Detecting influential observations in single-index Fréchet regression

Abdul-Nasah Soale

Department of Mathematics, Applied Mathematics & Statistics,
Case Western Reserve University, OH, USA

Seminar Talk

Department of Economics, University of Crete
Gallos Campus, Rethymnon, Greece

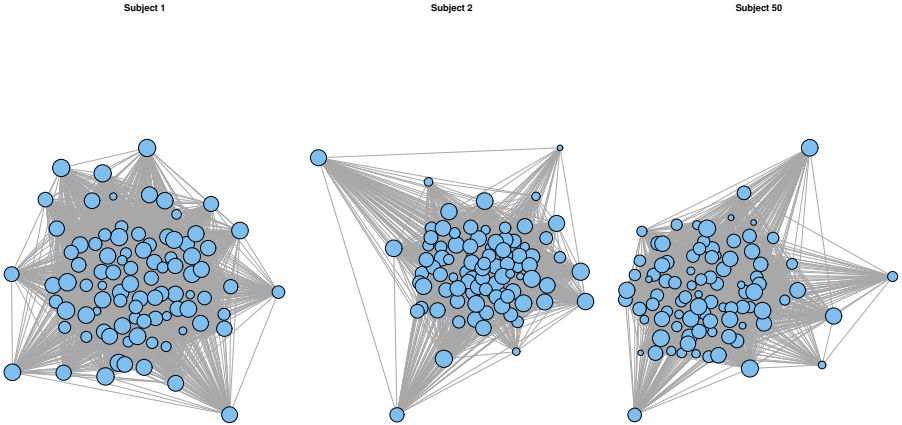
May 7, 2025

Table of Contents

Table of Contents

Body mass and human brain structural connectivity

- Consider the following brain structural connectivity networks of three healthy subjects:



- Does any of these networks look anomalous?

Body mass and human brain structural connectivity

- Suppose for each subject, we have additional information as shown in the Table below:

Subject	Age (years)	Weight (kg)	Height (m)
1	24.8	58	1.70
2	27.7	57	1.65
50	39.9	108	1.71

- In terms of network density, which of the networks is an outlier based on the subjects weight?

Body mass and human brain structural connectivity

- Suppose for each subject, we have additional information as shown in the Table below:

Subject	Age (years)	Weight (kg)	Height (m)
1	24.8	58	1.70
2	27.7	57	1.65
50	39.9	108	1.71

- In terms of network density, which of the networks is an outlier based on the subjects weight?

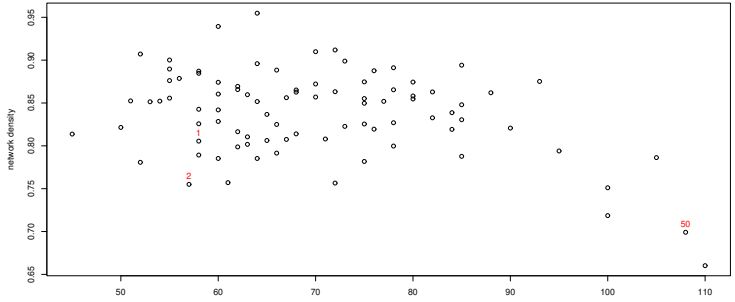


Table of Contents

Brief introduction to Fréchet regression

- Let $Y \in (\mathbf{\Omega}_Y, d)$ denote a random response, where the metric $d : \Omega_Y \times \Omega_Y \to \mathbb{R}$, and $\mathbf{X} \in \mathbb{R}^p$ be the predictor.

- Let F_X and F_Y be the respective marginal distributions of $\mathbf{X}$ and $\mathbf{Y}$.

- For a random object $(\mathbf{X}, Y) \in \mathbb{R}^p \times \Omega_Y$, assume the conditional distribution $F_{Y|X}$ exists and is well-defined.

- Petersen and Müller (2019) defined the conditional Fréchet mean of Y given $\mathbf{X}$ as

$$m_{\oplus}(\boldsymbol{x}) = \underset{\omega \in \Omega_Y}{\arg\min}\, \mathrm{E}\left(d^2(Y, \omega)|\mathbf{X} = \boldsymbol{x}\right). \tag{1}$$

Common issues in Fréchet regression

- Equation (1) is analogous to $\mathrm{E}(Y|\mathbf{X} = \boldsymbol{x})$.
- Thus, Fréchet regression inherits most of the common problems associated with classical regression such as the "curse of dimensionality" as p grows and the influence of outliers.
- Sufficient dimension reduction (SDR) methods have been proposed to addressed the dimensionality of $\mathbf{X}$. See [Dong and Wu, 2022], [Zhang et al., 2023], and [Soale and Dong, 2023].
- The SDR methods for Fréchet regression are also susceptible to outliers.

Table of Contents

Single-index Fréchet regression

- For $Y \in \mathbf{\Omega}_Y$ and $\mathbf{X} \in \mathbb{R}^p, p \geq 1$, consider the single-index model:

$$Y = f(\beta^\top \mathbf{X}, \epsilon), \tag{2}$$

  where $f(.)$ is some unknown link function, $\beta \in \mathbb{R}^p$ with $\|\beta\|_2 = 1$, and $\|.\|$ denotes the ℓ_2 norm. ϵ is some noise with $\mathrm{E}(\epsilon|\mathbf{X}) = 0$.

- In model 2, Y depends on $\mathbf{X}$ only through $\beta^\top \mathbf{X}$.

- Our goal is to find the mean space, i.e., a β that satisfies

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\beta^\top \mathbf{X}, \text{ or equivalently, } \mathrm{E}(Y|\mathbf{X}) = \mathrm{E}(Y|\beta^\top \mathbf{X}). \tag{3}$$

- β is not unique but $\mathcal{S}_{E(Y|\mathbf{X})} = \mathrm{span}(\beta)$ if it exists, is identifiable, and we call it the central mean space.

Responses in non-Euclidean spaces

- The ordinary least squares (OLS) is one of the most popular methods for estimating the central mean space in single-index models.
- OLS is a great tool for detecting *influential* observations.
- Y may belong to a non-Euclidean space where the classical OLS does not apply.
- We propose a surrogate OLS based on Euclidean embedding for detecting influential observations in Fréchet regression.

Surrogate OLS

Lemma

Let $\tilde{Y}$ be a random copy of $Y \in (\mathbf{\Omega}_Y, d)$. Define the surrogate $S_Y = \phi\big(d(Y, \tilde{Y})\big)$, for some measurable function $\phi : d(Y, \tilde{Y}) \to \mathbb{R}$. Then, $\mathcal{S}_{S_Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

Theorem

Suppose the link function f is monotone and $\mathsf{E}(\mathbf{X}|\beta^\top\mathbf{X})$ is a linear function of $\beta^\top\mathbf{X}$. If $\mathbf{\Sigma} = Var(\mathbf{X})$ is invertible, then

$$\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{XS_Y} \in \mathcal{S}_{E(Y|\mathbf{X})}, \text{ where } \mathbf{\Sigma}_{XS_Y} = Cov(\mathbf{X}, S_Y). \tag{4}$$

Metric Cook's Distance algorithm

Algorithm 1 Metric Cook's Distance

1. Input: Predictor $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response $Y_n = \{y_1, \ldots, y_n\}$ as list
2. Compute $\mathbf{D}_Y \in \mathbb{R}^{n \times n}$, where $\mathbf{D}_{Y_{ij}} \leftarrow d(y_i, y_j),\ \forall i, j = 1 \ldots, n$
3. Compute the first MDS factor score based on $\mathbf{D}_Y$ and set as S_Y
4. Compute the OLS estimate $\tilde{\beta}$ from the regression between S_Y and $\mathbf{X}$
5. Repeat Steps 2-4 after deleting the ith observation and compute the ith Cook's distance.

The ith Cook's distance is defined as

$$\delta_i = \frac{(\tilde{\beta}^{(-i)} - \tilde{\beta})^\top (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})(\tilde{\beta}^{(-i)} - \tilde{\beta})}{(p+1)s^2}, \text{ where } s^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (S_{Y_i} - \tilde{\beta}^\top \tilde{\mathbf{X}}_i)^2,$$

and $\tilde{\mathbf{X}} = [\mathbf{1}, \mathbf{X}]$.

Table of Contents

Metric Cook's distance examples

- We will consider responses in different spaces.
- An observation is considered influential if $\delta_i > \dfrac{4}{n-p-1}$.
- For each case, we show the effect of omitting the influential observations on the accuracy of β estimates using

$$\Delta = \|\mathbf{P}_{\beta} - \mathbf{P}_{\hat{\beta}}\|_F, \tag{5}$$

  where $\mathbf{P_A} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$ and $\|.\|_F$ is the matrix Frobenius norm.

- Smaller values of Δ indicate better performance.

Regression with Euclidean response

- Fix $(n, p) = (100, 5)$ and then generate the predictor as $\mathbf{X} = (X_1, \ldots, X_5) \sim t_{20}(\mathbf{0}, \mathbf{I}_5)$ and set $\beta^\top = (1, 1, 0, 0, 0)/\sqrt{2}$.
- Generate each response as $\boldsymbol{Y}_i = \sin(\pi/2 + \mathbf{B}^\top \mathbf{X}_i) + \boldsymbol{\epsilon}_i$, for $i = 1, \ldots, n$, where $\mathbf{B}^\top = (\beta, -2\beta)^\top$ and noise $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with $\boldsymbol{\Sigma}_\epsilon = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
- The effect of omitting influential observations on the surrogate-assisted OLS of [Soale and Dong, 2023].

Table: * denotes estimates without influential observations

	x_1	x_2	x_3	x_4	x_5	Δ
sa-OLS	-0.6113	-0.5810	0.1662	-0.0724	0.3617	0.6125
sa-OLS*	-0.5149	-0.7949	-0.0996	-0.2098	0.2345	0.5428

Regression with Euclidean response

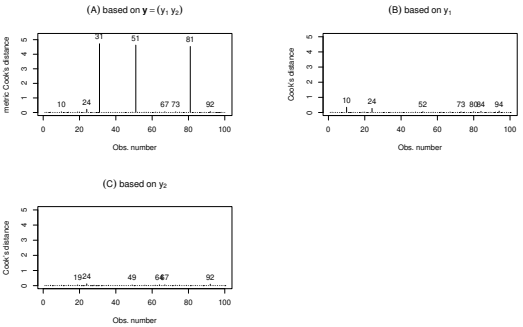


Figure: Cook's distances in (A) are based on OLS regression with S^Y as response while those for exhibits (B) and (C) are based on separate regressions with y_1 and y_2, respectively.

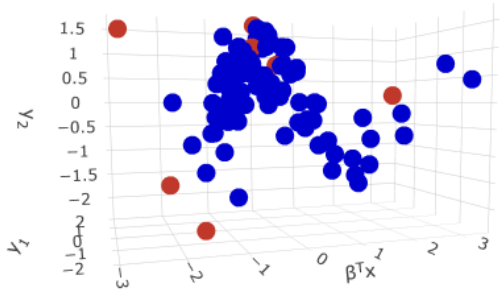


Figure: Scatter plot of the bivariate response $\boldsymbol{y}$ versus the sufficient direction $\beta^\top \boldsymbol{x}$. Influential observations based on the metric Cook's distance are colored red.

Regression with distribution as response

- We use the Wasserstein metric to find the pairwise distances.
- We focus on only univariate distributions.
- The kth Wasserstein is given by

$$W_k(y, y') = \|F_y^{-1} - F_{y'}^{-1}\|_k = \left(\int_0^1 \left| F_y^{-1}(s) - F_{y'}^{-1}(s) \right|^k ds \right)^{1/k}, \tag{6}$$

  where F_y^{-1} and $F_{y'}^{-1}$ are the respective quantile functions of distributions y and y'.

- We use $W_1(.)$ instead of $W_2(.)$ used in previous studies.

Regression with distribution as response

- Fix $(n, p) = (100, 5)$ and then generate the predictor as $\mathbf{X} = (X_1, \ldots, X_5) \sim t_{20}(\mathbf{0}, \mathbf{I}_5)$ and set $\beta^\top = (1, 1, 0, 0, 0)/\sqrt{2}$.
- Generate each response as a mixture of normal distributions as follows:

$$Y_i \in \mathbb{R}^{100} \overset{i.i.d.}{\sim} 0.6N(\beta^\top \mathbf{X}_i, 1) + 0.4N(-\beta^\top \mathbf{X}_i, 2), \text{ for } i = 1, \ldots, n.$$

- The effect of omitting influential observations on the Fréchet OLS of [Zhang et al., 2023] and the surrogate-assisted OLS of [Soale and Dong, 2023].

Table: * denotes estimates without influential observations

	x_1	x_2	x_3	x_4	x_5	Δ
FOLS	0.7752	0.5521	-0.1268	0.0155	0.0135	0.2995
FOLS*	0.8237	0.7880	-0.0164	-0.0087	-0.0289	0.0529
sa-OLS	-0.7673	-0.5326	0.1395	-0.0634	0.0151	0.3382
sa-OLS*	0.8284	0.7707	-0.0325	0.0203	-0.0466	0.0909

Regression with distribution as response

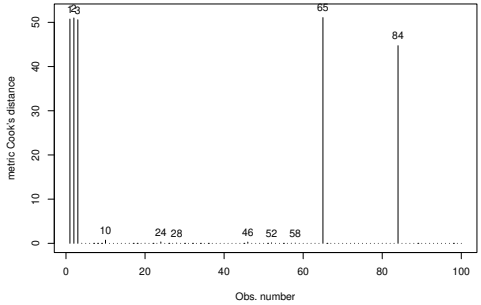


Figure: Metric Cook's distances for distributional response regression

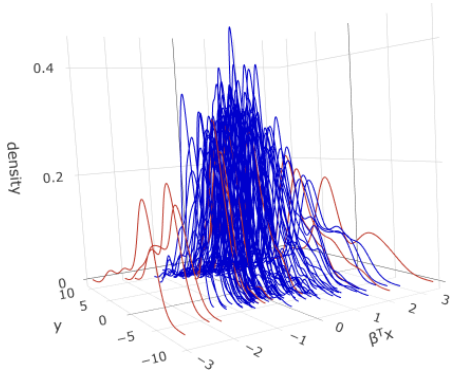


Figure: 3D plot of distributions vs sufficient predictor $\beta^\top \boldsymbol{x}$. The red lines indicate the influential observations based on the metric Cook's distance.

Regression with network as response

- We denote the response network or graph as $G = (V, E)$, where V is the set of vertices or nodes and E is the set of edges or links
- We employ two distance measures: the *centrality distance (CD)* of [Roy et al., 2014] and the *graph diffusion distance (DD)* of [Hammond et al., 2013].
- The centrality distance is based on unweighted edges and is given by

$$d_{CD}(G_1, G_2) = \sum_{v \in V} |C(G_1, v) - C(G_2, v)|. \tag{7}$$

- The diffusion distance incorporate the weight of the edges and is given by

$$d_{DD,t}(G_1, G_2) = \max_{t \in (0,1)} \left(\| e^{-t\boldsymbol{L}_{G_1}} - e^{-t\boldsymbol{L}_{G_2}} \|_F^2 \right)^{1/2}, \tag{8}$$

where for graph G_i, the Laplacians $\boldsymbol{L}_{G_i} = \mathbf{D}_{G_i} - \mathbf{A}_{G_i}$, $\mathbf{D}_{G_i}$ is the diagonal matrix of degrees and $\mathbf{A}_{G_i}$ is the adjacency matrix.

Regression with network as response

- Fix $(n, p) = (100, 5)$ and then generate the predictor as $\mathbf{X} = (X_1, \ldots, X_5) \sim t_{20}(\mathbf{0}, \mathbf{I}_5)$ and set $\beta^\top = (1, 1, 0, 0, 0)/\sqrt{2}$.
- Set the number of nodes to 20 and generate the random network response y_i based on an Erdös–Rényi model with probability $p = \mathit{plogis}\big(\sin(\beta^\top \mathbf{X}_i)\big),\ i = 1, \ldots, n$, where $\mathit{plogis}(x) = \dfrac{1}{1 + e^{-x}}$.
- The effect of omitting influential observations on the surrogate-assisted OLS of [Soale and Dong, 2023]. We did not consider FOLS as it was not implemented with network responses in [Zhang et al., 2023].

Table: * denotes estimates without influential observations

	x_1	x_2	x_3	x_4	x_5	Δ
sa-OLS (cd)	0.7049	0.7442	-0.0042	-0.0022	-0.1193	0.1680
sa-OLS (cd) *	0.8422	0.8164	-0.0246	-0.0101	-0.0897	0.1146
sa-OLS (dd)	-0.7043	-0.7372	0.0163	-0.0106	0.0835	0.1228
sa-OLS (dd)*	0.8515	0.8075	-0.0230	0.0129	-0.0635	0.0908

Regression with network as response

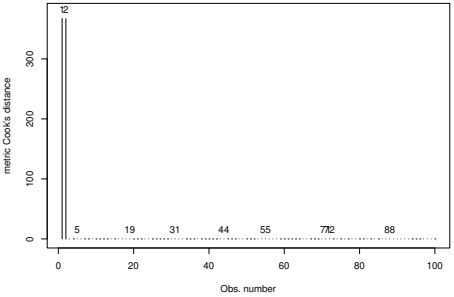


Figure: Metric Cook's distances for network response regression

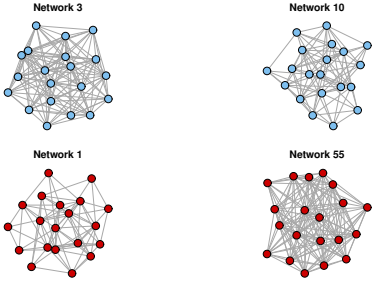


Figure: Four sample networks out of the 100 networks. The networks with red nodes indicate the influential/anomalous networks based on the metric Cook's distance.

Regression with functional (time-varying) response

- Fix $(n, p) = (100, 5)$ and then generate the predictor as $\mathbf{X} = (X_1, \ldots, X_5) \sim t_{20}(\mathbf{0}, \mathbf{I}_5)$ and set $\beta^\top = (1, 1, 0, 0, 0)/\sqrt{2}$.
- Let $\alpha(t) = 2\sin(\pi + \pi t/5)$, where $t \in \mathbb{R}^{30} \overset{i.i.d.}{\sim} Unif(0, 10)$. Generate the response as $Y_i(t) = \alpha(t) + 2\sin\left(\pi t/2 + \beta_1^\top \mathbf{X}_i\right) + \epsilon_i(t)$ for $i = 1, \ldots, n$, where $\epsilon_i(t) \overset{i.i.d.}{\sim} N(0, 1)$.
- We compute the distance between responses as the Euclidean distance between the discrete Fourier coefficients.
- The effect of omitting influential observations on the surrogate-assisted OLS of [Soale and Dong, 2023].

Table: * denotes estimates without influential observations

	x_1	x_2	x_3	x_4	x_5	Δ
sa-OLS	0.7533	0.5970	-0.0687	0.0197	0.0470	0.2047
sa-OLS*	0.7866	0.7817	0.0251	0.0293	-0.0506	0.0812

Regression with functional (time-varying) response

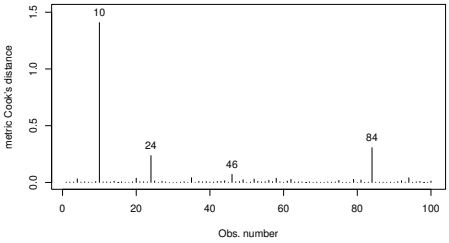


Figure: Metric Cook's distances for functional response regression

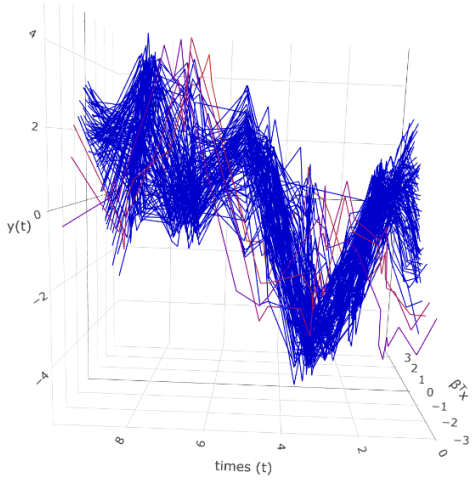


Figure: 3D plot of functional responses vs sufficient predictor $\beta^\top \boldsymbol{x}$. The red lines indicate the influential observations based on the metric Cook's distance.

Table of Contents

Demographics and COVID-19 transmission

- In this application, the responses are the distributions of total new COVID-19 cases per 100,000 persons in the last 7 days between 08/1/2021 and 02/21/2022 in the U.S.
- We focus on the county level transmissions for the State of Texas, which consists of 254 counties.
- The predictors are nine demographic characteristics of the counties from the 2020 American Community Survey.
- Both data sets are publicly available at [CDC, 2023] and [US Census Bureau, 2022], respectively.
- The following observations: 73, 110, 151, 171, and 181 corresponding to counties: Falls, Hockley, Loving, Moore, and Orange were found to be the most influential.

Demographics and COVID-19 transmission

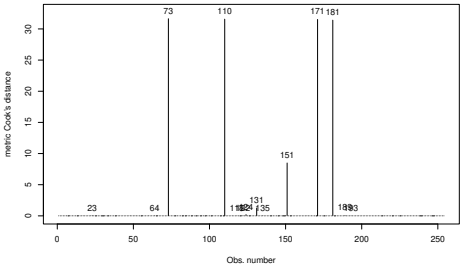


Figure: Influential observations based on metric Cook's distance for the regression between COVID-19 transmission distributions vs demographic characteristics.

Table: estimated bases of the central mean space. * denotes estimates after omitting influential observations.

	FOLS	FOLS*	sa-OLS	sa-OLS*
% Non-Hispanic Blacks	0.0612	0.1860	0.1728	0.3102
% Hispanics	-0.2771	-0.0339	-0.3974	-0.3431
% Adults 65+	-0.7254	-0.5820	-0.8304	-0.8659
% No high school diploma	0.3848	-0.0789	0.3949	-0.1724
% Living below poverty line	0.2294	-0.5378	0.0284	-0.1615
% Unemployed	-0.0225	0.3190	0.0455	0.2875
% Renter-occupied homes	-0.8263	0.6574	-0.6710	0.0731
% On public assistance	0.0359	-0.1459	0.0969	0.1479

Demographics and COVID-19 transmission

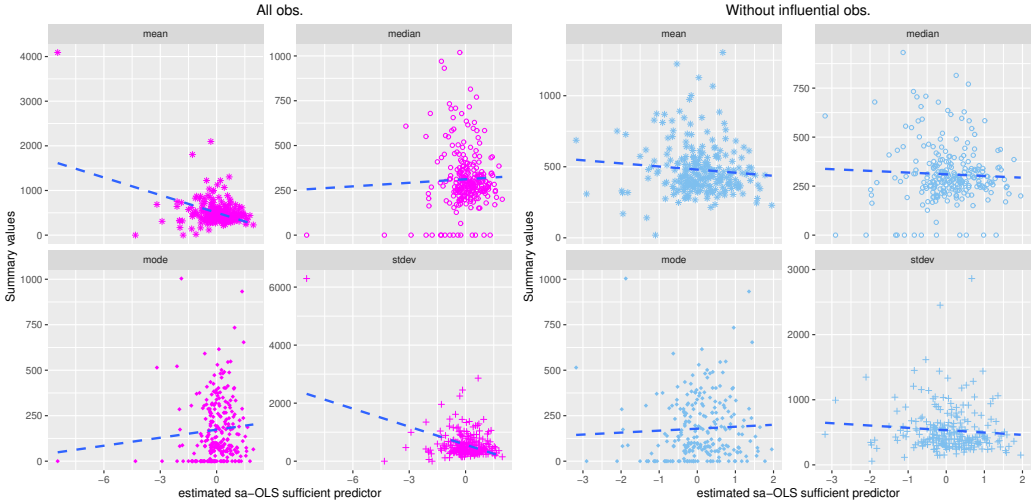


Figure: Summary plots for the regression between COVID-19 transmission distributions vs demographic characteristics with and without the influential observations. The blue dash lines indicate the fitted lines for the simple linear regression.

Body mass and human brain structural connectivity

- The study sample consists of 88 healthy individuals.
- The weighted connectivity matrices among 90 cortical regions of interests in the brain of the study subjects are taken as the responses.
- The predictors are age, weight, and height of the subjects.
- This data is also publicly available on the Open Science Framework (OSF) at https://osf.io/yw5vf/ and published in [Škoch et al., 2022].
- Subjects 7, 8, and 54 were found to have the most anomalous brain networks. Their respective (age, weight, heights) are (48.7yrs, 67kg, 1.76m), (37.7yrs ,62kg, 1.85m), and (45.3yrs, 60kg, 1.62m).

Body mass and human brain structural connectivity

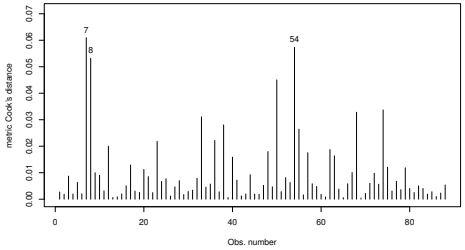


Figure: Influential observations for the regression between structural brain connectivity networks vs age, height, and weight of subjects.

Table: estimated bases of the central mean space. * denotes estimates after omiting influential observations.

	sa-OLS (CD)	sa-OLS (CD)*	sa-OLS (DD)	sa-OLS (DD)*
Age	-0.0805	-0.1258	-0.2234	-0.4726
Weight	-1.3446	-1.3182	-1.2794	-1.1156
Height	0.6452	0.6116	1.2263	1.1323

Body mass and human brain structural connectivity

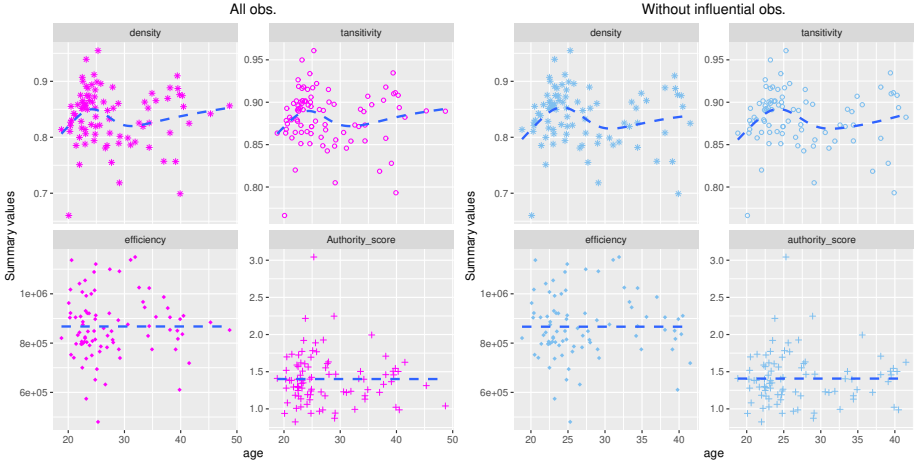


Figure: Summary plots for the regression between some properties of the structural brain connectivity networks vs age of subjects with and without the influential observations. The blue dash lines represent the LOESS curve.

Table of Contents

Conclusion

- The metric Cook's distance is applicable in a wide variety of applications including responses in Euclidean and non-Euclidean spaces.
- Influential observations can seriously impact the central mean space estimates.
- Rather than relying on hard thresholds for categorizing observations as influential, we advise investigating observations with large Cook's distances further.
- Omitting influential observations may not be the best way to handle outliers/anomalies. We advise investigators to delve deeper.
- The paper is published in Technometrics [Soale, 2025].
- More information on choosing the optimal metrics can be found in [Soale et al.,].

Thank You

CDC (2023).
United States COVID-19 County Level of Community Transmission Historical Changes. [Online; accessed 13-March-2023].

Dong, Y. and Wu, Y. (2022).
Fréchet kernel sliced inverse regression.
Journal of Multivariate Analysis, 191:105032.

Hammond, D. K., Gur, Y., and Johnson, C. R. (2013).
Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel.
In *2013 IEEE global conference on signal and information processing*, pages 419–422. IEEE.

Roy, M., Schmid, S., and Tredan, G. (2014).
Modeling and measuring graph similarity: The case for centrality distance.
In *Proceedings of the 10th ACM international workshop on Foundations of mobile computing*, pages 47–52.

Škoch, A., Rehák Bučková, B., Mareš, J., Tintěra, J., Sanda, P., Jajcay, L., Horáček, J., Španiel, F., and Hlinka, J. (2022).
Human brain structural connectivity matrices–ready for modelling.

Scientific Data, 9(1):486.

Soale, A.-N. (2025).
Detecting influential observations in single-index fréchet regression.
Technometrics, pages 1–12.

Soale, A.-N. and Dong, Y. (2023).
Data visualization and dimension reduction for metric-valued response regression.
arXiv preprint arXiv:2310.12402.

Soale, A.-N., Ma, C., Chen, S., and Koomson, O.
On metric choice in dimension reduction for fréchet regression.
International Statistical Review.

US Census Bureau (2022).
United States Census Bureau.
`https://www.census.gov/data.html`.
Accessed: 2022-10-26.

Zhang, Q., Xue, L., and Li, B. (2023).
Dimension reduction for fréchet regression.
Journal of the American Statistical Association, pages 1–15.