Integrative analysis of non-Euclidean data

James Buenfil

November 13, 2025

University of Washington

Integrative Data Analysis

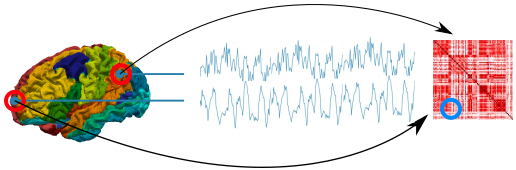
- **Goal:** understand relationships between two heterogeneous *data views*.

Integrative Data Analysis

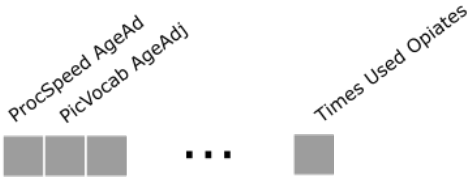
- **Goal:** understand relationships between two heterogeneous *data views*.
- **Motivating example:** fMRI (brain connectivity) + multivariate lifestyle/demographic/psychometric data.

Integrative Data Analysis

- **Goal:** understand relationships between two heterogeneous *data views*.
- **Motivating example:** fMRI (brain connectivity) + multivariate lifestyle/demographic/psychometric data.



Brain imaging data y_i for each patient, $i = 1, \ldots N$.



Multivariate data $x_i \in \mathbb{R}^p$ for each patient $i = 1, \ldots N$.

- Random variables: $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$; both centered.

Introduction to Canonical Correlation Analysis

- Random variables: $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$; both centered.
- Regression problem:

$$\beta_{\mathsf{OLS}} = \arg\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^2].$$

Introduction to Canonical Correlation Analysis

- Random variables: $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$; both centered.
- Regression problem:

$$\beta_{\mathsf{OLS}} = \arg\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^2].$$

- Canonical correlation analysis (CCA):

$$\theta_{\mathsf{CCA}} = \arg\max_{\theta} \operatorname{Corr}(Y, \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}\left(\theta^\top X\right) = 1.$$

Introduction to Canonical Correlation Analysis

- Random variables: $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$; both centered.
- Regression problem:

$$\beta_{\mathsf{OLS}} = \arg\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^2].$$

- Canonical correlation analysis (CCA):

$$\theta_{\mathsf{CCA}} = \arg\max_{\theta} \operatorname{Corr}(Y, \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}\left(\theta^\top X\right) = 1.$$

- CCA and regression find same direction: $\theta_{\mathsf{CCA}} \propto \beta_{\mathsf{OLS}}$

Multivariate CCA

- Univariate CCA:

$$\theta_{\mathsf{CCA}} = \arg\max_{\theta} \operatorname{Corr}(Y, \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}\left(\theta^\top X\right) = 1.$$

Solve

$$(\eta, \theta) = \arg\max_{\eta \in \mathbb{R}^q, \theta \in \mathbb{R}^p} \operatorname{Corr}(\eta^\top Y,\ \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}(\eta^\top Y) = \operatorname{Var}(\theta^\top X) = 1.$$

Multivariate CCA

- Univariate CCA:

$$\theta_{\mathsf{CCA}} = \arg\max_{\theta} \operatorname{Corr}(Y, \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}\left(\theta^\top X\right) = 1.$$

- Multivariate CCA: let Y be multivariate ($Y \in \mathbb{R}^q$).
  Solve

$$(\eta, \theta) = \arg\max_{\eta \in \mathbb{R}^q, \theta \in \mathbb{R}^p} \operatorname{Corr}(\eta^\top Y,\ \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}(\eta^\top Y) = \operatorname{Var}(\theta^\top X) = 1.$$

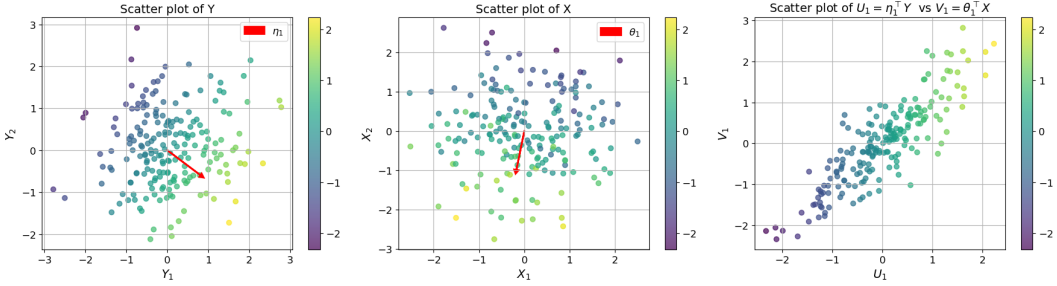
Multivariate CCA

- Univariate CCA:

$$\theta_{\text{CCA}} = \arg\max_{\theta} \operatorname{Corr}(Y, \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}\left(\theta^\top X\right) = 1.$$

- Multivariate CCA: let Y be multivariate ($Y \in \mathbb{R}^q$).
  Solve

$$(\eta, \theta) = \arg\max_{\eta \in \mathbb{R}^q, \theta \in \mathbb{R}^p} \operatorname{Corr}(\eta^\top Y,\ \theta^\top X) \quad \text{s.t.} \quad \operatorname{Var}(\eta^\top Y) = \operatorname{Var}(\theta^\top X) = 1.$$



Multivariate CCA - multiple canonical vectors

- Estimating *multiple* canonical vectors:

$$\max_{\{\eta_i,\theta_i\}_{i=1}^d} \sum_{i=1}^{d} \mathrm{Corr}\left(\eta_i^\top Y,\ \theta_i^\top X\right) \quad \text{s.t.} \quad \mathrm{Corr}(\eta_i^\top Y, \eta_j^\top Y) = \mathrm{Corr}(\theta_i^\top X, \theta_j^\top X) = \delta_{ij}.$$

Multivariate CCA - multiple canonical vectors

- Estimating *multiple* canonical vectors:

$$\max_{\{\eta_i,\theta_i\}_{i=1}^{d}} \sum_{i=1}^{d} \mathrm{Corr}\left(\eta_i^\top Y,\ \theta_i^\top X\right) \quad \text{s.t.} \quad \mathrm{Corr}(\eta_i^\top Y, \eta_j^\top Y) = \mathrm{Corr}(\theta_i^\top X, \theta_j^\top X) = \delta_{ij}.$$

- **Canonical vectors**, $H \equiv [\eta_1 \dots \eta_d] \in \mathbb{R}^{q\times d}$ and $T \equiv [\theta_1, \dots \theta_d] \in \mathbb{R}^{p\times d}$.

Multivariate CCA - multiple canonical vectors

- Estimating *multiple* canonical vectors:

$$\max_{\{\eta_i,\theta_i\}_{i=1}^{d}} \sum_{i=1}^{d} \mathrm{Corr}\left(\eta_i^\top Y,\ \theta_i^\top X\right) \quad \text{s.t.} \quad \mathrm{Corr}(\eta_i^\top Y, \eta_j^\top Y) = \mathrm{Corr}(\theta_i^\top X, \theta_j^\top X) = \delta_{ij}.$$

- **Canonical vectors**, $H \equiv [\eta_1 \dots \eta_d] \in \mathbb{R}^{q\times d}$ and $T \equiv [\theta_1, \dots \theta_d] \in \mathbb{R}^{p\times d}$.
- **Scores** $U = H^\top Y$, $V = T^\top X$.

Multivariate CCA - multiple canonical vectors

- Estimating *multiple* canonical vectors:

$$\max_{\{\eta_i,\theta_i\}_{i=1}^d} \sum_{i=1}^{d} \mathrm{Corr}\left(\eta_i^\top Y,\ \theta_i^\top X\right) \quad \text{s.t.} \quad \mathrm{Corr}(\eta_i^\top Y, \eta_j^\top Y) = \mathrm{Corr}(\theta_i^\top X, \theta_j^\top X) = \delta_{ij}.$$

- **Canonical vectors**, $H \equiv [\eta_1 \dots \eta_d] \in \mathbb{R}^{q\times d}$ and $T \equiv [\theta_1, \dots \theta_d] \in \mathbb{R}^{p\times d}$.
- **Scores** $U = H^\top Y$, $V = T^\top X$.
- Equivalent reformulation:

$$\max_{H\in\mathbb{R}^{q\times d},\, T\in\mathbb{R}^{p\times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d.$$

Why choose CCA for integrative data analysis

Recall that CCA solves

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d,$$

where $U = H^\top Y$, $V = T^\top X$.

Why choose CCA for integrative data analysis

Recall that CCA solves

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d,$$

where $U = H^\top Y$, $V = T^\top X$.

- Characterizes the dependence structure between the data views Y and X.

Why choose CCA for integrative data analysis

Recall that CCA solves

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d,$$

where $U = H^\top Y$, $V = T^\top X$.

- Characterizes the dependence structure between the data views Y and X.
- Provides a **separate** supervised embedding for Y and X, $H^\top : \mathbb{R}^q \to \mathbb{R}^d$ and $T^\top : \mathbb{R}^p \to \mathbb{R}^d$.

Why choose CCA for integrative data analysis

Recall that CCA solves

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d,$$

where $U = H^\top Y$, $V = T^\top X$.

- Characterizes the dependence structure between the data views Y and X.
- Provides a **separate** supervised embedding for Y and X, $H^\top : \mathbb{R}^q \to \mathbb{R}^d$ and $T^\top : \mathbb{R}^p \to \mathbb{R}^d$.
- Scores $U = H^\top Y$ and $V = T^\top X$ maximally correlated: embeddings informed by one another.

Why choose CCA for integrative data analysis

Recall that CCA solves

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[U^\top V\right] \quad \text{s.t.} \quad \Sigma_U = \Sigma_V = I_d,$$

where $U = H^\top Y$, $V = T^\top X$.

- Characterizes the dependence structure between the data views Y and X.
- Provides a **separate** supervised embedding for Y and X, $H^\top : \mathbb{R}^q \to \mathbb{R}^d$ and $T^\top : \mathbb{R}^p \to \mathbb{R}^d$.
- Scores $U = H^\top Y$ and $V = T^\top X$ maximally correlated: embeddings informed by one another.
- U shares information across the components of Y: not true for regression matrix B in $\mathbb{E}[\|Y - B^\top X\|_2^2]$.

Where classical CCA fails – high-dimensional data

- X and Y *high-dimensional*: number of samples N smaller than dimension of X and Y.
- CCA **overfits**, and estimated canonical correlations $\to 1$ even when X and Y are unrelated.

Where classical CCA fails – high-dimensional data

- X and Y *high-dimensional*: number of samples N smaller than dimension of X and Y.
- CCA **overfits**, and estimated canonical correlations $\to 1$ even when X and Y are unrelated.
- Canonical vectors (η_i, θ_i) **hard to interpret**.

Where classical CCA fails – high-dimensional data

- X and Y *high-dimensional*: number of samples N smaller than dimension of X and Y.
- CCA **overfits**, and estimated canonical correlations $\to 1$ even when X and Y are unrelated.
- Canonical vectors (η_i, θ_i) **hard to interpret**.
- Solution: **Sparse CCA**:

$$\begin{aligned} & \max_{\eta,\theta} \ \mathrm{Corr}(\eta^\top Y,\ \theta^\top X) \\ \text{s.t.} \quad & \|\eta\|_1 \le c_1,\ \|\theta\|_1 \le c_2. \end{aligned}$$

Where classical CCA fails – high-dimensional data

- X and Y *high-dimensional*: number of samples N smaller than dimension of X and Y.
- CCA **overfits**, and estimated canonical correlations $\to 1$ even when X and Y are unrelated.
- Canonical vectors (η_i, θ_i) **hard to interpret**.
- Solution: **Sparse CCA**:

$$\max_{\eta,\theta} \ \mathrm{Corr}(\eta^\top Y,\ \theta^\top X)$$
$$\text{s.t.} \quad \|\eta\|_1 \le c_1,\ \|\theta\|_1 \le c_2.$$

- Drawback: non-trivial to enforce the orthogonality constraints $\Sigma_U = \Sigma_V = I_d$ while maintaining sparsity.

Where classical CCA fails - functional data

- $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are *random functions*.

Where classical CCA fails - functional data

- $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are *random functions*.
- **Multivariate CCA**, with $\langle a, b \rangle \equiv a^\top b$, is

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

Where classical CCA fails - functional data

- $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are *random functions*.
- **Multivariate CCA**, with $\langle a, b \rangle \equiv a^\top b$, is

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

- **Functional CCA**, with $\langle a, b \rangle \equiv \int_0^1 a(t) b(t)\, dt$, is

$$(\theta_1, \eta_1) = \underset{\theta : [0,1] \to \mathbb{R}, \eta : [0,1] \to \mathbb{R}, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

Where classical CCA fails - functional data

- $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are *random functions*.
- **Multivariate CCA**, with $\langle a, b \rangle \equiv a^\top b$, is

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

- **Functional CCA**, with $\langle a, b \rangle \equiv \int_0^1 a(t) b(t)\, dt$, is

$$(\theta_1, \eta_1) = \underset{\theta : [0,1] \to \mathbb{R}, \eta : [0,1] \to \mathbb{R}, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

- Directions θ_1, η_1 are infinite dimensional objects.

Where classical CCA fails - functional data

- $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are *random functions*.
- **Multivariate CCA**, with $\langle a, b \rangle \equiv a^\top b$, is

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

- **Functional CCA**, with $\langle a, b \rangle \equiv \int_0^1 a(t) b(t)\, dt$, is

$$(\theta_1, \eta_1) = \underset{\theta : [0,1] \to \mathbb{R}, \eta : [0,1] \to \mathbb{R}, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}\left(\langle \theta, X \rangle, \langle \eta, Y \rangle\right).$$

- Directions θ_1, η_1 are infinite dimensional objects.
- **Issue**: This generalization has subtleties. Why?

Where classical CCA fails - functional data cont.

Functional CCA, with $\langle a, b \rangle \equiv \int_0^1 a(t)b(t)\,dt$, solves

$$(\theta_1, \eta_1) = \underset{\theta:[0,1]\to\mathbb{R},\eta:[0,1]\to\mathbb{R},\mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}{\arg\max} \mathrm{Corr}\left(\langle\theta, X\rangle, \langle\eta, Y\rangle\right).$$

Where classical CCA fails - functional data cont.

Functional CCA, with $\langle a, b\rangle \equiv \int_0^1 a(t)b(t)\,dt$, solves

$$(\theta_1, \eta_1) = \underset{\theta:[0,1]\to\mathbb{R},\eta:[0,1]\to\mathbb{R},\mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}{\arg\max} \mathrm{Corr}\left(\langle\theta, X\rangle, \langle\eta, Y\rangle\right).$$

- Scores $U_1 = \langle\eta, Y\rangle$ and $V_1 = \langle\theta, X\rangle$ **well-defined**.

Where classical CCA fails - functional data cont.

Functional CCA, with $\langle a, b\rangle \equiv \int_0^1 a(t)b(t)\,dt$, solves

$$(\theta_1, \eta_1) = \underset{\theta:[0,1]\to\mathbb{R},\eta:[0,1]\to\mathbb{R},\mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}{\arg\max} \mathrm{Corr}\left(\langle\theta, X\rangle, \langle\eta, Y\rangle\right).$$

- Scores $U_1 = \langle\eta, Y\rangle$ and $V_1 = \langle\theta, X\rangle$ **well-defined**.
- Canonical directions (θ_1, η_1) may not exist: maximizers may **not** be attained.

Where classical CCA fails - functional data cont.

Functional CCA, with $\langle a, b\rangle \equiv \int_0^1 a(t)b(t)\,dt$, solves

$$(\theta_1, \eta_1) = \underset{\theta:[0,1]\to\mathbb{R},\eta:[0,1]\to\mathbb{R},\mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}{\arg\max} \mathrm{Corr}\left(\langle\theta, X\rangle, \langle\eta, Y\rangle\right).$$

- Scores $U_1 = \langle\eta, Y\rangle$ and $V_1 = \langle\theta, X\rangle$ **well-defined**.
- Canonical directions (θ_1, η_1) may not exist: maximizers may **not** be attained.
- Additional population-level assumptions necessary: e.g., finite-dimensional $X(t)$, $Y(t)$.

Where classical CCA fails – nonlinear data

- CCA uses linear projections of X and Y: no inner product $\langle \cdot, \cdot \rangle$ structure for nonlinear data

Where classical CCA fails – nonlinear data

- CCA uses linear projections of X and Y: no inner product $\langle \cdot, \cdot \rangle$ structure for nonlinear data
- CCA can fail to capture nonlinear relationships.

Where classical CCA fails – nonlinear data

- CCA uses linear projections of X and Y: no inner product $\langle \cdot, \cdot \rangle$ structure for nonlinear data
- CCA can fail to capture nonlinear relationships.
- Example: let X be symmetric about 0, and $Y = X^2$.

Where classical CCA fails – nonlinear data

- CCA uses linear projections of X and Y: no inner product $\langle \cdot, \cdot \rangle$ structure for nonlinear data
- CCA can fail to capture nonlinear relationships.
- Example: let X be symmetric about 0, and $Y = X^2$.
  - Perfectly dependent, but $\mathrm{Corr}\left(Y, X^2\right) = 0$.

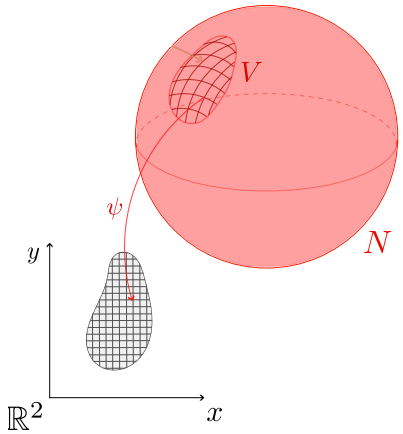
Nonlinear data: known vs unknown geometry

Geometry known a priori:

Nonlinear data: known vs unknown geometry

Geometry known a priori:

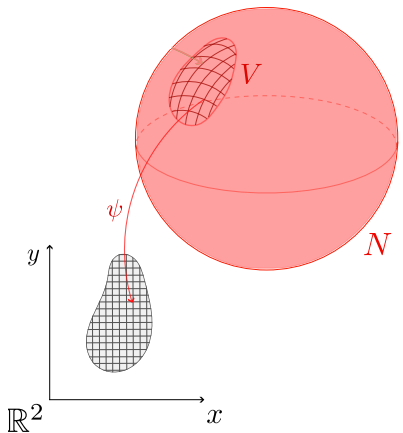
- Manifold representation (e.g., SPD matrices, spheres, densities).



Nonlinear data: known vs unknown geometry

Geometry known a priori:

- Manifold representation (e.g., SPD matrices, spheres, densities).
- Gain access to tools of differential geometry: log/exp maps, parallel transport, Riemannian metric.

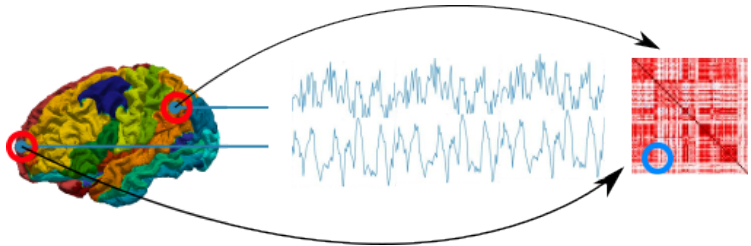


Known geometry for brain imaging data

- Example: **static functional connectivity**.

Known geometry for brain imaging data

- Example: **static functional connectivity**.
- $Y \in \mathbb{R}^{m \times m}$: Covariance matrices belonging to the **manifold of symmetric positive definite (SPD) matrices**.



Brain connectivity naturally represented by SPD matrix.

Nonlinear data: known vs unknown geometry

Geometry unknown (manifold hypothesis):

Nonlinear data: known vs unknown geometry

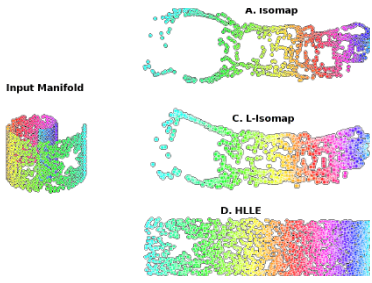
Geometry unknown (manifold hypothesis):

- Manifold structure of the data is **not known in advance** — must be **learned from data**.

Nonlinear data: known vs unknown geometry

Geometry unknown (manifold hypothesis):

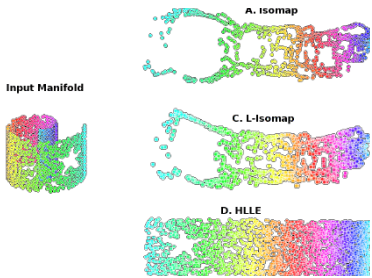
- Manifold structure of the data is **not known in advance** — must be **learned from data**.
- Classical approaches: **dimensionality reduction** that preserves geometric structure (e.g., Isomap, LLE, diffusion maps).



Nonlinear data: known vs unknown geometry

Geometry unknown (manifold hypothesis):

- Manifold structure of the data is **not known in advance** — must be **learned from data**.
- Classical approaches: **dimensionality reduction** that preserves geometric structure (e.g., Isomap, LLE, diffusion maps).
- Deep learning approaches: **neural encoders/decoders** jointly learn nonlinear representations.



This talk: two settings

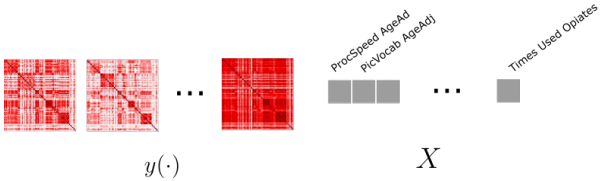
- This work: **integrative data analysis** of **non-linear** and **high-dimensional data**

This talk: two settings

- This work: **integrative data analysis** of **non-linear** and **high-dimensional data**

  Two settings:

  Project 1: **Known geometry**

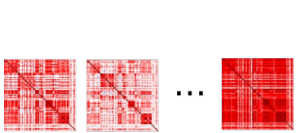


This talk: two settings

- This work: **integrative data analysis** of **non-linear** and **high-dimensional data**

  Two settings:

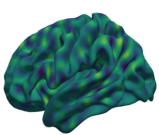
  Project 1: **Known geometry**



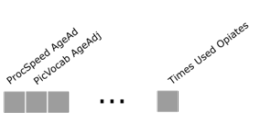
$y(\cdot)$

X

  Project 2: **Unknown geometry**



Y



X

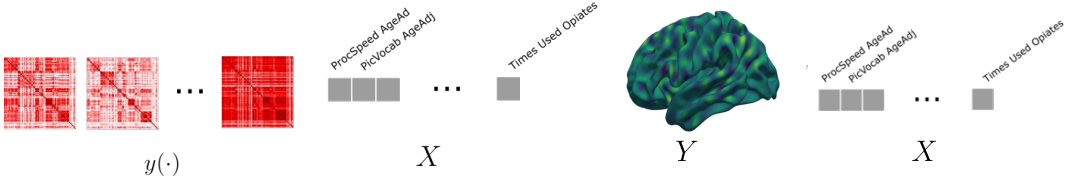
This talk: two settings

- This work: **integrative data analysis** of **non-linear** and **high-dimensional data**

  Two settings:

  Project 1: **Known geometry**

  Project 2: **Unknown geometry**



$y(\cdot)$ X Y X

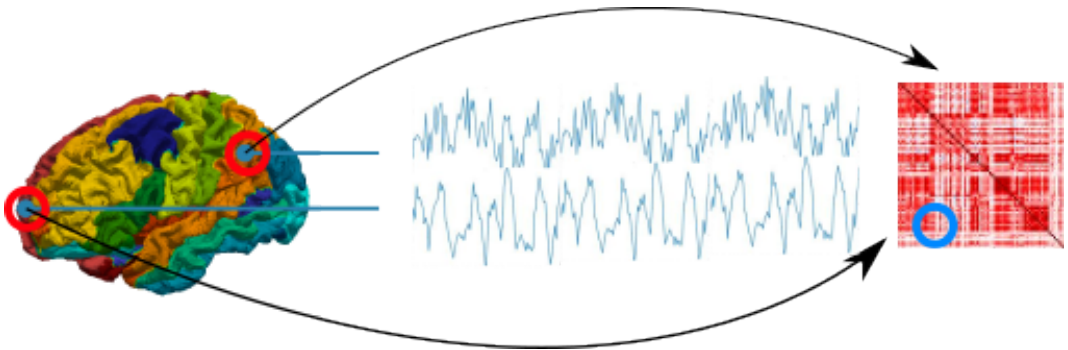
Aim: uncover shared structure between heterogeneous data views

Project 1:
Asymmetric canonical correlation analysis of Riemannian and high-dimensional data

Imaging Data - Static functional connectivity

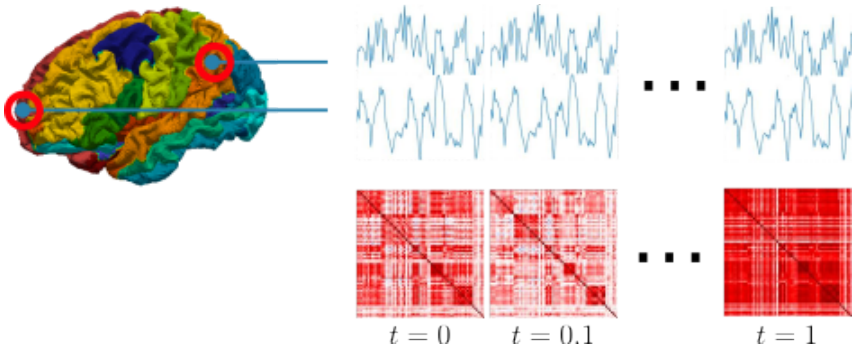
- **Static functional connectivity**: covariance matrix based on signals from m different brain regions.

- $y_i \in \mathbb{R}^{m \times m}$ for each patient $i = 1, \ldots N$.



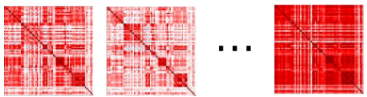
Imaging data - Dynamic functional connectivity

- **Dynamic functional connectivity**: sequence of covariance matrices obtained by partitioning time
- $y_i(t_1), \dots y_i(t_L) \in \mathbb{R}^{m \times m}$ for each patient $i = 1, \dots N$.



Setup: Study relationship between different data views

- $y : [0, 1] \to \mathcal{M}$ is a **random manifold-valued function**, represents dynamic brain imaging data.

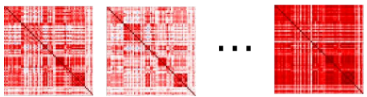


$y(\cdot)$

X

Setup: Study relationship between different data views

- $y : [0, 1] \to \mathcal{M}$ is a **random manifold-valued function**, represents dynamic brain imaging data.
- $X \in \mathbb{R}^p$ is a **multivariate random vector**, represents high-dimensional data.

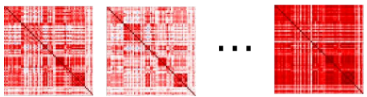


$y(\cdot)$

X

Setup: Study relationship between different data views

- $y : [0, 1] \to \mathcal{M}$ is a **random manifold-valued function**, represents dynamic brain imaging data.
- $X \in \mathbb{R}^p$ is a **multivariate random vector**, represents high-dimensional data.
- In practice, we observe i.i.d. pairs (X_i, y_i) for $i = 1, \ldots N$, where each y_i is observed is $\{y_i(t_l) : l = 1, \ldots L\}$.



$y(\cdot)$

X

Can we generalize classical CCA?

- The classical, multivariate CCA model solves

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle \theta, X \rangle, \langle \eta, Y \rangle \right).$$

Can we generalize classical CCA?

- The classical, multivariate CCA model solves

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle \theta, X \rangle, \langle \eta, Y \rangle \right).$$

- Do we have an analogue of $\langle \eta, y \rangle$ for $y : [0, 1] \to \mathcal{M}$?

Can we generalize classical CCA?

- The classical, multivariate CCA model solves

$$(\theta_1, \eta_1) = \underset{\theta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \mathrm{Var}(\langle \theta, X \rangle) = \mathrm{Var}(\langle \eta, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle \theta, X \rangle, \langle \eta, Y \rangle \right).$$

- Do we have an analogue of $\langle \eta, y \rangle$ for $y : [0, 1] \to \mathcal{M}$?

- No, since we don't necessarily have an **inner product structure** on a non-Euclidean geometry $\mathcal{M}$.

Machinery of Riemannian manifolds

Geodesic distance:

- $d_{\mathcal{M}}(\cdot,\cdot) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$

Fréchet mean:

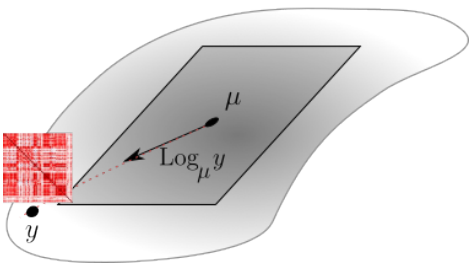
- For a random element $y \in \mathcal{M}$, the average value of y, $\arg\min_{x \in \mathcal{M}} \mathbb{E}\left[d^2_{\mathcal{M}}(x, y)\right]$

Tangent space at $x \in \mathcal{M}$:

- Vector space $T_x\mathcal{M}$ equipped with Riemannian metric $\langle \cdot, \cdot \rangle_x$

Logarithmic and Exponential maps:

- $\mathrm{Log}_x(\cdot) : \mathcal{M} \to T_x\mathcal{M}$
- $\mathrm{Exp}_x(\cdot) : T_x\mathcal{M} \to \mathcal{M}$ (Inverse of Logarithmic map)



Riemannian manifold of positive definite matrices

Affine-invariant metric on set of $m \times m$ positive definite matrices:

- **Tangent spaces $T_P\mathcal{M}$:** space of symmetric matrices (unconstrained).
- **Riemannian metric:** $P \in \mathcal{M}$ between $W, Z \in T_P\mathcal{M}$ is defined as $\langle W, Z \rangle_{\mathcal{M}} = \operatorname{tr}\left(P^{-1} W P^{-1} Z\right)$.
- **Logarithmic map:** $\operatorname{Log}_P(Q) = P^{1/2} \log\left(P^{-1/2} Q P^{-1/2}\right) P^{1/2}$
  - Maps manifold representation to tangent space representation.
- **Exponential map:** $\operatorname{Exp}_P(W) = P^{1/2} \exp\left(P^{-1/2} W P^{-1/2}\right) P^{1/2}$
  - Maps tangent space representation to manifold representation.

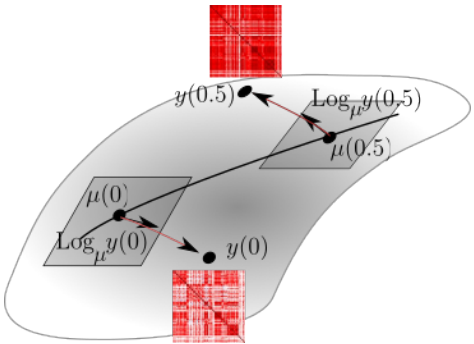
Riemannian-valued functional data

- Fréchet mean μ and Logarithmic map allow us to move from:

$$y : [0,1] \to \mathcal{M}$$

to unconstrained tangent space representation

$$\mathrm{Log}_\mu\, y : [0,1] \to T_\mu \mathcal{M}.$$



Riemannian-valued functional data

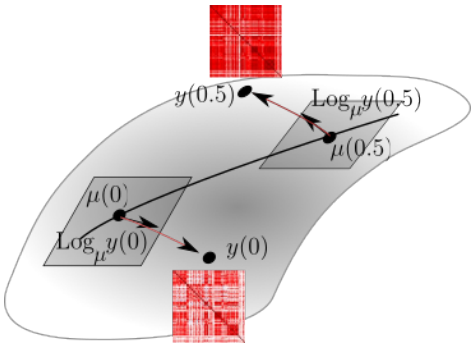
- Fréchet mean μ and Logarithmic map allow us to move from:

$$y : [0,1] \to \mathcal{M}$$

to unconstrained tangent space representation

$$\mathrm{Log}_\mu\, y : [0,1] \to T_\mu \mathcal{M}.$$

- Tangent space is a **Hilbert space** we denote by $L^2(T\mu)$, equipped with the inner product $\langle\!\langle U, V \rangle\!\rangle_\mu := \int_{[0,1]} \langle V(t), U(t) \rangle_{\mu(t)} \mathrm{d}t$.



Population CCA Problem

- The canonical correlation problem we end up with is the following: for $y : [0,1] \to \mathcal{M}$ and $X \in \mathbb{R}^p$, solve

$$\underset{\substack{\theta \in \mathbb{R}^p,\ \psi \in L^2(T\mu) \\ \mathrm{Var}\left(\langle \theta, X \rangle\right)=1,\ \mathrm{Var}\left(\langle\!\langle \psi, \mathrm{Log}_\mu y \rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\Big(\langle\!\langle \psi, \mathrm{Log}_\mu y \rangle\!\rangle_\mu,\ \langle \theta, X \rangle\Big) \tag{1}$$

Population CCA Problem

- The canonical correlation problem we end up with is the following: for $y : [0,1] \to \mathcal{M}$ and $X \in \mathbb{R}^p$, solve

$$\underset{\substack{\theta \in \mathbb{R}^p,\ \psi \in L^2(T\mu) \\ \mathrm{Var}\left(\langle \theta, X \rangle\right)=1,\ \mathrm{Var}\left(\langle\!\langle \psi, \mathrm{Log}_\mu\, y \rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\Big(\langle\!\langle \psi, \mathrm{Log}_\mu\, y \rangle\!\rangle_\mu,\ \langle \theta, X \rangle\Big) \tag{1}$$

- Two additional issues to handle: ψ lives in an **infinite dimensional space** $L^2(T\mu)$, and θ is **high dimensional**.

Non-linearity	Functional	High-dimensional
✓	✗	✗

Handling infinite-dimensional data

- $\mathrm{Log}_\mu y$ lives in an **infinite-dimensional** space, but its variation is often along only a few directions.

Handling infinite-dimensional data

- $\mathrm{Log}_\mu y$ lives in an **infinite-dimensional** space, but its variation is often along only a few directions.
- **Approach:** use data-driven **functional PCA (FPCA)** for dimensionality reduction.

Non-linearity	Functional	High-dimensional
✓	✓	✗

Handling infinite-dimensional data

- $\mathrm{Log}_\mu y$ lives in an **infinite-dimensional** space, but its variation is often along only a few directions.
- **Approach:** use data-driven **functional PCA (FPCA)** for dimensionality reduction.
- FPCA finds **principal components** $\{\phi_j\}_{j=1}^d$ such that

$$\mathrm{Log}_\mu y \approx \sum_{j=1}^{d} \phi_j Y_j.$$

Non-linearity	Functional	High-dimensional
✓	✓	✗

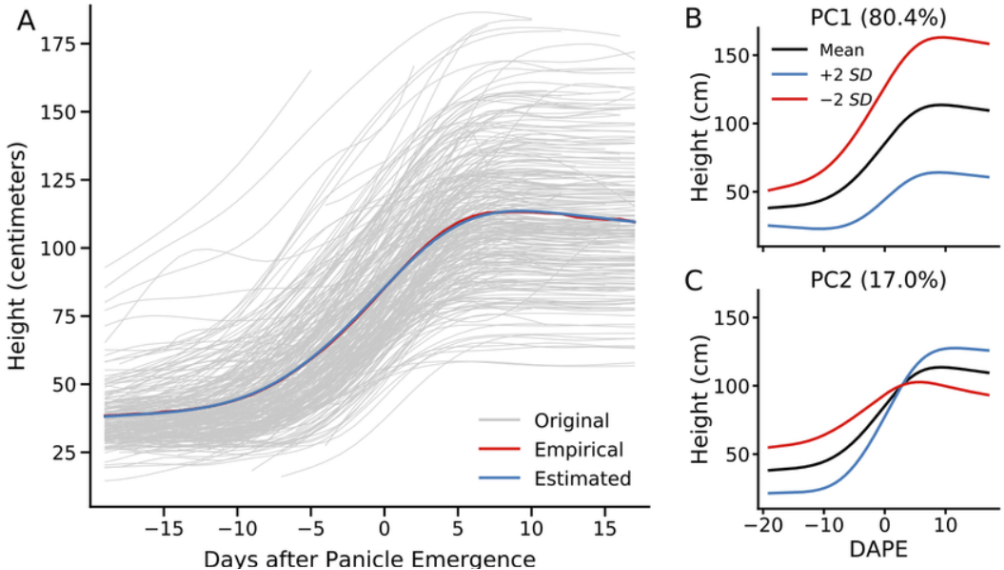
Handling infinite-dimensional data

- $\mathrm{Log}_\mu y$ lives in an **infinite-dimensional** space, but its variation is often along only a few directions.
- **Approach:** use data-driven **functional PCA (FPCA)** for dimensionality reduction.
- FPCA finds **principal components** $\{\phi_j\}_{j=1}^d$ such that

$$\mathrm{Log}_\mu y \approx \sum_{j=1}^{d} \phi_j Y_j.$$

- This yields a low-dimensional multivariate representation $Y = (Y_1, \ldots, Y_d) \in \mathbb{R}^d$.

Non-linearity	Functional	High-dimensional
✓	✓	✗



$\mathrm{Log}_\mu\, y \approx \sum_{j=1}^{d} \phi_j Y_j$. PCs ϕ_j, and scores Y_j

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu y$ is represented by PC scores $Y \in \mathbb{R}^d$.

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu\, y$ is represented by PC scores $Y \in \mathbb{R}^d$.
- **Functional CCA:**

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \psi\in L^2(T\mu)\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}\left(\langle\!\langle\psi,\mathrm{Log}_\mu\, y\rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\left(\langle\!\langle\psi, \mathrm{Log}_\mu\, y\rangle\!\rangle_\mu,\, \langle\theta, X\rangle\right)$$

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu\, y$ is represented by PC scores $Y \in \mathbb{R}^d$.
- **Functional CCA:**

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \psi\in L^2(T\mu)\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}\left(\langle\!\langle\psi,\mathrm{Log}_\mu\, y\rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\left(\langle\!\langle\psi,\mathrm{Log}_\mu\, y\rangle\!\rangle_\mu,\ \langle\theta,X\rangle\right)$$

- **Reformulation**: Multivariate CCA between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$:

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \eta\in\mathbb{R}^d\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}}{\text{maximize}} \quad \mathrm{Corr}^2(\langle\eta,Y\rangle,\ \langle\theta,X\rangle)$$

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu\, y$ is represented by PC scores $Y \in \mathbb{R}^d$.
- **Functional CCA:**

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \psi\in L^2(T\mu)\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}\left(\langle\!\langle\psi,\mathrm{Log}_\mu\, y\rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\left(\langle\!\langle\psi,\mathrm{Log}_\mu\, y\rangle\!\rangle_\mu,\ \langle\theta,X\rangle\right)$$

- **Reformulation**: Multivariate CCA between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$:

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \eta\in\mathbb{R}^d\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}}{\text{maximize}} \quad \mathrm{Corr}^2(\langle\eta,Y\rangle,\ \langle\theta,X\rangle)$$

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu y$ is represented by PC scores $Y \in \mathbb{R}^d$.
- **Functional CCA:**

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \psi\in L^2(T\mu)\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}\left(\langle\!\langle\psi,\mathrm{Log}_\mu y\rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\left(\langle\!\langle\psi,\mathrm{Log}_\mu y\rangle\!\rangle_\mu,\ \langle\theta,X\rangle\right)$$

- **Reformulation**: Multivariate CCA between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$:

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \eta\in\mathbb{R}^d\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}}{\text{maximize}} \quad \mathrm{Corr}^2(\langle\eta,Y\rangle,\ \langle\theta,X\rangle)$$

- Since X is high-dimensional, can appeal to **sparse CCA**.

Problem reformulation: Multivariate CCA

- After FPCA, $\mathrm{Log}_\mu y$ is represented by PC scores $Y \in \mathbb{R}^d$.
- **Functional CCA:**

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \psi\in L^2(T\mu)\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}\left(\langle\!\langle\psi,\mathrm{Log}_\mu y\rangle\!\rangle_\mu\right)=1}}{\text{maximize}} \quad \mathrm{Corr}^2\left(\langle\!\langle\psi,\mathrm{Log}_\mu y\rangle\!\rangle_\mu,\ \langle\theta,X\rangle\right)$$

- **Reformulation**: Multivariate CCA between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$:

$$\underset{\substack{\theta\in\mathbb{R}^p,\ \eta\in\mathbb{R}^d\\ \text{s.t. } \mathrm{Var}(\langle\theta,X\rangle)=\mathrm{Var}(\langle\eta,Y\rangle)=1}}{\text{maximize}} \quad \mathrm{Corr}^2(\langle\eta,Y\rangle,\ \langle\theta,X\rangle)$$

- Since X is high-dimensional, can appeal to **sparse CCA**.

Non-linearity	Functional	High-dimensional
✓	✓	✓

Estimation

1. **Data**: $(X_i, y_i)_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i : [0,1] \to \mathcal{M}$.
2. **Tangent space representations**: $\mathrm{Log}_{\hat{\mu}}\, y_i$
3. **FPCA**: $\mathrm{Log}_{\hat{\mu}}\, y \approx \sum_{j=1}^{d} \hat{\phi}_j Y_j$

Estimation

1. **Data**: $(X_i, y_i)_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i : [0,1] \to \mathcal{M}$.
2. **Tangent space representations**: $\operatorname{Log}_{\hat{\mu}} y_i$
3. **FPCA**: $\operatorname{Log}_{\hat{\mu}} y \approx \sum_{j=1}^d \hat{\phi}_j Y_j$
4. Solve **group lasso regression** from data matrices $\mathbb{Y} \in \mathbb{R}^{N \times d}$, $\mathbb{X} \in \mathbb{R}^{N \times p}$

$$\hat{B} = \underset{B \in \mathbb{R}^{p \times d}}{\arg\min} \frac{2}{N} \left\| \mathbb{Y} \hat{\Sigma}_Y^{-1/2} - \mathbb{X} B \right\|_F^2 + \lambda \|B\|_{\ell_1, \ell_2}$$

Estimation

1. **Data**: $(X_i, y_i)_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i : [0,1] \to \mathcal{M}$.
2. **Tangent space representations**: $\operatorname{Log}_{\hat{\mu}} y_i$
3. **FPCA**: $\operatorname{Log}_{\hat{\mu}} y \approx \sum_{j=1}^d \hat{\phi}_j Y_j$
4. Solve **group lasso regression** from data matrices $\mathbb{Y} \in \mathbb{R}^{N \times d}$, $\mathbb{X} \in \mathbb{R}^{N \times p}$

$$\hat{B} = \underset{B \in \mathbb{R}^{p \times d}}{\arg\min} \frac{2}{N} \left\| \mathbb{Y} \hat{\Sigma}_Y^{-1/2} - \mathbb{X} B \right\|_F^2 + \lambda \|B\|_{\ell_1, \ell_2}$$

5. Find the **eigenvector decomposition** $ED^2E^\top = \hat{B}^\top \hat{\Sigma}_X \hat{B}$.
6. **Compute** $\hat{H} = [\hat{\eta}_1, \ldots, \hat{\eta}_d]$ and $\hat{T} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_d\right]$ via $\hat{H} = \hat{\Sigma}_Y^{-1/2} E$ and $\hat{T} = \hat{B} E D^{-1}$. Then $\hat{\psi}_j = \sum_{k=1}^d \hat{\phi}_k \hat{\eta}_{jk} \in L^2(T\hat{\mu})$.
7. **Return** canonical directions $(\hat{\theta}_j, \hat{\psi}_j)_{j=1,\ldots d}$.

Theory

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

- $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$ are subgaussian, with invertible covariance matrices.

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

- $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$ are subgaussian, with invertible covariance matrices.
- $d \log(p) = o(N)$

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

- $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$ are subgaussian, with invertible covariance matrices.
- $d \log(p) = o(N)$
- Lasso parameter $\lambda = O\left(\sqrt{\frac{d \log(p)}{N}}\right)$

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

- $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$ are subgaussian, with invertible covariance matrices.
- $d \log(p) = o(N)$
- Lasso parameter $\lambda = O\left(\sqrt{\frac{d \log(p)}{N}}\right)$

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\left(\frac{d}{N} \log p\right)^{1/2} \frac{\left\|\Sigma_X^{-1}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

Simpler case: Multivariate low-dimensional Y

Main assumptions (slow-rate bound):

- $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$ are subgaussian, with invertible covariance matrices.
- $d \log(p) = o(N)$
- Lasso parameter $\lambda = O\left(\sqrt{\frac{d \log(p)}{N}}\right)$

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\left(\frac{d}{N} \log p\right)^{1/2} \frac{\left\|\Sigma_X^{-1}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

$$\left\|\eta_k - \hat{\eta}_k\right\|_2^2 = O_P\left(\left(\frac{d}{N} \log p\right)^{1/2} \frac{\left\|\Sigma_Y^{-1}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

Nonlinear functional $y(\cdot)$

Assumptions:

Nonlinear functional $y(\cdot)$

Assumptions:

- The manifold $\mathcal{M}$ is a complete simply-connected Riemannian manifold with nonpositive sectional curvature.

Nonlinear functional $y(\cdot)$

Assumptions:

- The manifold $\mathcal{M}$ is a complete simply-connected Riemannian manifold with nonpositive sectional curvature.
- The functional data are such that $\sup_{t \epsilon \mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.

Nonlinear functional $y(\cdot)$

Assumptions:

- The manifold $\mathcal{M}$ is a complete simply-connected Riemannian manifold with nonpositive sectional curvature.
- The functional data are such that $\sup_{t \in \mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.
- The θ_k satisfy a group s-sparsity condition.

Nonlinear functional $y(\cdot)$

Assumptions:

- The manifold $\mathcal{M}$ is a complete simply-connected Riemannian manifold with nonpositive sectional curvature.
- The functional data are such that $\sup_{t \in \mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.
- The θ_k satisfy a group s-sparsity condition.

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\frac{ds\log(p)}{N}\,\frac{1}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

Nonlinear functional $y(\cdot)$

Assumptions:

- The manifold $\mathcal{M}$ is a complete simply-connected Riemannian manifold with nonpositive sectional curvature.
- The functional data are such that $\sup_{t \in \mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.
- The θ_k satisfy a group s-sparsity condition.

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\frac{ds\log(p)}{N}\,\frac{1}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

$$\left\|\psi_k - \Gamma_{\hat{\mu},\mu}\hat{\psi}_k\right\|_{\mu}^2 = O_P\left(\frac{d^2 s\log(p)}{N}\,\frac{1}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2}\right)$$

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

- **Population variables** (U_k, V_k):
  - Solution to the population CCA problem **without** finite-dimensional assumption

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

- **Population variables** (U_k, V_k):
  - Solution to the population CCA problem **without** finite-dimensional assumption

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

- **Population variables** (U_k, V_k):
  - Solution to the population CCA problem **without** finite-dimensional assumption

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

- **Population variables** (U_k, V_k):
  - Solution to the population CCA problem **without** finite-dimensional assumption
- **Consistency metric**:

Consistency of canonical variables

- Given a new independent test point $(X_{\text{test}}, y_{\text{test}})$
- **Estimated canonical variables**:

$$\hat{U}_k = \left\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \right\rangle_{\hat{\mu}}, \qquad \hat{V}_k = \langle X_{\text{test}}, \hat{\theta}_k \rangle.$$

- **Population variables** (U_k, V_k):
  - Solution to the population CCA problem **without** finite-dimensional assumption

- **Consistency metric**:

$$\text{MSE}(\hat{U}_k) = \mathbb{E}\big[(U_k - \hat{U}_k)^2 \mid \{(X_i, y_i)\}\big], \quad \text{MSE}(\hat{V}_k) = \mathbb{E}\big[(V_k - \hat{V}_k)^2 \mid \{(X_i, y_i)\}\big]$$

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_\mu(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_\mu(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

- **Interpretation:**

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

- **Interpretation:**
  - Increasing d:

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_\mu(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

- **Interpretation:**
  - Increasing d:
    - Decreases bias term $\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|$

Canonical variables: bias–variance trade-off in d

- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

- **Interpretation:**
  - Increasing d:
    - Decreases bias term $\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|$
    - Increases variance term $\frac{ds\log p}{N}$

Canonical variables: bias–variance trade-off in d

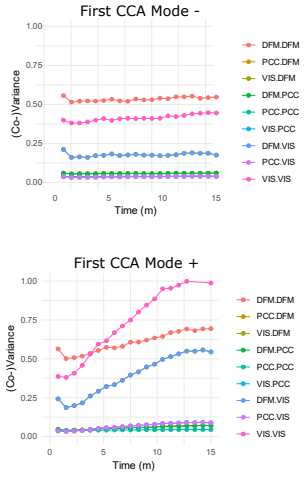
- Cross-covariance operator $\mathscr{C}_{12}$ encodes correlation between $\mathrm{Log}_{\mu}(y)$ and X
- We use a rank-d approximation $\mathscr{C}_{12}^{(d)}$
- Main result (Informal):

$$\max\{\mathsf{MSE}(\hat{U}_k), \mathsf{MSE}(\hat{V}_k)\} = O_P\Big(\underbrace{\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}_{\text{bias}} + \underbrace{\frac{ds\log p}{N}}_{\text{variance}} \Big).$$

- **Interpretation:**
  - Increasing d:
    - Decreases bias term $\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|$
    - Increases variance term $\frac{ds\log p}{N}$
- Classical bias–variance trade-off in selecting the number of principal components

Application

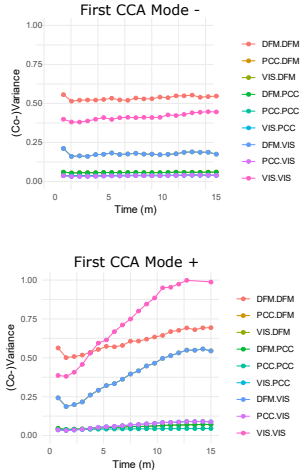
Connectivity



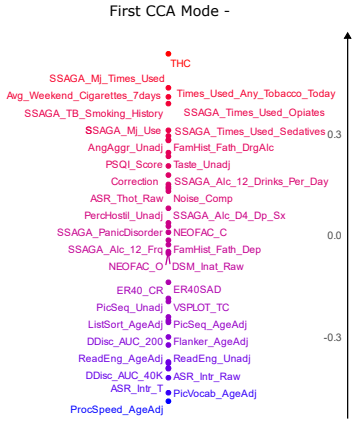
Behaviour

Application

Connectivity



Behaviour

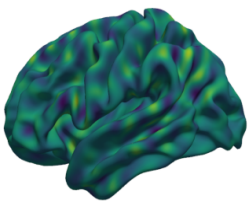


Accepted to the Electronic Journal of Statistics!

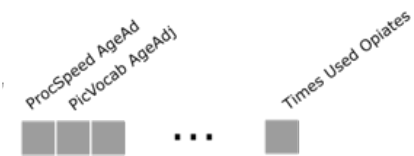
Project 2:
A correlation analysis approach to supervised disentanglement

Supervised disentanglement via CCA

- **Goal**: Apply CCA to the **supervised disentanglement problem**: construct a low-dimensional representation of a target view Y, guided by an auxiliary view X,



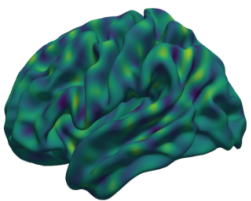
Y



X

Supervised disentanglement via CCA

- **Goal**: Apply CCA to the **supervised disentanglement problem**: construct a low-dimensional representation of a target view Y, guided by an auxiliary view X,
- $Y \in \mathbb{R}^q$ represents **cortical thickness** imaging data
- $X \in \mathbb{R}^p$ **multivariate and interpretable data** (cognitive test, etc.).

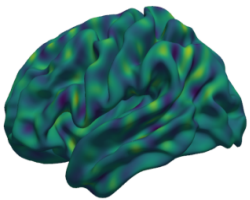


Y

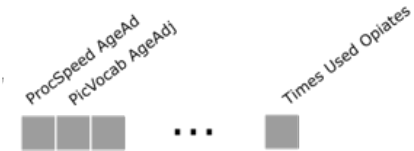
X

Supervised disentanglement via CCA

- **Goal**: Apply CCA to the **supervised disentanglement problem**: construct a low-dimensional representation of a target view Y, guided by an auxiliary view X,
- $Y \in \mathbb{R}^q$ represents **cortical thickness** imaging data
- $X \in \mathbb{R}^p$ **multivariate and interpretable data** (cognitive test, etc.).
- **Assumption**: Y lies on a abstract **unknown** manifold



Y



X

Supervised disentanglement via Linear CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\Big[\big(H^\top Y\big)^\top \big(T^\top X\big)\Big] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

Supervised disentanglement via Linear CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\Big[\big(H^\top Y\big)^\top \big(T^\top X\big)\Big] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

- When $q = d$, H is $d \times d$ and *invertible*: no information is lost moving from Y to $U = H^\top Y$

Supervised disentanglement via Linear CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d},\, T \in \mathbb{R}^{p \times d}} \mathbb{E}\Big[\big(H^\top Y\big)^\top \big(T^\top X\big)\Big] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

- When $q = d$, H is $d \times d$ and *invertible*: no information is lost moving from Y to $U = H^\top Y$
- **Disentanglement:** Scores in U are uncorrelated with one another

Supervised disentanglement via Linear CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\Big[\big(H^\top Y\big)^\top \big(T^\top X\big)\Big] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

- When $q = d$, H is $d \times d$ and *invertible*: no information is lost moving from Y to $U = H^\top Y$
- **Disentanglement:** Scores in U are uncorrelated with one another
- **Supervised disentanglement:** Scores ordered by correlation with $T^\top X$, i.e. linear combinations of X

Supervised disentanglement via Linear CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d},\, T \in \mathbb{R}^{p \times d}} \mathbb{E}\Big[\big(H^\top Y\big)^\top \big(T^\top X\big)\Big] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

- When $q = d$, H is $d \times d$ and *invertible*: no information is lost moving from Y to $U = H^\top Y$
- **Disentanglement:** Scores in U are uncorrelated with one another
- **Supervised disentanglement:** Scores ordered by correlation with $T^\top X$, i.e. linear combinations of X
- **Issue**: H provides a *linear* embedding

From linear CCA to partially linear invertible CCA

Linear CCA:

For $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, solve

$$\max_{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[\left(H^\top Y\right)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d.$$

From linear CCA to partially linear invertible CCA

Partially Linear CCA:

Replace $H^\top Y$ with $g(Y)$, where $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{g:\mathbb{R}^q \to \mathbb{R}^d,\, T \in \mathbb{R}^{p\times d}} \mathbb{E}\left[g(Y)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d.$$

From linear CCA to partially linear invertible CCA

Partially Linear CCA:

Replace $H^\top Y$ with $g(Y)$, where $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{g:\mathbb{R}^q \to \mathbb{R}^d,\, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[g(Y)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d.$$

- **Issue**: $g(Y)$ is not an invertible representation of Y.

From linear CCA to partially linear invertible CCA

Partially Linear CCA:

Replace $H^\top Y$ with $g(Y)$, where $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{g:\mathbb{R}^q \to \mathbb{R}^d,\, T \in \mathbb{R}^{p \times d}} \mathbb{E}\left[g(Y)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d.$$

- **Issue**: $g(Y)$ is not an invertible representation of Y.
- We need a constraint on g that formalizes invertibility.

From linear CCA to partially linear invertible CCA

Partially Linear invertible CCA:

Replace $H^\top Y$ with $g(Y)$, where $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\, T \in \mathbb{R}^{p\times d} \\ g \in \mathcal{C}}} \mathbb{E}\left[g(Y)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d.$$

- Issue: even if we find large correlations, $g(Y)$ is not a representation of Y.
- We need a constraint on g that formalizes invertibility.

From linear CCA to partially linear invertible CCA

Partially Linear invertible CCA (PLiCCA):

Replace $H^\top Y$ with $g(Y)$, where $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{\substack{g:\mathbb{R}^q \to \mathbb{R}^d, T \in \mathbb{R}^{p \times d} \\ g \in \mathcal{C}}} \mathbb{E}\left[g(Y)^\top \left(T^\top X\right)\right] \quad \text{s.t.} \quad \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d.$$

- Issue: even if we find large correlations, $g(Y)$ is not a representation of Y.
- We need a constraint on g that formalizes invertibility.
- $\mathcal{C}$ must formalize a notion of invertibility for g.

How do we formalize invertibility?

- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.

How do we formalize invertibility?

- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.
- True invertibility would require Y to lie *exactly* on a d-dimensional manifold $\mathcal{M}$.

How do we formalize invertibility?

- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.
- True invertibility would require Y to lie *exactly* on a d-dimensional manifold $\mathcal{M}$.
- This is unrealistic in practice: data are noisy and only approximately low-dimensional.

How do we formalize invertibility?

- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.
- True invertibility would require Y to lie *exactly* on a d-dimensional manifold $\mathcal{M}$.
- This is unrealistic in practice: data are noisy and only approximately low-dimensional.
- We instead enforce an **autoencoder-type constraint**:

How do we formalize invertibility?

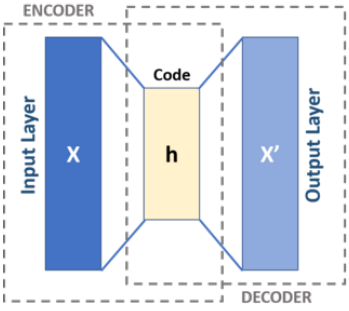
- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.
- True invertibility would require Y to lie *exactly* on a d-dimensional manifold $\mathcal{M}$.
- This is unrealistic in practice: data are noisy and only approximately low-dimensional.
- We instead enforce an **autoencoder-type constraint**:

$$g : \exists f : \mathbb{R}^d \to \mathbb{R}^q \quad \text{s.t.} \quad \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] < \varepsilon.$$

How do we formalize invertibility?

- g maps from $\mathbb{R}^q$ to $\mathbb{R}^d$ with $d \ll q$.
- True invertibility would require Y to lie *exactly* on a d-dimensional manifold $\mathcal{M}$.
- This is unrealistic in practice: data are noisy and only approximately low-dimensional.
- We instead enforce an **autoencoder-type constraint**:

$$g : \exists f : \mathbb{R}^d \to \mathbb{R}^q \quad \text{s.t.} \quad \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] < \varepsilon.$$



PLiCCA with autoencoder constraint

- Formally, we define the constraint set:

$$\mathcal{C}_{\text{VAE}} = \left\{ g : \mathbb{R}^q \to \mathbb{R}^d : \exists f : \mathbb{R}^d \to \mathbb{R}^q,\ \mathbb{E}\left[\|Y - f(g(Y))\|_2^2 \right] < \varepsilon \right\}.$$

PLiCCA with autoencoder constraint

- Formally, we define the constraint set:

$$\mathcal{C}_{\text{VAE}} = \left\{ g : \mathbb{R}^q \to \mathbb{R}^d : \exists f : \mathbb{R}^d \to \mathbb{R}^q,\ \mathbb{E}\left[\|Y - f(g(Y))\|_2^2 \right] < \varepsilon \right\}.$$

- Denoted $\mathcal{C}_{\text{VAE}}$ anticipating its connection to variational autoencoders.
- This ensures g is approximately invertible through the decoder f.

PLiCCA with autoencoder constraint

- Formally, we define the constraint set:

$$\mathcal{C}_{\text{VAE}} = \left\{ g : \mathbb{R}^q \to \mathbb{R}^d : \exists f : \mathbb{R}^d \to \mathbb{R}^q,\ \mathbb{E}\left[\|Y - f(g(Y))\|_2^2 \right] < \varepsilon \right\}.$$

- Denoted $\mathcal{C}_{\text{VAE}}$ anticipating its connection to variational autoencoders.
- This ensures g is approximately invertible through the decoder f.
- Partially Linear invertible CCA (PLiCCA):

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}_{\text{VAE}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) \theta_i^\top X \right]^2.$$

Existence of PLiCCA with autoencoder constraint

Theorem (Existence with $\mathcal{C}_{\mathrm{VAE}}$)

Assume Σ_X invertible and $\mathrm{supp}(Y) \subseteq K \subset \mathbb{R}^q$ compact. Fix $M, m > 0$ and $\varepsilon > 0$. Define

$$\mathcal{C}_{\mathrm{VAE}} = \Big\{ g : K \to \mathbb{R}^d : \ \mathbb{E}\left[g(Y)\right] = 0, \ g \text{ is } M\text{-Lipschitz}, \\ \exists f : \mathbb{R}^d \to \mathbb{R}^q \text{ s.t. } \mathbb{E}\left[f(g(Y))\right] = 0, \ f \text{ is } m\text{-Lipschitz}, \\ \mathbb{E}\left[\|Y - f(g(Y))\|_2^2\right] \le \varepsilon \ \ (\text{autoencoder constraint}) \Big\}.$$

If $\mathcal{C}_{\mathrm{VAE}} \cap \{g : \Sigma_{g(Y)} = I_d\} \neq \varnothing$, then there exists (g, T) solving

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}_{\mathrm{VAE}}}}{\text{maximize}} \ \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2.$$

Special case: a priori dimension reduction

- PLiCCA simplifies when Y lies on a **known or previously estimated low-dimensional manifold** $\mathcal{M} \subseteq \mathbb{R}^q$.

Special case: a priori dimension reduction

- PLiCCA simplifies when Y lies on a **known or previously estimated low-dimensional manifold** $\mathcal{M} \subseteq \mathbb{R}^q$.
- E.g., we have access to an unsupervised neural net (foundation model) $\phi : \mathbb{R}^q \to \mathbb{R}^d$ trained to represent Y

Special case: a priori dimension reduction

- PLiCCA simplifies when Y lies on a **known or previously estimated low-dimensional manifold** $\mathcal{M} \subseteq \mathbb{R}^q$.
- E.g., we have access to an unsupervised neural net (foundation model) $\phi : \mathbb{R}^q \to \mathbb{R}^d$ trained to represent Y
- Define the dimension-reduced variable:

$$W \equiv \phi(Y) \in \mathbb{R}^d.$$

Special case: a priori dimension reduction

- PLiCCA simplifies when Y lies on a **known or previously estimated low-dimensional manifold** $\mathcal{M} \subseteq \mathbb{R}^q$.
- E.g., we have access to an unsupervised neural net (foundation model) $\phi : \mathbb{R}^q \to \mathbb{R}^d$ trained to represent Y
- Define the dimension-reduced variable:

$$W \equiv \phi(Y) \in \mathbb{R}^d.$$

- In this setting, we define PLiCCA in terms of W instead of Y.

PLiCCA with a priori dimension reduction

The population problem becomes:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)} = \Sigma_{T^\top X} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}[\tilde{g}_i(W)\, \theta_i^\top X]^2,$$

where $W = \phi(Y)$.

- W only provides unsupervised embeddings

PLiCCA with a priori dimension reduction

The population problem becomes:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)} = \Sigma_{T^\top X} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}[\tilde{g}_i(W)\, \theta_i^\top X]^2,$$

where $W = \phi(Y)$.

- W only provides unsupervised embeddings
- Invertibility of ϕ implies equivalence to the original problem via

$$g = \tilde{g} \circ \phi.$$

PLiCCA with a priori dimension reduction

The population problem becomes:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^\top X}=I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}[\tilde{g}_i(W)\,\theta_i^\top X]^2,$$

where $W = \phi(Y)$.

- W only provides unsupervised embeddings
- Invertibility of ϕ implies equivalence to the original problem via

$$g = \tilde{g} \circ \phi.$$

- No dimension reduction required: Assuming $\tilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ is strictly invertible is a reasonable assumption.

PLiCCA with a priori dimension reduction

The population problem becomes:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^\top X}=I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}[\tilde{g}_i(W)\,\theta_i^\top X]^2,$$

where $W = \phi(Y)$.

- W only provides unsupervised embeddings
- Invertibility of ϕ implies equivalence to the original problem via

$$g = \tilde{g} \circ \phi.$$

- No dimension reduction required: Assuming $\tilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ is strictly invertible is a reasonable assumption.
- Issue: Invertibility of $g : \mathbb{R}^d \to \mathbb{R}^d$ not sufficient for existence.

Invertibility constraint for existence to PLiCCA

- We need to enforce a stricter form of invertibility by restricting $\tilde{g}$ to bilipschitz functions:

$$m\|x-y\|_2 \leq \|\tilde{g}(x)-\tilde{g}(y)\|_2 \leq M\|x-y\|_2.$$

Invertibility constraint for existence to PLiCCA

- We need to enforce a stricter form of invertibility by restricting $\tilde{g}$ to bilipschitz functions:

$$m\|x-y\|_2 \leq \|\tilde{g}(x)-\tilde{g}(y)\|_2 \leq M\|x-y\|_2.$$

- Bilipschitz functions are closed under limits in the $\sup$ norm.

Invertibility constraint for existence to PLiCCA

- We need to enforce a stricter form of invertibility by restricting $\tilde{g}$ to bilipschitz functions:

$$m\|x-y\|_2 \leq \|\tilde{g}(x)-\tilde{g}(y)\|_2 \leq M\|x-y\|_2.$$

- Bilipschitz functions are closed under limits in the $\sup$ norm.
- Define the constraint set:

$$\mathcal{C}_{\mathrm{NF}} = \{\tilde{g} : K \to \mathbb{R}^d : \mathbb{E}[\tilde{g}(W)] = 0,\ \tilde{g} \text{ bilipschitz } (m, M)\}.$$

Invertibility constraint for existence to PLiCCA

- We need to enforce a stricter form of invertibility by restricting $\tilde{g}$ to bilipschitz functions:

$$m\|x-y\|_2 \leq \|\tilde{g}(x)-\tilde{g}(y)\|_2 \leq M\|x-y\|_2.$$

- Bilipschitz functions are closed under limits in the $\sup$ norm.
- Define the constraint set:

$$\mathcal{C}_{\mathrm{NF}} = \{\tilde{g} : K \to \mathbb{R}^d : \mathbb{E}[\tilde{g}(W)] = 0,\ \tilde{g} \text{ bilipschitz } (m, M)\}.$$

- Denoted $\mathcal{C}_{\mathrm{NF}}$ anticipating its connection to normalizing flows.

Existence of PLiCCA under invertibility constraint

Theorem

Suppose Σ_X *is invertible, and that* $\operatorname{supp}(W) \subseteq K$*, where* K *is a compact subset of* $\mathbb{R}^d$*. Fix constants* $M, m > 0$*. Then, if* $\mathcal{C}$ *is chosen to be*

$$\mathcal{C}_{\mathrm{NF}} \equiv \{\tilde{g} : K \to \mathbb{R}^d : \mathbb{E}\left[\tilde{g}(W)\right] = 0, \tilde{g} \text{ is bilipschitz with parameters } (m, M)\},$$

there exists a solution $(\tilde{g}, T)$ *to the problem*

$$\underset{\substack{\tilde{g}:\mathbb{R}^d \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)} = \Sigma_{T^\top X} = I_d \\ \tilde{g} \in \mathcal{C}_{\mathrm{NF}}}}{\operatorname{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\tilde{g}_i(W)\theta_i^\top X\right], \tag{2}$$

where the $\theta_i \in \mathbb{R}^p$ *are the columns of* $T \in \mathbb{R}^{p\times d}$*.*

Recap: Population PLiCCA problem

Aim: Provide supervised embeddings of Y

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\, g_i(Y)\, \theta_i^\top X \,\right]^2 .$$

- $Y \in \mathcal{M} \subseteq \mathbb{R}^q$: target.

Recap: Population PLiCCA problem

Aim: Provide supervised embeddings of Y

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\, g_i(Y)\, \theta_i^\top X \,\right]^2 .$$

- $Y \in \mathcal{M} \subseteq \mathbb{R}^q$: target.
- $X \in \mathbb{R}^p$: high-dimensional auxiliary data.

Recap: Population PLiCCA problem

Aim: Provide supervised embeddings of Y

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d \\ g\in\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\, g_i(Y)\, \theta_i^\top X \,\right]^2 .$$

- $Y \in \mathcal{M} \subseteq \mathbb{R}^q$: target.
- $X \in \mathbb{R}^p$: high-dimensional auxiliary data.
- g: nonlinear encoder; T: sparse linear embedding.

Recap: Population PLiCCA problem

Aim: Provide supervised embeddings of Y

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d \\ g\in\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\, g_i(Y)\, \theta_i^\top X \,\right]^2 .$$

- $Y \in \mathcal{M} \subseteq \mathbb{R}^q$: target.
- $X \in \mathbb{R}^p$: high-dimensional auxiliary data.
- g: nonlinear encoder; T: sparse linear embedding.

Two choices for the constraint set $\mathcal{C}$:

Approximate invertibility

$$\mathcal{C}_{\text{VAE}} = \Big\{ g : \mathbb{R}^q \to \mathbb{R}^d \ : \ \exists f : \mathbb{R}^d \to \mathbb{R}^q$$

$$\text{with} \quad \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] < \varepsilon. \Big\}$$

Recap: Population PLiCCA problem

Aim: Provide supervised embeddings of Y

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d \\ g\in\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\, g_i(Y)\, \theta_i^\top X \,\right]^2 .$$

- $Y \in \mathcal{M} \subseteq \mathbb{R}^q$: target.
- $X \in \mathbb{R}^p$: high-dimensional auxiliary data.
- g: nonlinear encoder; T: sparse linear embedding.

Two choices for the constraint set $\mathcal{C}$:

Approximate invertibility

$$\mathcal{C}_{\text{VAE}} = \Big\{ g : \mathbb{R}^q \to \mathbb{R}^d \ : \ \exists f : \mathbb{R}^d \to \mathbb{R}^q$$

$$\text{with} \quad \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] < \varepsilon. \Big\}$$

Strict invertibility

$$\mathcal{C}_{\text{NF}} = \Big\{ \tilde{g} : \mathbb{R}^d \to \mathbb{R}^d \ : \ \tilde{g} \text{ is bilipschitz} \Big\}.$$

Connection to regression models

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\, \theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\, \theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Moreover, let $\Sigma_{g'(Y)}^{-1/2}\, B^\top \Sigma_X\, B\, \Sigma_{g'(Y)}^{-1/2} = \tilde{H}\, \Lambda^2\, \tilde{H}^\top$ *with* $\tilde{H}$ *orthogonal and* Λ *diagonal, and define* $H \equiv \Sigma_{g'(Y)}^{-1/2} \tilde{H}$. *Then the canonical directions are recovered by*

$$T = B\, H\, \Lambda^{-1}, \qquad g(y) = H^\top g'(y).$$

Connection to regression models

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\, \theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Moreover, let $\Sigma_{g'(Y)}^{-1/2}\, B^\top \Sigma_X\, B\, \Sigma_{g'(Y)}^{-1/2} = \tilde{H}\, \Lambda^2\, \tilde{H}^\top$ *with* $\tilde{H}$ *orthogonal and* Λ *diagonal, and define* $H \equiv \Sigma_{g'(Y)}^{-1/2} \tilde{H}$. *Then the canonical directions are recovered by*

$$T = B\, H\, \Lambda^{-1}, \qquad g(y) = H^\top g'(y).$$

- Optimize over unconstrained B rather than constrained T.

Connection to regression models

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\,\theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Moreover, let $\Sigma_{g'(Y)}^{-1/2}\, B^\top \Sigma_X\, B\, \Sigma_{g'(Y)}^{-1/2} = \tilde{H}\, \Lambda^2\, \tilde{H}^\top$ *with* $\tilde{H}$ *orthogonal and* Λ *diagonal, and define* $H \equiv \Sigma_{g'(Y)}^{-1/2} \tilde{H}$. *Then the canonical directions are recovered by*

$$T = B\, H\, \Lambda^{-1}, \qquad g(y) = H^\top g'(y).$$

- Optimize over unconstrained B rather than constrained T.
- Recover T and g via $T = BH\Lambda^{-1}$, $g = H^\top g'$. Sparsity structure of B carries over to T.

Connection to regression models

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\,\theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Moreover, let $\Sigma_{g'(Y)}^{-1/2}\, B^\top \Sigma_X\, B\, \Sigma_{g'(Y)}^{-1/2} = \tilde{H}\, \Lambda^2\, \tilde{H}^\top$ *with* $\tilde{H}$ *orthogonal and* Λ *diagonal, and define* $H \equiv \Sigma_{g'(Y)}^{-1/2} \tilde{H}$. *Then the canonical directions are recovered by*

$$T = B\, H\, \Lambda^{-1}, \qquad g(y) = H^\top g'(y).$$

- Optimize over unconstrained B rather than constrained T.
- Recover T and g via $T = BH\Lambda^{-1}$, $g = H^\top g'$. Sparsity structure of B carries over to T.
- Enforce $\Sigma_{g'(Y)} \succeq a I_d$ over $\Sigma_{g(Y)} = I_d$ during training.

Connection to regression models

Theorem (Regression formulation of PLiCCA)

Finding (g, T) *that solves*

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d,\ g \in a^{-1/2}\mathcal{C}}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}\big[g_i(Y)\, \theta_i^\top X\big]^2$$

is equivalent to finding (g', B) *that solves*

$$\underset{\substack{g' \in \mathcal{C},\ B \in \mathbb{R}^{p\times d} \\ \Sigma_{g'(Y)} \succeq a I_d}}{\text{minimize}} \ \mathbb{E}\big[\|g'(Y) - B^\top X\|_2^2\big], \quad \textit{for any } a > 0.$$

Moreover, let $\Sigma_{g'(Y)}^{-1/2}\, B^\top \Sigma_X\, B\, \Sigma_{g'(Y)}^{-1/2} = \tilde{H}\, \Lambda^2\, \tilde{H}^\top$ *with* $\tilde{H}$ *orthogonal and* Λ *diagonal, and define* $H \equiv \Sigma_{g'(Y)}^{-1/2} \tilde{H}$. *Then the canonical directions are recovered by*

$$T = B\, H\, \Lambda^{-1}, \qquad g(y) = H^\top g'(y).$$

- Optimize over unconstrained B rather than constrained T.
- Recover T and g via $T = BH\Lambda^{-1}$, $g = H^\top g'$. Sparsity structure of B carries over to T.
- Enforce $\Sigma_{g'(Y)} \succeq aI_d$ over $\Sigma_{g(Y)} = I_d$ during training.
- How do we enforce the $\mathcal{C}_{\text{VAE}}$ and $\mathcal{C}_{NF}$ constraints?

Regression formulation of PLiCCA:

$\mathcal{C}_{\mathrm{VAE}}$: solve

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big]$$

subject to $\Sigma_{g(Y)} \succeq a I_d$.

How do we solve PLiCCA?

Regression formulation of PLiCCA:

$\mathcal{C}_{\mathrm{VAE}}$: solve

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big]$$

subject to $\Sigma_{g(Y)} \succeq aI_d$.

$\mathcal{C}_{\mathrm{NF}}$: solve

$$\min_{\tilde{g},f,B} \mathbb{E}\big[\|\tilde{g}(W) - B^\top X\|_2^2\big]$$

subject to $\Sigma_{\tilde{g}(Y)} \succeq aI_d$.

Regression formulation of PLiCCA:

$\mathcal{C}_{\mathrm{VAE}}$: solve

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big]$$

subject to $\Sigma_{g(Y)} \succeq aI_d$.

$\mathcal{C}_{\mathrm{NF}}$: solve

$$\min_{\tilde{g},f,B} \mathbb{E}\big[\|\tilde{g}(W) - B^\top X\|_2^2\big]$$

subject to $\Sigma_{\tilde{g}(Y)} \succeq aI_d$.

- **Issue**: global constraint $\Sigma_{g(Y)} \succeq aI_d$, not compatible with gradient descent.

Regression formulation of PLiCCA:

$\mathcal{C}_{\mathrm{VAE}}$: solve

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big]$$

subject to $\Sigma_{g(Y)} \succeq aI_d$.

$\mathcal{C}_{\mathrm{NF}}$: solve

$$\min_{\tilde{g},f,B} \mathbb{E}\big[\|\tilde{g}(W) - B^\top X\|_2^2\big]$$

subject to $\Sigma_{\tilde{g}(Y)} \succeq aI_d$.

- **Issue**: global constraint $\Sigma_{g(Y)} \succeq aI_d$, not compatible with gradient descent.
- **Aim**: relax this constraint (slightly): latent variable models.

PLiCCA and latent variable models

Conditional Variational Autoencoder

- Introduce a latent $Z \in \mathbb{R}^d$ and specify:

$$Y \mid Z \sim \mathcal{N}\big(f(Z), I_d\big)$$
$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big).$$

Conditional Variational Autoencoder

- Introduce a latent $Z \in \mathbb{R}^d$ and specify:

$$Y \mid Z \sim \mathcal{N}\big(f(Z), I_d\big)$$
$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big).$$

- Determines the joint distribution $p(Y, Z|X)$.

Conditional Variational Autoencoder

- Introduce a latent $Z \in \mathbb{R}^d$ and specify:

$$Y \mid Z \sim \mathcal{N}\big(f(Z), I_d\big)$$
$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big).$$

- Determines the joint distribution $p(Y, Z|X)$.
- Variational inference: approximate posterior $p(Z|Y)$ with $q(z \mid y) = \mathcal{N}\big(g(y), I_d\big)$.

Conditional Variational Autoencoder

- Introduce a latent $Z \in \mathbb{R}^d$ and specify:

$$Y \mid Z \sim \mathcal{N}\big(f(Z), I_d\big)$$
$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big).$$

- Determines the joint distribution $p(Y, Z|X)$.
- Variational inference: approximate posterior $p(Z|Y)$ with $q(z \mid y) = \mathcal{N}\big(g(y), I_d\big)$.

Evidence Lower Bound (ELBO)

Rather than maximize the likelihood $p(Y|X)$ directly, more tractable to minimize

$$\min_{g,f,B} \underbrace{\mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big]}_{\text{regression}} + \beta_{\text{VAE}} \underbrace{\mathbb{E}\Big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\Big]}_{\text{reconstruction}}$$

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\begin{aligned}
&\min_{g,f,B} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right] \\
&\quad + \lambda\, \mathbb{E}\left[\|Y - f(g(Y))\|_2^2\right] \\
&\text{s.t. } \Sigma_{g(Y)} \succeq a I_d
\end{aligned}$$

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda\, \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big]$$
$$\text{s.t. } \Sigma_{g(Y)} \succeq a I_d$$

Conditional VAE

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \beta_{\text{VAE}}\, \mathbb{E}\Big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\Big]$$

$$\text{with } q(z \mid Y) = \mathcal{N}\big(g(Y), I_d\big)$$

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda\, \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] \quad \text{s.t. } \Sigma_{g(Y)} \succeq a I_d$$

Conditional VAE

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \beta_{\text{VAE}}\, \mathbb{E}\big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\big]$$

with $q(z \mid Y) = \mathcal{N}\big(g(Y), I_d\big)$

- Proposal: **use conditional VAE as a proxy regression for PLiCCA.**

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda\, \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] \quad \text{s.t. } \Sigma_{g(Y)} \succeq a I_d$$

Conditional VAE

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \beta_{\text{VAE}}\, \mathbb{E}\big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\big]$$

with $q(z \mid Y) = \mathcal{N}\big(g(Y), I_d\big)$

- Proposal: **use conditional VAE as a proxy regression for PLiCCA.**
- First term: equivalent.

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda\, \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] \quad \text{s.t. } \Sigma_{g(Y)} \succeq a I_d$$

Conditional VAE

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \beta_{\text{VAE}}\, \mathbb{E}\Big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\Big]$$

with $q(z \mid Y) = \mathcal{N}\big(g(Y), I_d\big)$

- Proposal: **use conditional VAE as a proxy regression for PLiCCA.**
- First term: equivalent.
- Second term: similar, except with noise; $\mathbb{E}\Big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\Big]$ is enforcing a weaker version of $\Sigma_{g(Y)} \succeq a I_d$.

Objectives side by side

PLiCCA with $\mathcal{C}_{\text{VAE}}$ constraint

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \lambda \mathbb{E}\big[\|Y - f(g(Y))\|_2^2\big] \quad \text{s.t. } \Sigma_{g(Y)} \succeq aI_d$$

Conditional VAE

$$\min_{g,f,B} \mathbb{E}\big[\|g(Y) - B^\top X\|_2^2\big] + \beta_{\text{VAE}}\, \mathbb{E}\Big[\mathbb{E}_\varepsilon\big[\|Y - f(g(Y) + \varepsilon)\|_2^2\big]\Big]$$

with $\varepsilon \sim \mathcal{N}(0, I_d)$

- Proposal: **use conditional VAE as a proxy regression for PLiCCA.**
- First term: equivalent.
- Second term: similar, except with noise; $\mathbb{E}\Big[\mathbb{E}_{q(z|Y)}\big[\|Y - f(z)\|_2^2\big]\Big]$ is enforcing a weaker version of $\Sigma_{g(Y)} \succeq aI_d$.

Theorem

Fix positive constants δ and σ^2_{enc}. Suppose g and f are such that the reconstruction error $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right] < \delta$, and suppose that we model $q(z|y) \sim \mathcal{N}\left(g(y), \sigma^2_{\text{enc}} I_d\right)$. Then,

$$\operatorname{tr}\left(\Sigma_{g(Y)}\right) \geq \sigma^2_{\text{enc}} C(\delta), \tag{3}$$

where $C(\delta)$ is non-increasing in δ.

How to formalize similarity?

Theorem

Fix positive constants δ and σ^2_{enc}. Suppose g and f are such that the reconstruction error $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right] < \delta$, and suppose that we model $q(z|y) \sim \mathcal{N}\left(g(y), \sigma^2_{\text{enc}} I_d\right)$. Then,

$$\operatorname{tr}\left(\Sigma_{g(Y)}\right) \geq \sigma^2_{\text{enc}} C(\delta), \tag{3}$$

where $C(\delta)$ is non-increasing in δ.

- Summary: the proxy problem relaxes $\Sigma_{g(Y)} \succeq a I_d$ to $\operatorname{tr}\left(\Sigma_{g(Y)}\right)$ instead.

How to formalize similarity?

Theorem

Fix positive constants δ and σ^2_{enc}. Suppose g and f are such that the reconstruction error $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right] < \delta$, and suppose that we model $q(z|y) \sim \mathcal{N}\left(g(y), \sigma^2_{\text{enc}} I_d\right)$. Then,

$$\operatorname{tr}\left(\Sigma_{g(Y)}\right) \geq \sigma^2_{\text{enc}} C(\delta), \tag{3}$$

where $C(\delta)$ is non-increasing in δ.

- Summary: the proxy problem relaxes $\Sigma_{g(Y)} \succeq a I_d$ to $\operatorname{tr}\left(\Sigma_{g(Y)}\right)$ instead.
- Solving conditional VAE amenable to gradient descent.

Conditional Normalizing Flow

- Assume a dimension-reduced representation $W \in \mathbb{R}^d$ such that $Y = \psi(W) + \varepsilon$, with encoder $\phi : \mathbb{R}^q \to \mathbb{R}^d$ (e.g., a pretrained VAE).

Conditional Normalizing Flow

- Assume a dimension-reduced representation $W \in \mathbb{R}^d$ such that $Y = \psi(W) + \varepsilon$, with encoder $\phi : \mathbb{R}^q \to \mathbb{R}^d$ (e.g., a pretrained VAE).
- Introduce latent $Z \in \mathbb{R}^d$ and specify:

$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big)$$

Conditional Normalizing Flow

- Assume a dimension-reduced representation $W \in \mathbb{R}^d$ such that $Y = \psi(W) + \varepsilon$, with encoder $\phi : \mathbb{R}^q \to \mathbb{R}^d$ (e.g., a pretrained VAE).
- Introduce latent $Z \in \mathbb{R}^d$ and specify:

$$Z \mid X \sim \mathcal{N}\big(B^\top X, I_d\big)$$

Proxy Conditional NF Objective

Maximizing the conditional likelihood is equivalent to minimizing

$$\min_{\tilde{g},B} \mathbb{E}\big[\|\tilde{g}(W) - B^\top X\|_2^2\big] \; - \; \beta_{\mathrm{NF}}\, \mathbb{E}[\ln|\det J_{\tilde{g}}(W)|]\,,$$

where $J_{\tilde{g}}$ is the Jacobian of $\tilde{g}$.

Objectives side by side

PLiCCA regression form

$$\min_{\tilde{g},B} \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right]$$
$$\text{s.t. } \Sigma_{\tilde{g}(Y)} \succeq aI_d,\ \tilde{g} \in \mathcal{C}_{\mathrm{NF}}$$

Objectives side by side

PLiCCA regression form

$$\min_{\tilde{g},B} \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right]$$

$$\text{s.t. } \Sigma_{\tilde{g}(Y)} \succeq aI_d,\ \tilde{g} \in \mathcal{C}_{\text{NF}}$$

Proxy NF objective

$$\min_{\tilde{g},B} \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right] \tag{4}$$

$$\text{s.t. } \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b,\ \tilde{g} \in \mathcal{C}_{\text{NF}} \tag{5}$$

Objectives side by side

PLiCCA regression form

$$\min_{\tilde{g},B} \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right]$$

$$\text{s.t. } \Sigma_{\tilde{g}(Y)} \succeq aI_d,\ \tilde{g} \in \mathcal{C}_{\text{NF}}$$

Proxy NF objective

$$\min_{\tilde{g},B} \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right] \tag{4}$$

$$\text{s.t. } \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b,\ \tilde{g} \in \mathcal{C}_{\text{NF}} \tag{5}$$

- Proposal: **Use NF as a proxy regression for PLiCCA**.

Objectives side by side

PLiCCA regression form

$$\min_{\tilde{g},B} \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right]$$

$$\text{s.t. } \Sigma_{\tilde{g}(Y)} \succeq aI_d,\ \tilde{g} \in \mathcal{C}_{\text{NF}}$$

Proxy NF objective

$$\min_{\tilde{g},B} \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right] \tag{4}$$

$$\text{s.t. } \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b,\ \tilde{g} \in \mathcal{C}_{\text{NF}} \tag{5}$$

- Proposal: **Use NF as a proxy regression for PLiCCA**.
- First term: identical to regression fit.

Objectives side by side

PLiCCA regression form

$$\min_{\tilde{g},B} \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right]$$

$$\text{s.t. } \Sigma_{\tilde{g}(Y)} \succeq aI_d,\ \tilde{g} \in \mathcal{C}_{\text{NF}}$$

Proxy NF objective

$$\min_{\tilde{g},B} \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right] \tag{4}$$

$$\text{s.t. } \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b,\ \tilde{g} \in \mathcal{C}_{\text{NF}} \tag{5}$$

- Proposal: **Use NF as a proxy regression for PLiCCA**.
- First term: identical to regression fit.
- Second term: $\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b$ a *local* constraint replacing $\Sigma_{\tilde{g}(Y)} \succeq aI_d$.

Theorem

Let (m, M) be the bilipschitz constant bounds of $\mathcal{C}_{\mathrm{NF}}$ and $W \in \mathbb{R}^d$ be an isotropic Gaussian random vector. Then, if $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, we have, for $a > 0$,

$$\Sigma_{\tilde{g}(W)} \succeq a I_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a), \tag{6}$$

where $b(a) \equiv \frac{d}{2}\ln(a) - C$ where C is a constant that depends on m, M, as well as the Hessian matrices of the coordinates $\tilde{g}_i$ of $\tilde{g}$.

How to formalize similarity?

Theorem

Let (m, M) be the bilipschitz constant bounds of $\mathcal{C}_{\mathrm{NF}}$ and $W \in \mathbb{R}^d$ be an isotropic Gaussian random vector. Then, if $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, we have, for $a > 0$,

$$\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a), \tag{6}$$

where $b(a) \equiv \frac{d}{2}\ln(a) - C$ where C is a constant that depends on m, M, as well as the Hessian matrices of the coordinates $\tilde{g}_i$ of $\tilde{g}$.

- Log-determinant term enforces a weaker, local version of $\Sigma_{\tilde{g}(W)} \succeq aI_d$.

How to formalize similarity?

Theorem

Let (m, M) be the bilipschitz constant bounds of $\mathcal{C}_{\mathrm{NF}}$ and $W \in \mathbb{R}^d$ be an isotropic Gaussian random vector. Then, if $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, we have, for $a > 0$,

$$\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a), \tag{6}$$

where $b(a) \equiv \frac{d}{2}\ln(a) - C$ where C is a constant that depends on m, M, as well as the Hessian matrices of the coordinates $\tilde{g}_i$ of $\tilde{g}$.

- Log-determinant term enforces a weaker, local version of $\Sigma_{\tilde{g}(W)} \succeq aI_d$.
- **Takeaway:** conditional NF can be viewed as a relaxation of the PLiCCA problem.

Corollary

Fix $a > 0$ and let $W \in \mathbb{R}^d$ be an isotropic Gaussian random vector. Then the PLiCCA problem,

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in a^{-1/2}\mathcal{C}_{\mathrm{NF}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{7}$$

can be relaxed into a NF problem

$$\underset{\substack{\tilde{g} \in \mathcal{C}_{\mathrm{NF}}, B \in \mathbb{R}^{p\times d} \\ \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a)}}{\text{minimize}} \quad \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right] \tag{8}$$

via the previous result.

Further connection between NF and PLiCCA

- We have shown $\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a)$.

Further connection between NF and PLiCCA

- We have shown $\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a)$.
- We can further relax the Jacobian constraint.

Further connection between NF and PLiCCA

- We have shown $\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a)$.
- We can further relax the Jacobian constraint.

Lemma

Let $W \in \mathbb{R}^d$ be an isotropic Gaussian random vector. Then, if $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, we have

$$\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b \implies \det\left(\Sigma_{\tilde{g}(W)}\right) \geq c(b). \tag{9}$$

where $c(b) \equiv e^{2b}$.

Conditional NF $\implies$ geometric PLiCCA

Theorem

Suppose that $W \in \mathbb{R}^d$ is an isotropic Gaussian random vector. Then the alternate NF problem,

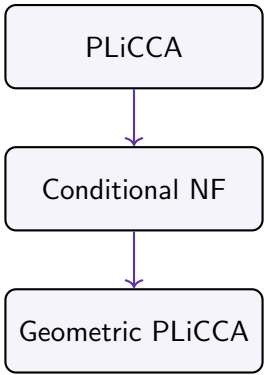
$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\mathrm{NF}},B\in\mathbb{R}^{p\times d}\\ \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right]\geq b}}{\text{minimize}} \quad \mathbb{E}\left[\left\|\tilde{g}(W)-B^{\top}X\right\|_2^2\right], \tag{10}$$

can be relaxed into a 'geometric' PLiCCA problem via the previous result,

$$\sup_{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^{\top}X}=I_d\\ \tilde{g}\in c(b)^{-1/2}\mathcal{C}_{\mathrm{NF}}}} \left(\prod_{i=1}^{d} h\left(\rho_i\right)\right)^{1/d}, \tag{11}$$

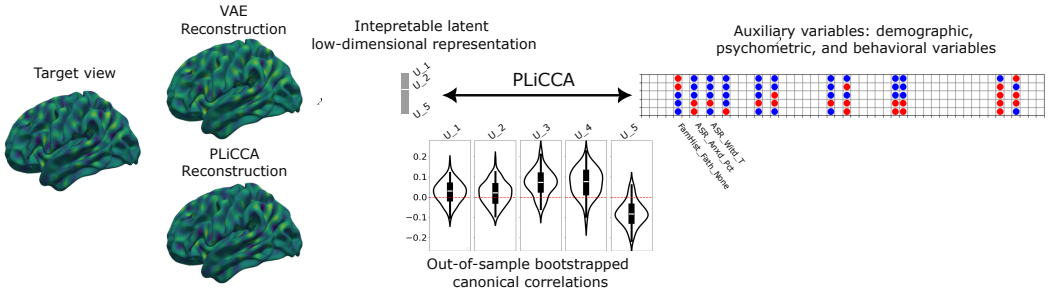
where $h(x) = \frac{1}{1-x^2}$, $\rho_i = \mathbb{E}\left[\tilde{g}_i(W)\theta_i^{\top}X\right]$, and the $\theta_i \in \mathbb{R}^p$ are the columns of T.

Recap on conditional NF and PLiCCA

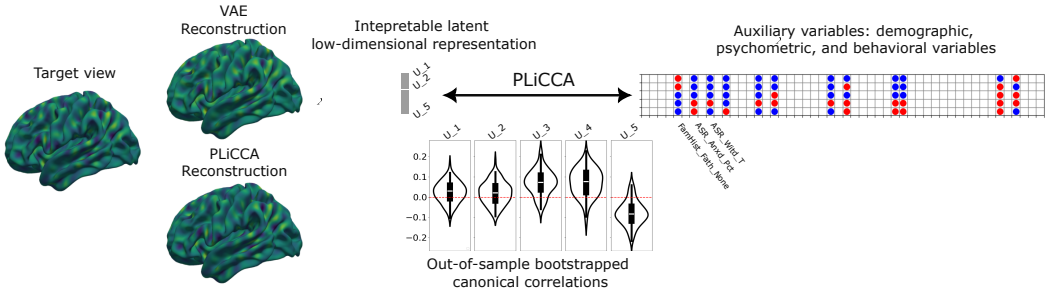


NF is *sandwiched* between PLiCCA and geometric PLiCCA.

Application



Application



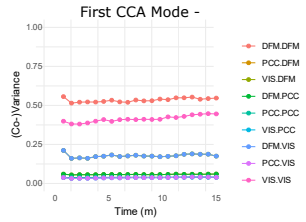
Currently in review at AISTATS 2026

Thank you!

Questions?

Application

Connectivity



Behaviour