta_p) - V_n(\theta_2, \dots, \theta_r + h, \dots, \theta_s, \dots, \theta_p) \\ \qquad\qquad - V_n(\theta_2, \dots, \theta_r, \dots, \theta_s + a, \dots, \theta_p) + V_n(\theta_2, \dots, \theta_r, \dots, \theta_s, \dots, \theta_p) \Big], \end{cases}$$

for $r, s, = 2, \dots, p$, where $h = h(n)$ is a tuning parameter depending on n such that when $n \to \infty$, $h(n) \to 0$, and $Mh^2(n) \to \infty$ as in assumption (A6).

In the same vein, we define the derivatives for the intermediate objective function $\tilde{V}_n(\cdot)$ for $r, s = 2, \ldots, p$,

(S.12)

$$\boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta}) := \left(\frac{\partial \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_2}, \ldots \frac{\partial \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_p}\right)^\top, \ \boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}) := \left(\left(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}\right)\right)_{r,s=2,\ldots,p}, \text{ with}$$

$$\begin{cases} \frac{\partial \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r} := \frac{\tilde{V}_n(\theta_2,\ldots,\theta_r+h,\ldots,\theta_p) - \tilde{V}_n(\theta_2,\ldots,\theta_r,\ldots,\theta_p)}{h}, \\ \frac{\partial \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r \theta_s} = \frac{1}{h^2}\Big[\tilde{V}_n(\theta_2,\ldots,\theta_r+h,\ldots,\theta_s+h,\ldots\theta_p) - \tilde{V}_n(\theta_2,\ldots,\theta_r+h,\ldots,\theta_s,\ldots,\theta_p) \\ \qquad\qquad -\tilde{V}_n(\theta_2,\ldots,\theta_r,\ldots,\theta_s+a,\ldots,\theta_p) + \tilde{V}_n(\theta_2,\ldots,\theta_r,\ldots,\theta_s,\ldots,\theta_p)\Big] \end{cases}$$

Using the notations defined in (S.8) we can rewrite

$$\frac{\partial \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r} = \frac{1}{h}\frac{1}{M}\sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r), \ r = 2, \ldots, p,$$

$$\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} = \frac{1}{h^2}\frac{1}{M}\sum_{l=1}^{M} \nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s), \ r, s = 2, \ldots, p.$$

The relevant limits are assumed to exist.

PROPOSITION S.1.1. *Under assumptions (A2) and (A6),*

$$\|\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}) - \boldsymbol{\Delta^2}H(\boldsymbol{\theta})\| \xrightarrow{P} 0 \text{ for any } \boldsymbol{\theta} \in \Theta.$$

PROOF. $\mathbb{E}(\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta})) = \left(\mathbb{E}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial^2 \theta_2}), \ldots \mathbb{E}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial^2 \theta_p})\right)^\top$. For $r = 2, \ldots, p$,

$$\mathbb{E}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}) = \mathbb{E}\left(\frac{1}{h^2}\frac{1}{M}\sum_{l=1}^{M} \nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) = \frac{1}{h^2}\mathbb{E}\left(\frac{1}{M}\sum_{l=1}^{M} \nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right)$$

$$= \frac{1}{h^2}\mathbb{E}\Big(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\Big) \to \frac{\partial^2 H(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \text{ as } h \to 0.$$

Similarly, for $r, s = 2, \ldots, p$,

$$\mathrm{Var}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial^2 \theta_r \partial \theta_s}) = \mathrm{Var}\left(\frac{1}{h^2}\frac{1}{M}\sum_{l=1}^{M} \nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) = \frac{1}{h^4 M}\mathrm{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right)$$

Now from the definition of $\nabla_h^2(\tilde{Y}_l, \theta_r, \theta_s)$ in (S.8),

(S.13)
$$\mathrm{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) \leqslant \mathbb{E}\left(\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right)^2\right)$$

$$= E\Big[d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r + h, \ldots, \theta_s + h, \ldots \theta_p)))$$

$$- d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r + h, \ldots, \theta_s, \ldots, \theta_p)))$$

$$- d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r, \ldots, \theta_s + h, \ldots, \theta_p)))$$

$$+ d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r, \ldots, \theta_s, \ldots, \theta_p)))\Big]^2$$

Using the fact that for any two random variable U and V, $\mathbb{E}(U+V)^2 \leqslant 2\mathbb{E}(U^2) + 2\mathbb{E}(V^2)$,

$$
\begin{aligned}
\text{(S.14)} \qquad & \operatorname{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) \\
& \leqslant 2E\Big[d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r + h, \ldots, \theta_s + h, \ldots, \theta_p))) \\
& \quad - d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r + h, \ldots, \theta_s, \ldots, \theta_p)))\Big]^2 \\
& \quad + 2E\Big[d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r, \ldots, \theta_s + h, \ldots, \theta_p))) \\
& \quad - d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top(\theta_1, \ldots, \theta_r, \ldots, \theta_s, \ldots, \theta_p)))\Big]^2 \\
& \leqslant 16D^2L^2h^2.
\end{aligned}
$$

The last inequality follows using the same technique as in the proof of Proposition 2, employing the triangle inequality and fact that for any $u, v \in \Omega$ one has $d(u,v) < D$ for some $D > 0$ due to the total boundedness of the space (Ω, d) with diameter D; L is the Lipschitz constant for $m_\oplus$ from assumption (A2). Then

$$
\operatorname{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) = \frac{1}{h^4 M}\operatorname{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) \leqslant \frac{16D^2L^2}{h^2 M}.
$$

As long as $h = h(n) \to 0$ such that $h^2 M =\to \infty$, as $n \to \infty$ (assumption (A6)), we have $\operatorname{Var}\left(\nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r, \theta_s)\right) \to 0$ for any $\boldsymbol{\theta}$. Combining these we have $\operatorname{Var}(\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta})) = A(h(n)) \to 0$ for $h = h(n) \to 0$ such that $h^2 M \to \infty$, as $n \to \infty$, where $A = ((a_{rs}))_{r,s=2,\ldots,p}$; with

$$
a_{rs} = \begin{cases} \operatorname{Var}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r^2}) \text{ if } r = s, \\ \operatorname{Var}(\frac{\partial^2 \tilde{V}_n(\boldsymbol{\theta})}{\partial \theta_r \theta_s}) \text{ if } r \neq s. \end{cases} .
$$

The result follows. □

PROPOSITION S.1.2. *Under assumptions (A2) and(A6)*

$$
\text{(S.15)} \qquad \sqrt{M}(\boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta_0}) - \boldsymbol{\Delta}H(\boldsymbol{\theta_0})) \xrightarrow{D} N(0, \Sigma(\boldsymbol{\theta_0}))
$$

where $\Sigma(\boldsymbol{\theta_0}) = ((\sigma_{rs}(\boldsymbol{\theta_0})))_{r,s=2,\ldots,p}$ *with*

$$
\sigma_{rs}(\boldsymbol{\theta_0}) = \begin{cases} \lim\limits_{\varepsilon \to 0} \operatorname{Var}\left(\frac{1}{\varepsilon}\nabla_\varepsilon(\boldsymbol{X}, Y, \theta_{0r})\right) \text{ if } r = s \\ \lim\limits_{\varepsilon \to 0} \operatorname{Cov}\left(\frac{1}{\varepsilon}\nabla_\varepsilon(\boldsymbol{X}, Y, \theta_{0r}),\ \frac{1}{\varepsilon}\nabla_\varepsilon(\boldsymbol{X}, Y, \theta_{0s})\right) \text{ if } r \neq s \end{cases}
$$

PROOF. Writing $\boldsymbol{\theta_0} = (\theta_{02}, \ldots, \theta_{0p})^\top \in \Theta$ and recalling

$$
\frac{\partial \tilde{V}_n(\theta_{0r})}{\partial \theta_{0r}} = \frac{1}{M}\sum_{l=1}^{M}\frac{1}{h}\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}),
$$

we observe that $\frac{\partial \tilde{V}_n(\theta_{0r})}{\partial \theta_{0r}}$ is an i.i.d sum of M terms $\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}))_{l=1,\ldots,M}$. Note that

$$
\mathbb{E}\left(\frac{\partial \tilde{V}_n(\theta_{0r})}{\partial \theta_{0r}}\right) = \mathbb{E}\left(\frac{1}{h}\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\right) \to \frac{\partial H(\theta_{0r})}{\partial \theta_{0r}} \text{ as } h = h(n) \to 0, \text{ under assumption (A6).}
$$

Thus, $\mathbb{E}(\boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta_0})) \to \boldsymbol{\Delta}H(\boldsymbol{\theta_0})$ as $n \to \infty$ and $h = h(n) \to 0$. Further, under assumption (A6), as $n \to \infty$ and $h \to 0$,

$$\text{Cov}\left(\frac{1}{h}\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}), \frac{1}{h}\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s})\right) \to \begin{cases}\sigma_{rr}(\boldsymbol{\theta}_0), \text{ if } r = s,\\ \sigma_{rs}(\boldsymbol{\theta}_0), \text{ otherwise.}\end{cases}$$

□

PROPOSITION S.1.3. *Under assumptions (A0)-(A3),*

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| = O_P(M^{-1/2}).$$

PROOF. Consider the probability $P(\sqrt{M}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| > 2^L)$ for a large L. We aim to show that this probability can be made arbitrarily small as L grows large. Let $r_n = M^{-1/2}$. For any $\eta > 0$,

$$P\Big(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| > 2^L r_n\Big) \leqslant \sum_{j>L, 2^{j-1}r_n \leqslant \eta} P(2^{j-1}r_n < \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| \leqslant 2^j r_n) + P\left(2\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| > \eta\right).$$

The second term goes to zero by the consistency of $\tilde{\boldsymbol{\theta}}$ to $\boldsymbol{\theta_0}$ according to Theorem 1. As for the first term, define "shells" $S_j := \{\boldsymbol{\theta} \in \Theta : 2^{j-1}r_n < \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \leqslant 2^j r_n\}$ so that

$$P(2^{j-1}r_n < \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| \leqslant 2^j r_n) = P(\tilde{\boldsymbol{\theta}} \in S_j).$$

As $\tilde{\boldsymbol{\theta}}$ minimizes $\tilde{V}_n(\boldsymbol{\theta})$ it follows that

$$P(\tilde{\boldsymbol{\theta}} \in S_j) \leqslant P\left(\sup_{\boldsymbol{\theta}\in S_j}(\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta_0})) \geqslant 0\right).$$

Now, $\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| > 2^{j-1}r_n$ for $\boldsymbol{\theta} \in S_j$ implies by assumption (A1), that

$$H(\boldsymbol{\theta}) - H(\boldsymbol{\theta_0}) \gtrsim \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2 \gtrsim 2^{2j-2}r_n^2 \quad \text{for } \boldsymbol{\theta} \in S_j, \tag{S.16}$$

which implies $\sup_{\boldsymbol{\theta}\in S_j}|H(\boldsymbol{\theta}) - H(\boldsymbol{\theta_0})| \geqslant 2^{2j-2}r_n^2$. Thus, the event $\sup_{\boldsymbol{\theta}\in S_j}|\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta_0})| \geqslant 0$ can only happen if $\tilde{V}_n$ and H are not too close. Let

$$U_n(\boldsymbol{\theta}) := \tilde{V}_n(\boldsymbol{\theta}) - H(\boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \Theta.$$

It follows from (S.16) that

(S.17)

$$\begin{aligned}P\left(\sup_{\boldsymbol{\theta}\in S_j}(\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta_0})) \geqslant 0\right) &\leqslant P\left(\sup_{\boldsymbol{\theta}\in S_j}(U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0})) \geqslant 2^{2j-2}M^{-1}\right)\\ &\leqslant P\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|\leqslant 2^j r_n}(U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0})) \geqslant 2^{2j-2}M^{-1}\right)\\ &\lesssim \frac{1}{2^{2j-2}r_n^2}\mathbb{E}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|\leqslant 2^j r_n}(U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0}))\right].\end{aligned}$$

The last inequality follows using Markov's inequality. Next, to control the term on the right-hand side of (S.17) uniformly over small $\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|$, define the functions $g_\theta :=$

$d^2(y, m_\oplus(\mathbf{x}^\top\boldsymbol{\theta}))$ and the function class $\mathcal{M}_\delta = \{g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| < \delta\}$. Since by assumption (A2), for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,

$$g_{\boldsymbol{\theta}_1} - g_{\boldsymbol{\theta}_2} \leqslant 2\text{diam}(\Omega)L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \text{ for some constant } L > 0,$$

an envelope function for $\mathcal{M}_\delta$ is $\mathcal{G}_\delta = 2\text{diam}(\Omega)L\delta$. Note that $\mathbb{E}(\mathcal{G}_\delta^2) = O(\delta^2)$. Let $N(\varepsilon, B_\delta(\boldsymbol{\theta_0}), \|\cdot\|)$ be the $\varepsilon-$ covering number for the ball $B_\delta(\boldsymbol{\theta_0}) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \leqslant \delta\}$ of radius δ centered at $\boldsymbol{\theta_0}$, that is, the minimal number of balls of radius ε needed to cover the set $B_\delta(\boldsymbol{\theta_0})$ is $N(\varepsilon, B_\delta(\boldsymbol{\theta_0}), \|\cdot\|)$. Since $\Theta \subset \mathbb{R}^p$, we have [11],

$$N(\varepsilon, B_\delta(\boldsymbol{\theta_0}), \|\cdot\|) \leqslant \left(\frac{C\delta}{\varepsilon}\right)^p.$$

Thus the entropy integral

$$J = J(\delta) = \int_0^1 \sqrt{\log N(\varepsilon, B_\delta(\boldsymbol{\theta_0}), \|\cdot\|)}d\varepsilon = O(1).$$

Using Theorems 2.7.11 and 2.14.2 of [11] we have for small enough δ,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|\leqslant\delta}(U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0}))\right] \leqslant \frac{J\left[\mathbb{E}(\mathcal{G}_\delta^2)\right]^{1/2}}{\sqrt{M}}. \tag{S.18}$$

Thus

$$\begin{aligned} &P(2^{j-1}r_n < \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| \leqslant 2^j r_n) \\ =&P(\tilde{\boldsymbol{\theta}} \in S_j) \\ \leqslant&P\left(\sup_{\boldsymbol{\theta}\in S_j}\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta_0}) \geqslant 0\right) \\ \leqslant&P\left(\sup_{\boldsymbol{\theta}\in S_j}(U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0})) \geqslant 2^{2j-2}r_n^2\right) \\ \lesssim&\frac{2^j r_n}{\sqrt{M}2^{2j-2}r_n^2} = (const.)2^{-2j}. \end{aligned} \tag{S.19}$$

The last equality is obtained by setting $r_n = M^{-1/2}$. As a consequence,

$$P\left(\sqrt{M}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\| > 2^L\right) \leqslant \sum_{j>L, 2^{j-1}r_n\leqslant\eta}\left(\frac{1}{4}\right)^j.$$

The sum converges to zero as $L \to \infty$ and the result follows. □

PROPOSITION S.1.4. *Under assumptions (A2), (A3), (A4), and(A6)*

$$\sqrt{M}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{D} N\left(0, \Lambda(\boldsymbol{\theta_0})\right),$$

where $\Lambda(\boldsymbol{\theta_0}) := \left(\boldsymbol{\Delta^2}H(\boldsymbol{\theta_0})\right)^{-1}\Sigma(\boldsymbol{\theta_0})\left(\boldsymbol{\Delta^2}H(\boldsymbol{\theta_0})\right)^{-1}$.

PROOF. Consider a Taylor expansion of $\tilde{V}_n(\cdot)$, with derivatives as defined in (S.8)– (S.12). Under assumption (A6), for a suitable choice of the tuning parameter $h(n) \to 0$ such that $Mh^2(n) \to \infty$, as $n \to \infty$, the Taylor expansion of the first order difference $\boldsymbol{\Delta}\tilde{V}_n$ approaches

a smooth limit. The following arguments are similar to those demonstrating asymptotic normality of a multivariate Maximum Likelihood Estimator [4], using linearization [11]. As $\tilde{\boldsymbol{\theta}}$ minimizes $\tilde{V}_n$,

$$0 = \boldsymbol{\Delta}\tilde{V}_n(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta_0}) + \boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$$

$$\sqrt{M}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = -\sqrt{M}\left(\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}_*)\right)^{-1}\left(\boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta_0}) - \boldsymbol{\Delta}H(\boldsymbol{\theta_0})\right), \tag{S.20}$$

where $\boldsymbol{\theta}_* \in \mathbb{R}^{p-1}$ is such that, $\|\boldsymbol{\theta}_* - \boldsymbol{\theta_0}\| \leqslant \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\|$. The inverse $\left(\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}_*)\right)^{-1}$ is assumed to exist since $\tilde{\boldsymbol{\theta}}$ is the minimizer of $\tilde{V}_n(\cdot)$ and, $\tilde{V}_n(\cdot)$ being continuous, $\left(\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}_*)\right)$ is non-zero in a sufficiently small ball around $\tilde{\boldsymbol{\theta}}$. From Propositions S.1.1 and S.1.2,

$$\sqrt{M}(\boldsymbol{\Delta}\tilde{V}_n(\boldsymbol{\theta_0}) - \boldsymbol{\Delta}H(\boldsymbol{\theta_0})) \xrightarrow{D} N(0, \Sigma(\boldsymbol{\theta_0})),$$

and $\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta})$ is asymptotically consistent for $\boldsymbol{\Delta^2}H(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$. Thus,

$$\boldsymbol{\Delta^2}\tilde{V}_n(\boldsymbol{\theta}_*) - \boldsymbol{\Delta^2}H(\boldsymbol{\theta}_*) \xrightarrow{P} 0,$$

for $\boldsymbol{\theta}_*$ as described in (S.20). Note that $H(\cdot)$ is continuous and $\boldsymbol{\Delta^2}H(\boldsymbol{\theta})$ assumed to be non-zero in a small ball around $\boldsymbol{\theta_0}$. Now, since $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta_0}$ and $\|\boldsymbol{\theta}_* - \boldsymbol{\theta_0}\| \leqslant \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\|$, under the assumption that $\boldsymbol{\Delta^2}H(\cdot)$ is continuous, the result follows. □

We now proceed to show that the intermediate objective function $\tilde{V}_n(\cdot)$ has a positive curvature near its minimizer.

PROPOSITION S.1.5. *Under assumptions (A0) - (A5), there exist $c_1 > 0$ and $\eta > 0$ such that whenever $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| < \eta$,*

$$P[\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) - c_1\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|^2 \geqslant 0] \to 1,$$

PROOF. We apply assumption (A5) to each term in the summand in the definition of $\tilde{V}_n(\cdot)$. First note that, since $V_n(\tilde{\boldsymbol{\theta}}) \leqslant V_n(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Theta$, we have

$$\tilde{V}_n(\tilde{\boldsymbol{\theta}} + 2a) - 2\tilde{V}_n(\tilde{\boldsymbol{\theta}} + a) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) \leqslant \tilde{V}_n(\tilde{\boldsymbol{\theta}} + 2a) - 2\tilde{V}_n(\tilde{\boldsymbol{\theta}}) + \tilde{V}_n(\tilde{\boldsymbol{\theta}}) = \tilde{V}_n(\tilde{\boldsymbol{\theta}} + 2a) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}).$$

Setting $z_0 = \tilde{\mathbf{X}}_l^\top \tilde{\boldsymbol{\theta}}$, $a = \tilde{\mathbf{X}}_l^\top a$, $u = \tilde{Y}_l$ in assumption (A5) and considering $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} + 2a$ to be any point in the neighborhood of $\tilde{\boldsymbol{\theta}}$ such that $2a < \eta$, we have, using assumption (A5),

$$\begin{aligned}
\tilde{V}_n(\tilde{\boldsymbol{\theta}} + 2a) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) &\geqslant \tilde{V}_n(\tilde{\boldsymbol{\theta}} + 2a) - 2\tilde{V}_n(\tilde{\boldsymbol{\theta}} + a) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) \\
&\geqslant \frac{1}{M}\sum_{l=1}^{M} \kappa(\tilde{\mathbf{X}}_l^\top a)^2 = \frac{1}{M}\sum_{l=1}^{M} \frac{\kappa}{4}(\tilde{\mathbf{X}}_l^\top(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}))^2 \\
&= \frac{\kappa}{4}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top \left(\frac{1}{M}\sum_{l=1}^{M} \tilde{\mathbf{X}}_l\tilde{\mathbf{X}}_l^\top\right)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})
\end{aligned}$$

By the WLLN for the i.i.d sum of $\tilde{\mathbf{X}}_l\tilde{\mathbf{X}}_l^\top$, $l = 1, \ldots, M$,

$$\begin{aligned}
\frac{\kappa}{4}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top \left(\frac{1}{M}\sum_{l=1}^{M} \tilde{\mathbf{X}}_l\tilde{\mathbf{X}}_l^\top\right)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) &\xrightarrow{P} \frac{\kappa}{4}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top \mathbb{E}(XX^\top)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
&\geqslant \frac{\kappa}{4}\lambda_1(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) > 0.
\end{aligned}$$

The last two inequalities hold under the assumption that the matrix $\mathbb{E}(XX^\top)$ is positive definite with the smallest eigen value λ_1 bounded away from 0. Thus, for $c_1 = \frac{\kappa}{4}\lambda_1$, $P[\tilde{V}_n(\boldsymbol{\theta}) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) - c_1\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|^2 \geqslant 0] \to 1$. □

It remains to show that

PROPOSITION S.1.6. *Under assumptions (A0)-(A5), (U1)-(U3), and (R1)-(R2),*

$$\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| = O_P(a_n^{1/2}).$$

where a_n is as defined in for Lemma 1 [1].

PROOF. Define $\Theta_\epsilon := \{\boldsymbol{\theta} : \tilde{V}_n(\boldsymbol{\theta}) - \epsilon \leqslant \tilde{V}_n(\tilde{\boldsymbol{\theta}}) + \epsilon\}$ for some $\varepsilon > 0$. Then under proposition S.1.5, for any $\boldsymbol{\theta} \in \Theta_\varepsilon$, $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| = O_P(\epsilon^{1/2})$.

Now, from Lemma 1 [1], for any $\delta > 0$, there exists $M_0 > 0$ such that $P\left(a_n^{-1}|V_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta})| > M_0\right) \to 0$, for all $M > M_0$ and for any $\boldsymbol{\theta}$. Define

$$E_1 := \sup_{\boldsymbol{\theta}\in\Theta}\{a_n^{-1}|V_n(\boldsymbol{\theta}) - \tilde{V}_n(\boldsymbol{\theta})| \leqslant M_0\} \text{ and } E_2 = E_1^c$$

and observe $P(E_2) \to 0$ for large enough M_0. Choosing $\epsilon = M_0 a_n$ in the above definition of Θ_ϵ, one finds that $\Theta_{M_0 a_n}$ is nonempty and bounded above, since $\tilde{\boldsymbol{\theta}} \in \Theta_{M_0 a_n}$ and $\Theta_{M_0 a_n} \subset \Theta$. We claim that $\hat{\boldsymbol{\theta}} \in \Theta_{M_0 a_n}$ on E_1. Suppose this is not the case. Then $\tilde{V}_n(\hat{\boldsymbol{\theta}}) - M_0 a_n > \tilde{V}_n(\tilde{\boldsymbol{\theta}}) + M_0 a_n$. On E_1,

$$V_n(\hat{\boldsymbol{\theta}}) > \tilde{V}_n(\hat{\boldsymbol{\theta}}) - M_0 a_n > \tilde{V}_n(\tilde{\boldsymbol{\theta}}) + M_0 a_n > V_n(\tilde{\boldsymbol{\theta}}) > V_n(\hat{\boldsymbol{\theta}}),$$

which is a contradiction, since $\hat{\boldsymbol{\theta}}$ minimizes $V_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta$. Thus, $\hat{\boldsymbol{\theta}} \in \Theta_{M_0 a_n}$ on E_1, that is $\tilde{V}_n(\hat{\boldsymbol{\theta}}) - \tilde{V}_n(\tilde{\boldsymbol{\theta}}) \leqslant 2M_0 a_n$. Based on the positive curvature condition on $\tilde{V}_n(\cdot)$ around $\tilde{\boldsymbol{\theta}}$ given in proposition S.1.5, on E_1 we have,

$$P(c_1\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|^2 \leqslant 2M_0 a_n) \geqslant P(c_1\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|^2 \leqslant |\tilde{V}_n(\hat{\boldsymbol{\theta}}) - \tilde{V}_n(\tilde{\boldsymbol{\theta}})|) \to 1.$$

For $L > 2M_0/c_1$,

$$\begin{aligned}
&P\left(a_n^{-1/2}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| > L\right) \\
&= P\left(a_n^{-1/2}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| > L|E_1\right) P(E_1) + P\left(a_n^{-1/2}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| > L|E_2\right) P(E_2) \\
&= O_P(1), \text{ since } P(E_2) \to 0 \text{ for large enough } M_0.
\end{aligned}$$

Thus $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| = O_P(a_n^{1/2})$. □

PROOF OF THEOREM 2. Decompose

$$\sqrt{M}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \sqrt{M}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + \sqrt{M}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$$

From assumption (A4), $M = M(n)$ such that $Ma_n \to 0$, as $n \to \infty$, with a_n as defined in equation (3.1) in the main manuscript for Lemma 1 [1]. Thus, using Proposition S.1.6, the first term is $o_P(1)$ and by Proposition S.1.4 the second term converges in distribution to Z, where Z is a Gaussian random variable with mean 0 and variance $\Lambda(\boldsymbol{\theta_0}) = \left(\boldsymbol{\Delta^2} H(\boldsymbol{\theta_0})\right)^{-1} \Sigma(\boldsymbol{\theta_0}) \left(\boldsymbol{\Delta^2} H(\boldsymbol{\theta_0})\right)^{-1}$. The result follows. □

PROOF OF PROPOSITION 3. The estimator $\widehat{\Sigma}(\boldsymbol{\theta_0})$ for $\Sigma(\boldsymbol{\theta_0})$ has elements $\widehat{\Sigma}(\boldsymbol{\theta_0}) = ((\widehat{\sigma}_{rs}(\boldsymbol{\theta_0})))_{r,s=2,\ldots,p}$ with

$$\widehat{\sigma}_{rs}(\boldsymbol{\theta_0}) = \begin{cases} \frac{1}{hM} & \sum_{l=1}^{M} \widehat{\nabla^2}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}) - \left(\frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\right)^2, \text{ if } r = s, \\ \frac{1}{hM} & \sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s}) \\ & - \left(\frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\right)\left(\frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s})\right), \text{ if } r \neq s, \end{cases}$$

where the auxiliary quantities are defined as in (S.10). Given a $p \times q$ random matrix $R_n = ((R_{ij}^{(n)}))$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, R_n converges to a normal limit if $\mathbf{vec}(R_n)$ converges to a multivariate normal distribution in the standard sense. We denote the asymptotic expectation of $R_{ij}^{(n)}$ as μ_{ij} for all i, j and the asymptotic covariance between $R_{ij}^{(n)}$ and $R_{rs}^{(n)}$ as Σ_{ijrs}. If R_n has an asymptotic normal distribution, the distribution is characterized by the mean $\mu = ((\mu_{ij}))$ and covariance $\Sigma = \{\Sigma_{ijrs}\}$ for any $i, r = 1, \ldots, p$ and $j, s = 1, \ldots, q$. Here μ is $p \times q$ matrix and Σ is a four-tensor.

The convergence in distribution described here can easily be understood in the standard multivariate sense by vectorizing the matrices in question. Standard results from multivariate statistics, specifically the delta method, extend immediately to the matrix case. Define the following auxiliary quantities:

$$A_{rs}(\boldsymbol{\theta_0}) = \lim_{h\to 0} \frac{1}{h} E\left(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s})\right),$$

$$B_r(\boldsymbol{\theta_0}) = \lim_{h\to 0} \frac{1}{h} \mathbb{E}\Big(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\Big),$$

$$C_s(\boldsymbol{\theta_0}) = \lim_{h\to 0} \frac{1}{h} \mathbb{E}\Big(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s})\Big),$$

$$\widehat{A}_{rs}(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s}),$$

$$\widehat{B}_r(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}),$$

$$\widehat{C}_s(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s}),$$

$$\tilde{A}_{rs}(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r})\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s}),$$

$$\tilde{B}_r(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0r}),$$

$$\tilde{C}_s(\boldsymbol{\theta_0}) = \frac{1}{hM}\sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_{0s}).$$

We denote the collection $\mathcal{C} = \{A_{rs}, B_r, C_s\}$ for $r, s = 2, \ldots, p$. Let g be a matrix-valued function on the space of such collections with component functions $g_{rs}(\mathcal{C}) = A_{rs} - B_r C_s$

for all $r, s = 2, \ldots, p$. Note that g is clearly differentiable. For the collections

$$\widehat{\mathcal{C}}(\boldsymbol{\theta_0}) = \{\widehat{A}_{rs}(\boldsymbol{\theta_0}), \widehat{B}_r(\boldsymbol{\theta_0}), \widehat{C}_s(\boldsymbol{\theta_0})\},$$
$$\mathcal{C}(\boldsymbol{\theta_0}) = \{A_{rs}(\boldsymbol{\theta_0}), B_r(\boldsymbol{\theta_0}), C_s(\boldsymbol{\theta_0})\}, \text{ and}$$
$$\tilde{\mathcal{C}}(\boldsymbol{\theta_0}) = \{\tilde{A}_{rs}(\boldsymbol{\theta_0}), \tilde{B}_r(\boldsymbol{\theta_0}), \tilde{C}_s(\boldsymbol{\theta_0})\},$$

respectively, $\widehat{\Sigma}(\boldsymbol{\theta_0}) = g(\widehat{\mathcal{C}}(\boldsymbol{\theta_0}))$, $\Sigma(\boldsymbol{\theta_0}) = g(\mathcal{C}(\boldsymbol{\theta_0}))$, and $\tilde{\Sigma}(\boldsymbol{\theta_0}) = g(\tilde{\mathcal{C}}(\boldsymbol{\theta_0}))$. It is sufficient to show that the collection $\sqrt{M}(\widehat{\mathcal{C}}(\boldsymbol{\theta_0}) - \mathcal{C}(\boldsymbol{\theta_0}))$, appropriately vectorized, follows a multivariate normal distribution, whence an application of the delta method gives the final result.

Now,

$$\sqrt{M}(\widehat{\mathcal{C}}(\boldsymbol{\theta_0}) - \mathcal{C}(\boldsymbol{\theta_0})) = \sqrt{M}(\widehat{\mathcal{C}}(\boldsymbol{\theta_0}) - \tilde{\mathcal{C}}(\boldsymbol{\theta_0})) + \sqrt{M}(\tilde{\mathcal{C}}(\boldsymbol{\theta_0}) - \mathcal{C}(\boldsymbol{\theta_0})).$$

The first term in the sum is $o_P(1)$ and the second term converges in distribution to $N(0, D)$ by applying the CLT to the i.i.d. random vectors in the appropriately vectorized collection $\tilde{\mathcal{C}}(\boldsymbol{\theta_0})$, with D as the covariance matrix of the collection of vectors. Denoting the Jacobian of $\mathbf{vec}(g)$ evaluated at $\mathcal{C}(\boldsymbol{\theta_0})$ as $J(\mathcal{C}(\boldsymbol{\theta_0}))$, applying the delta method leads to $J(\mathcal{C}(\boldsymbol{\theta_0}))D(J(\mathcal{C}(\boldsymbol{\theta_0})))^\top$ as the asymptotic variance of $\mathbf{vec}(\widehat{\Sigma}(\boldsymbol{\theta_0}))$. □

The above Proposition 3 implies that $\widehat{\Sigma}(\boldsymbol{\theta_0}) - \Sigma(\boldsymbol{\theta_0}) = o_P(1)$. Furthermore,

PROPOSITION S.1.7. *Under assumptions (A0) and (A2)-(A6),*

$$\sup_{\boldsymbol{\theta}\in\Theta} (\hat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta})) \xrightarrow{P} 0.$$

PROOF. Recall the definition of $\Sigma(\boldsymbol{\theta}) = ((\sigma_{rs}(\boldsymbol{\theta})))_{r,s=2,\ldots,p}$, where, for any $\boldsymbol{\theta} \in \Theta$,

$$\sigma_{rs}(\boldsymbol{\theta}) = \begin{cases} \lim_{\varepsilon\to 0} \operatorname{Var}\left(\frac{1}{\varepsilon}\nabla_\varepsilon(\mathbf{X}, Y, \theta_r)\right), \text{ if } r = s \in \{2, \ldots, p\} \\ \lim_{\varepsilon\to 0} \operatorname{Cov}\left(\frac{1}{\varepsilon}\nabla_\varepsilon(\mathbf{X}, Y, \theta_r), \frac{1}{\varepsilon}\nabla_\varepsilon(\mathbf{X}, Y, \theta_s)\right), \text{ if } r \neq s, r, s \in \{2, \ldots, p\}. \end{cases}$$

The estimator of $\Sigma(\boldsymbol{\theta})$ is given by $\hat{\Sigma}(\boldsymbol{\theta}) = ((\hat{\sigma}_{rs}(\boldsymbol{\theta})))_{r,s=2,\ldots,p}$ with

$$\hat{\sigma}_{rs}(\boldsymbol{\theta}) = \begin{cases} \frac{1}{h^2 M}\sum_{l=1}^M \widehat{\nabla}_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \left(\frac{1}{hM}\sum_{l=1}^M \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\right)^2, \text{ if } r = s \in \{2, \ldots, p\} \\ \frac{1}{h^2 M}\sum_{l=1}^M \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) \\ -\left(\frac{1}{hM}\sum_{l=1}^M \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\right)\left(\frac{1}{hM}\sum_{l=1}^M \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\right), \text{ if } r \neq s \in \{2, \ldots, p\}. \end{cases}$$

To obtain elementwise convergence, we introduce an intermediate version, where for any $\boldsymbol{\theta} \in \Theta$, as $\tilde{\Sigma}(\boldsymbol{\theta}) = ((\tilde{\sigma}_{rs}(\boldsymbol{\theta})))_{r,s=2,\ldots,p}$ with

$$\tilde{\sigma}_{rs}(\boldsymbol{\theta}) = \begin{cases} \frac{1}{h^2 M}\sum_{l=1}^M \nabla_h^2(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \left(\frac{1}{hM}\sum_{l=1}^M \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\right)^2, \text{ if } r = s \in \{2, \ldots, p\} \\ \frac{1}{h^2 M}\sum_{l=1}^M \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) \\ -\left(\frac{1}{hM}\sum_{l=1}^M \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\right)\left(\frac{1}{hM}\sum_{l=1}^M \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\right), \text{ if } r \neq s \in \{2, \ldots, p\}, \end{cases}$$

where all auxiliary quantities are defined in (S.8)- (S.12). We focus on each element of the covariance matrix separately. For $r \in \{2, \ldots, p\}$ and any $\boldsymbol{\theta} \in \Theta$,

$$\sup_{\theta_r} \left| \frac{1}{M}\sum_{l=1}^M \frac{1}{h}\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\mathbb{E}(\nabla_\varepsilon(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) \right|$$

$$\leqslant \sup_{\theta_r} \left| \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \frac{1}{h} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) \right|$$

$$+ \sup_{\theta_r} \left| \frac{1}{h} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) - \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}(\nabla_\varepsilon(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) \right|$$

The second term on the r.h.s. goes to zero as $h \to 0$. As for the first term on the r.h.s., consider

$$\sup_{\theta_r} \left| \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \frac{1}{h} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) \right| \tag{S.21}$$

$$\leqslant \sup_{\theta_r} \left| \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \right|$$

$$+ \sup_{\theta_r} \left| \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \frac{1}{h} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)) \right|$$

The second term on the r.h.s. of (S.21) goes to zero in probability by a uniform LLN, as $|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)| \leqslant 2DLh$, where L is the Lipschitz constant for $m_\oplus$ in assumption (A2), and we note that Θ is compact and $\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \cdot)$ is continuous. For the first term on the r.h.s., by the uniform convergence of $\hat{m}_\oplus$ to $m_\oplus$ from Lemma 1,

$$\sup_{\theta_r} \left| \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \right| \leqslant 4D \sup_{\boldsymbol{\theta}} \sup_{t} d(\hat{m}_\oplus(t, \boldsymbol{\theta}), m_\oplus(t, \boldsymbol{\theta})) = O_P(a_n), \tag{S.22}$$

where a_n is as in equation (3.1) in the main manuscript and the rate is uniform for all $1 \leqslant l \leqslant M$ and does not depend on M. Then

$$\sup_{\theta_r} \left| \frac{1}{hM} \sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \frac{1}{hM} \sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \right| \xrightarrow{P} 0,$$

since $Mh \to \infty$, as per assumption (A6), and $Ma_n \to 0$, as per assumption (A4), as $n \to \infty$.

Next we consider the product terms in the element-wise covariance. For $r, s \in \{2, \ldots, p\}$, and any $\boldsymbol{\theta} \in \Theta$,

$$\sup_{\theta_r, \theta_s} \left| \frac{1}{h^2 M} \sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \frac{1}{h^2} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)) \right|$$

$$\leqslant \sup_{\theta_r, \theta_s} \left| \frac{1}{h^2 M} \sum_{l=1}^{M} \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \frac{1}{h^2 M} \sum_{l=1}^{M} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) \right|$$

$$+ \sup_{\theta_r, \theta_s} \left| \frac{1}{M} \sum_{l=1}^{M} \frac{1}{h^2} \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \frac{1}{h^2} \mathbb{E}(\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)) \right|.$$

Again, by a similar argument as above, the second term on the r.h.s. goes to zero in probability. As for the first term on the r.h.s.,

$$\sup_{\theta_r, \theta_s} \left| \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) \right| \tag{S.23}$$

$$= \sup_{\theta_r, \theta_s} \left| [\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)][\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)] \right|$$

$$+ \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) + \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)$$

$$- 2\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big| \tag{S.24}$$

The first term on the r.h.s. of (S.23) is $O_P(a_n^2)$ by (S.22), observing that

$$\sup_{\theta_r,\theta_s}\Big|[\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)][\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)]\Big|$$

$$\leqslant \sup_{\theta_r}\Big|\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\Big| \sup_{\theta_l}\Big|\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big|.$$

As for the second term on the r.h.s. of (S.23),

$$\sup_{\theta_r,\theta_s}\Big|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) + \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)$$

$$- 2\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big| \tag{S.25}$$

$$\leqslant \sup_{\theta_r,\theta_s}\Big|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big|$$

$$+ \sup_{\theta_r,\theta_s}\Big|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\widehat{\nabla}_h(\tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big|$$

$$= \sup_{\theta_r,\theta_s}\Big|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)[\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)]\Big|$$

$$+ \sup_{\theta_r,\theta_s}\Big|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)[\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)]\Big|.$$

From (S.22), and using the fact that $\sup_{\theta_j}|\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_j)| < 2DLh$ is bounded for $j = r, s$, for any $r, s \in \{2, \ldots, p\}$, both terms in (S.25) are $O_P(ha_n)$. Combining these results,

$$\sup_{\theta_r,\theta_s}\left|\frac{1}{h^2M}\sum_{l=1}^{M}\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \frac{1}{h^2M}\sum_{l=1}^{M}\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\right|$$

$$= \sup_{\theta_r,\theta_s}\Big|\frac{1}{h^2M}\sum_{l=1}^{M}[\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)][\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) - \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)]$$

$$+\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s) + \nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\widehat{\nabla}_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r) - 2\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_r)\nabla_h(\tilde{\mathbf{X}}_l, \tilde{Y}_l, \theta_s)\Big|$$

$$= O_P(a_n^2/h^2) + O_P(a_n/h) \xrightarrow{P} 0,$$

where assumptions (A4) and (A6) imply $hM \to \infty$ and $Ma_n \to 0$ as $n \to \infty$. Finally, plugging into the elementwise definitions, by decomposing

$$\sup_{\theta_r,\theta_s}|\hat{\sigma}_{rs}(\boldsymbol{\theta}) - \sigma_{rs}(\boldsymbol{\theta})| \leqslant \sup_{\theta_r,\theta_s}|\hat{\sigma}_{rs}(\boldsymbol{\theta}) - \tilde{\sigma}_{rs}(\boldsymbol{\theta})| + \sup_{\theta_r,\theta_s}|\tilde{\sigma}_{rs}(\boldsymbol{\theta}) - \sigma_{rs}(\boldsymbol{\theta})| \xrightarrow{P} 0, \tag{S.26}$$

the result follows. □

PROPOSITION S.1.8. *Under assumptions (A2), (A4), and (A6),*

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta}) - \boldsymbol{\Delta}^{\boldsymbol{2}}H(\boldsymbol{\theta})\| \xrightarrow{P} 0.$$

PROOF. Observe

(S.27)
$$\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta})-\boldsymbol{\Delta}^{\boldsymbol{2}}H(\boldsymbol{\theta})\|\leqslant\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta})-\boldsymbol{\Delta}^{\boldsymbol{2}}\tilde{V}_n(\boldsymbol{\theta})\|+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Delta}^{\boldsymbol{2}}\tilde{V}_n(\boldsymbol{\theta})-\boldsymbol{\Delta}^{\boldsymbol{2}}H(\boldsymbol{\theta})\|$$

and for the second term on the r.h.s. of (S.27), consider the difference between the second order partial derivatives

$$\begin{aligned}
&\sup_{\theta_r,\theta_s}\left|\frac{\partial^2\tilde{V}_n(\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}-\frac{\partial^2 H(\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}\right|\\
=&\sup_{\theta_r,\theta_s}\left|\frac{1}{M}\sum_{l=1}^{M}\frac{1}{h^2}\nabla_h^2(\tilde{\mathbf{X}}_l,\tilde{Y}_l,\theta_r,\theta_s)-\lim_{\varepsilon\to0}\frac{1}{\varepsilon^2}\mathbb{E}(\nabla_\varepsilon^2(\mathbf{X},Y,\theta_r,\theta_s))\right|\\
\leqslant&\sup_{\theta_r,\theta_s}\left|\frac{1}{M}\sum_{l=1}^{M}\frac{1}{h^2}\nabla_h^2(\tilde{\mathbf{X}}_l,\tilde{Y}_l,\theta_r,\theta_s)-\frac{1}{h^2}\mathbb{E}(\nabla_h^2(\mathbf{X},Y,\theta_r,\theta_s))\right|\\
&+\sup_{\theta_r,\theta_s}\left|\frac{1}{h^2}\mathbb{E}(\nabla_h^2(\mathbf{X},Y,\theta_r,\theta_s))-\lim_{\varepsilon\to0}\frac{1}{\varepsilon^2}\mathbb{E}(\nabla_\varepsilon^2(\mathbf{X},Y,\theta_r,\theta_s))\right|
\end{aligned}$$

The second term on the r.h.s. goes to zero as $h\to0$. As for the first term on the r.h.s., we note that for each $r,s=2,\ldots,p$, $|\nabla_h^2(\mathbf{X},Y,\theta_r,\theta_s)|\leqslant Kh^2$ for some constant $K>0$. Further, Θ is a compact set, and from the proof of Proposition 2, $\nabla_h^2(\mathbf{X},Y,\theta_r,\theta_s)/h^2$ is continuous in θ_r,θ_s. Thus, by a uniform LLN,

$$\sup_{\theta_r,\theta_s}\left|\frac{1}{M}\sum_{l=1}^{M}\frac{1}{h^2}\nabla_h^2(\tilde{\mathbf{X}}_l,\tilde{Y}_l,\theta_r,\theta_s)-\frac{1}{h^2}\mathbb{E}(\nabla_h^2(\mathbf{X},Y,\theta_r,\theta_s))\right|\xrightarrow{P}0,$$

as $h\to0$. As for the first term on the r.h.s. of (S.27). Consider the elementwise differences

$$\begin{aligned}
&\sup_{\theta_r,\theta_s}\left|\frac{\partial^2V_n(\theta)}{\partial\theta_r\partial\theta_s}-\frac{\partial^2\tilde{V}_n(\theta)}{\partial\theta_r\partial\theta_s}\right|\leqslant\frac{1}{h^2}\frac{1}{M}\sum_{l=1}^{M}\sup_{\theta_r,\theta_s}\left|\widehat{\nabla^2}h(\tilde{\mathbf{X}}_l,\tilde{Y}_l,\theta_r,\theta_s)-\nabla_h^2(\tilde{\mathbf{X}}_l,\tilde{Y}_l,\theta_r,\theta_s)\right|\\
&\leqslant8D\,\frac{1}{h^2}\frac{1}{M}\sum_{l=1}^{M}\sup_{\boldsymbol{\theta}}\sup_{t}\,d(\hat{m}_\oplus(t,\boldsymbol{\theta}),m_\oplus(t,\boldsymbol{\theta}))=8D\,\frac{1}{h^2}O_P(a_n),
\end{aligned}$$

where the O_P term is uniform in $l=1,\ldots,M$ and doesn't depend on M. By assumptions (A4) and (A6), the above term goes to zero in probability, implying that

$$\sup_{\boldsymbol{\theta}\in\Theta}|\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta})-\boldsymbol{\Delta}^{\boldsymbol{2}}\tilde{V}_n(\boldsymbol{\theta})|=o_P(1)$$

and the result follows. □

PROOF OF PROPOSITION 4. Recall, for any $\boldsymbol{\theta}\in\Theta$, $\Lambda(\boldsymbol{\theta})=(\boldsymbol{\Delta}^{\boldsymbol{2}}H(\boldsymbol{\theta}))^{-1}\Sigma(\boldsymbol{\theta})(\boldsymbol{\Delta}^{\boldsymbol{2}}H(\boldsymbol{\theta}))^{-1}$ and $\widehat{\Lambda}(\boldsymbol{\theta})=\left(\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta})\right)^{-1}\widehat{\Sigma}(\boldsymbol{\theta})\left(\boldsymbol{\Delta}^{\boldsymbol{2}}V_n(\boldsymbol{\theta})\right)^{-1}$. Writing

(S.28)
$$\widehat{\Lambda}(\hat{\boldsymbol{\theta}})-\Lambda(\boldsymbol{\theta_0})=(\widehat{\Lambda}(\hat{\boldsymbol{\theta}})-\Lambda(\hat{\boldsymbol{\theta}}))+(\Lambda(\hat{\boldsymbol{\theta}})-\Lambda(\boldsymbol{\theta_0})),$$

we need to show that both terms on the RHS converge to zero in probability. For the first term, it suffices to show that

$$\sup_{\boldsymbol{\theta}\in\Theta_{0\delta}}\widehat{\Lambda}(\boldsymbol{\theta})-\Lambda(\boldsymbol{\theta})\xrightarrow{P}0,$$

where $\Theta_{0\delta}$ is a small δ- neighborhood of the true parameter $\boldsymbol{\theta_0}$ i.e., for any $\delta > 0$, $\Theta_{0\delta} := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \leqslant \delta\}$. By Theorem 1, since $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta_0}$, $\hat{\boldsymbol{\theta}} \in \Theta_{0\delta}$ for large enough sample size n with high probability. Observe that

$$
\begin{aligned}
&\widehat{\Lambda}(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}) \\
&= \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} \widehat{\Sigma}(\boldsymbol{\theta}) \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} - (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1} \Sigma(\boldsymbol{\theta}) (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1} \\
&= \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} \left[\widehat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta})\right] \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} \\
&\quad + (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1} \Sigma(\boldsymbol{\theta}) \left[\left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} - (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1}\right] \\
&\quad + \left[\left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} - (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1}\right] \Sigma(\boldsymbol{\theta}) \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1}.
\end{aligned}
$$

From the proof of Propositions S.1.7 and S.1.8, we have $\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}) \xrightarrow{P} 0$ and $\sup_{\boldsymbol{\theta} \in \Theta} \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right) - (\boldsymbol{\Delta^2} H(\boldsymbol{\theta})) \xrightarrow{P} 0$. Also, $\boldsymbol{\Delta^2} H(\boldsymbol{\theta})$ is upper-bounded, and since $\boldsymbol{\theta_0}$ is the global minimizer of $H(\boldsymbol{\theta})$, it is bounded away from 0 in any neighborhood $\Theta_{0\delta}$ of $\boldsymbol{\theta_0}$. Thus $(\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1}$ is bounded above on $\Theta_{0\delta}$ and uniformly continuous. Further, by virtue of continuity of $\boldsymbol{\Delta^2} H(\boldsymbol{\theta})$ and the uniform probability convergence of $\boldsymbol{\Delta^2} V_n(\cdot)$ to $\boldsymbol{\Delta^2} H(\cdot)$ on the parameter space Θ (Proposition S.1.8), one can show that $\left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1}$ such that $\boldsymbol{\theta} \in \Theta_{0\delta}$, is bounded above and uniformly continuous with high probability for a large enough sample size, therefore yielding $\sup_{\boldsymbol{\theta} \in \Theta_{0\delta}} \left(\boldsymbol{\Delta^2} V_n(\boldsymbol{\theta})\right)^{-1} - (\boldsymbol{\Delta^2} H(\boldsymbol{\theta}))^{-1} \xrightarrow{P} 0$. Combining the above arguments it follows that $\sup_{\boldsymbol{\theta} \in \Theta_{0\delta}} \widehat{\Lambda}(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}) \xrightarrow{P} 0$. Finally, noting that $P(\hat{\boldsymbol{\theta}} \in \Theta_{0\delta}) \to 1$ for any $\delta > 0$ and a large enough sample size, we have $\widehat{\Lambda}(\hat{\boldsymbol{\theta}}) - \Lambda(\hat{\boldsymbol{\theta}}) \xrightarrow{P} 0$.

For the second term in (S.28), under the assumption of the total boundedness of (Ω, d) and the continuity of the local linear Fréchet regression function (assumption (A2)), both $\Sigma(\cdot)$ and $\boldsymbol{\Delta^2} H(\cdot)$ are continuous functions of $\boldsymbol{\theta}$. Thus applying continuous mapping theorem and using the consistency of $\hat{\boldsymbol{\theta}}$ (Theorem 1) we have $\Lambda(\hat{\boldsymbol{\theta}}) - \Lambda(\boldsymbol{\theta_0}) \xrightarrow{P} 0$. The result follows combining the two terms in (S.28) using Slutsky's theorem. □

PROOF OF COROLLARY 4. For any $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$, we observe that, by the triangle inequality of the metric,

$$
\begin{aligned}
&d(\hat{m}_{\oplus}(\mathbf{x}^\top \widehat{\bar{\boldsymbol{\theta}}}, \widehat{\bar{\boldsymbol{\theta}}}), m_{\oplus}(\mathbf{x}^\top \bar{\boldsymbol{\theta}}_{\mathbf{0}}, \bar{\boldsymbol{\theta}}_{\mathbf{0}})) \\
&\leqslant d(\hat{m}_{\oplus}(\mathbf{x}^\top \widehat{\bar{\boldsymbol{\theta}}}, \widehat{\bar{\boldsymbol{\theta}}}), m_{\oplus}(\mathbf{x}^\top \widehat{\bar{\boldsymbol{\theta}}}, \widehat{\bar{\boldsymbol{\theta}}})) + d(m_{\oplus}(\mathbf{x}^\top \widehat{\bar{\boldsymbol{\theta}}}, \widehat{\bar{\boldsymbol{\theta}}}), m_{\oplus}(\mathbf{x}^\top \bar{\boldsymbol{\theta}}_{\mathbf{0}}, \bar{\boldsymbol{\theta}}_{\mathbf{0}})).
\end{aligned} \tag{S.29}
$$

From Lemma 1, we know that

$$
\sup_{\bar{\boldsymbol{\theta}}} \sup_{t} d(\hat{m}_{\oplus}(t, \bar{\boldsymbol{\theta}}), m_{\oplus}(t, \bar{\boldsymbol{\theta}})) = O_P(a_n),
$$

where a_n is as defined in equation (3.1) in the main manuscript. Thus, the first term of (S.29) is $O_P(a_n)$. Now by assumption (A2), $m_{\oplus}$ is continuous and, by Theorem 2, $\widehat{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}_{\mathbf{0}} = O_P(M^{-1/2})$. Using continuous mapping theorem, the second term of (S.29) is $O_P(M^{-1/2})$. The result follows using Slutsky's theorem under assumption (A4). □

PROOF OF COROLLARY 5. We partition $\hat{\boldsymbol{\theta}}$ into sub-vectors as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{(\mathbf{r})}, \hat{\boldsymbol{\theta}}^{(\mathbf{r})})^\top$, where $\hat{\boldsymbol{\theta}}_{(\mathbf{r})} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)^\top$ and $\hat{\boldsymbol{\theta}}^{(\mathbf{r})} = (\hat{\theta}_{r+1}, \ldots, \hat{\theta}_p)^\top$. Similarly, we can partition the true direction

as $\boldsymbol{\theta_0} = (\boldsymbol{\theta}_{\mathbf{0(r)}}, \boldsymbol{\theta}_{\mathbf{0}}^{(\mathbf{r})})^\top$, where $\boldsymbol{\theta}_{\mathbf{0(r)}} = (\theta_{01}, \ldots, \theta_{0r})^\top$ and $\boldsymbol{\theta}_{\mathbf{0}}^{(\mathbf{r})} = (\theta_{0r+1}, \ldots, \theta_{0p})^\top$. Since, from Corollary 3 we know that, under suitable assumptions

$$\sqrt{M}(\widehat{\Lambda}(\hat{\boldsymbol{\theta}}))^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{D} N(0, I_{p-1}),$$

applying the the linear transformation $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})}) \mapsto \boldsymbol{\theta}^{(\mathbf{r})}$ for any $\boldsymbol{\theta} : \boldsymbol{\theta}^\top\boldsymbol{\theta} < 1$, we have

$$\sqrt{M}(\widehat{\Lambda}(\hat{\boldsymbol{\theta}}^{(\mathbf{r})}))^{-1/2}\hat{\boldsymbol{\theta}}^{(\mathbf{r})} \xrightarrow{D} N(0, I_{p-r}),$$

where $\widehat{\Lambda}(\hat{\boldsymbol{\theta}}^{(\mathbf{r})})$ is the $(p-r)$ dimensional sub-matrix of the asymptotic covariance matrix $\widehat{\Lambda}(\hat{\boldsymbol{\theta}})$. Thus, under the null hypothesis $H_0 : B\boldsymbol{\theta}^{(\mathbf{r})} = \zeta$, for some $q \times (p-r)$ matrix B with $1 \leqslant q \leqslant p-r$ of rank q, we can form a Wald-type test statistic

$$M(B\hat{\boldsymbol{\theta}}^{(\mathbf{r})} - \zeta)^\top(B(\widehat{\Lambda}(\hat{\boldsymbol{\theta}}^{(\mathbf{r})}))^{-1}B^\top)^{-1}(B\hat{\boldsymbol{\theta}}^{(\mathbf{r})} - \zeta) \xrightarrow{D} \chi^2_q.$$

For the particular choice of $B = I_{p-r}$, a simultaneous confidence region for $\boldsymbol{\theta}_{\mathbf{0}}^{(\mathbf{r})}$ can be computed, as is given in Corollary 5. □

PROOF OF PROPOSITION 5. Recall the population and sample versions of the index parameters as M-estimation problems

$$\boldsymbol{\theta_0} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, H(\boldsymbol{\theta}),\ H(\boldsymbol{\theta}) = \mathbb{E}[d^2(Y, m_\oplus(\mathbf{X}^\top\boldsymbol{\theta}, \boldsymbol{\theta}))],$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, V_n(\boldsymbol{\theta}),\ V_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{l=1}^{M} d^2(\tilde{Y}_l, \hat{m}_\oplus(\tilde{\mathbf{X}}_l^\top\boldsymbol{\theta}, \boldsymbol{\theta})),$$

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \tilde{V}_n(\boldsymbol{\theta}),\ \tilde{V}_n(\boldsymbol{\theta}) = \frac{1}{M}\sum_{l=1}^{M} d^2(\tilde{Y}_l, m_\oplus(\tilde{\mathbf{X}}_l^\top\boldsymbol{\theta}, \boldsymbol{\theta})),$$

where $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{p-1}, ^\top \boldsymbol{\theta}^\top\boldsymbol{\theta} < 1\} \subset \mathbb{R}^{p-1}$, as defined in equation (3.4) in the main manuscript. We want to show the consistency of the proposed bootstrap estimator $\hat{\Lambda}^* := \mathbb{E}\Big[M(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^\intercal | (\tilde{\mathbf{X}}_1, \tilde{Y}_1), \ldots, (\tilde{\mathbf{X}}_M, \tilde{Y}_M)\Big]$. Here $\hat{\boldsymbol{\theta}}^*$ denotes the M-estimator computed from a bootstrap sample $(\tilde{\mathbf{X}}_l^*, \tilde{Y}_l^*)$, $l = 1, \ldots, M$ for the objective function V_n. Define the auxiliary quantity
$\tilde{\Lambda}^* := \mathbb{E}\Big[M(\tilde{\boldsymbol{\theta}}^* - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}^* - \tilde{\boldsymbol{\theta}})^\intercal | (\tilde{\mathbf{X}}_1, \tilde{Y}_1), \ldots, (\tilde{\mathbf{X}}_M, \tilde{Y}_M)\Big]$, where $\tilde{\boldsymbol{\theta}}^*$ denotes the M-estimator computed from a bootstrap sample $(\tilde{\mathbf{X}}_l^*, \tilde{Y}_l^*)$, $l = 1, \ldots, M$ for the objective function $\tilde{V}_n$.

First note that by Proposition S.1.6, $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} \xrightarrow{P} 0$. Also, under similar assumptions required for Proposition S.1.6, one can show that $\hat{\boldsymbol{\theta}}^* - \tilde{\boldsymbol{\theta}}^* \xrightarrow{P} 0$, resulting in $\hat{\Lambda}^* - \tilde{\Lambda}^* \xrightarrow{P} 0$. Now, one can show the consistency of $\tilde{\Lambda}^*$ by applying Theorem 2.2 of [5], Define $g_{\boldsymbol{\theta}} = d^2(y, m_\oplus(\mathbf{x}^\top\boldsymbol{\theta}))$. One needs to show
(i) There exist a $\boldsymbol{\theta_0} \in \Theta$, and a positive constant c such that $H(\boldsymbol{\theta}) - H(\boldsymbol{\theta_0}) \geqslant c\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2$ for all $\boldsymbol{\theta} \in \Theta$.
(ii) The class of functions $\mathcal{M}_\delta := \{g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta_0}} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \leqslant \delta, \boldsymbol{\theta} \in \Theta\}$ has envelope $\mathcal{F}_\delta$ such that for $\varepsilon > 0$, $\mathbb{E}[\mathcal{F}_\delta^{2+\varepsilon}] \leqslant$ constant $\times\, \delta^{2+\varepsilon}$ for all $\delta > 0$. Further the class $\mathcal{M}_\delta$ satisfies the uniform entropy condition: $J(1, \mathcal{M}_\delta) \leqslant$ constant for all $\delta > 0$, where the constants are independent of δ. Here the entropy integral $J(1, \mathcal{M}_\delta) = \int_0^1 N(\varepsilon'\delta, \mathcal{F}_\delta, \|\cdot\|)d\varepsilon'$, where $N(\varepsilon'\delta, \mathcal{F}_\delta, \|\cdot\|)$ is the covering number for the set $\mathcal{F}_\delta$ using balls of size $\delta\varepsilon'$.

(i) follows immediately from assumption (A1). For showing (ii) note that

$$
\begin{aligned}
g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} =& d^2(y, m_\oplus(\mathbf{x}^\top \boldsymbol{\theta})) - d^2(y, m_\oplus(\mathbf{x}^\top \boldsymbol{\theta}_0)) \\
\leqslant& 2\mathrm{diam}(\Omega) d(m_\oplus(\mathbf{x}^\top \boldsymbol{\theta}), m_\oplus(\mathbf{x}^\top \boldsymbol{\theta}_0)) \\
\leqslant& 2\mathrm{diam}(\Omega) L \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|.
\end{aligned} \tag{S.30}
$$

The last inequality follows using the assumption (A2) on $m_\oplus$ being Lipschitz continuous with the Lipschitz constant L and assumption (A3) on $\mathbf{X}$ having a bounded support. Thus the function class $\mathcal{M}_\delta$ has the envelope $\mathcal{F}_\delta = 2\mathrm{diam}(\Omega) L\delta$ with $\mathbb{E}[\mathcal{F}_\delta^{2+\varepsilon}] = \text{constant} \times \delta^{2+\varepsilon}$, since Ω is totally bounded. Now, since $\Theta \subset \mathbb{R}^{p-1}$, and from (S.30) the function class $\mathcal{M}_\delta$ is a class of Lipschitz functions in $\boldsymbol{\theta} \in \Theta$, we have $N(\varepsilon' \delta, \mathcal{F}_\delta, \|\cdot\|) \leqslant C \left(\frac{1}{\varepsilon'}\right)^{p-1}$ [11]. As a consequence, $J(1, \mathcal{M}_\delta) = O(1)$ since $\int_0^1 \log\left(\frac{1}{\varepsilon'}\right) d\varepsilon' < \infty$. Thus the result follows. □

PROOF OF LEMMA 1. First recall that, for any given unit direction $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, $\mathcal{T}_{\bar{\boldsymbol{\theta}}}$ denotes the support of the random variable $T = \mathbf{X}^\top \bar{\boldsymbol{\theta}}$, where $\bar{\Theta}$ is defined in equation (2.5) in Section 2. For bounded random variables $\mathbf{X}$, we can write $\mathcal{T}_{\bar{\boldsymbol{\theta}}} \subset \mathcal{T}$ for some bounded $\mathcal{T} \subset \mathbb{R}$. Since all possible values of $\mathbf{x}^\top \bar{\boldsymbol{\theta}}$ is contained in $\mathcal{T}$ uniformly for $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, one has

$$
\begin{aligned}
&\sup_{\bar{\boldsymbol{\theta}} \in \bar{\Theta}} \sup_{t \in \mathcal{T}_{\bar{\boldsymbol{\theta}}}} d(\hat{m}_\oplus(t, \bar{\boldsymbol{\theta}}), m_\oplus(t, \bar{\boldsymbol{\theta}})) \\
\leqslant& \sup_{t \in \mathcal{T}} d(\hat{m}_\oplus(t, \bar{\boldsymbol{\theta}}), m_\oplus(t, \bar{\boldsymbol{\theta}})) = O_P(a_n),
\end{aligned}
$$

where a_n is the appropriate sequence described in assumption (A4). The last result follows from Theorem 1 of [1] using the technical assumptions (U1)- (U3) and (R1)- (R2). □

S.2. Technical assumptions (U1)- (U3), (R1)- (R2) .

In this section, we describe the technical assumptions needed to establish the uniform rate of convergence for the local linear Fréchet regression estimator in Lemma 1 in Section 3 of the main manuscript. We also provide motivation and discuss suitable examples regarding the assumptions.

The assumptions required to obtain the technical results are essentially the same as those used before in the Fréchet regression literature, specifically in [1]. To adapt these assumptions to the present situation, we require the curvature and entropy conditions to hold uniformly across all index values and direction parameters. The curvature and entropy conditions can be verified for commonly observed objects such as univariate probability distributions, positive definite matrices, or data on the surface of a sphere, as well as other random objects under suitable metrics.

Denote by $\mathcal{T}_{\bar{\boldsymbol{\theta}}}$ the support of the random variable $T = \mathbf{X}^\top \bar{\boldsymbol{\theta}}$ for any given unit direction $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, where $\bar{\Theta}$ is defined in equation (2.5) of the main manuscript. Under assumption (A3), for bounded random variables $\mathbf{X}$, we can write $\mathcal{T}_{\bar{\boldsymbol{\theta}}} \subset \mathcal{T}$ for some bounded subset $\mathcal{T}$ of $\mathbb{R}$. For a given direction $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$ such that $\mathbf{X}^\top \bar{\boldsymbol{\theta}} = t$, where $\bar{\Theta}$ is as given in equation (2.5), the conditional Fréchet mean is given by

$$
m_\oplus(t, \bar{\boldsymbol{\theta}}) = \operatorname*{argmin}_{\omega \in \Omega} M(\omega, t, \bar{\boldsymbol{\theta}}); \quad M(\omega, t, \bar{\boldsymbol{\theta}}) := \mathbb{E}(d^2(Y, \omega) | \mathbf{X}^\top \bar{\boldsymbol{\theta}} = t), \tag{S.31}
$$

and the local linear Fréchet regression estimate by

$$
\hat{m}_\oplus(t, \bar{\boldsymbol{\theta}}) = \operatorname*{argmin}_{\omega \in \Omega} \hat{L}_n(\omega, t, \bar{\boldsymbol{\theta}}); \quad \hat{L}_n(\omega, t, \bar{\boldsymbol{\theta}}) := \frac{1}{n} \sum_{i=1}^n \widehat{S}(\mathbf{X}_i^\top \bar{\boldsymbol{\theta}}, t, b) d^2(Y_i, \omega)), \tag{S.32}
$$

where $\widehat{S}$ is the empirical estimate (from equation (2.10)) of the nonparametric weight function (described in equation (2.8)) in Section 2 of the main manuscript and b is the bandwidth

parameter for the kernel involved in the localized Fréchet mean. We also define the intermediate localized weighted Fréchet means as

$$\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}) = \operatorname*{argmin}_{\omega\in\Omega} \tilde{L}_b(\omega,t,\bar{\boldsymbol{\theta}}); \quad \tilde{L}_b(\omega,t,\bar{\boldsymbol{\theta}}) := \mathbb{E}(S(\mathbf{X}^\top\bar{\boldsymbol{\theta}},t,b)d^2(Y,\omega)), \tag{S.33}$$

where the nonparametric weight function is described in equation (2.8) in the main manuscript. The following additional assumptions are required, which are analogous versions of the assumptions in [1].

(U1) For all $t\in\mathcal{T}$ and $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, the minimizers $m_{\oplus}(t,\bar{\boldsymbol{\theta}})$, $\hat{m}_{\oplus}(t,\bar{\boldsymbol{\theta}})$, and $\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}})$ exist and are unique, the latter two almost surely. In addition, for any $\varepsilon>0$,

$$\begin{aligned} &\inf_{t\in\mathcal{T}} \inf_{d(m_{\oplus}(t,\bar{\boldsymbol{\theta}}),\omega)>\varepsilon} [M(\omega,t,\bar{\boldsymbol{\theta}}) - M(m_{\oplus}(t,\bar{\boldsymbol{\theta}}),t,\bar{\boldsymbol{\theta}})] > 0, \\ &\liminf_{b\to 0} \inf_{t\in\mathcal{T}} \inf_{d(\omega,\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}))>\varepsilon} [\tilde{L}_b(\omega,t,\bar{\boldsymbol{\theta}}) - \tilde{L}_b(\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}),t,\bar{\boldsymbol{\theta}})] > 0, \end{aligned} \tag{S.34}$$

and there exists $c=c(\varepsilon)>0$ such that

$$P\left(\inf_{t\in\mathcal{T}} \inf_{d(\hat{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}),\omega)>\varepsilon} [\hat{L}_n(\omega,t,\bar{\boldsymbol{\theta}}) - \hat{L}_n(\hat{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}),t,\bar{\boldsymbol{\theta}})] \geqslant c\right) \to 1. \tag{S.35}$$

(U2) Let $\mathcal{B}_r(m_{\oplus}(t,\bar{\boldsymbol{\theta}}))\subset\Omega$ be a ball of radius r centered at $m_{\oplus}(t,\bar{\boldsymbol{\theta}})$ and $\mathcal{N}(\varepsilon,\mathcal{B}_r(m_{\oplus}(t,\bar{\boldsymbol{\theta}})),d)$ be its covering number using balls of radius ϵ. Then

$$\lim_{r\to 0+} \int_0^1 \sup_{t\in\mathcal{T}} \sqrt{1+\log\mathcal{N}(r\varepsilon,\mathcal{B}_r(m_{\oplus}(t,\bar{\boldsymbol{\theta}})),d)}d\epsilon = O(1). \tag{S.36}$$

(U3) There exists $r_1,r_2>0$, $c_1,c_2>0$, and $\beta_1,\beta_2>1$ such that

$$\begin{aligned} &\inf_{t\in\mathcal{T}} \inf_{d(m_{\oplus}(t,\bar{\boldsymbol{\theta}}),\omega)<r_1} [M(\omega,t,\bar{\boldsymbol{\theta}}) - M(m_{\oplus}(t,\bar{\boldsymbol{\theta}}),t,\bar{\boldsymbol{\theta}}) - c_1 d^2(\omega,m_{\oplus}(t,\bar{\boldsymbol{\theta}}))^{\beta_1}] \geqslant 0, \\ &\liminf_{b\to 0} \inf_{t\in\mathcal{T}} \inf_{d(\omega,\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}))<r_2} [\tilde{L}_b(\omega,t,\bar{\boldsymbol{\theta}}) - \tilde{L}_b(d(\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}),t,\bar{\boldsymbol{\theta}}) - c_2 d^2(\omega,\tilde{m}_{\oplus}(t,\bar{\boldsymbol{\theta}}))^{\beta_2}] \geqslant 0. \end{aligned} \tag{S.37}$$

Furthermore, we require the following assumptions for kernels and distributions.

(R1) The kernel K is a probability density function, symmetric around zero, uniformly continuous on $\mathbb{R}$ such that $\int_{\mathbb{R}} K(x)^j x^k<\infty$, for $j,k=1,\dots 6$. The derivative K' exists and is bounded on the support of K, i.e., $\sup_{x:K(x)>0}|K'(x)|<\infty$. Additionally, $\int_{\mathbb{R}} x^2|K'(x)|\sqrt{|x\log|x||}dx<\infty$.

(R2) For any given unit direction $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, the marginal density $f_{T,\bar{\boldsymbol{\theta}}}$ of $T=\mathbf{X}^\top\bar{\boldsymbol{\theta}}$ and the conditional densities $f_{T,\bar{\boldsymbol{\theta}}|Y}(\cdot,y)$ of T given $Y=y$ exist and are twice continuously differentiable in the interior of $\mathcal{T}$ for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, the latter for all $y\in\Omega$. The marginal density $f_{T,\boldsymbol{\theta}}$ is bounded away from zero on its support $\mathcal{T}$ for all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ i.e., $\inf_{t\in\mathcal{T}} f_{\mathbf{X}^\top\bar{\boldsymbol{\theta}}}(t)>0$.

The second-order derivative $f''_{T,\bar{\boldsymbol{\theta}}}$ is uniformly bounded for all $t\in\mathcal{T}$, $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, that is, $\sup_{t\in\mathcal{T}}|f''_{T,\bar{\boldsymbol{\theta}}}(t)|<\infty$. The second-order partial derivatives $(\partial^2 f_{T,\bar{\boldsymbol{\theta}}|Y}/\partial t^2)(\cdot,y)$ are uniformly bounded, uniform over all $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, i.e., $\sup_{t\in\mathcal{T}}\sup_{y\in\Omega}|(\partial^2 f_{T,\bar{\boldsymbol{\theta}}|Y}/\partial t^2)(\cdot,y)|<\infty$.

Additionally, for any open set $B\subset\Omega$, $P(Y\in B|\mathbf{X}^\top\boldsymbol{\theta}=t)$ is continuous as a function of t and $\bar{\boldsymbol{\theta}}$. For any $t\in\mathcal{T}$ and $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$, $M(\omega,t,\bar{\boldsymbol{\theta}})$ is equicontinuous, that is,

$$\limsup_{\bar{\boldsymbol{\theta}}_1\to\bar{\boldsymbol{\theta}}_2} \sup_{t\in\mathcal{T}} \sup_{\omega\in\Omega} \left|M(\omega,t,\bar{\boldsymbol{\theta}}_1) - M(\omega,t,\bar{\boldsymbol{\theta}}_2)\right| = 0.$$

Similar yet weaker assumptions have been made in [7] for pointwise rates of convergence for local linear Fréchet regression estimators. [1] made stronger assumptions in this regard to establish uniform convergence results over univariate predictor values. In the above assumptions (U1)- (U3) we adapt those in [1], incorporating uniform bounds over the index parameter as well as over the values of the single index. Since the objective function for the local Fréchet regression involves both the index value $\mathbf{x}^\top\bar{\boldsymbol{\theta}} = t$ and the index parameter $\bar{\boldsymbol{\theta}}$, conditions on the well-separatedness, entropy, and curvature needs to be extended for all values of t and $\bar{\boldsymbol{\theta}}$. These assumptions are adapted from empirical process theory, guarantee the asymptotic uniform equicontinuity of $\tilde{L}_b$, and control the behavior of $\tilde{L}_b - M$ and $\hat{L}_n - \tilde{L}_b$ near the minimizers $m_\oplus(t, \bar{\boldsymbol{\theta}})$ and $\tilde{m}_\oplus(t, \bar{\boldsymbol{\theta}})$, respectively, uniformly over t and $\bar{\boldsymbol{\theta}}$. assumption (U1) is commonly used to establish the uniform consistency of M-estimators [11] by showing the weak convergence of the respective empirical processes. In conjunction with the assumption that the metric space Ω is totally bounded, this implies the pointwise convergence of the minimizers for any given t and $\bar{\boldsymbol{\theta}}$; it also ensures that the asymptotic uniform equicontinuity of $\tilde{L}_b$ and $\hat{L}_n$, and implies the (asymptotic) uniform equicontinuity of $\tilde{m}_\oplus$ and $\hat{m}_\oplus$, whence the uniform convergence of the minimizers follows as the support of $\mathbf{x}^\top\bar{\boldsymbol{\theta}}$ is compact for any $\bar{\boldsymbol{\theta}}$.

Assumptions (U1)- (U3) are easily verified for specific metric space-valued objects.

Example 1 Let Ω be the set of probability distributions on a closed interval of $\mathbb{R}$ with finite second moments, endowed with the Wasserstein-2 distance d_W, i.e., for any two distributional objects Y_1 and Y_2 with cdfs F_{Y_1} and F_{Y_2} respectively,

$$d_W(Y_1, Y_2) = \int_0^1 (F_{Y_1}^{-1}(z) - F_{Y_2}^{-1}(z))^2 dz,$$

where $F_{Y_j}^{-1}(z)$ is the quantile function for Y_j, $j = 1, 2$. The Wasserstein space (Ω, d_W) satisfies assumptions (U1)- (U3) with $\beta_1 = \beta_2 = 2$.

Example 2 Let Ω be the space of r-dimensional correlation matrices, i.e., symmetric, positive semidefinite matrices in $\mathbb{R}^{r\times r}$ with diagonal elements equal to 1, endowed with the Frobenius metric d_F . Specifically for any two elements $Y_1, Y_2 \in \Omega$,

$$d_F(Y_1, Y_2) = \sqrt{\text{trace}((Y_1 - Y_2)^\intercal(Y_1 - Y_2))}.$$

The space (Ω, d_F) satisfies assumptions (U1)- (U3) with $\beta_1 = \beta_2 = 2$.

For Examples 1-2, we note that since the Wasserstein space for one-dimensional distributions and the space of correlation matrices are Hadamard spaces, there exists a unique minimizer of $M(\cdot, t, \bar{\boldsymbol{\theta}})$ for any $t \in \mathcal{T}$ and $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$ [10]. Examples 1-2 follow from similar arguments as those in the proofs of Propositions 1-2 of [7] by observing that the arguments hold uniformly across t and $\bar{\boldsymbol{\theta}}$. Assumptions (R1) and (R2) are standard distributional assumptions for local nonparametric regression and are needed to show the convergence of the bias and stochastic parts for the local linear Fréchet estimator uniformly over all t and $\bar{\boldsymbol{\theta}}$. In particular, Assumption (R1) can be verified for a general class of kernel functions given by

$$c_\kappa(1 - x^2)^\kappa \mathbb{I}([-1, 1]), \ \kappa \in \mathcal{Z},$$

where $c_\kappa = \frac{\Gamma(k+\frac{3}{2})}{\sqrt{\pi}\Gamma(k+1)}$ is such that $\int_{-1}^1 c_\kappa(1 - x^2)^\kappa dx = 1$ and the indicator function is defined as $\mathbb{I}(A) = 1$ if $\mathbf{X} \in A$, and 0 otherwise. The Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}([-1, 1])$ belongs to this class of kernel functions for $\kappa = 1$ with $c_\kappa = 3/4$.

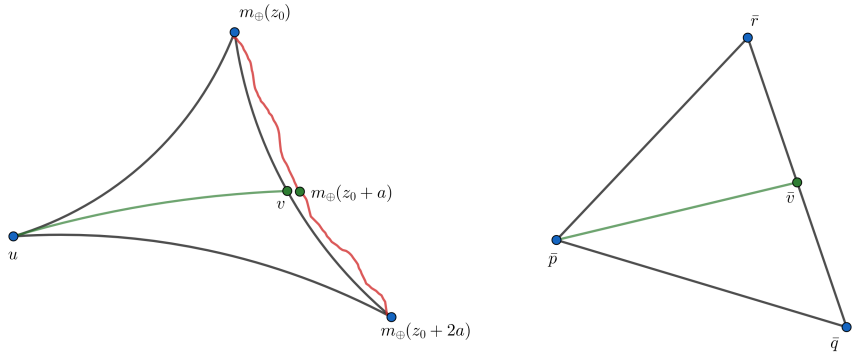

Fig 1: The left figure shows the geodesic triangle formed by the three points u, $m_\oplus(z_0)$, $m_\oplus(z_0+2a)$, where v is the midpoint of the geodesic connecting the points $m_\oplus(z_0)$ and $m_\oplus(z_0+2a)$. The red line depicts the true regression function $m_\oplus$. $m_\oplus(z_0+a)$ is closely approximated by v lying on a geodesic that connects $m_\oplus(z_0)$ with $m_\oplus(z_0+2a)$. The right hand side shows the reference triangle in $\mathbb{R}^2$ as an illustration of the CAT(0) inequality.

S.3. Further discussion of assumption (A5).

Assumption (A5) in Section 3 of the main manuscript intuitively means that $m_\oplus$ can be locally approximated by straight lines in Euclidean space and geodesics in geodesic spaces. In the Euclidean case, it is satisfied for twice differentiable functions $m_\oplus$, a common assumption for classical single index modeling. Beyond the Euclidean special case, assumption (A5)

Consider first the Euclidean case, where Ω is a compact subset $\mathcal{M} \subset \mathbb{R}$ and denote the link function $m_\oplus$ by m. Noting that the map $h : \boldsymbol{\theta} \mapsto \bar{\boldsymbol{\theta}}$ is continuous, and $m_\oplus(\mathbf{z}^\top\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}) := \phi(\bar{\boldsymbol{\theta}}) = \phi(h(\boldsymbol{\theta}))$, for some function ϕ of $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$ and for any given $\mathbf{z} \in \mathcal{X} \subset \mathbb{R}^p$, with a slight abuse of notation, we write $m_\oplus(\mathbf{z}^\top\boldsymbol{\theta}, \boldsymbol{\theta})$ instead of $m_\oplus(\mathbf{z}^\top\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})$. For any given $z \in \mathcal{X} \subset \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Theta$ such that $\boldsymbol{\theta}^\top\boldsymbol{\theta} < 1$, denote $m\left(z^\top\boldsymbol{\theta}, \boldsymbol{\theta}\right) = m(z_0, \boldsymbol{\theta})$ by $m(z_0)$, where $z_0 = z^\top\boldsymbol{\theta} \in \mathbb{R}$ and for a small enough $a \in (0, a_0)$, such that $z_0, z_0 + 2a \in \mathcal{T}$, we have $m(z_0), m(z_0 + a), m(z_0 + 2a) \in \mathcal{M}$. If $m(\cdot)$ is twice continuously differentiable in any open subset containing z_0 such that the derivatives are uniformly bounded, the midpoint on the straight line (geodesic path) connecting $m(z_0)$ and $m(z_0 + 2a)$ is given by $v = \frac{1}{2}[m(z_0) + m(z_0 + 2a)]$. Using a second-order Taylor expansion for the function m around z_0, we have

$$
\begin{aligned}
&\|v - m_\oplus(z_0 + a)\|_E \\
=&\|\frac{1}{2}[m(z_0) + m(z_0 + 2a)] - m_\oplus(z_0 + a)\|_E \\
=&\|[\frac{1}{2}m(z_0) + \frac{1}{2}m(z_0) + am'(z_0) + \frac{1}{2}\frac{(2a)^2}{2}m''(\zeta_1)] - [m(z_0) + am'(z_0) + \frac{a^2}{2!}m''(\zeta_2)]\|_E \\
=&\|a^2[m''(\zeta_1) - \frac{1}{2}m''(\zeta_2)]\|_E,
\end{aligned}
$$

where $z_0 < \zeta_1 < z_0 + 2a$, and $z_0 < \zeta_2 < z_0 + a$. Assuming a uniform bound on the second derivative of m, such that $|m''(z)| \leqslant C$ for some $C > 0$ and for all $z \in \mathcal{T}$, we have that $\|v - m_\oplus(z_0 + a)\|_E \leqslant \frac{3C}{2}a^2$. Thus, assumption (K2) holds for $C_* = 3C/2$, as long as the bound C on the second derivative of m is sufficiently small.

Next, we consider Ω to be the space of univariate distributions, $\mathcal{F}$, endowed with the Wasserstein-2 metric d_W. The quantile functions for the distributional objects $m_\oplus(z_0)$, $m_\oplus(z_0 + a)$, and $m_\oplus(z_0 + 2a)$ are denoted by $Q(m_\oplus(z_0))(\cdot)$, $Q(m_\oplus(z_0 + a))(\cdot)$, and $Q(m_\oplus(z_0 + 2a))(\cdot)$, respectively. Similarly, the quantile function of the midpoint v of the

geodesic path connecting $m_\oplus(z_0)$ and $m_\oplus(z_0+2a)$ is given by

$$Q(v)(\cdot) = \frac{1}{2}[Q(m_\oplus(z_0))(\cdot) + Q(m_\oplus(z_0+2a))(\cdot)].$$

We write $q(z_0)(\cdot) = Q(m_\oplus(z_0))(\cdot) = q(z_0)(\cdot)$, analogously for related quantities. The Wasserstein distance between v and $m_\oplus(z_0+a)$ is then given by

$$\begin{aligned} d_W^2(v, m_\oplus(z_0+a)) &= \int_0^1 \left(Q(v)(t) - Q(m_\oplus(z_0+2a))(t)\right)^2 dt \\ &= \int_0^1 \left(\frac{q(z_0)(t) + q(z_0+2a)(t)}{2} - q(z_0+2a)(t)\right)^2 dt \end{aligned}$$

We assume that for every $t \in [0,1]$, $q(z)(t)$ is twice continuously differentiable as a function of z, for any z in an open subset containing z_0 such that derivatives of $q(z)(t)$ are uniformly bounded for each $t \in [0,1]$. Using a second-order Taylor expansion of $q(\cdot)(t)$ pointwise $t \in [0,1]$, and following a similar argument as in the Euclidean case, we have

$$d_W^2(v, m_\oplus(z_0+a)) = \int_0^1 \left(a^2[q''(\zeta_1)(t) - \frac{1}{2}q''(\zeta_2)(t)]\right)^2 dt,$$

Lastly, under the assumption that the $|q''(z)(t)| \leqslant r(t)$, such that $\int_0^1 r^2(t) < C$, assumption (K2) holds for $C_* = 3/2C$, as long as the bound C is sufficiently small.

We further illustrate the argument for assumption (K2) for distributional objects in the specific context of a location-scale family of univariate distributions, $\mathcal{F}$, endowed with the Wasserstein-2 metric d_W. Denoting the location and scale parameters as $\mu(\cdot)$ and $\sigma(\cdot)$ respectively, the quantile function corresponding to the distribution object $m_\oplus(z_0) \in \mathcal{F}$ will be given by

$$Q(m_\oplus(z_0))(\cdot) = \mu(z_0) + \sigma(z_0)F^{-1}(\cdot),$$

where $F^{-1}(\cdot)$ is the quantile function for the distribution object $m_\oplus(z_0)$. The quantile functions for $m_\oplus(z_0+a)$ and $m_\oplus(z_0+2a)$ can be similarly defined. Also, the quantile function of the midpoint of the geodesic path connecting $m_\oplus(z_0)$ and $m_\oplus(z_0+2a)$ is given by

$$Q(v)(\cdot) = \frac{1}{2}[\mu(z_0) + \mu(z_0+2a)] + \frac{1}{2}[\sigma(z_0) + \sigma(z_0+2a)]F^{-1}(\cdot).$$

The Wasserstein distance between v and $m_\oplus(z_0+a)$ is given by

$$\begin{aligned} d_W^2(v, m_\oplus(z_0+a)) &= \left|\frac{\mu(z_0) + \mu(z_0+2a)}{2} - \mu(z_0+a)\right|^2 \\ &\quad + \left|\frac{\sigma(z_0) + \sigma(z_0+2a)}{2} + \sigma(z_0+a) - 2\left(\frac{\sigma(z_0) + \sigma(z_0+2a)}{2}\sigma(z_0+a)\right)^{1/2}\right|^2 \\ &\leqslant \left|\frac{\mu(z_0) + \mu(z_0+2a)}{2} - \mu(z_0+a)\right|^2 + \left|\frac{\sigma(z_0) + \sigma(z_0+2a)}{2} - \sigma(z_0+a)\right|^2, \end{aligned}$$

where the last inequality holds because $\frac{1}{2}\sigma(z_0) + \sigma(z_0+2a)$ and $\sigma(z_0+a)$ are both positive. Assuming $\mu(\cdot)$ and $\sigma(\cdot)$ are twice continuously differentiable in any open subset containing z_0 such that their derivatives are uniformly bounded, the result follows in a similar manner to the Euclidean case.

We next show that assumption (A5) holds under the sufficient conditions (K1), (K2),and (K3), that is, for any $u \in \Omega$, and $z_0 \in \mathcal{T}$, there exists some $\kappa > 0$, such that, for any small $a > 0$,

$$\frac{1}{a^2}[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, m_\oplus(z_0 + a)) + d^2(u, m_\oplus(z_0))] \geqslant \kappa \tag{S.38}$$

Observe that

$$\begin{aligned} &\frac{1}{a^2}[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, m_\oplus(z_0 + a)) + d^2(u, m_\oplus(z_0))] \\ =&\frac{1}{a^2}[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, v) + d^2(u, m_\oplus(z_0))] \\ +&\frac{1}{a^2}[2d^2(u, v) - 2d^2(u, m_\oplus(z_0 + a))]. \end{aligned} \tag{S.39}$$

Assumption (K3) in conjunction with assumption (A2) implies that $m_\oplus$ is bi-Lipschitz with constants $0 \leqslant L_* \leqslant L$. We have

$$2aL_* \leqslant d(m_\oplus(z_0 + 2a), m_\oplus(z_0)) \leqslant 2La. \tag{S.40}$$

Thus the first term of (S.39) becomes

$$\begin{aligned} &\frac{1}{a^2}\left[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, v) + d^2(u, m_\oplus(z_0))\right] \\ \geqslant &\frac{4L_*^2}{d^2(m_\oplus(z_0 + 2a), m_\oplus(z_0))}\left[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, v) + d^2(u, m_\oplus(z_0))\right], \end{aligned} \tag{S.41}$$

where this inequality follows from assumptions (A2), using (S.40). Assuming Ω is a geodesic CAT(0) space, the geodesic triangle $\triangle(u, m_\oplus(z_0), m_\oplus(z_0 + 2a))$, formed by the vertices u, $m_\oplus(z_0)$, and $m_\oplus(z_0 + 2a)$, will have a comparison triangle $\triangle(\bar{p}, \bar{q}, \bar{r})$ in the reference space $\mathbb{R}^2$ for some points $\bar{p}, \bar{q}, \bar{r} \in \mathbb{R}^2$. This implies

$$\begin{aligned} d(u, m_\oplus(z_0)) = ||\bar{p} - \bar{q}||_E, \quad d(u, m_\oplus(z_0 + 2a)) = ||\bar{p} - \bar{r}||_E, \\ d(m_\oplus(z_0), v) = ||\bar{q} - \bar{v}||_E, \quad d(m_\oplus(z_0 + 2a), v) = ||\bar{r} - \bar{v}||_E. \end{aligned} \tag{S.42}$$

By virtue of assumption (K1),

$$d(u, v) \leqslant ||\bar{p} - \bar{v}||_E. \tag{S.43}$$

Thus combining (S.41)– (S.43) one obtains

$$\begin{aligned} &\frac{1}{a^2}\left[d^2(u, m_\oplus(z_0 + 2a)) - 2d^2(u, v) + d^2(u, m_\oplus(z_0))\right] \\ \geqslant &2L_*^2 \frac{\frac{\|\bar{p}-\bar{r}\|_E^2 - \|\bar{p}-\bar{v}\|_E^2}{||\bar{r}-\bar{v}||_E} - \frac{\|\bar{p}-\bar{v}\|_E^2 - \|\bar{p}-\bar{q}\|_E^2}{||\bar{q}-\bar{v}||_E}}{||\bar{r} - \bar{q}||_E} = 2L_*^2 > 0. \end{aligned} \tag{S.44}$$

This uses the fact that $\bar{r}$, $\bar{v}$, $\bar{q}$ are co-linear in the Euclidean space with $\bar{v}$ being the midpoint between $\bar{r}$ and $\bar{q}$, and hence the second order difference is just 1. Thus the first term of (S.39) is seen to be greater than or equal to $2L_*^2$.

As for the second term of (S.39), by simple algebra and the triangle inequality,

$$\begin{aligned} &\left|\frac{2}{a^2}[d^2(u, v) - d^2(u, m_\oplus(z_0 + a))]\right| \\ =&\frac{2}{a^2}\left|(d(u, v) + d(u, m_\oplus(z_0 + a)))\right| \left|(d(u, v) - d(u, m_\oplus(z_0 + a)))\right| \\ \leqslant&\frac{4D}{a^2} d(v, m_\oplus(z_0 + a)) \leqslant 4DC_*. \end{aligned} \tag{S.45}$$

The last inequality follows from equation (B.2) in assumption (K2). In assumption (K2), given L and D, C_* can be chosen sufficiently small such that $2L_*^2 > 4DC_*$. Thus, combining (S.44) and (S.45) with (S.39), the result follows for $\kappa = 2L_*^2 - 4DC_* > 0$.

S.4. Additional data illustrations and simulations.

This section provides further illustrations of data applications and simulations. Random objects considered in the additional data demonstrations discussed in this section are univariate probability distributions with compact support endowed with the Wasserstein-2 metric (applied to human mortality data) and compositional data that are mapped to the positive segment of a sphere, endowed with the geodesic distance and applied to the mood compositional data. Further illustrations of the proposed method include an additional plot for the ADNI study and a simulation study with Euclidean responses.

S.4.1. *Human mortality and age-at-death distributional object responses.*

The performance of the proposed model is demonstrated with an application to human mortality data across countries. We view the age-at-death distributions as random object responses of interest and aim to find their association with Euclidean predictors such as economic, social, and healthcare indices among other relevant factors, aiming at a comprehensive understanding of human longevity and health conditions.

For this analysis, we used the lifetables for males aggregated yearly in age groups varying from age 0 to 110 for 40 countries in the calendar year 2010. The data consist of period lifetables for each country and each calendar year and were obtained from the Human Mortality Database (https://www.mortality.org/). We computed histograms of age-at-death from the lifetables for each country and calendar year, which were then smoothed with local least squares to obtain smooth estimated probability density functions for age-at-death using the R package frechet [2]. After this preprocessing step, the data are a sample of univariate probability distributions for $n = 40$ countries was obtained, shown in the left panel of Figure 2. We equipped the sample of age-at-death distributions with the Wasserstein-2 metric (Ω, d_W) and selected the following six socio-economic predictors measured at the calendar year 2010: $X_1 =$ Population density (people per sq. km of land area), $X_2 =$ Fertility rate, total (births per woman), $X_3 =$ GDP per capita, at Purchasing Power Parity (PPP), $X_4 =$ Access to electricity (% of the population), $X_5 =$ Current health expenditure (% of GDP), and $X_6 =$ Unemployment, total (% of the total labor force) (national estimate). The data were obtained from the World Bank Database at https://data.worldbank.org.

We first standardized all predictors separately, then applied the proposed Index Fréchet Regression (IFR) method to obtain the estimated unit direction parameter (rounded to 4 decimal places)

$$\widehat{\widehat{\boldsymbol{\theta}}} = (0.0173, 0.7875, 0.5879, 0.0167, 0.1646, -0.0807)^\intercal.$$

The estimated coefficient for the predictor Fertility Rate (X_2) has the highest absolute value, indicating its heavy influence relative to the other five predictors on the index $\mathbf{X}^\top \widehat{\widehat{\boldsymbol{\theta}}}$, and hence on the fitted value for the IFR model. The estimated index $\mathbf{X}^\top \widehat{\widehat{\boldsymbol{\theta}}}$ can be also perceived as the first sufficient predictor, which reduces the dimension of the predictor space without losing the information about the response. This aligns with the sufficient dimension reduction methods for Fréchet regression [12] and provides an insight into the overall dependence of the predictors on the object response.

In the right panel of Figure 2, the age-at-death densities are plotted against the estimated index values, aka the first sufficient predictors, $\mathbf{X}^\top \widehat{\widehat{\boldsymbol{\theta}}}$. It is evident that countries with low

index values have modes of the distribution at lower ages, while for countries with high values of the index, the modes of mortality distributions are significantly higher. Further, the countries with higher index values indicate very low infant mortality rates.

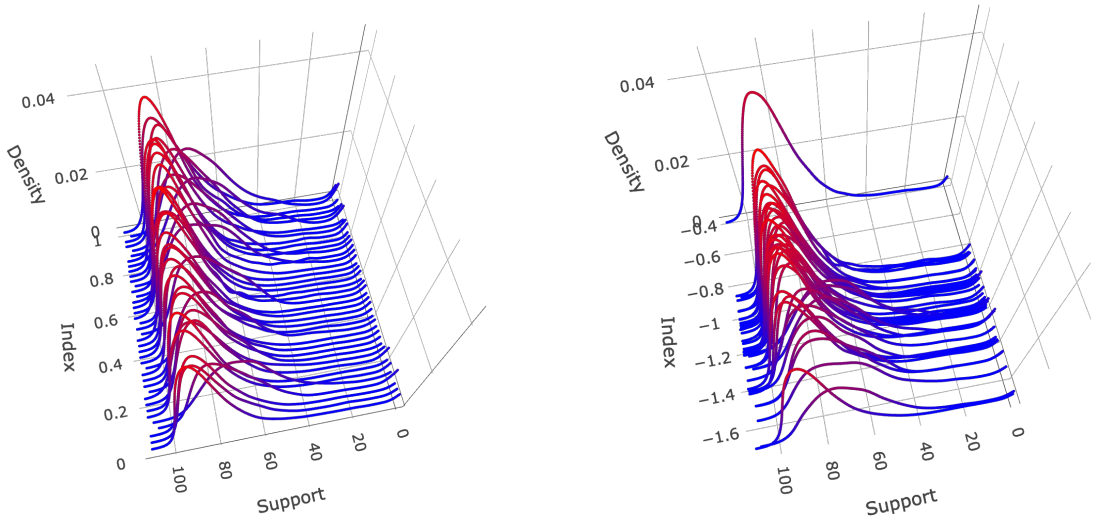

Fig 2: Data visualization for age-at-death densities for 40 countries at the calendar year 2010. The left panel shows the observed densities at random order while the right panel plots the observed densities against the estimated index values from the proposed Index Fréchet Regression (IFR) model.

The plots of the observed and estimated age-at-death densities over the support of age $[0, 110]$ and against the estimated index values, aka the first estimated sufficient predictor, are shown in Figure 3. It is interesting to observe that the estimated index values are associated with the location and variation features of the age-at-death distributions. Specifically, with the increase in the values of the index, the mean of the mortality distribution increases nonlinearly while the standard deviation diminishes, indicating the death age more concentrates between 70 and 80. This finding is in line with the observations of [12], who employed several sufficient dimension reduction (SDR) techniques to the mortality distributions.

Further, the importance of various predictors can be inferred from the estimated coefficients $\widehat{\widetilde{\boldsymbol{\theta}}}$. As before we keep the first predictor ($X_1 =$ Population density) with the corresponding coefficient $\hat{\theta}_1 = 0.0173 > 0$ in the model and test for the following hypothesis: $H_0 : \theta_{02} = \cdots = \theta_{0p} = 0$ vs. H_1, the complement of H_0, which is the test for overall regression effect for object responses. Writing $\hat{\boldsymbol{\theta}} = (\hat{\theta}_2, \ldots, \hat{\theta}_6)$, the test statistic is constructed as $\tilde{T}_n = \hat{\boldsymbol{\theta}}^\top (\widehat{\Lambda}_B^*)^{-1} \hat{\boldsymbol{\theta}} \overset{approx.}{\sim} \chi_5^2$ under H_0 (see Section 5.1), where $\widehat{\Lambda}_B^*$ is the bootstrap estimator for asymptotic covariance matrix as described in Proposition 5. The null hypothesis is rejected at level α if $\tilde{T}_n > \chi_5^2(1-\alpha)$. From our analysis, $\tilde{T}_n = 18.883 > 11.0705 = \chi_5^2(1-\alpha)$ for the level $\alpha = 0.05$. The p-value is actually 0.002 and the null hypothesis is thus clearly rejected, demonstrating there is a regression effect. Upon further analysis it is found that the most significant predictors, in order, are $X_2 =$ Fertility rate, total (births per woman), $X_3 =$ GDP per capita, at Purchasing Power Parity (PPP), and $X_5 =$ Current health expenditure (% of GDP).

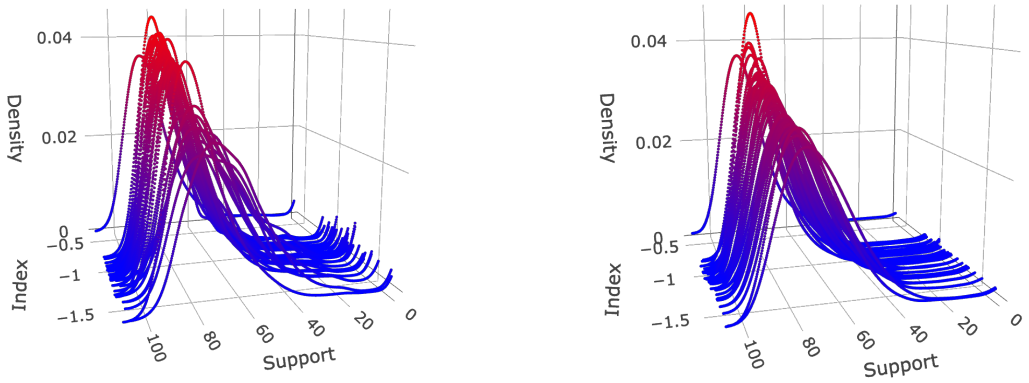

Fig 3: The observed and estimated age-at-death distributions for 40 countries at the calendar year 2010 are displayed in the left and right panel of figure, respectively. The distributions are plotted over the support of the age interval $[0, 110]$ against the index values estimated by the IFR model.

We proceed to compare fits for the year 2010 from the IFR model with the Global Fréchet Regression (GFR) model with the $6-$dimensional predictors, as well as with three separate Local Fréchet Regression (LFR) models, where the three important predictors Fertility Rate, GDP per capita and Health Expenditure are considered in each LFR model separably as univariate predictors. The global Fréchet model suffers from model-induced bias, while the local linear Fréchet Regression models with individual univariate predictors lack relevant information from other variables. The IFR model is a semiparametric approach that combines the strengths of both of these models. Figure 4 displays the observed as well as the fitted distributions (as densities) for these five models. The superiority of the IFR model compared to the local linear Fréchet fits, using only the relatively important predictor variables individually indicates that all predictors simultaneously play an important role in the overall prediction through the estimated index $\mathbf{x}^\top \widehat{\widehat{\boldsymbol{\theta}}}$. To study the effect of the most important predictors, GDP per capita, fertility rate, and Health expenditure percentage on the age-of-death densities, we fitted the IFR model when varying the value of one predictor, while keeping the other two fixed at their mean levels. For example, the left-most panel of Figure 5 illustrates how the age-at-death density changes with increasing levels of GDP per capita, while the other two predictors are kept fixed. The fitted densities are color coded such that blue to red indicates a smaller to a larger value of GDP. We find that smaller values of GDP are associated with left-shifted age-at-death distributions for the population. For increasing levels of health expenditure per capita and fertility rates, the age-at-death densities also shift rightwards, but to a lesser extent.

Finally, to illustrate the out-of-sample prediction performance of the proposed IFR model, we randomly split the dataset into a training set with sample size $n_{\text{train}} = 20$ and a test set with the remaining $n_{\text{test}} = 20$ subjects. The IFR method was implemented as follows: For any given unit direction $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, we partition the domain of the projections into M equal-width non-overlapping bins and consider the representative observations $\tilde{\mathbf{X}}_l$ and $\tilde{Y}_l$ for the data points belonging to the $l-$th bin. The "true" index parameter is estimated as $\widehat{\widehat{\boldsymbol{\theta}}}$ as per equation (2.11). We then take the fitted objects obtained from the training set and predict the responses in the test set using the covariates present in the test set. As a measure of the

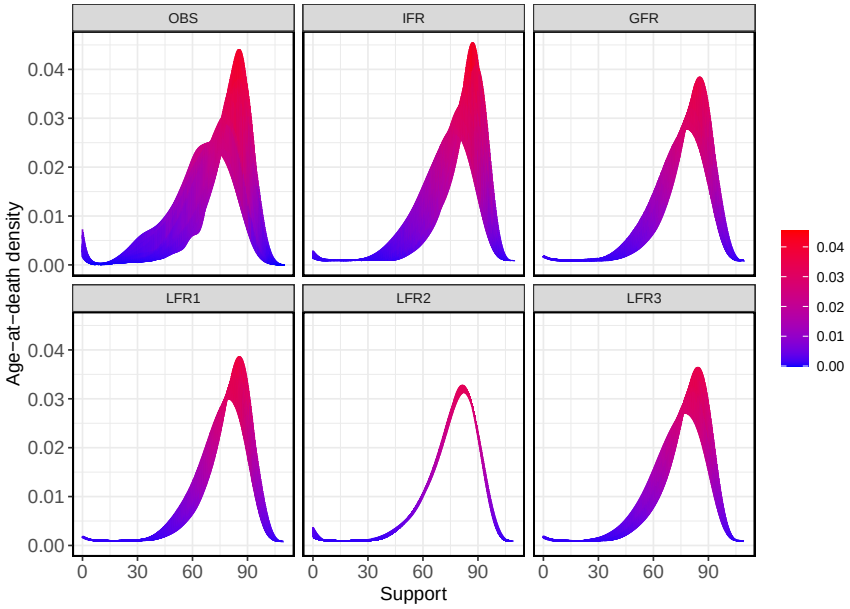

Fig 4: Figure displaying the observed and predicted smooth densities. Clockwise, from top-left the observed densities (OBS), the fitted densities using Index Frechet Regression (IFR), Global Fréchet Regression (GFR), and Local Fréchet Regression (LFR). The predictors used for the LFR fits are Fertility Rate (LFR1), GDP per capita (LFR2) and Health Expenditures (LFR3), respectively. Densities are color-coded (blue to red indicating low to high) by the mode of the age-at-death distribution.

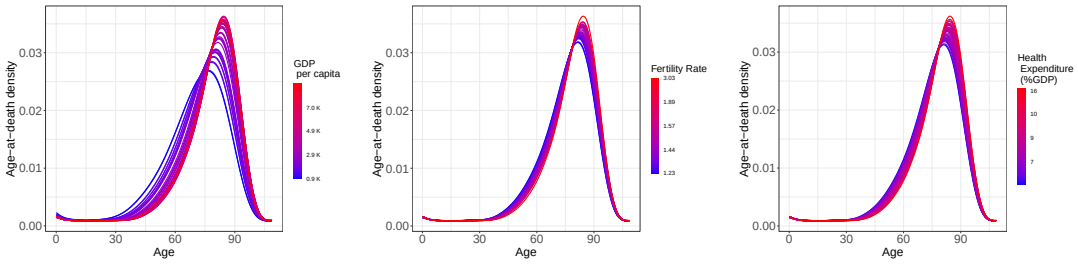

Fig 5: Figure showing the effects of the significant predictors X_3 = GDP per capita, X_2 = Fertility rate, and X_5 = Current health expenditure. The left panel shows the change in density with changing value of X_3 from low (blue) to high (red), when X_2 and X_5 are fixed at their mean level, and analogously for middle and right panels.

efficacy of the fitted model, we compute the root mean squared prediction error (RMPE) as

$$\text{RMPE} = \left[\frac{1}{M_{n_{\text{test}}}} \sum_{i=1}^{M_{n_{\text{test}}}} d_W^2 \left(\tilde{Y}_l^{\text{test}}, \hat{m}_{\oplus}(\tilde{\mathbf{X}}_l^{\text{test}\intercal} \hat{\widehat{\boldsymbol{\theta}}}, \hat{\widehat{\boldsymbol{\theta}}}) \right) \right]^{1/2}, \tag{S.46}$$

where $\tilde{Y}_l^{\text{test}}$ and $\hat{m}_{\oplus}(\tilde{\mathbf{X}}_l^{\text{test}\intercal} \hat{\widehat{\boldsymbol{\theta}}}, \hat{\widehat{\boldsymbol{\theta}}})$ denote, respectively, the l^{th} observed and predicted responses in the test set, evaluated at the binned observation $\tilde{\mathbf{X}}_l^{\text{test}}$. For any two distribution objects

$F, G \in (\Omega, d_W)$, the Wasserstein-2 distance is given by

$$d_W(F,G) = \int_0^1 (F^{-1}(s) - G^{-1}(s))^2 ds,$$

where F^{-1} and G^{-1} are the quantile functions corresponding to F and G respectively. We repeat this process 500 times, and compute RMPE for each split for the subjects separately. The mean and sd of the RMPE over the repetitions are shown in Table 1 for the IFR method, as well as for the GFR and individual LFR fits.

TABLE 1

Mean and sd (in parenthesis) of the RMPE as given in (S.46) comparing the performance of various Fréchet regression models: Index Frćhet Regression (IFR), Global Fréchet Regression (GFR), Local Fréchet Regression (LFR). The predictors used for the three individual LFR fits are Fertility Rate, GDP per capita at PPP, and Health Expenditure, respectively, as indicated in parentheses.

IFR	GFR	LFR1 (on Fertility Rate)	LFR2 (on GDP per Capita-PPP)	LFR3 (on Health Expenditure)
0.178 (0.0552)	0.287 (0.0671)	0.491 (0.0605)	0.603 (0.0654)	0.339 (0.0565)

Using out-of-sample performance, the IFR model emerges as the best model, as the average RMPE of 0.178 is much lower than that of any of the other models.

S.4.2. *Emotional well-being for unemployed workers: Compositional data as random object responses.*

We demonstrate the proposed IFR method for the analysis of mood compositional data. Compositional data are random vectors with non-negative components, where the components of these vectors sum to 1. With a square-root transformation of the components, compositional vectors can be transformed to unit vectors that lie on the positive segment of a sphere S^{p-1} if the compositional vectors are $p-$dimensional [8, 9]. Thus one can represent compositional data as manifold-valued objects that lie on the surface of a sphere. The data used for this application were collected in the Survey of Unemployed Workers in New Jersey [6] conducted in the fall of 2009 and the beginning of 2010, during which the unemployment rate in the US peaked at 10% after the financial crisis of 2007 – 2008; similar data were used to illustrate longitudinal compositional methods in [3]. We note that here the object-valued responses lie on a manifold (sphere) with positive curvature. Thus the sufficient (but not necessary) condition for assumption (A5) that the underlying metric space behaves like a CAT(0) space is not satisfied. This example thus provides a check on the behavior of IFR when the random objects are situated in a positively curved space.

Unemployed workers belonging to a stratified random sample were surveyed at entry into the study, where we analyzed the data for $n = 3301$ workers with complete measurements. A key variable in the survey was the proportion of time the workers spent in each of the four moods: bad, low/irritable, mildly pleasant, and very good while at home; we use this 4-dimensional compositional vector as the response. Formally, the composition measurement of interest is $Z = (Z_1, Z_2, Z_3, Z_4)^\intercal$, where Z_j is the proportion of time a worker spent in the j-th mood when at home, $j = 1, \ldots, 4$. The square-root transformed compositional data

$$Y = (Y_1, Y_2, Y_3, Y_4)^\intercal = (\sqrt{Z}_1, \sqrt{Z}_2, \sqrt{Z}_3, \sqrt{Z}_4)^\top,$$

lie on the sphere $\mathcal{S}^3$. We adopt the geodesic metric on this sphere $d_g(y, y^*) = \arccos(y^\top y^*)$.

These square root transformed compositional data are treated as the object responses in a regression model with the following 10 baseline predictors obtained from the questionnaire, reflecting various socio-economic and demographic information: (1) life satisfaction

(discrete with levels 0-3, 3 meaning most satisfied) (2) highest education level (discrete with levels 0-5, indicating high school or less, high school diploma or equivalent, college education, college diploma, graduate school, and graduate degree, respectively), (3) marital status (discrete with levels 0-5, indicating single (never married), married, separated, divorced, widowed, and domestic partnership (living together but not married), respectively), (4) number of children (discrete), (5) the number of people in the household (discrete), (6) total annual household income (continuous), (7) hours per week working at the last job (continuous), (8) how the last job ended (discrete with levels 0-2 lost job, quit job, and temporary job ended, respectively), (9) weeks spent looking for work (continuous), and (10) credit card balance (continuous).

For these data, the IFR model produces the coefficient estimates

$$\widehat{\widetilde{\boldsymbol{\theta}}} = (0.483, 0.134, -0.166, -0.190, 0.042, 0.303, 0.075, 0.230, 0.662, -0.307)^{\intercal}.$$

The estimated coefficients can be used to obtain interpretable visualizations of the effect of the individual predictors on the compositional response through the (estimated) single index link function, which can further lead to effective inference for the proposed IFR model. For example, we illustrate below (Figure 6) the effect of the predictor "life satisfaction" on the mood compositional data. To this end, the IFR model is fitted over varying levels of life satisfaction, from low (0) to high (3), while the other predictors are fixed at their median levels. We observe an association between a lower life satisfaction level with a higher proportion of bad mood, while a higher value of life satisfaction is associated with a better mood when all of the other predictors are fixed.

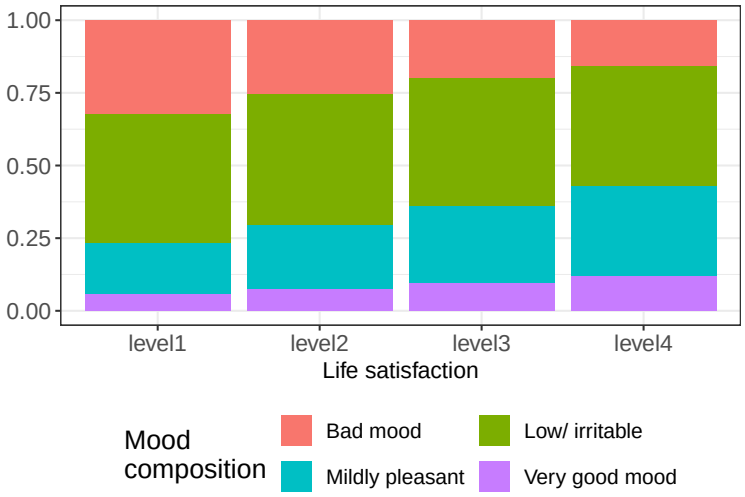

Fig 6: A stacked barplot showing the effect of life satisfaction, from Level 1 (0) to Level 4 (3), on the mood composition, when all the other predictor levels are kept fixed. A higher life satisfaction level is associated with a larger proportion of good mood.

The predictive performance of the model is computed based on the root mean prediction error (RMPE) as

$$\text{RMPE} = \left[\frac{1}{M_{n_{\text{test}}}} \sum_{i=1}^{M_{n_{\text{test}}}} d_g^2 \left(\tilde{Y}_l^{\text{test}}, \hat{m}_{\oplus}(\tilde{\mathbf{X}}_l^{\top} \widehat{\widetilde{\boldsymbol{\theta}}}, \widehat{\widetilde{\boldsymbol{\theta}}}) \right) \right]^{1/2},$$

TABLE 2

Mean and sd (in parenthesis) of root mean prediction error (RMPE) over 200 *repetitions, as obtained from the local fits of the index Fréchet regression (IFR) model, the global Fréchet regression (GFR) model, and four individual local linear Fréchet regression (LFR) models incorporating univariate continuous predictors. Here,* n_{train} *and* n_{test} *denote the sample sizes for the split training and testing datasets respectively.*

n_{train}	n_{test}	IFR	GFR	LFR_1	LFR_2	LFR_3	LFR_4
2201	1100	0.4779 (0.0720)	0.7661 (0.0418)	0.6771 (0.0021)	0.7220 (0.0450)	1.1127 (0.0910)	1.0122 (0.0810)

where $\tilde{Y}_l^{\text{test}}$ and $\hat{m}_{\oplus}(\tilde{\mathbf{X}}_l^{\top}\widehat{\widehat{\boldsymbol{\theta}}})$ denote, respectively, the l^{th} observed and predicted responses in the test set, evaluated at the binned average $\tilde{\mathbf{X}}_l$. We repeat this process 200 times, and compute RMPE for each split for the subjects separately. For comparison purposes, we fit the data with the other applicable object regression methods, namely, the global Fréchet regression (GFR) method with the four-dimensional mood-compositional data as the response residing on the surface of the sphere $S^3 \subset \mathbb{R}^4$, coupled with the 10-dimensional predictors; and individual local linear Fréchet regression (LFR) methods accommodating the afore-mentioned object response, while incorporating the continuous predictors total annual household income, hours per week working at the last job, weeks spent looking for work and credit card balance as univariate predictors. Like nonparametric regression, the LFR method does not work for discrete/ categorical predictors. We denote the results from the four individual univariate local regression by LFR_j, $j = 1, 2, 3, 4$, respectively. Table 2 summarizes the results.

We observe that the out-of-sample prediction error is quite low. In fact, it is very close to the average fitting error (0.351), calculated as the average distance between the observed training sample and the predicted objects based on the covariates in the training sets, which supports the validity of the proposed IFR models.

Since in this example the object-valued responses lie on a manifold (sphere) with positive curvature, the sufficient (but not necessary) condition for assumption (A5) that the underlying metric space behaves like a CAT(0) space is not satisfied. However, the numerical performance of the IFR method is quite good, suggesting a certain degree of model robustness of the IFR method.

S.4.3. *Additional results for the analysis of ADNI neuroimaging data.*
Continuing from Section 5.1 in the main manuscript, we illustrate the 95% confidence region for the coefficients $(\theta_1, \theta_2, \theta_4)$ of the predictors: stages of the disease, age, and total score in a 3-dimensional plot in Figure 7.

S.4.4. *Additional simulations for Euclidean responses.*
Here the object response of interest is assumed to lie in the Euclidean space. For generating the predictor vectors we consider a $5-$dimensional vector distributed as truncated multivariate normal distributions, where each of the components is truncated to lie between $[-10, 10]$. The components are assumed to be correlated such that X_1 correlates with X_2 and X_3 with $r = 0.5$, and X_2 and X_3 correlate with $r = 0.25$. The variances for each of the five components are 0.1. The empirical power against the sequence of alternatives in equation (3.10) increases steeply (see Figure 8) as we deviate from the null hypothesis in equation (3.9) in Section 3 of the main manuscript, especially corresponding to higher sample size and under identity link.

The empirical power function, as we deviate from the null hypothesis in equation (3.9) is computed and illustrated in the left panel in Figure 8. Empirical evidence suggests that the proposed test is consistent for a higher sample size of $n = 1000$, and leads to the correct nominal level of the test.

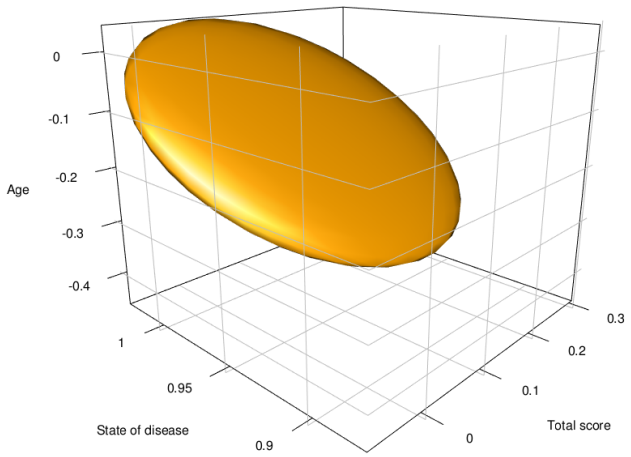

Fig 7: The figure shows the 3-dimensional plot for the 95% confidence region of $(\theta_1, \theta_2, \theta_4)$: the coefficients of the effects of the predictors- age, total score, and stage of the disease, respectively.

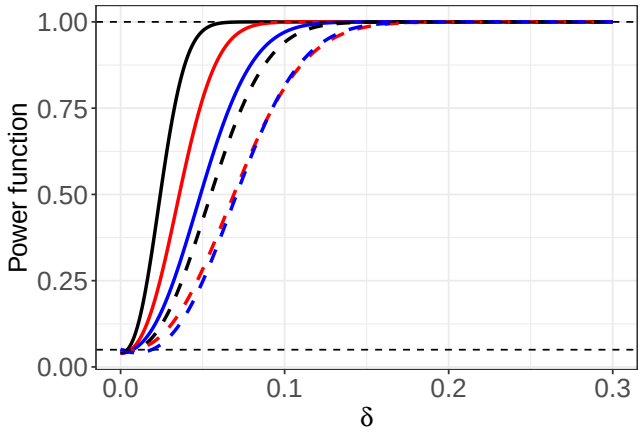

Fig 8: Simulation for Euclidean response using different link functions showing the empirical power function for Euclidean responses. The black, red, and blue curves correspond to the identity, square, and exponential link functions used in the data-generating mechanism, respectively, while the dashed and solid patterns correspond to the varying sample sizes $n = 100$ and $n = 1000$, respectively. The level of the tests is $\alpha = 0.05$ and is indicated by the dashed line parallel to the x-axis.

The consistency of the estimates is illustrated in Table 3 based on 500 replications of the simulation scenario. Further, the performance of the proposed method is compared to the classical Euclidean single index model fits. To this end, the R package *np* was called from Julia, for fitting the classical single index regression to the simulated Euclidean responses. The prediction performance of the classical single index fits, denoted by NP, is compared with that of the IFR method, as well as with a Global Fréchet Regression (GFR) method and four separate Local Frécet Regression (LFR) fits. The GFR method utilizes the multivariate predictors while the four LFR methods treat each of the four-dimensional predictor components as a univariate predictor individually. Note that in all of the methods- NP, GFR,

TABLE 3

Table showing bias and variance of $\widehat{\widehat{\boldsymbol{\theta}}}$ (measured in radians) based on 500 replications for a Euclidean vector response. The predictors $X_1, \ldots, X_5$ are generated from a truncated multivariate normal distribution.

	link1 ($x \mapsto x$)		link2 ($x \mapsto x^2$)		link3 ($x \mapsto e^x$)	
	bias	dev	bias	dev	bias	dev
$n = 100$	0.013	0.061	0.025	0.048	0.037	0.029
$n = 1000$	0.006	0.021	0.014	0.019	0.013	0.009

LFR - binning is not required. The mean and sd of the root mean prediction error (RMPE) over 200 Monte Carlo simulation runs are reported in Table 4. The data is simulated using

TABLE 4

Table showing the mean (sd in parenthesis) RMPE for various regression methods for simulated Euclidean responses. The methods compared are index Fréchet regression (IFR), classical Euclidean single index regression using the R package "np" (NP), global Fréchet Regression (GFR) with the 4-dimensional predictor, and four individual local linear Fréchet regression (LFR) models that treat each predictor components as a univariate predictor. The sample size is fixed at $n = 1000$ and the RMPE are computed over 200 Monte Carlo simulation runs.

	Identity link	Square link	Exponential link
IFR	0.0255 (0.0110)	0.1383 (0.1031)	0.1972 (0.1205)
NP	0.0187 (0.0201)	0.1117 (0.1077)	0.1578 (0.0442)
GFR	0.0003 (0.0018)	0.1465 (0.0299)	0.2181 (0.0748)
LFR1	0.0788 (0.0208)	0.2686 (0.0558)	0.3342 (0.1882)
LFR2	0.0784 (0.0204)	0.2627 (0.0540)	0.3237 (0.1912)
LFR3	0.0617 (0.0209)	0.2774 (0.0555)	0.3162 (0.1892)
LFR4	0.0730 (0.0197)	0.2694 (0.0561)	0.3664 (0.1888)

three different generating mechanisms - the identity, squared, and exponential link functions, and the sample size $n = 1000$ is considered. For the identity link function, i.e., when the simulated data is generated according to a linear model, the GFR method gives the lowest prediction error. This is indeed expected since the GFR boils down to a linear regression model when the object data are Euclidean. For other situations the NP method for the classical single index model outperforms the other methods, however, the proposed IFR method proves competitive with a comparable magnitude of the prediction error. The boxplot of the RMPEs for the above situations is shown in Figure 9.

REFERENCES

[1] CHEN, Y. and MÜLLER, H.-G. (2022). Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *The Annals of Statistics* **50** 1573–1592.

[2] CHEN, Y., GAJARDO, A., FAN, J., ZHONG, Q., DUBEY, P., HAN, K., BHATTACHARJEE, S. and MÜLLER, H.-G. (2020). frechet: Statistical Analysis for Random Objects and Non-Euclidean Data R package version 0.2.0.

[3] DAI, X., LIN, Z. and MÜLLER, H.-G. (2021). Modeling sparse longitudinal data on Riemannian manifolds. *Biometrics* **77** 1328–1341.

[4] DAVISON, A. C. (2003). *Statistical Models* **11**. Cambridge University Press.

[5] KATO, K. (2011). A note on moment convergence of bootstrap M-estimators. *Statistics & Decisions* **28** 51–61.

[6] KRUEGER, A. B., MUELLER, A., DAVIS, S. J. and ŞAHIN, A. (2011). Job search, emotional well-being, and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data [with comments and discussion]. *Brookings Papers on Economic Activity* 1–81.

[7] PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics* **47** 691–719. https://doi.org/10.1214/17-AOS1624

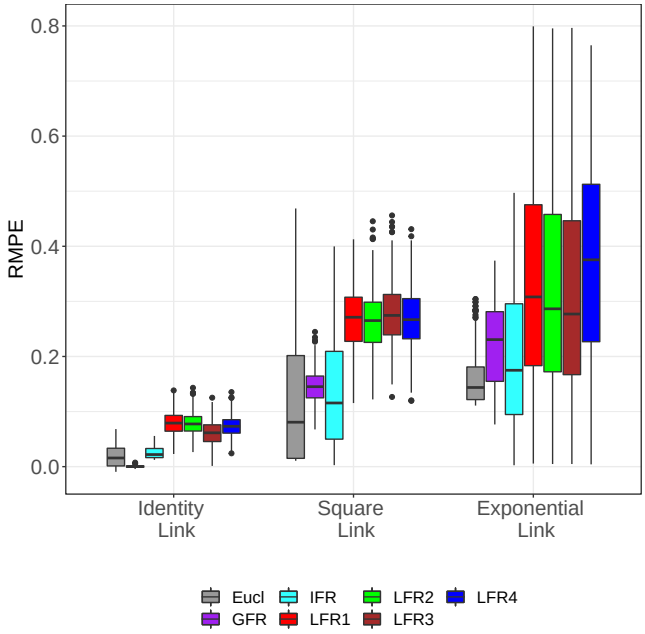

Fig 9: Figure showing boxplot of RMPEs for various regression methods for simulated Euclidean responses. The methods compared are index Fréchet regression (IFR), classical Euclidean single index regression using the R package "np" (NP), global Fréchet Regression (GFR) with the 4-dimensional predictor, and four individual local linear Fréchet regression (LFR) models that treat each predictor components as a univariate predictor. The sample size is fixed at $n = 1000$ and the RMPE are computed over 200 Monte Carlo simulation runs.

[8] SCEALY, J. and WELSH, A. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 351–375.

[9] SCEALY, J. and WELSH, A. (2014). Colours and cocktails: Compositional data analysis. *Australian & New Zealand Journal of Statistics* **56** 145–169.

[10] STURM, K.-T. (2003). Probability measures on metric spaces of nonpositive. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France* **338** 357.

[11] VAN DER VAART, A. and WELLNER, J. (2000). *Weak Convergence and Empirical Processes: with Applications to Statistics (Springer Series in Statistics)*. Springer.

[12] ZHANG, Q., XUE, L. and LI, B. (2021). Dimension reduction and data visualization for Fréchet regression. *arXiv preprint arXiv:2110.00467*.