Supplement to “Variable Screening and Spatial Smoothing in Fréchet Regression with Application to Diffusion Tensor Imaging” by

Lei Yan[1], Xin Zhang[1], Lan Zhou[2], Dipankar Bandyopadhyay[3], Yichao Wu[4], and For the Alzheimer’s Disease Neuroimaging Initiative[5]

[1] *Department of Statistics, Florida State University, Tallahassee, FL*
[2] *Dept. of Radiology, Brigham and Women’s Hospital, Boston, MA*
[3] *Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA*
[4] *Dept. of Mathematics, Statistics, and Computer Science, The University of Illinois, Chicago, IL*

Supplement A: Definition of Distance Covariance in Metric Spaces

Lyons (2013) introduced distance covariance (DC) defined on general metric spaces. Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two separable metric spaces, θ be the probability measure on the product space $\mathcal{X} \times \mathcal{Y}$, and $\mu := \pi_1(\theta), \nu := \pi_2(\theta)$ be marginal probability measures on $\mathcal{X}$ and $\mathcal{Y}$. Here, π_1 and π_2 are the coordinate projections onto the marginal spaces. For any $k > 0$, a finite signed Borel measure on $(\mathcal{X}, d_{\mathcal{X}})$ is said to have finite k-th moment if $\int d_{\mathcal{X}}^k(x, o) d|\mu|(x) < \infty$, for some $o \in \mathcal{X}$. Let $M^k(\mathcal{X})$ denotes the space of all finite signed measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with finite k-th moment, $M^{k,k}(\mathcal{X}, \mathcal{Y})$ be the space of all finite signed measures θ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}))$ for which $\pi_1(|\theta|) \in M^k(\mathcal{X})$ and $\pi_2(|\theta|) \in M^k(\mathcal{Y})$, $M_1(\mathcal{X})$ be the space of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and $M_1(\mathcal{X} \times \mathcal{Y})$ be the space of all probability measures on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}))$. When we put both a subscript and superscript, it denotes the intersection. For example. $M_1^{1,1}(\mathcal{X} \times \mathcal{Y})$ is the space of all probability measures θ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}))$, such that $\pi_1(|\theta|) \in M^1(\mathcal{X})$ and $\pi_2(|\theta|) \in M^1(\mathcal{Y})$. For any $\mu \in M^1(\mathbf{X})$, the mapping $a_\mu : \mathcal{X} \to \mathbb{R}$ is given by

$$a_\mu(x) = \int d_{\mathcal{X}}(x, y) d\mu(y),$$

and the mapping $D : M^1(\mathcal{X}) \to \mathbb{R}$ by

$$D(\mu) = \int a_\mu(x) d\mu(x) = \int d_{\mathcal{X}}(x_1, x_2) d\mu \times \mu(x_1, x_2).$$

Finally, we define $d_\mu : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$d_\mu(x_1, x_2) = d_{\mathcal{X}}(x_1, x_2) - a_\mu(x_1) - a_\mu(x_2) + D(\mu).$$

We define a metric space $(\mathcal{X}, d_{\mathcal{X}})$ is of negative type, if $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j d_{\mathcal{X}}(x_i, x_j) \leq 0$ for any $n \in \mathbb{N}, x \in \mathcal{X}^n, \alpha \in \mathbb{R}^n$ with $\sum_{i=1}^n \alpha_i = 0$. A metric space $(\mathcal{X}, d_{\mathcal{X}})$ of negative type, is of strong negative type if $D(\mu_1 - \mu_2) = 0 \iff \mu_1 = \mu_2$, for all probability measures μ_1, μ_2 with finite first moment. When both $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are metric spaces of strong negative type, $\mathrm{dcov}(\theta)$ equals to zero if and only if $\theta = \mu \times \nu$.

The following spaces are of strong negative type.

- Separable Hilbert space (Theorem 3.16 Lyons (2013)).
- Separable Hilbert spaces with distance $\|\cdot\|^{2r}$ for $0 < r < 1$ (Theorem 4.2 Dehling et al. (2020)).
- Non-separable Hilbert space that the dimension is a cardinal of measure zero (Remark 3.17 Lyons (2013)).
- Real hyperbolic space (Lyons, 2014).
- Separable spaces $(L^p, \|\cdot\|^{pr})$ with $0 < p \leq 2$ and $0 < r < 1$ (Lyons, 2021).
- Subsets of the unit-radius sphere S^n, centered at the origin of $\mathbb{R}^{n+1}$, which contains at most one pair of antipodal points (Lyons, 2020).

Theorem 3.6 in Meckes (2013) shows the following metric spaces have negative type.

- A^α, where A is a metric space of negative type and $0 < \alpha \leq 1$.
- Two-dimensional real normed spaces.

- Metric spaces with at most four points.
- Ultrametric spaces.
- Round spheres (with the geodesic distance).
- Complex hyperbolic space.
- Weighted trees.

According to Lyons (2013), if (X, d) has negative type, then (X, d^r) has strong negative type when $0 < r < 1$. Therefore, all the above metric spaces have strong negative type when we change their distance from d to d^r.

Supplement B: Proof of Lemmas

B.1 Lemma 1

Following the proof of statement (2.18) in Székely et al. (2007), it is easy to prove the result.

B.2 Lemma 2

Let Y_i be $M \times M$ SPD matrices, w_i be positive numbers representing the weights for Y_i. We are interested in the following optimization problem:

$$\underset{\omega \in \Omega}{\operatorname{argmin}} \sum_{i=1}^{n} w_i d_C^2(Y_i, \omega). \tag{B.1}$$

Without loss of generality, we assume $\sum_i^n w_i = 1$. Define $\widehat{B} = \sum_{i=1}^n w_i Y_i^{1/2}$. Then we have

$$\begin{aligned}
\sum_{i=1}^{n} w_i d_C^2(Y_i, \omega) &= \sum_{i=1}^{n} w_i \mathrm{tr}\left((Y_i^{1/2})^T Y_i^{1/2} - (\omega^{1/2})^T Y_i^{1/2} - (Y_i^{1/2})^T \omega^{1/2} + (\omega^{1/2})^T \omega^{1/2} \right) \\
&= \sum_{i=1}^{n} w_i \mathrm{tr}\left((Y_i^{1/2})^T Y_i^{1/2} - \widehat{B}^T Y_i^{1/2} - (Y_i^{1/2})^T \widehat{B} + \widehat{B}^T \widehat{B} \right) \\
&+ \mathrm{tr}\left((\omega^{1/2})^T \omega^{1/2} - (\omega^{1/2})^T \widehat{B} - \widehat{B}^T \omega^{1/2} + \widehat{B}^T \widehat{B} \right) \\
&= \sum_{i=1}^{n} w_i d_C^2(Y_i, \widehat{B}^T \widehat{B}) + d_C^2(\omega, \widehat{B}^T \widehat{B})
\end{aligned}$$

Therefore, we obtain

$$\underset{\omega \in \Omega}{\operatorname{argmin}} \sum_{i=1}^{n} w_i d_C^2(Y_i, \omega) = \underset{\omega \in \Omega}{\operatorname{argmin}}\, d_C^2(\omega, \widehat{B}^T \widehat{B}) \tag{B.2}$$

Note that Ω is the set of SPD matrices and $\widehat{B}^T \widehat{B}$ is a SPD matrix. Thus, the solution to (B.1) is simply $\omega = \widehat{B}^T \widehat{B}$. Let $w_i = w_i(\mathbf{x})$ and $Y_i = Y_{i,v}$. We immediately obtain Lemma 2.

B.3 Lemma 3

Let $n-1$ be the number of neighbors and we label neighbors from 1 to $n-1$. Let $w_{v'} = w_i, i = 1, 2, \ldots, n-1$, and $w_v = w_n = 1$. Then the optimization (2.7) can be rewritten in the form of (B.1). Then according to the proof of Lemma 2, we proved Lemma 3.

Supplement C: Proof of Theorem 1

Define

$$S_{k1} = Ed_{\mathcal{X}_k}(X_k, X_k')d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}'),$$
$$S_{k2} = Ed_{\mathcal{X}_k}(X_k, X_k')Ed_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}'),$$
$$S_{k3} = Ed_{\mathcal{X}_k}(X_k, X_k')d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}''),$$

and

$$\widehat{S}_{k1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j),$$
$$\widehat{S}_{k2} = \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{X}_k}(x_{ki}, x_{kj})\right)\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j)\right),$$
$$\widehat{S}_{k3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l).$$

Then, we have

$$\omega_k = S_{k1} + S_{k2} - 2S_{k3}, \quad \widehat{\omega}_k = \widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}.$$

To find the convergence rate of $\widehat{\omega}_k$, we study each term separately. Let us start with S_{k1}. We define

$$\widehat{S}^*_{k1} = \frac{1}{n(n-1)}\sum_{i\neq j} d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j),$$

which is a U-statistic. By Cauchy-Schwarz inequality, we have

$$S_{k1} \leq \sqrt{Ed^2_{\mathcal{X}_k}(X_k, X_k')Ed^2_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}')}.$$

According to assumption 1 and Jensen's inequality, we know $Ed^2_{\mathcal{X}_k}(X_k, X_k') < \infty$ and $Ed^2_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}') < \infty$, which implies $S_{k1} < \infty$ for $k = 1, \dots, p$. For any given $\epsilon > 0$, there exists a larger enough n such that $S_{k1}/n < \epsilon$. Then, we have

$$\begin{aligned}
P(|\widehat{S}_{k1} - S_{k1}| \geq 2\epsilon) &= P\left(\left|\widehat{S}^*_{k1}\frac{n-1}{n} - S_{k1}\frac{n-1}{n} - S_{k1}\frac{1}{n}\right| \geq 2\epsilon\right) \\
&\leq P\left(\left|\widehat{S}^*_{k1} - S_{k1}\right|\frac{n-1}{n} \geq 2\epsilon - \frac{S_{k1}}{n}\right) \\
&\leq P\left(\left|\widehat{S}^*_{k1} - S_{k1}\right| \geq \epsilon\right).
\end{aligned}$$

Thus, to establish the uniform consistency of $\widehat{S}_{k1}$, it suffices to show the consistency of $\widehat{S}^*_{k1}$. Let $h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) = d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j)$ be the kernel of the U-statistic $\widehat{S}^*_{k1}$. We can decompose h_1 into two parts: $h_1 = h_1\mathbf{1}(h_1 > M) + h_1\mathbf{1}(h_1 \leq M)$, where M is specified later. Then we rewrite $\widehat{S}^*_{k1}$ as

$$\begin{aligned}
\widehat{S}^*_{k1} = &\frac{1}{n(n-1)}\sum_{i\neq j} h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \times \mathbf{1}(h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \leq M) \\
&+ \frac{1}{n(n-1)}\sum_{i\neq j} h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \times \mathbf{1}(h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) > M) \\
=: &\widehat{S}^*_{k1,1} + \widehat{S}^*_{k1,2}.
\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
S_{k1} &= E d_{\mathcal{X}_k}(X_k, X_k') d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}') \\
&= E\left[h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \times \mathbf{1}(h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \leq M)\right] \\
&\quad + E\left[h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \times \mathbf{1}(h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) > M)\right] \\
&=: S_{k1,1} + S_{k1,2}.
\end{aligned}
$$

It is easy to see that $\widehat{S}^*_{k1,1}, \widehat{S}^*_{k1,1}$ are unbiased estimator of $S_{k1,1}$ and $S_{k1,2}$. Using Markov's inequality, for any $t > 0$, we have

$$
\begin{aligned}
P(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \epsilon) &= P(t(\widehat{S}^*_{k1,1} - S_{k1,1}) \geq t\epsilon) \\
&= P(\exp(t(\widehat{S}^*_{k1,1} - S_{k1,1})) \geq \exp(t\epsilon)) \\
&\leq \frac{E[\exp(t(\widehat{S}^*_{k1,1} - S_{k1,1}))]}{\exp(t\epsilon)} \\
&= \exp(-t\epsilon)\exp(-tS_{k1,1})E[\exp(t\widehat{S}^*_{k1,1})].
\end{aligned} \tag{C.3}
$$

Serfling (2009, section 5.1.6) showed a U-statistic with symmetric kernel can be represented as an average of averages of independent and identically distributed random variables. To be precise, let $x_1, \ldots, x_n$ be a sample, and $h(x_1, \ldots, x_u)$ be a symmetric kernel of a U-statistic U_n. Define $m = \lfloor n/u \rfloor$ to be the greatest integer smaller than n/u and $W(x_1, \ldots, x_n) = 1/m[h(x_1, \ldots, x_u) + h(x_{u+1}, \ldots, x_{2u}) + \cdots + h(x_{mu-u+1}, \ldots, x_{mu})]$. Then, we have $U_n = 1/n! \sum_{\Pi_n} W(x_{i_1}, \ldots, x_{i_n})$, where Π_n is the set of all permutations $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$. Apply this result to $\widehat{S}^*_{k1,1}$ and let $m = \lfloor n/2 \rfloor$, we have

$$
\begin{aligned}
\widehat{S}^*_{k1,1} &= \frac{1}{n!}\sum_{\Pi_n} W_1((x_{i_1}, \mathbf{y}_{i_1}), \ldots, (x_{i_n}, \mathbf{y}_{i_n})) \\
&= \frac{1}{n!}\sum_{\Pi_n} \frac{1}{m}[h_1((x_{i_1}, \mathbf{y}_{i_1}), (x_{i_2}, \mathbf{y}_{i_2}))\mathbf{1}(h_1 \leq M) + \cdots \\
&\quad + h_1((x_{i_{2m-1}}, \mathbf{y}_{i_{2m-1}}), (x_{i_{2m}}, \mathbf{y}_{i_{2m}}))\mathbf{1}(h_1 \leq M)] \\
&=: \frac{1}{n!}\sum_{\Pi_n} \frac{1}{m}\sum_{r=1}^{m} h_1^{(r)}\mathbf{1}(h_1^{(r)} \leq M).
\end{aligned}
$$

We know $\exp(\cdot)$ is a convex function. It follows from Jensen's inequality that

$$
\exp\left(\frac{1}{n!}\sum_{\Pi_n} tW_1((x_{i_1}, \mathbf{y}_{i_1}), \ldots, (x_{i_n}, \mathbf{y}_{i_n}))\right) \leq \frac{1}{n!}\sum_{\Pi_n} E[\exp(tW_1((x_{i_1}, \mathbf{y}_{i_1}), \ldots, (x_{i_n}, \mathbf{y}_{i_n})))].
$$

Then, we have

$$
\begin{aligned}
E[\exp(t\widehat{S}^*_{k1,1})] &\leq \frac{1}{n!}\sum_{\Pi_n} E[\exp(tW_1((x_{i_1}, \mathbf{y}_{i_1}), \ldots, (x_{i_n}, \mathbf{y}_{i_n})))] \\
&= \frac{1}{n!}\sum_{\Pi_n} E\left[\exp\left(t\frac{1}{m}\sum_{r=1}^{m} h_1^{(r)}\mathbf{1}(h_1^{(r)} \leq M)\right)\right] \\
&= E\left[\exp\left(t\frac{1}{m}\sum_{r=1}^{m} h_1^{(r)}\mathbf{1}(h_1^{(r)} \leq M)\right)\right] \\
&= E^m\left[\exp\left(t/m h_1^{(r)}\mathbf{1}(h_1^{(r)} \leq M)\right)\right], \text{ for any } 1 \leq r \leq m.
\end{aligned}
$$

Serfling (2009, section 5.6.1, A) proved the following lemma.

Lemma 1. *Let $\mu = E(Y)$, if $P(a \le Y \le b) = 1$, then*

$$E[\exp(s(Y-\mu))] \le \exp\left(s^2 \frac{(b-a)^2}{8}\right), \forall s > 0.$$

Using Lemma 1, we have

$$E[\exp(\underbrace{t/m}_{s}[\underbrace{h_1^{(r)}\mathbf{1}(h_1^{(r)} \le M)}_{Y} - \underbrace{S_{k1,1}}_{\mu}])] \le \exp\left(\frac{t^2M^2}{8m^2}\right).$$

Continue (C.3),

$$\begin{aligned} P(\widehat{S}^*_{k1,1} - S_{k1,1} \ge \epsilon) &\le \exp(-t\epsilon)\exp(-tS_{k1,1})E^m\left[\exp\left(t/mh_1^{(r)}\mathbf{1}(h_1^{(r)} \le M)\right)\right] \\ &= \exp(-t\epsilon)E^m[\exp(-t/mS_{k1,1})]E^m\left[\exp\left(t/mh_1^{(r)}\mathbf{1}(h_1^{(r)} \le M)\right)\right] \\ &= \exp(-t\epsilon)E^m\left[\exp\left(t/m[h_1^{(r)}\mathbf{1}(h_1^{(r)} \le M) - S_{k1,1}]\right)\right] \\ &\le \exp\left(-t\epsilon + \frac{t^2M^2}{8m}\right). \end{aligned}$$

Let $t = 4\epsilon m/M^2$, we have $P(\widehat{S}^*_{k1,1} - S_{k1,1} \ge \epsilon) \le \exp\left(-2\epsilon^2 m/M^2\right)$. By symmetry of $\widehat{S}^*_{k1,1}$, the convergence rate is given by

$$P(|\widehat{S}^*_{k1,1} - S_{k1,1}| \ge \epsilon) \le 2\exp\left(-\frac{2\epsilon^2 m}{M^2}\right). \tag{C.4}$$

Now let's study $\widehat{S}^*_{k1,2}$. By Cauchy-Schwartz inequality,

$$\begin{aligned} S^2_{k1,2} &= E^2\left[h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) \times \mathbf{1}(h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) > M)\right] \\ &\le E[h_1^2((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j))]E[\mathbf{1}^2(h_1^2((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) > M)] \\ &= E[h_1^2((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j))]P(h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) > M). \end{aligned}$$

With Markov's inequality, for any $0 < s' \le 2s_0$,

$$P(h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) > M) \le \frac{E[\exp(s'h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)))]}{\exp(s'M)}.$$

Thus, we have

$$S^2_{k1,2} \le E[h_1^2((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j))]\frac{E[\exp(s'h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)))]}{\exp(s'M)}. \tag{C.5}$$

According to assumption 1, $E[h_1^2((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j))]$ is finite. Note that for some $o_k \in \mathcal{X}_k, o \in \mathcal{Y}$

$$\begin{aligned} h_1((x_{ki},\mathbf{y}_i),(x_{kj},\mathbf{y}_j)) &= d_{\mathcal{X}_k}(x_{ki},x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i,\mathbf{y}_j) \\ &\le [(d_{\mathcal{X}_k}(x_{ki},o_k) + d_{\mathcal{X}_k}(x_{kj},o_k))^2(d_{\mathcal{Y}}(\mathbf{y}_i,o)d_{\mathcal{Y}}(\mathbf{y}_j,o))^2]^{1/2} \\ &\le 2\{[d^2_{\mathcal{X}_k}(x_{ki},o_k) + d^2_{\mathcal{X}_k}(x_{kj},o_k)][d^2_{\mathcal{Y}}(\mathbf{y}_i,o) + d^2_{\mathcal{Y}}(\mathbf{y}_j,o)]\}^{1/2} \\ &\le 2\{[d^2_{\mathcal{X}_k}(x_{ki},o_k) + d^2_{\mathcal{X}_k}(x_{kj},o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i,o) + d^2_{\mathcal{Y}}(\mathbf{y}_j,o)]^2 1/4\}^{1/2} \\ &= d^2_{\mathcal{X}_k}(x_{ki},o_k) + d^2_{\mathcal{X}_k}(x_{kj},o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i,o) + d^2_{\mathcal{Y}}(\mathbf{y}_j,o). \end{aligned}$$

In the three inequalities, we first use the triangle inequality, then $(a+b)^2 \leq 2(a^2+b^2)$, and last $|ab| \leq (a+b)^2/4$. Using this result and Cauchy-Schwartz inequality,

$$\begin{aligned}
E[h_1^2((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j))] &\leq E[\exp(s'[d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{X}_k}(x_{kj}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) + d^2_{\mathcal{Y}}(\mathbf{y}_j, o)])] \\
&\leq \{E[\exp(2s' d^2_{\mathcal{X}_k}(x_{ki}, o_k))] E[\exp(2s' d^2_{\mathcal{X}_k}(x_{kj}, o_k))]\}^{1/2} \\
&\quad \times \{E[\exp(2s' d^2_{\mathcal{Y}}(\mathbf{y}_i, o))] E[\exp(2s' d^2_{\mathcal{Y}}(\mathbf{y}_j, o))]\}^{1/2} \\
&= E[\exp(2s' d^2_{\mathcal{X}_k}(x_{ki}, o_k))] E[\exp(2s' d^2_{\mathcal{Y}}(\mathbf{y}_i, o))] \\
&< \infty,
\end{aligned}$$

where in the last inequality, we use the assumption 1. Let $M = cn^\gamma$, where $0 < \gamma < 0.5 - \kappa, 0 \leq \kappa < 0.5$. According to (C.5), $S_{k1,2} \leq \epsilon/2$ when n is large enough. This implies that $P(|\widehat{S}^*_{k1,2} - S_{k1,2}| > \epsilon) \leq P(|\widehat{S}^*_{k1,2}| \geq \epsilon/2)$. Therefore, it suffices to bound $P(|\widehat{S}^*_{k1,2}| \geq \epsilon/2)$. We state that

$$\left\{ |\widehat{S}^*_{k1,2}| > \frac{\epsilon}{2} \right\} \subseteq \left\{ d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) > \frac{M}{2}, \text{ for some } i \right\}.$$

To prove this, suppose $d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) \leq \frac{M}{2}$ for all i. Then $h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) \leq d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{X}_k}(x_{kj}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) + d^2_{\mathcal{Y}}(\mathbf{y}_j, o) \leq M$. By the definition of $\widehat{S}^*_{k1,2} = 1/(n(n-1) \sum_{i \neq j} h_1 \mathbf{1}(h_1 > M)$, we have $\widehat{S}^*_{k1,2} = 0$, which contradicts to the event $\left\{ |\widehat{S}^*_{k1,2}| > \epsilon/2 \right\}$. By Markov's inequality, $P\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) \geq \sqrt{M}/2 \right) \leq E[\exp(s d^2_{\mathcal{X}_k}(x_{ki}, o_k))]/\exp(sM/4)$, for any $s > 0$. According to assumption 1, there exists a constant C such that
$E[\exp(s d^2_{\mathcal{X}_k}(x_{ki}, o_k))] \leq C$. Thus, $P\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) \geq \sqrt{M}/2 \right) \leq C\exp(-sM/4)$. Using similar argument, it is easy to show $P\left(d^2_{\mathcal{Y}}(\mathbf{y}_i, o) \geq \sqrt{M}/2 \right) \leq C\exp(-sM/4)$. Therefore, we have

$$\begin{aligned}
P\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) > M/2 \right) &\leq P\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) \geq \sqrt{M}/2 \right) + P\left(d^2_{\mathcal{Y}}(\mathbf{y}_i, o) \geq \sqrt{M}/2 \right) \\
&\leq 2C\exp(-sM/4).
\end{aligned}$$

It follows that

$$\begin{aligned}
P(|\widehat{S}^*_{k1,2}| \geq \epsilon/2) &\leq nP\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) > M/2 \right) \\
&\leq 2nC\exp(-sM/4).
\end{aligned} \tag{C.6}$$

Combining (C.4) and (C.6), we have

$$\begin{aligned}
P(|\widehat{S}_{k1} - S_{k1}| \geq 4\epsilon) &\leq P(|\widehat{S}^*_{k1} - S_{k1}| \geq 2\epsilon) \\
&= P(|(\widehat{S}^*_{k1,1} - S_{k1,1}) + (\widehat{S}^*_{k1,2} - S_{k1,2})| \geq 2\epsilon) \\
&\leq P(|\widehat{S}^*_{k1,1} - S_{k1,1}| \geq \epsilon) + P(|\widehat{S}^*_{k1,2} - S_{k1,2}| \geq \epsilon) \\
&\leq 2\exp(-2\epsilon^2 m/M^2) + 2nC\exp(-sM/4).
\end{aligned}$$

Recall that $m = \lfloor n/2 \rfloor$ and $M = cn^\gamma$,

$$P(|\widehat{S}_{k1} - S_{k1}| \geq 4\epsilon) \leq 2\exp(-\epsilon^2 c^{-2} n^{1-2\gamma}) + 2nC\exp(-scn^\gamma/4). \tag{C.7}$$

Next, we study $\widehat{S}_{k2}$. Recall that

$$\begin{aligned}
\widehat{S}_{k2} &= \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{\mathcal{X}_k}(x_{ki}, x_{kj}) \right) \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) \right) \\
&=: \widehat{S}_{k2,1} \widehat{S}_{k2,2}.
\end{aligned}$$

Similarly,

$$Ed_{\mathcal{X}_k}(X_k, X_k')Ed_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}') =: S_{k2,1}S_{k2,2}.$$

We wish to bound $P(|\widehat{S}_{k2,1} - S_{k2,1}| \geq \epsilon)$ and $P(|\widehat{S}_{k2,2} - S_{k2,2}| \geq \epsilon)$ as we did before. For $S_{k2,1}$, we will follow the same argument for $S_{k1,1}$. Define $\widehat{S}^*_{k2,1} = 1/(n(n-1)) \sum_{i \neq j} d_{\mathcal{X}_k}(x_{ki}, x_{kj})$. It is easy to see $\widehat{S}^*_{k2,1}$ is a U-statistic with a symmetric kernel. By assumption 1, $S_{k2,1}/n < \epsilon$ for sufficiently large n. Then,

$$\begin{aligned} P(|\widehat{S}_{k2,1} - S_{k2,1}| \geq 2\epsilon) &= P\left(\left|\widehat{S}_{k2,1}\frac{n-1}{n} - S_{k2,1}\frac{n-1}{n} - S_{k2,1}\frac{1}{n}\right| \geq 2\epsilon\right) \\ &\leq P\left(\left|\widehat{S}^*_{k2,1} - S_{k2,1}\right|\frac{n-1}{n} \geq 2\epsilon - \frac{S_{k2,1}}{n}\right) \\ &\leq P(|\widehat{S}^*_{k2,1} - S_{k2,1}| \geq \epsilon). \end{aligned}$$

Let $h_2(x_{ki}, x_{kj}) = d_{\mathcal{X}_k}(x_{ki}, x_{kj})$ be the kernel of the U-statistic $\widehat{S}^*_{k2,1}$. Similarly, we divide h_2 into two parts $h_2 = h_2\mathbf{1}(h_2 > M) + h_2\mathbf{1}(h_2 \leq M)$. Then we can rewrite $\widehat{S}^*_{k2,1}$

$$\begin{aligned} \widehat{S}^*_{k2,1} &= \frac{1}{n(n-1)}\sum_{i \neq j} h_2(x_{ki}, x_{kj})\mathbf{1}(h_2(x_{ki}, x_{kj}) \leq M) \\ &\quad + \frac{1}{n(n-1)}\sum_{i \neq j} h_2(x_{ki}, x_{kj})\mathbf{1}(h_2(x_{ki}, x_{kj}) > M) \\ &=: \widehat{S}^*_{k2,1,1} + \widehat{S}^*_{k2,1,2}. \end{aligned}$$

Similarly, we divide $S_{k2,1}$ into two parts

$$\begin{aligned} S_{k2,1} &= E[h_2(x_{ki}, x_{kj})\mathbf{1}(h_2(x_{ki}, x_{kj}) \leq M)] \\ &\quad + E[h_2(x_{ki}, x_{kj})\mathbf{1}(h_2(x_{ki}, x_{kj}) > M)] \\ &=: S_{k2,1,1} + S_{k2,1,2}. \end{aligned}$$

For $\widehat{S}^*_{k2,1,1}$, we can obtain the exact same bound for $\widehat{S}^*_{k1,1}$ as in (C.4) since we only use some properties of U-statistic in the derivation. It is easy to show

$$P(|\widehat{S}^*_{k2,1,1} - S_{k2,1,1}| \geq \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2 m}{M^2}\right). \tag{C.8}$$

Using Cauchy-Schwarz inequality and Markov's inequality, for any $0 < s' \leq 2s_0$,

$$\begin{aligned} S^2_{k2,1,2} &\leq E[h_2^2(x_{ki}, x_{kj})]E[\mathbf{1}^2(h_2(x_{ki}, x_{kj}) > M)] \\ &= E[h_2^2(x_{ki}, x_{kj})]P(h_2(x_{ki}, x_{kj}) > M) \\ &\leq E[h_2^2(x_{ki}, x_{kj})]E[\exp(s'h_2(x_{ki}, x_{kj}))]/\exp(s'M). \end{aligned}$$

Note that $h_2((x_{ki}, x_{kj})) = d_{\mathcal{X}_k}(x_{ki}, x_{kj}) \leq d_{\mathcal{X}_k}(x_{ki}, o_k) + d_{\mathcal{X}_k}(x_{kj}, o_k)$, where $o_k \in \mathcal{X}_k$. According to assumption 1 and the fact that $\exp(s'd_{\mathcal{X}_k}(x_{ki}, o_k)) \leq \exp(s'\max(1, d^2_{\mathcal{X}_k}(x_{ki}, o_k)))$, we have $E[\exp(s'd_{\mathcal{X}_k}(x_{ki}, o_k))] < \infty$. Therefore, $E[\exp(s'h_2(x_{ki}, x_{kj}))] < \infty$. Then, for any $\epsilon > 0$, $S_{k2,1,2} \leq \epsilon/2$ for large enough n when we let $M = cn^\gamma$. This means

$$P(|\widehat{S}^*_{k2,1,2} - S_{k2,1,2}| \geq \epsilon) \leq P(|\widehat{S}^*_{k2,1,2}| \geq \epsilon/2).$$

Thus, it suffices to bound $P(|\widehat{S}^*_{k2,1,2}|\geq \epsilon/2)$. We state that

$$\left\{|\widehat{S}^*_{k2,1,2}|\geq \frac{\epsilon}{2}\right\} \subseteq \left\{d_{\mathcal{X}_k}(x_{ki}, o_k) > \frac{M}{2},\ \text{for some } i\right\}.$$

To see this, suppose $d_{\mathcal{X}_k}(x_{ki}, o_k) \leq M/2$ for all i. Then $h_2(x_{ki}, x_{kj}) \leq d_{\mathcal{X}_k}(x_{ki}, o_k)+d_{\mathcal{X}_k}(x_{kj}, o_k) \leq M$, which means $\widehat{S}^*_{k2,1,2} = 0$. This contradicts to the event $\{|\widehat{S}^*_{k2,1,2}|\geq \epsilon/2\}$. By Markov's inequality and assumption 1, for any $0 < s \leq 2s_0$, there exists a constant C such that

$$P(d_{\mathcal{X}_k}(x_{ki}, o_k) > M/2) \leq E[\exp(sd^2_{\mathcal{X}_k}(x_{ki}, o_k))]/\exp(sM^2/4) \leq C\exp(-sM^2/4).$$

It follows that

$$P(|\widehat{S}^*_{k2,1,2}|> \epsilon/2) \leq nP(d_{\mathcal{X}_k}(x_{ki}, o_k) > M/2) \leq 2nC\exp(-sM^2/4). \tag{C.9}$$

Combining (C.8) and (C.9), we have

$$\begin{aligned}
P(|\widehat{S}_{k2,1} - S_{k2,1}|\geq 4\epsilon) &\leq P(|\widehat{S}^*_{k2,1} - S_{k2,1}|\geq 2\epsilon)\\
&\leq P(|\widehat{S}^*_{k2,1,1} - S_{k2,1,1}|\geq \epsilon) + P(|\widehat{S}^*_{k2,1,2} - S_{k2,1,2}|\geq \epsilon)\\
&\leq 2\exp(-2\epsilon^2 m/M^2) + 2nC\exp(-sM^2/4).
\end{aligned}$$

Since $m = \lfloor n/2\rfloor, M = cn^\gamma$,

$$P(|\widehat{S}_{k2,1} - S_{k2,1}|\geq 4\epsilon) \leq 2\exp(-\epsilon^2c^{-2}n^{1-2\gamma}) + 2nC\exp(-sc^2n^{2\gamma}/4). \tag{C.10}$$

Using the same argument, we can get the convergence rate of $\widehat{S}_{k2,2}$,

$$P(|\widehat{S}_{k2,2} - S_{k2,2}|\geq 4\epsilon) \leq 2\exp(-\epsilon^2c^{-2}n^{1-2\gamma}) + 2nC\exp(-sc^2n^{2\gamma}/4). \tag{C.11}$$

According to Jakobsen (2017, Lemma 2.5), $S_{k2,1} = Ed_{\mathcal{X}}(X_k, X'_k) < \infty$ and $S_{k2,2} = Ed_{\mathcal{Y}}(Y, Y') < \infty$. Thus, there exists a constant C such that $\max\{\max_k S_{k2,1}, S_{k2,2}\} \leq C$. Then,

$$\begin{aligned}
P(|(\widehat{S}_{k2,1} - S_{k2,2})S_{k2,2}|\geq \epsilon) &\leq P(|\widehat{S}_{k2,1} - S_{k2,2}|\geq \epsilon/C)\\
&\leq 2\exp(-C^{-2}\epsilon^2c^{-2}n^{1-2\gamma}/16) + 2nC\exp(-sc^2n^{2\gamma}/4),
\end{aligned}$$

$$\begin{aligned}
P(|(\widehat{S}_{k2,2} - S_{k2,2})S_{k2,1}|\geq \epsilon) &\leq P(|\widehat{S}_{k2,2} - S_{k2,2}|\geq \epsilon/C)\\
&\leq 2\exp(-C^{-2}\epsilon^2c^{-2}n^{1-2\gamma}/16) + 2nC\exp(-sc^2n^{2\gamma}/4),
\end{aligned}$$

and

$$\begin{aligned}
P(|(\widehat{S}_{k2,1} - S_{k2,1})(\widehat{S}_{k2,1} - S_{k2,1})|\geq \epsilon) &\leq P(|\widehat{S}_{k2,1} - S_{k2,1}|\geq \sqrt{\epsilon}) + P(|\widehat{S}_{k2,1} - S_{k2,1}|\geq \sqrt{\epsilon})\\
&\leq 4\exp(-\epsilon c^{-2}n^{1-2\gamma}/16) + 4nC\exp(-sc^2n^{2\gamma}/4)
\end{aligned}$$

Note that

$$\widehat{S}_{k2,1}\widehat{S}_{k2,2} - S_{k2,1}S_{k2,2} = (\widehat{S}_{k2,1} - S_{k2,1})S_{k2,2} + S_{k2,1}(\widehat{S}_{k2,2} - S_{k2,2}) + (\widehat{S}_{k2,1} - S_{k2,1})(\widehat{S}_{k2,1} - S_{k2,1}).$$

Then we have

$$\begin{aligned}
P(|\widehat{S}_{k2} - S_{k2}|\geq 3\epsilon) &= P(|\widehat{S}_{k2,1}\widehat{S}_{k2,2} - S_{k2,1}S_{k2,2}|\geq 3\epsilon)\\
&\leq P(|(\widehat{S}_{k2,1} - S_{k2,1})S_{k2,2}|\geq \epsilon)\\
&\quad + P(|S_{k2,1}(\widehat{S}_{k2,2} - S_{k2,2})|\geq \epsilon)\\
&\quad + P(|(\widehat{S}_{k2,1} - S_{k2,1})(\widehat{S}_{k2,1} - S_{k2,1})|\geq \epsilon)\\
&\leq 8\exp(-C^2\epsilon^2c^{-2}n^{1-2\gamma}/16) + 8nC\exp(-sc^2n^{2\gamma}/4),
\end{aligned} \tag{C.12}$$

where the last inequality holds when ϵ is sufficiently small and C is sufficiently large.

Last, let's study $\widehat{S}_{k3}$. Recall $\widehat{S}_{k3} = \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l)$. Note that $d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l)$ is not symmetric. Thus, our first step is to define a U-statistic with a symmetric kernel, which can be done by averaging all permutations of indices of i, j, l. To be precise, we define

$$\begin{aligned}
\widehat{S}^*_{k3} &= \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq l} d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) \\
&= \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq l}\frac{1}{3!}[d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l) + d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_l, \mathbf{y}_j) \\
&\qquad + d_{\mathcal{X}_k}(x_{kj}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_l) + d_{\mathcal{X}_k}(x_{kj}, x_{ki})d_{\mathcal{Y}}(\mathbf{y}_l, \mathbf{y}_i) \\
&\qquad + d_{\mathcal{X}_k}(x_{kl}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) + d_{\mathcal{X}_k}(x_{kl}, x_{ki})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_i)] \\
&=: \frac{6}{n(n-1)(n-2)}\sum_{i\leq j\leq l} h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l),
\end{aligned}$$

where h_3 is defined as the average inside the summation in the second equality. Then $\widehat{S}^*_{k3}$ is a U-statistic with a symmetric kernel. Following the same argument for $\widehat{S}^*_{k1}$, we decompose h_3 as $h_3 = h_3\mathbf{1}(h_3 \leq M) + h_3\mathbf{1}(h_3 > M)$. Thus, we have

$$\begin{aligned}
\widehat{S}_{k3} &= \frac{6}{n(n-1)(n-2)}\sum_{i<j<l} h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)\mathbf{1}(h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) \leq M) \\
&\quad + \frac{6}{n(n-1)(n-2)}\sum_{i<j<l} h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)\mathbf{1}(h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) > M) \\
&=: \widehat{S}^*_{k3,1} + \widehat{S}^*_{k3,2},
\end{aligned}$$

and

$$\begin{aligned}
S_{k3} &= E[h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)\mathbf{1}(h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) \leq M)] \\
&\quad + E[h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)\mathbf{1}(h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) > M)] \\
&=: S_{k3,1} + S_{k3,2}.
\end{aligned}$$

Following the same argument for deriving bound of $P(|\widehat{S}_{k1,1} - S_{k1,1}| \geq \epsilon)$, we have

$$P(|\widehat{S}^*_{k3,1} - S_{k3,1}| \geq \epsilon) \leq 2\exp(-\epsilon^2 m'/M^2), \tag{C.13}$$

where $m' = \lfloor n/3 \rfloor$. Next, we study $\widehat{S}^*_{k3,2}$. As before, we wish to bound h_3. Recall that $h_1((x_{ki}, \mathbf{y}_i), (x_{kj}, \mathbf{y}_j)) = d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) \leq d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{X}_k}(x_{kj}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) + d^2_{\mathcal{Y}}(\mathbf{y}_j, o)$. Then,

$$\begin{aligned}
h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) &= \frac{1}{3!}[d_{\mathcal{X}_k}(x_{ki}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l) + d_{\mathcal{X}_k}(x_{ki}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_l, \mathbf{y}_j) \\
&\quad + d_{\mathcal{X}_k}(x_{kj}, x_{kl})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_l) + d_{\mathcal{X}_k}(x_{kj}, x_{ki})d_{\mathcal{Y}}(\mathbf{y}_l, \mathbf{y}_i) \\
&\quad + d_{\mathcal{X}_k}(x_{kl}, x_{kj})d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) + d_{\mathcal{X}_k}(x_{kl}, x_{ki})d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_i)] \\
&\leq \frac{4}{6}[d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{X}_k}(x_{kj}, o_k) + d^2_{\mathcal{X}_k}(x_{kl}, o_k) \\
&\quad + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) + d^2_{\mathcal{Y}}(\mathbf{y}_j, o) + d^2_{\mathcal{Y}}(\mathbf{y}_l, o)].
\end{aligned}$$

By Cauchy-Schwarz inequality, for any $0 < s' \leq 2s_0$,

$$\begin{aligned}
S^2_{k3,2} &\leq E[h_3^2(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)]P(h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) > M) \\
&\leq E[h_3^2(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l)]E[\exp(s' h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l))]/\exp(s'M).
\end{aligned}$$

According to assumption 1, $E[\exp(s'h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l))] < \infty$. Let $M = cn^\gamma$, we know $S_{k3,2} \le \epsilon/2$, for large n. Then, $P(|\widehat{S}^*_{k3,2} - S_{k3,2}| > \epsilon) \le P(|\widehat{S}^*_{k3,2}| \ge \epsilon/2)$. Note that

$$\left\{\left|\widehat{S}^*_{k3,2}\right| > \frac{\epsilon}{2}\right\} \subseteq \left\{d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d_{\mathcal{Y}}(\mathbf{y}_i, o) > \frac{M}{2}, \text{ for some } i\right\}.$$

If $d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d_{\mathcal{Y}}(\mathbf{y}_i, o) \le M/2$ for all i, $h_3(x_{ki}, \mathbf{y}_i; x_{kj}, \mathbf{y}_j; x_{kl}, \mathbf{y}_l) \le 4/6(M/2 \times 3) \le M$, which contradicts to $|\widehat{S}^*_{k3,2}| \ge \epsilon/2$. We know $P\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) > M/2\right) \le 2C\exp(-sM/4)$. Thus,

$$P(|\widehat{S}^*_{k3,2}| \ge \epsilon/2) \le nP\left(d^2_{\mathcal{X}_k}(x_{ki}, o_k) + d^2_{\mathcal{Y}}(\mathbf{y}_i, o) > M/2\right) \le 2nC\exp(-sM/4),$$

which means

$$P(|\widehat{S}^*_{k3,2} - S_{k3,2}| > \epsilon) \le 2nC\exp(-sM/4). \tag{C.14}$$

Combining (C.13) and (C.14), we have

$$\begin{aligned}P(|\widehat{S}^*_{k3} - S_{k3}| \ge 2\epsilon) &\le P(|\widehat{S}^*_{k3,1} - S_{k3,1}| \ge \epsilon) + P(|\widehat{S}^*_{k3,2} - S_{k3,2}| \ge \epsilon)\\ &\le 2\exp(-2c^{-2}\epsilon^2 n^{1-2r}/3) + 2nC\exp(-scn^\gamma/4).\end{aligned}$$

It is easy to show that

$$\widehat{S}_{k3} = \frac{1}{n^3}\sum_{i,j,l} d_{\mathcal{X}_k}(x_{ki}, x_{kl}) d_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l) = \frac{(n-1)(n-2)}{n^2}\left(\widehat{S}^*_{k3} + \frac{1}{n-2}\widehat{S}^*_{k-1}\right).$$

Then,

$$\begin{aligned}P(|\widehat{S}_{k3} - S_{k3}| \ge 4\epsilon) &= P\left(\left|\frac{(n-1)(n-2)}{n^2}\left(\widehat{S}^*_{k3} + \frac{1}{n-2}\widehat{S}^*_{k1}\right) - S_{k3}\right| \le 4\epsilon\right)\\ &= P(|(n-1)(n-2)/n^2(\widehat{S}^*_{k3} - S_{k3}) - (3n-2)/n^2 S_{k2}\\ &\qquad + (n-1)/n^2(\widehat{S}^*_{k1} - S_{k1}) + (n-1)/n^2 S_{k1}| \ge 4\epsilon)\\ &\le P(|\widehat{S}^*_{k3} - S_{k3}| \ge \epsilon) + P(|\widehat{S}^*_{k1} - S_{k1}| \ge \epsilon)\\ &\le 2\exp(-c^{-2}\epsilon^2 n^{1-2\gamma}/6) + 2nC\exp(-scn^\gamma/4)\\ &\quad + 2\exp(-c^2\epsilon^2 n^{1-2\gamma}/4) + 2nC\exp(-scn^\gamma/4)\\ &\le 4\exp(-c^{-2}\epsilon^2 n^{1-2\gamma}/6) + 4nC\exp(-scn^\gamma/4),\end{aligned} \tag{C.15}$$

where the first inequality holds when n is large enough such that $(3n-2)/n^2 S_{k3} \le \epsilon$ and $(n-1)/n^2 S_{k1} \le \epsilon$.

Combining (C.7), (C.12) and (C.15), we have

$$\begin{aligned}P(|\widehat{\omega}_k - \omega| \ge \epsilon) &= P(|(\widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}) - (S_{k1} + S_{k2} - 2S_{k3})| \ge \epsilon)\\ &\le P(|\widehat{S}_{k1} - S_{k1}| \ge \epsilon/4) + P(|\widehat{S}_{k2} - S_{k2}| \ge \epsilon) + P(|\widehat{S}_{k3} - S_{k3}| \ge \epsilon/4)\\ &\le 2\exp(-\epsilon^2 c^{-2} n^{1-2\gamma}/4^4) + 2nC\exp(-scn^\gamma/4)\\ &\quad + 8\exp(-C^2\epsilon^2 c^{-2} n^{1-2\gamma}/(16\cdot 3^2\cdot 4^2)) + 8nC\exp(-sc^2 n^{2\gamma}/4)\\ &\quad + 4\exp(-c^{-2}\epsilon^2 n^{1-2\gamma}/(6\cdot 4^2\cdot 4^2)) + 4nC\exp(-scn^\gamma/4).\end{aligned}$$

Therefore,

$$P(|\widehat{\omega}_k - \omega| \ge \epsilon) \le O(\exp(-c_1\epsilon^2 n^{1-2\gamma}) + n\exp(-c_2 n^\gamma)), \tag{C.16}$$

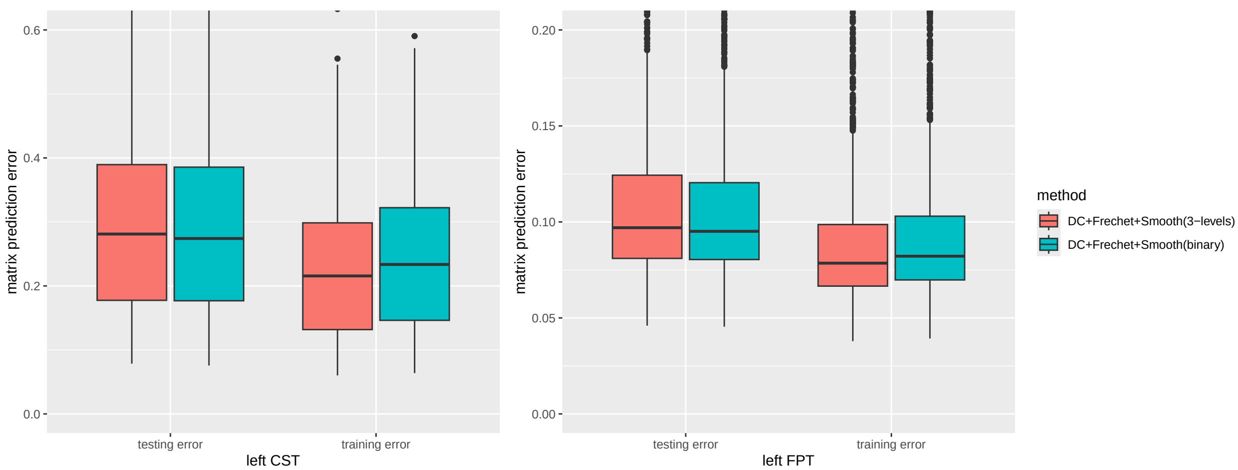


Figure E.1: Boxplots of mean matrix prediction errors measured in Cholesky distances, for the $10 \times 10 \times 10$ regions of left CST and left FPT. The red and blue boxplots correspond to SNPs, coded as a 3-level categorical variable, and binary variable, respectively. The method used is DC+Fréchet+Smooth with the model size selected by BIC.

for some constants c_1 and c_2. Let $\epsilon = cn^{-\kappa}$, where $0 < \kappa + \gamma < 0.5$. We have

$$\begin{aligned} P\left(\max_{1\le k\le p} |\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\right) &\le p \max_{1\le k\le p} P(|\widehat{\omega}_k - \omega| \ge cn^{-\kappa}) \\ &\le O(p[\exp(-c_1 n^{1-2(\gamma+\kappa)}) + n\exp(-c_2 n^{\gamma})]). \end{aligned} \tag{C.17}$$

Thus, we proved the first part of Theorem 1.

Now we prove the second part. Under assumption 3, distance covariance can be used to test independence. According to definition 1, if $\mathcal{I} \not\subseteq \widehat{\mathcal{I}}$, there exist some $k \in \mathcal{I}$ such that $\widehat{\omega}_k < cn^{\kappa}$. It follows from assumption 2 that $|\widehat{\omega}_k - \omega_k| \ge cn^{\kappa}$. Then, we have

$$\{\mathcal{I} \not\subseteq \widehat{\mathcal{I}}\} \subseteq \{|\widehat{\omega}_k - \omega_k| > cn^{-\kappa} \text{ for some } k \in \mathcal{I}\}$$

Let $A_n = \{\max_{k\in\mathcal{I}} |\widehat{\omega}_k - \omega_k| \le cn^{-\kappa}\} \subseteq \{\mathcal{I} \subseteq \widehat{\mathcal{I}}\}$. Thus,

$$\begin{aligned} P(\mathcal{I} \subseteq \widehat{\mathcal{I}}) &\ge P(A_n) \\ &= 1 - P(A_n^c) \\ &\ge 1 - |\mathcal{I}| P(|\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}) \\ &\ge 1 - O(|\mathcal{I}|[\exp(-c_1 n^{1-2(\gamma+\kappa)}) + n\exp(-c_2 n^{\gamma})]). \end{aligned} \tag{C.18}$$

Supplement D: Simulation Results

Supplement E: Tables and Figures from additional analysis of the ADNI data

Table D.1: The selection rates of each variables $\mathcal{P}_s \times 100$ and selection rates of all active variables $\mathcal{P}_a \times 100$. The d_1, d_2 are model sizes when we use DC. For sphere data, we let $n = 100, p = 9$ due to the inefficiency of FRiSO, and the important variables are (X_1, X_5, X_9). For all the other models, $n = 200, p = 20$. In the Time column, we present selection time + fitting time for DC method.

Model	Size/Method	X_1	X_5	X_9	X_{13}	X_{17}	X_{20}	All	Time(sec)
SPD Continuous	$d_1 = 6$	98	99	81	67	79	81	22	0.79+0.45
	$d_2 = 9$	100	100	97	92	97	95	82	0.79+0.46
	FRiSO $\bar{d} = 6$	100	100	100	100	100	100	100	1193
SPD Categorical	$d_1 = 6$	99	100	84	89	90	79	48	0.76+0.45
	$d_2 = 8$	100	100	95	96	98	93	82	0.76+0.46
	FRiSO $\bar{d} = 6$	100	100	100	100	100	100	100	1180
SPD Mixed	$d_1 = 6$	100	100	91	89	90	91	63	0.74+0.45
	$d_1 = 8$	100	100	97	100	98	99	94	0.74+0.46
	FRiSO $\bar{d} = 6.1$	100	100	100	100	100	100	100	1114
Probability Density	$d_1 = 6$	100	72	73	73	79	78	11	9.81+0.04
	$d_1 = 12$	100	95	98	99	97	97	86	9.81+0.06
	FRiSO $\bar{d} = 12.5$	100	100	100	100	100	100	100	5841
Sphere Data	$d_1 = 3$	99	97	100				96	0.03+4.66
	$d_2 = 4$	100	100	100				100	0.03+5.03
	FRiSO $\bar{d} = 3.4$	100	100	100				100	24695

Table E.1: Frequencies (expressed as percentages) of the top 20 most frequently selected genes in addition to other covariates (age, group and gender) corresponding to the left CST region, obtained from the DC method. SNPs are coded as a 3-level categorical variable. The underlined SNPs are those that are related to the progression of AD from previous research, while the bolded SNPs (corresponding to the columns for the competing LASSO+Cholesky, and LASSO+FA methods) imply that they were also selected by the DC method.

DC		LASSO + Cholesky		LASSO + FA	
Name	frequency(%)	Name	frequency(%)	Name	frequency(%)
age	73.5	age	79.0	age	34.4
group	35.6	gender	35.9	gender	9.7
rs3737964	28.0	**rs7095427**	19.6	**rs3737964**	6.6
gender	26.9	**rs3737964**	19.4	**rs2756271**	4.7
rs6584778	26.5	**rs13288561**	19.0	**rs3093660**	4.4
rs11594752	26.3	**rs2756271**	18.7	**rs6584778**	4.1
rs11814145	24.0	rs28455654	18.1	**rs2245123**	4.0
rs2245123	21.5	rs10502262	17.8	**rs13288561**	3.8
rs7095427	20.4	**rs3093660**	17.7	**rs10746816**	3.7
rs11193188	20.1	rs2239942	17.5	rs383902	3.7
<u>rs10491052</u>	19.5	**rs1475524**	17.4	rs11117959	3.4
rs13288561	19.4	rs1941375	16.7	rs1427282	3.4
<u>rs1133174</u>	19.1	rs2025677	16.3	**rs11594752**	3.3
rs4846051	18.0	rs2279339	16.1	rs1883235	3.0
rs1475524	17.8	**rs11594752**	15.9	**rs4846051**	3.0
rs10746816	17.3	rs383902	15.8	**rs7095427**	2.9
rs2756271	17.3	rs2418834	15.8	**rs1475524**	2.9
<u>rs4918255</u>	17.0	rs4936637	15.3	rs3026845	2.8
rs3093660	16.7	rs3026845	15.2	**<u>rs10491052</u>**	2.8
rs2569537	16.5	rs7101373	15.2	rs11817518	2.8
		group	5.7	group	0.6

Table E.2: Frequencies (expressed as percentages) of the top 20 most frequently selected genes in addition to other covariates (age, group and gender) corresponding to the left FPT region, obtained from the DC method. SNPs are coded as a 3-level categorical variable. The underlined SNPs are those that are related to the progression of AD from previous research, while the bolded SNPs (corresponding to the columns for the competing LASSO+Cholesky, and LASSO+FA methods) imply that they were also selected by the DC method.

DC		LASSO + Cholesky		LASSO + FA	
Name	frequency(%)	Name	frequency(%)	Name	frequency(%)
age	74.0	age	79.9	age	40.8
gender	41.5	gender	47.8	gender	16.1
group	38.4	**rs28455654**	20.0	**<u>rs11218301</u>**	3.9
rs6584778	23.2	rs2025677	17.9	**rs2279339**	3.8
rs1883235	23.2	**rs7095427**	17.7	rs7769173	3.7
rs2245123	21.3	**<u>rs7945931</u>**	17.6	rs1941375	3.7
rs4846051	19.8	rs1941375	17.5	**<u>rs7945931</u>**	3.6
rs2279339	18.4	**rs2279339**	17.3	**rs1883235**	3.5
<u>rs7945931</u>	18.3	rs2239942	16.9	rs13288561	3.5
rs2274606	17.7	rs13288561	16.9	rs383902	3.3
<u>rs11218301</u>	17.4	**rs1883235**	16.6	rs1473180	3.2
rs3118862	17.1	rs1475524	16.5	rs3093663	3.2
rs720273	16.6	**rs6584778**	16.3	rs1014306	3.1
rs84853	16.5	rs1473180	15.9	rs744970	3.1
<u>rs1133174</u>	16.4	rs13430599	15.6	**rs7095427**	3.0
rs10746816	16.2	**rs3118862**	15.5	rs11678252	3.0
<u>rs688</u>	16.2	**rs720273**	15.5	**rs4846051**	3.0
rs28455654	16.1	rs1049432	15.4	**rs28455654**	3.0
rs7095427	15.9	rs7101373	15.3	rs10502262	2.9
<u>rs5930</u>	15.4	**<u>rs5930</u>**	15.2	rs13430599	2.9
		group	6.3	group	0.9

Table E.3: Frequencies (expressed as percentages) of the top 20 most frequently selected genes in addition to other covariates (age, group and gender) corresponding to the right CST region, obtained from the DC method. SNPs are coded as a 3-level categorical variable. The underlined SNPs are those that are related to the progression of AD from previous research, while the bolded SNPs (corresponding to the columns for the competing LASSO+Cholesky, and LASSO+FA methods) imply that they were also selected by the DC method.

DC		LASSO + Cholesky		LASSO + FA	
Name	frequency(%)	Name	frequency(%)	Name	frequency(%)
age	73.9	**age**	80.0	**age**	32.7
gender	31.9	**gender**	39.2	**gender**	11.2
group	31.1	**rs7095427**	21.4	rs3737964	5.1
rs7095427	24.5	rs3093660	19.4	rs13430599	4.4
<u>rs1133174</u>	24.0	rs13430599	18.5	rs3093660	4.0
rs11594752	23.4	rs2756271	18.3	<u>rs2756271</u>	3.9
rs3737964	22.1	rs2239942	17.8	**rs7095427**	3.9
rs10746816	21.8	rs1427282	17.5	rs11117959	3.7
rs6584778	21.4	rs10502262	17.4	rs3026845	3.5
rs3026845	20.3	<u>rs1133174</u>	17.3	rs720273	3.4
rs11814145	19.6	rs10503814	17.3	rs10503814	3.3
rs2756271	17.5	rs3026845	17.1	rs383902	3.3
rs11193188	17.4	rs28455654	16.8	rs2239942	3.1
rs2245123	17.1	rs1941375	16.7	rs1427282	3.1
rs1427282	17.0	rs3737964	16.6	rs1475524	3.1
rs13430599	17.0	rs13288561	16.6	rs6584778	3.0
rs10503814	16.8	rs3118862	16.2	rs1883235	3.0
rs1475524	16.7	rs1475524	15.9	rs11594752	2.9
rs3093660	16.6	rs11117959	15.5	<u>rs16871253</u>	2.9
rs11817518	16.6	rs1473180	15.3	<u>rs1133174</u>	2.9
		group	4.4	group	0.3

Table E.4: Frequencies (expressed as percentages) of the top 20 most frequently selected genes in addition to other covariates (age, group and gender) corresponding to the right FPT region, obtained from the DC method. SNPs are coded as a 3-level categorical variable. The underlined SNPs are those that are related to the progression of AD from previous research, while the bolded SNPs (corresponding to the columns for the competing LASSO+Cholesky, and LASSO+FA methods) imply that they were also selected by the DC method.

DC		LASSO + Cholesky		LASSO + FA	
Name	frequency(%)	Name	frequency(%)	Name	frequency(%)
age	72.8	**age**	82.2	**age**	42.0
gender	43.5	**gender**	49.1	**gender**	17.2
group	35.6	rs7095427	20.3	rs7095427	4.4
rs6584778	28.6	rs1941375	19.5	11218301	4.1
rs2245123	24.0	rs7945931	19.3	rs1883235	3.9
rs2274606	23.4	rs6584778	18.4	rs7945931	3.9
rs1883235	22.3	rs2756271	18.3	rs7769173	3.9
rs1133174	20.6	rs4878104	17.6	rs1891385	3.8
rs10746816	20.4	rs5930	17.5	rs4846051	3.7
rs720273	18.4	rs2239942	17.5	rs10503814	3.7
rs5930	18.3	rs28455654	17.4	rs4878104	3.6
rs7095427	17.9	rs13430599	17.3	rs6584778	3.4
rs4846051	17.7	rs1427282	17.1	rs1941375	3.4
rs7945931	17.4	rs13288561	17.1	rs2274606	3.4
rs3118862	17.0	rs1883235	17.0	rs7901596	3.3
rs4878104	16.9	rs1571344	16.7	rs2239942	3.3
rs12245545	16.8	rs720273	16.7	rs720273	3.3
rs6457131	16.8	rs10502262	16.6	rs12034383	3.3
rs1571344	16.6	rs7901596	16.6	rs12001404	3.3
rs3026845	16.1	rs3118862	16.5	rs10502262	3.2
		group	5.6	group	0.8

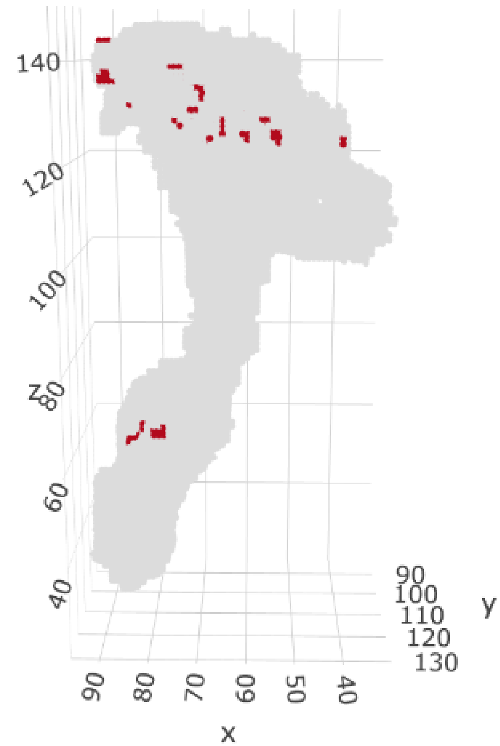


Figure E.2: Plots of the differences in prediction between a real subject (73 year old woman) in the Control group from the ADNI data and a fake subject in the AD group (with same covariates as the real subject), at selected voxels (denoted in red) in the right CST region, obtained by fitting our method (with/without smoothing).

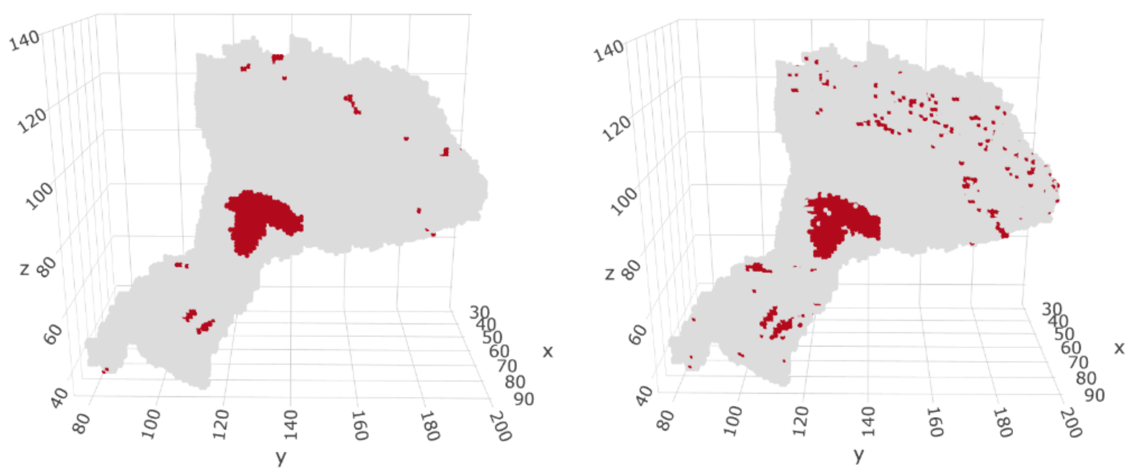
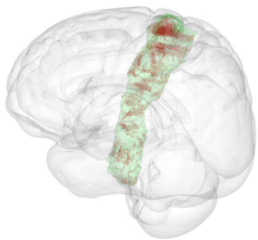
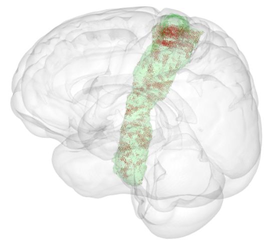


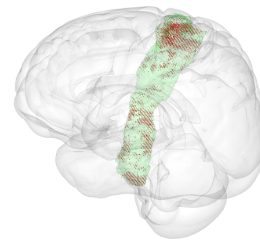
Figure E.3: Plots of the differences in prediction between a real subject (73 year old woman) in the Control group from the ADNI data and a fake subject in the AD group (with same covariates as the real subject), at selected voxels (denoted in red) in the right FPT region, obtained by fitting our method (with/without smoothing).



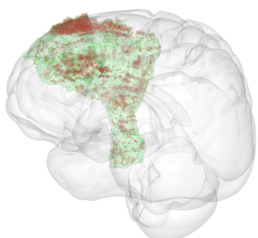
(a) left CST | rs6584778



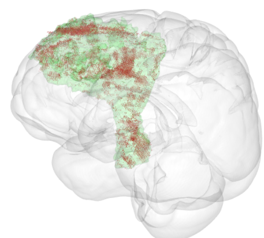
(b) left CST | rs7095427



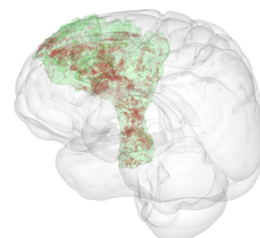
(c) left CST | rs11193188



(d) left FPT | rs2245123



(e) left FPT | rs4846051



(f) left FPT | rs7945931

Figure E.4: Plots of selected voxels (locations) corresponding to specific SNPs (`rs6584778`, `rs7095427`, `rs11193188`, `rs2245123`, `rs4846051` and `rs7945931`) for left CST (upper panel) and left FPT (lower panel), respectively, obtained from fitting our DC model. While the selected voxels appear in red, the brain ROI are colored green.

radius	neighborhood size	(a)	(b)	(c)	(d)	(e)	(f)
$r = 1$	26	17.1	9.4	8.4	13.1	9.7	5.0
$r = 2$	124	67.7	33.2	28.9	45.7	36.0	14.4
$r = 3$	342	160.6	74.7	62.4	100.8	84.9	29.2

Table E.5: The contiguity measure ρ for various radii, representing the average number of selected voxels within a neighborhood of radius r around each selected voxel. A larger ρ indicates more contiguous regions, with the maximum possible ρ equal to the number of voxels within the neighborhood of radius r, as shown in the second column. The columns correspond to the subfigures in Figure 3. Subfigures (a) and (d) correspond to regions selected by DC; (b) and (e) correspond to regions selected by LASSO+Cholesky; (c) and (f) correspond to regions selected by LASSO+FA.

References

Dehling, H., Matsui, M., Mikosch, T., Samorodnitsky, G., and Tafakori, L. (2020). Distance covariance for discretized stochastic processes. *Bernoulli* **26,** 2758 – 2789.

Jakobsen, M. E. (2017). *Distance Covariance in Metric Spaces: Non-Parametric Independence Testing in Metric Spaces.* Master's thesis, University of Copenhagen.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41,** 3284–3305.

Lyons, R. (2014). Hyperbolic space has strong negative type. *Illinois Journal of Mathematics* **58,** 1009–1013.

Lyons, R. (2020). Strong negative type in spheres. *Pacific Journal of Mathematics* **307,** 383–390.

Lyons, R. (2021). Second errata to "distance covariance in metric spaces". *The Annals of Probability* **49,** 2668–2670.

Meckes, M. W. (2013). Positive definite metric spaces. *Positivity* **17,** 733–757.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics* **35,** 2769–2794.