Integrative analysis of non-Euclidian data

James Buenfil

University of Washington

October 28th, 2024

Overview

Asymmetric Canonical Correlation Analysis of Riemannian and High-dimensional data

Introduction

Large-scale neuroimaging studies - asymmetric data

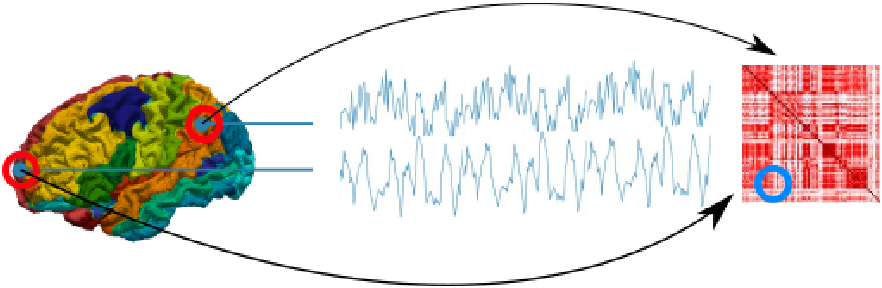
- A primary goal of studies like the Human Connectome Project, ABCD, and the UK Biobank is to **understand the relationship** between brain imaging data and non-imaging high-dimensional variables.

- **Imaging data** come from fMRI data which are summarized via a covariance matrix.

- **High dimensional variables** include measures of cognitive ability, neurodegenerative conditions, mental health disorders, psychometric test scores, and other external factors.

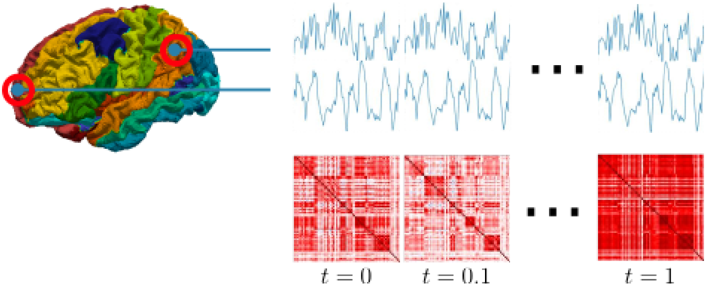
Imaging Data - Static functional connectivity

- For each patient, form a covariance matrix based on signals from m different brain regions.
- $y_i \in \mathbb{R}^{m \times m}$ for each patient $i = 1, \ldots N$.



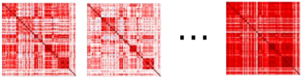
Imaging data - Dynamic functional connectivity

- In contrast to static functional connectivity is dynamic functional connectivity: if we partition the time interval into smaller parts, we can form several covariance matrices for each patient.
- $y_i(t_1), \ldots y_i(t_L) \in \mathbb{R}^{m \times m}$ for each patient $i = 1, \ldots N$.
- The set of $m \times m$ covariance matrices form a manifold $\mathcal{M}$, with several Riemannian metrics of interest for different applications.



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 $i = 1, \ldots N$ and where each y_i is observed at discrete timepoints t_l for $l = 1, \ldots L$: $y_i(t_l)$.



$y(\cdot)$

X

Generalizing Canonical Correlation Analysis

How can we study the relationship between X and y?

- Suppose $y \in \mathbb{R}^q$, multivariate data.
- We can use multivariate linear regression: $\underset{B \in \mathbb{R}^{q \times p}}{\text{minimize}}\ \mathbb{E}\left[\|y - BX\|_2^2\right]$
- Interpretation of B derives from the fact that B maps X onto the y scale.
- B contains 'joint' information about both X and Y.

Introduction to CCA

- Given random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, classical CCA solves the following problem:

$$\underset{a \in \mathbb{R}^p, b \in \mathbb{R}^q}{\text{maximize}} \ \text{Corr}^2\left(\langle a, X\rangle, \langle b, Y\rangle\right).$$

- $\langle a, X\rangle \equiv a^\top X$. Since the problem is invariant to scaling of a and b, a and b are rescaled so that $\text{Var}\left(\langle a, X\rangle\right) = \text{Var}\left(\langle b, Y\rangle\right) = 1$.

- The pair of random variables $U \equiv \langle a, X\rangle$ and $V \equiv \langle b, Y\rangle$ are called the *first pair of canonical scores* (or canonical variables).

- The solution pair (a, b) is called the *first pair of canonical directions* (or canonical vectors).

Introduction to CCA

- We can define subsequent pairs of canonical vectors

$$(a_1, b_1) = \underset{a \in \mathbb{R}^p, b \in \mathbb{R}^q, \mathrm{Var}(\langle a, X \rangle) = \mathrm{Var}(\langle b, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle a, X \rangle, \langle b, Y \rangle \right),$$

$$(a_k, b_k) = \underset{\substack{a \in \mathbb{R}^p, b \in \mathbb{R}^q, \mathrm{Var}(\langle a, X \rangle) = \mathrm{Var}(\langle b, Y \rangle) = 1 \\ \mathrm{Corr}(\langle a, X \rangle, \langle a_i, X \rangle) = 0, i = 1, \ldots, k-1 \\ \mathrm{Corr}(\langle b, Y \rangle, \langle b_i, Y \rangle) = 0, i = 1, \ldots, k-1}}{\arg\max} \mathrm{Corr}^2 \left(\langle a, X \rangle, \langle b, Y \rangle \right)$$

for $k = 2, \ldots \min(p, q)$

- When $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, there are always at most $\min(p, q)$ nontrivial canonical vector pairs (a_k, b_k).

- In practice, we observe i.i.d. pairs (X_i, Y_i) for $i = 1, \ldots N$.

Can we generalize classical CCA?

- Data: $X \in \mathbb{R}^p$, $y : [0,1] \to \mathcal{M}$
- The classical CCA model solves

$$(a_1, b_1) = \underset{a \in \mathbb{R}^p, b \in \mathbb{R}^q, \mathrm{Var}(\langle a, X \rangle) = \mathrm{Var}(\langle b, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle a, X \rangle, \langle b, Y \rangle \right).$$

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

Can we generalize classical CCA?

- Data: $X \in \mathbb{R}^p$, $y : [0,1] \to \mathcal{M}$
- The classical CCA model solves

$$(a_1, b_1) = \underset{a \in \mathbb{R}^p, b \in \mathbb{R}^q, \mathrm{Var}(\langle a, X \rangle) = \mathrm{Var}(\langle b, Y \rangle) = 1}{\arg\max} \mathrm{Corr}^2 \left(\langle a, X \rangle, \langle b, Y \rangle \right).$$

- Do we have an analogue of $\langle b, y \rangle$ for $y : [0,1] \to \mathcal{M}$?
- No, since we don't necessarily have an **inner product structure** on a non-Euclidian $\mathcal{M}$.

Geodesic distance:

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

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

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

Logarithmic map:

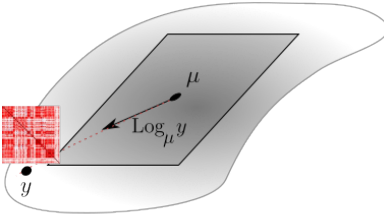
- $\mathrm{Log}_x(\cdot) : \mathcal{M} \to T_x\mathcal{M}$

Exponential map:

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

Frechet mean:

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



Geometry of positive definite matrices

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

- **Affine-invariant property:** $d_{\mathcal{M}}(\Sigma_X, \Sigma_Y) = d_{\mathcal{M}}(\Sigma_{RX}, \Sigma_{RY})$ for any orthogonal matrix $R \in \mathbb{R}^{m\times m}$, random vectors $X, Y \in \mathbb{R}^m$.
- **Tangent spaces $T_P\mathcal{M}$:** isomorphic to the set of $m \times m$ symmetric matrices.
- **Riemannian metric:** $P \in \mathcal{M}$ between $W, Z \in T_P\mathcal{M}$ is defined as $\langle W, Z\rangle_{\mathcal{M}} = \operatorname{tr}\left(P^{-1}WP^{-1}Z\right)$.
- **Logarithmic map:** $\operatorname{Log}_P(Q) = P^{1/2}\log\left(P^{-1/2}QP^{-1/2}\right)P^{1/2}$
  - Maps manifold representation to tangent space representation.
- **Exponential map:** $\operatorname{Exp}_P(W) = P^{1/2}\exp\left(P^{-1/2}WP^{-1/2}\right)P^{1/2}$
  - Maps tangent space representation to manifold representation.
- Log and Exp are global bijections.

Move $y : [0,1] \to \mathcal{M}$ to its tangent space representation

Define **Frechet mean** μ of y:

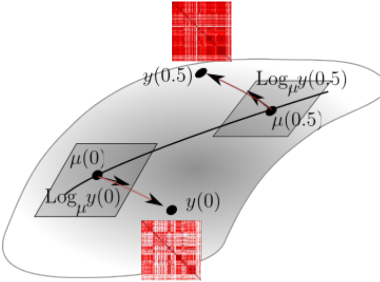
- $\mu(t) = \underset{x \in \mathcal{M}}{\arg\min}\ \mathbb{E}\left[d^2_{\mathcal{M}}(y(t), x)\right]$.

Define y's **tangent space representation:**

- $\mathrm{Log}_\mu\, y : t \mapsto \mathrm{Log}_{\mu(t)}\, y(t)$.
- $\forall t \in [0,1], \mathrm{Log}_{\mu(t)}\, y(t) \in T_{\mu(t)}\mathcal{M}$

Vector fields with this property $\{V(t) : \forall t \in [0,1], V(t) \in T_{\mu(t)}\mathcal{M}\}$ **form a vector space:**

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



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

$$\underset{a \in \mathbb{R}^p, b \in L^2(T\mu)}{\text{maximize}} \; \text{Corr}^2 \left(\langle\!\langle b, \text{Log}_\mu y \rangle\!\rangle_\mu, \langle a, X \rangle \right) \tag{1}$$

  subject to the constraints that

$$\text{Var}\left(\langle a, X \rangle\right) = \text{Var}\left(\langle\!\langle b, \text{Log}_\mu y \rangle\!\rangle_\mu\right) = 1.$$

- a and b called the canonical directions.

- In order to interpret the canonical direction b, we can **map it back to the manifold via the exponential map** along μ.
  $t \mapsto \text{Exp}_{\mu(t)}\left(b(t)\right)$

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

$$\underset{a\in\mathbb{R}^p, b\in L^2(T\mu)}{\text{maximize}} \ \text{Corr}^2\left(\langle\!\langle b, \text{Log}_\mu y\rangle\!\rangle_\mu, \langle a, X\rangle\right) \tag{1}$$

subject to the constraints that

$$\text{Var}\left(\langle a, X\rangle\right) = \text{Var}\left(\langle\!\langle b, \text{Log}_\mu y\rangle\!\rangle_\mu\right) = 1.$$

- a and b called the canonical directions.

- In order to interpret the canonical direction b, we can **map it back to the manifold via the exponential map** along μ.
$t \mapsto \text{Exp}_{\mu(t)}\left(b(t)\right)$

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

Approach to the problem

How to handle infinite dimensional $L^2(T\mu)$?

We don't expect $\operatorname{Log}_\mu y$ to vary in its infinite dimensional space along many directions.

- **Use dimensionality reduction.**

We can use the data-driven functional Principal component analysis (FPCA) to find the 'best' finite dimensional basis to project $\operatorname{Log}_\mu y$ into.

- **This reduces $\operatorname{Log}_\mu y$ to a multivariate $Y \in \mathbb{R}^d$.**

$\operatorname{Log}_\mu y \approx \sum_{j=1}^d \phi_j Y_j$ for functions $\phi_j \in L^2(T\mu)$ and a random vector $Y \in \mathbb{R}^d$.

- ϕ_j are the principal components.
- Y_j are the principal scores.

Problem Reformulation: Multivariate CCA

$$\underset{a \in \mathbb{R}^p, \eta \in \mathbb{R}^d}{\text{maximize}} \ \operatorname{Corr}^2(\langle\langle \eta, Y \rangle\rangle, \langle a, X \rangle) \tag{2}$$

b related to η via $b = \sum_{k=1}^d \eta_k \phi_k$.

Multivariate CCA with high-dimensional X

- $X \in \mathbb{R}^p$, p large and $Y \in \mathbb{R}^d$, d small.
- In practice we have N samples: if N is smaller than p, then classical CCA fails.
  - Classical CCA uses an estimate of the **precision matrix** Σ_X^{-1}.
  - **Estimation for $N < p$ is hard:** requires either strong assumptions on the form of Σ_X or Σ_X^{-1}.
- In the high-dimensional setting, we would like to do variable selection.
  - Even if $N > p$, classical CCA does not perform variable selection.
  - Ideally we would have **sparse** estimates a_k: then the canonical variable $a_k^\top X$ ignores the corresponding X variables where there are 0s in a_k.
  - Sparse canonical directions are much more interpretable.

Sparse regression implies sparse CCA

Theorem

Letting B be the solution to the multivariate least-squares problem

$$\underset{B\in\mathbb{R}^{p\times d}}{\text{minimize}}\ \mathbb{E}\left[\|\Sigma_Y^{-1/2}Y - B^\top X\|_2^2\right], \tag{3}$$

we can find all canonical vectors for both Y and X via B.
Denote the canonical vectors associated with X (the a_k) as the columns of A, and those with Y (the η_k) as the columns of H. Let the eigenvector decomposition of the matrix $B^\top\Sigma_X B$ be $ED^2E^\top$ where $E\in\mathbb{R}^{d\times d}$ is orthogonal, and $D\in\mathbb{R}^{d\times d}$ is diagonal. Then, $H = \Sigma_Y^{-1/2}E$ and $A = BED^{-1}$.

Sparse regression implies sparse CCA

Theorem

Letting B be the solution to the multivariate least-squares problem

$$\underset{B \in \mathbb{R}^{p \times d}}{\text{minimize}} \; \mathbb{E}\left[\|\Sigma_Y^{-1/2} Y - B^\top X\|_2^2 \right], \tag{3}$$

we can find all canonical vectors for both Y and X via B.
Denote the canonical vectors associated with X (the a_k) as the columns of A, and those with Y (the η_k) as the columns of H. Let the eigenvector decomposition of the matrix $B^\top \Sigma_X B$ be $ED^2E^\top$ where $E \in \mathbb{R}^{d \times d}$ is orthogonal, and $D \in \mathbb{R}^{d \times d}$ is diagonal. Then, $H = \Sigma_Y^{-1/2} E$ and $A = BED^{-1}$.

Sparsity in the regression matrix B is carried over into our estimates of the canonical vectors for X.

- If B has only s non-zero rows, then A has only s non-zero rows.

Methodology

1. We are **given** $(X_i, y_i(t_l))$ pairs for $i = 1, \ldots N$, $l = 1, \ldots L$, where $X_i \in \mathbb{R}^p$, $y_i : [0,1] \to \mathcal{M}$.
2. Estimate the **Frechet mean** of the $\{y_i(t_l)\}_{i=1,\ldots N}$ for every l, forming $\hat{\mu}(t_l)$.
3. **Compute** $\operatorname{Log}_{\hat{\mu}(t_l)} y_i(t_l) \in T_{\hat{\mu}(t_l)}\mathcal{M}$ for all i and l.
4. Summarize the $\operatorname{Log}_{\hat{\mu}} y_i$ as $Y_i \in \mathbb{R}^d$ with **FPCA**: $\operatorname{Log}_{\hat{\mu}} y \approx \sum_{j=1}^d \hat{\phi}_j Y_j$ for functions $\hat{\phi}_j \in L^2(T\hat{\mu})$, $j = 1, \ldots d$.
5. Compute $\hat{B}$ solving the **group lasso problem**

$$\hat{B} = \operatorname*{arg\,min}_{B \in \mathbb{R}^{p \times d}} \frac{2}{N} \left\| \mathbb{Y} \hat{\Sigma}_Y^{-1/2} - \mathbb{X} B \right\|_F^2 + \lambda \left\| B \right\|_{\ell_1, \ell_2} \tag{4}$$

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

Theory

Special case of multivariate 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 $\lambda = O\left(\sqrt{\frac{d \log(p)}{N}}\right)$ rate.

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

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

Bounds for full algorithm

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 a_k satsify an group s-sparsity condition.

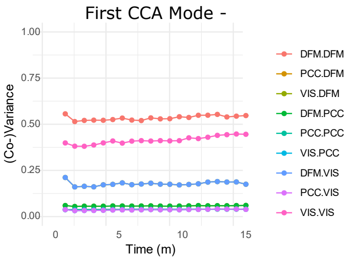
$$\|a_k - \hat{a}_k\|_2^2 = O_P\left(\frac{ds\log(p)}{N} \frac{1}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2} \right),$$

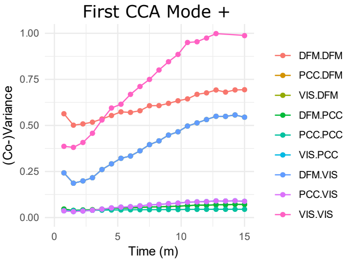
$$\|b_k - \Gamma_{\hat{\mu},\mu}\hat{b}_k\|_\mu^2 = O_P\left(\frac{d^2 s\log(p)}{N} \frac{1}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)^2} \right).$$

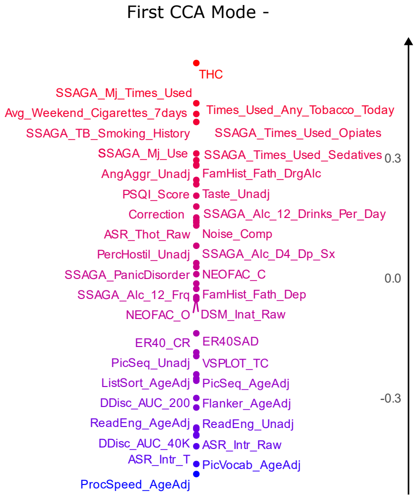
Application

Connectivity

Behaviour

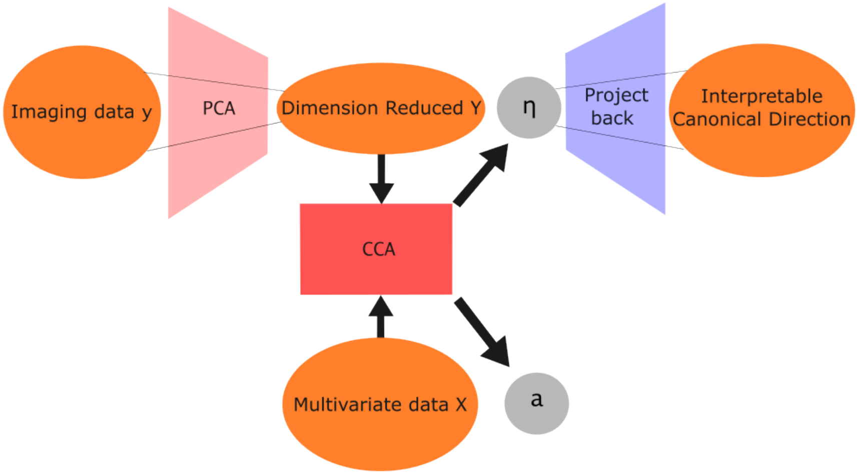






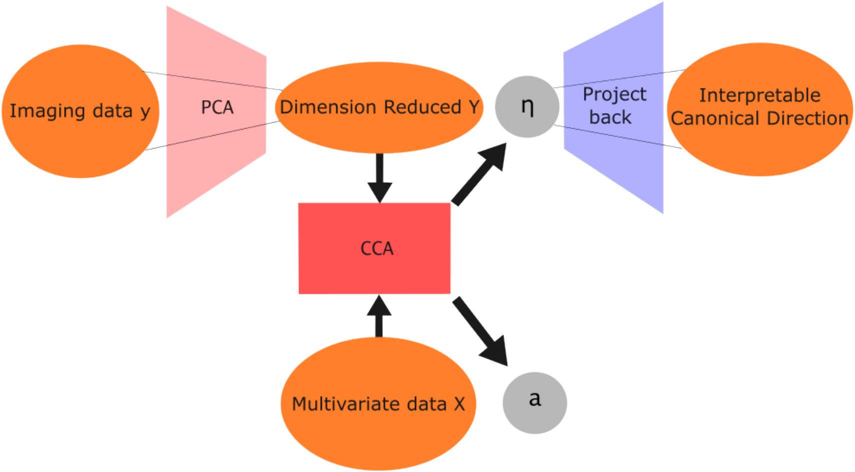
Canonical corrrelation analysis via Variational Autoencoders

Motivation



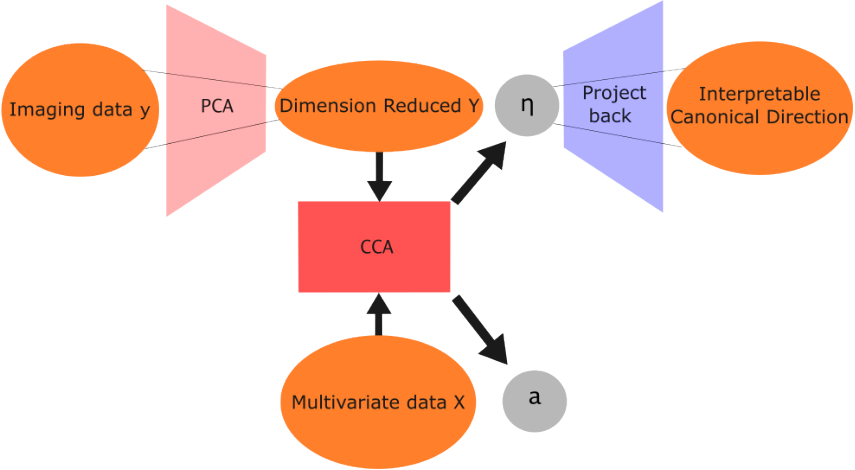
Imaging data y
PCA
Dimension Reduced Y
η
Project back
Interpretable Canonical Direction
CCA
Multivariate data X
a

Motivation



- Dimension reduction and CCA are not done jointly.

Motivation



- Dimension reduction and CCA are not done jointly.
- Nonlinear mapping (moving to tangent spaces) is prespecified.

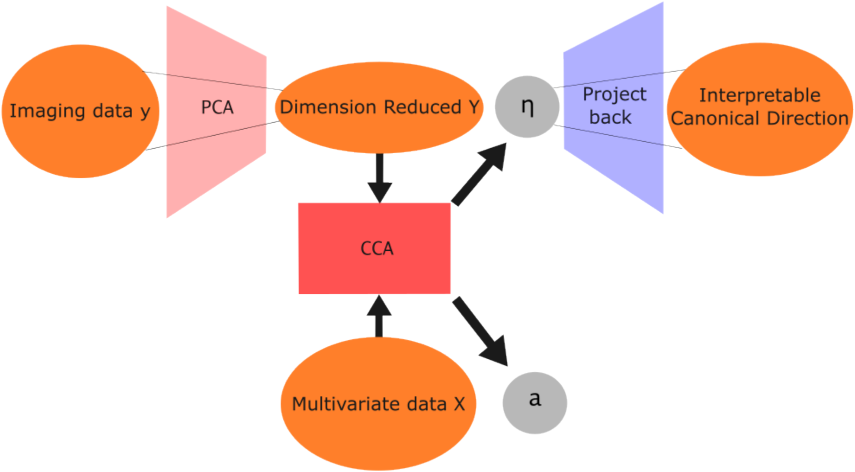
Motivation cont.

- On the other hand, we can think of the dimension reduced Y as a **latent variable.**
- Both PCA and CCA can be defined in terms of latent variable models.
- Probabilistic PCA: $Y \in \mathbb{R}^d, y \in \mathbb{R}^q$, $q > d$:

$$Y \sim \mathcal{N}(0, I_d) \tag{5}$$
$$y \sim \mathcal{N}\left(WY + \mu, \sigma^2 I_q\right) \tag{6}$$

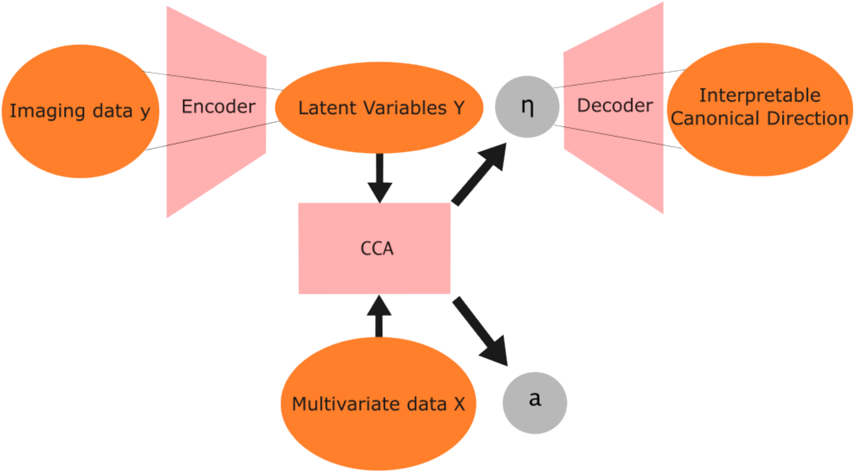
- Given a finite sample $\{Y_i\}_{i=1,\ldots N}$, the maximum likelihood solution for W **reduces to classical PCA** as σ approaches 0.

Original Model

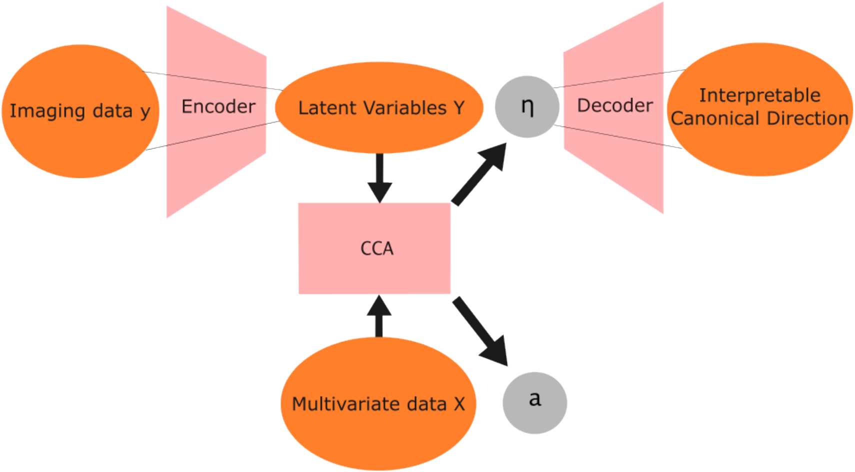


- Dimension reduction and CCA are not done jointly.
- Nonlinear mapping (moving to tangent spaces) is prespecified.

Model - Variational Autoencoder



Model - Variational Autoencoder



- We still perform CCA by learning the regression matrix B, but now the encoder and decoder (represented by a neural networks) is a learned nonlinear mapping.

Model - Variational Autoencoder

- Data $X_i \in \mathbb{R}^p$ and imaging data y_i, $i = 1, \dots N$.

$$\hat{\xi}, \hat{\gamma}, \hat{B} = \underset{B \in \mathbb{R}^{p \times d}, \xi, \gamma}{\arg\min} \sum_{i=1}^{N} \underbrace{\| y_i - D_{\xi}(Y_i) \|_2^2}_{\text{Image reconstruction error}} + \underbrace{D_{KL}\left(q_{\gamma}(\cdot|y_i), p_{\mathcal{N}}(\cdot)\right)}_{\text{Distribution of latent variables}} + \underbrace{\left\| Y_i - B^{\top} X_i \right\|_2^2}_{\text{CCA via Regression}}$$

- ξ are the parameters for the decoder, while γ control the parameters for the encoder.
- We can then apply the same eigenvector approach as before to learn the canonical vectors via B, relative to X and Y.
- The canonical vectors for y can then be mapped through the decoder: $b_k = D_{\xi}(\eta_k)$

Conclusions

Conclusions

- We define the CCA problem in the asymmetric setting of X multivariate and $y : \mathcal{T} \to \mathcal{M}$, by utilizing the Frechet mean and Logarithmic map on $\mathcal{M}$.
- Theoretical guarantees for manifold and multivariate cases.
- We use our methodology to find shared correlation structure between dynamical functional connectivity and subject traits.
- We generalize our model from the first project via variational autoencoders to automatically uncover non-linear structure.

Thank you!

Questions?