Integrative Analysis of Non-Euclidean Data

James Buenfil

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Eardi Lila, Chair
Marina Meila
Yen-Chi Chen

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Integrative Analysis of Non-Euclidean Data

James Buenfil

Chair of the Supervisory Committee:
Eardi Lila
Department of Biostatistics

In large-scale imaging studies, a primary goal is to understand the relationship between distinct data views of study participants. For example, one data view could consist of patients' brain MRI scans, while a second view includes their lifestyle, demographic, or psychometric measures. A significant challenge is that these views are often subject to complex non-Euclidean constraints. Two settings arise: in some cases, the geometric constraints are known a priori, such as brain functional connectivity data which lie on the manifold of positive definite matrices. In other cases, no explicit manifold representation is available, and the underlying geometry must be learned from the data. Additionally, the relationships between these views are often weak, further complicating the analysis.

Despite extensive work on data integration, most approaches fail to accommodate non-Euclidean constraints while providing interpretable embeddings. In this dissertation, we propose novel frameworks to identify interpretable relationships between heterogeneous data views, while accounting for their distinct underlying structures.

Specifically, in Chapter 2, we develop a canonical correlation analysis model to integrate time-varying, manifold-valued data with high-dimensional data. Our approach leverages tools from Riemannian geometry to handle non-Euclidean constraints and introduces a group-sparsity penalty to select important variables. The proposed method shows improved empirical performance over existing approaches and is applied to dynamic functional

connectivity data from the Human Connectome Project. Furthermore, we establish asymptotic consistency through both in-sample and out-of-sample error bounds for the estimated canonical directions and scores.

In Chapter 3, we extend the proposed model to automatically learn interpretable embeddings from the data, thereby estimating its underlying geometry. To achieve this, we formulate a *Partially Linear interpretable Canonical Correlation Analysis* model (PLiCCA) and prove the existence of population solutions. We establish formal connections between PLiCCA and conditional latent-variable models, specifically, conditional variational autoencoders and conditional normalizing flows. We show that these latent-variable models can be interpreted as relaxations of the PLiCCA problem, where difficult global constraints are replaced by tractable local ones. This perspective enables efficient solving of PLiCCA via 'proxy' problems derived from contemporary conditional generative models, providing an alternative to the models proposed in the first project when the underlying structure of the data is unknown.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I want to thank my research advisor, Eardi Lila, for his support and guidance throughout my PhD. I am especially grateful for the time and care he invested in me, which shaped both my work and the researcher I am today. It has truly been a pleasure working with him.

I am also grateful to Marina Meila, whose mentorship during my early years in the program helped define my approach to research, and whose continued support has been instrumental. I thank Marina, Yen-Chi Chen, Zeyu Wei, and Yikun Zhang for the opportunity to co-organize the geometric data analysis group, and all its members for the excellent discussions. I additionally thank my committee—Eardi, Marina, Yen-Chi, Bamdad Hosseini, and Ali Shojaie—for providing insightful feedback. I would also like to acknowledge my academic advisor, Thomas Richardson, for his positive outlook during our advising meetings.

I am fortunate to have been surrounded by friends and a PhD cohort who made this a wonderful journey: Alan Min, Alana McGovern, Alex Jiang, Alex Kokot, Anupreet Porwal, Apara Venkat, Jess Kunke, Jeffrey Li, Jillian Fisher, Kenny Zhang, Maia Lathrop, Medha Agarwal, Nobuaki Masaki, Ronak Mehta, Saksham Jain, Samson Koelle, Trinity Fan, and Vydhourie Thiyageswaran. Special thanks to my partner, Shreya Prakash, who has been my steady source of support and joy throughout this journey.

I want to thank Joris Roos, Benjamin Peherstorfer, and Garvesh Raskutti from my time at UW-Madison as an undergraduate; their time, mentorship, and teaching fostered my love of learning and inspired me to pursue the PhD.

Finally, I thank my family for their unwavering belief in me, which has made everything I've accomplished possible.

DEDICATION

To my family

Chapter 1

INTRODUCTION

Broadly defined, the problem of *Integrative Data Analysis* is to understand the relationship between two distinct *data views*. As a motivating example, consider a medical setting in which we have access to (1) brain imaging data from patients, such as functional magnetic resonance imaging (fMRI) scans, and (2) multivariate measurements on the same patients, including lifestyle, demographic, or psychometric variables. Because the two views differ in nature, we refer to this setting as *asymmetric*. This heterogeneity requires distinct regularization strategies to appropriately control estimation complexity. The goal of integrative data analysis is then to characterize the dependence structure between the two views, by properly accounting for such asymmetry.

We treat the brain imaging data as the 'target' view and seek an interpretable, low-dimensional embedding that captures its complex, non-Euclidean structure. The multivariate data serve as the 'auxiliary' view, from which we perform variable selection through a sparse linear embedding designed to produce a latent representation maximally correlated with that of the target view. The resulting methodologies enable efficient discovery of previously unknown associations between the two data views of interest.

1.1 Introduction to Canonical Correlation Analysis

The flagship method for integrative data analysis is canonical correlation analysis (CCA), which we introduce in its classical linear form. Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors, and assume that their covariance matrices Σ_X and Σ_Y are invertible.

The population CCA problem between Y and X is formulated as

$$\underset{\substack{\mathrm{Corr}\left(\eta_i^\top Y, \eta_j^\top Y\right)=\delta_{ij} \\ \mathrm{Corr}\left(\theta_i^\top X, \theta_j^\top X\right)=\delta_{ij}}}{\text{maximize}} \sum_{i=1}^{d} \mathrm{Corr}\left(\eta_i^\top Y, \theta_i^\top X\right), \tag{1.1}$$

where $\eta_j \in \mathbb{R}^q$ and $\theta \in \mathbb{R}^p$ are the *canonical vectors*, and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The resulting correlations $\mathrm{Corr}\left(\eta_i^\top Y, \theta_i^\top X\right)$ are called the *canonical correlations* and are denoted as $\gamma_1, \dots \gamma_d$. Collecting the canonical vectors into matrices, $H \equiv [\eta_1 \dots \eta_d] \in \mathbb{R}^{q \times d}$ and $T \equiv [\theta_1, \dots \theta_d] \in \mathbb{R}^{p \times d}$, the CCA problem can be equivalently expressed as

$$\underset{\substack{H \in \mathbb{R}^{q \times d}, T \in \mathbb{R}^{p \times d} \\ \Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d}}{\text{maximize}} \mathbb{E}\left[(H^\top Y)^\top T^\top X\right]. \tag{1.2}$$

Replacing the correlation constraint with a covariance constraint resolves the scaling indeterminacy arising from the fact that correlations are invariant to scaling. From this formulation, we observe that the CCA solution is not unique: if (T, H) is a solution, then so is (RT, RH) for any orthogonal matrix $R \in \mathbb{R}^{d \times d}$ so that $R^\top R = I_d$. This rotational ambiguity can be removed by simply requiring that the cross-covariance matrix $\Sigma_{T^\top X, H^\top Y}$ is diagonal. We are still left with a permutation ambiguity, where reordering the columns of H and T does not change the objective function, but provided that the canonical correlations $\gamma_1, \dots \gamma_d$ are distinct, this ambiguity is resolved by ordering the vectors according to decreasing $\gamma_1 > \gamma_2 \dots . \gamma_d$. Finally, there is a trivial sign ambiguity: each pair (η_k, θ_k) can be simultaneously multiplied by -1 without changing the objective, but this can be ignored as the canonical vectors are interpreted jointly.

More generally, we interpret the matrices H and T as view-specific embeddings $H^\top(y) : \mathbb{R}^q \to \mathbb{R}^d$ and $T^\top(x) : \mathbb{R}^p \to \mathbb{R}^d$, optimized to capture the information shared between the two data views Y and X. These embeddings are ordered according to the strength of the correlation found between the learned latent variables $H^\top Y$ and $T^\top X$, referred to as the canonical variables $U \equiv H^\top Y \in \mathbb{R}^d$ and $V \equiv T^\top X \in \mathbb{R}^d$. Since these are linear embeddings, each canonical variable admits a straightforward interpretation as a linear combination of the original features from its respective data view.

Linear CCA is appealing thanks to its simplicity and interpretability. However, it struggles to handle data that do not lie on linear subspaces as the learned embeddings are restricted to linear subspaces. Moreover, when the dimensionality of the data exceeds the number of available samples, linear CCA overfits the data, and generalizes poorly out-of-sample. In such high-dimensional settings, the canonical variables also lose their interpretability, as taking a linear combination of a large number of covariates is not informative.

In this dissertation, we address both of these limitations within a flexible framework that accommodates heterogeneous data structures. Specifically, we allow the target data view Y to exhibit an underlying non-Euclidean geometry, while the auxiliary data view X consists of high-dimensional multivariate measurements. The linear embedding $H^\top$ is generalized to a nonlinear mapping $g : \mathbb{R}^q \to \mathbb{R}^d$, whereas $T^\top$ remains linear but is learned with sparsity-promoting regularization to retain interpretability. We appeal to generalizations of CCA as they aim to characterize the dependence structure between multiple views, while also allowing for an asymmetric modeling approach by providing two separate embeddings for each data view. In the following, we discuss the challenges posed by high dimensionality and non-Euclidean geometry in greater detail.

1.2 Modeling high-dimensional data

When the number of covariates exceeds the sample size, we enter the high-dimensional regime, where classical statistical results and guarantees often fail. To make inference possible, it is common to assume sparsity, that is, only a small subset of covariates meaningfully contribute to given task. Practically, this assumption typically manifests itself as an ℓ_1 penalty on the quantities of interest, encouraging the model to fully ignore a large subset of the covariates. Indeed, the vanilla sparse CCA problem, for the first pair of canonical directions, is to solve

$$\max_{\eta,\theta} \ \mathrm{Corr}(\eta^\top Y,\ \theta^\top X) \tag{1.3}$$

$$\text{s.t.} \quad \|\eta\|_1 \le c_1,\ \|\theta\|_1 \le c_2, \tag{1.4}$$

where for $v \in \mathbb{R}^d$, $\|v\|_1$ is the sum of the absolute values of the entries of v. Such approaches are attractive for high-dimensional settings; however, enforcing the orthogonality constraints $\Sigma_{H^\top Y} = \Sigma_{T^\top X} = I_d$ while simultaneously maintaining sparsity is non-trivial.

There is ongoing work on sparse CCA methods that adapt linear CCA to handle high-dimensional data (Bykhovskaya & Gorin, 2023; Li et al., 2024). Our methodologies build on this work and borrow strategies from high-dimensional statistics in order to accommodate the potentially high-dimensional auxiliary data view X.

1.3 Modeling non-linear data

In the asymmetric framework studied in this thesis, one of the data views is assumed to lie on a nonlinear manifold; in this case, linear models are not sufficient. We consider two primary settings: one in which the nonlinear geometry of the data is known a priori, and another in which it must be learned directly from the data. The former can be formalized as having access to a *manifold representation* of the data in advance. Such representations arise naturally in many domains, for instance, when observations are probability densities (Cho et al., 2022), covariance or correlation matrices residing on the manifold of positive definite matrices (Kim et al., 2014), or points on a sphere (Zhang & Chen, 2023). Provided with a manifold representation, we gain access to geometric tools that enable the application of Euclidean statistical techniques to non-Euclidean data. Of particular importance for us are the tangent spaces of the manifold, which provide linearizations of the manifold in local neighborhoods. The exponential and logarithmic maps allow us to move between the manifold and its tangent spaces, and the Riemannian metric endows these spaces with an inner-product structure that reflects the manifold's geometry. In the context of CCA, related extensions have been developed when such manifold representations are available (Kim et al., 2014).

In other scenarios, the manifold structure of the data is not known in advance, and we must learn it from the data. This is a considerably more challenging but general setting, and has seen extensive work. Classical approaches include dimensionality reduction techniques

that, in an unsupervised manner, usually aim to ensure that a geometric quantity of interest present in the original data is preserved in the dimension-reduced representation (Meilă & Zhang, 2024). Deep learning methods for dimensionality reduction, especially autoencoder-based techniques, have gained significant traction in recent years; see Alberti et al. (2024) for contemporary examples. Adaptation of these unsupervised approaches to the supervised setting, that is, when auxiliary data is also available, have been explored. However, much of this work focuses on the so called multi-view data problem rather than the integrative data analysis problem that we are interested in (Lyu et al., 2022; Senellart et al., 2023). Multi-view approaches typically seek to find a joint representation of the data views, as opposed to identifying a representation of the target view that is informed by an auxiliary view.

This dissertation proceeds in two main parts. Chapter 2 addresses the case when manifold representations are known a priori for the target view and we can leverage the many tools of differential geometry to aid us in characterizing the dependence structure between the two data views. Chapter 3 considers the complementary setting, where no manifold representation for the target view is available in advance. In this case, we utilize conditional latent variable models in order to learn the geometry of the target view directly from the data. Across both chapters, the auxiliary view is high-dimensional, and by using sparsity-inducing regularization, we can interpret the target view in terms of a small subset of the auxiliary variables. The remainder of this section provides an overview of these works.

1.4 Asymmetric canonical correlation analysis of Riemannian and high-dimensional data

In Chapter 2, we introduce a novel statistical model for the integrative analysis of Riemannian-valued data and high-dimensional data. Our methodology also extends naturally to time-varying Riemannian-valued data. We apply this model to explore the dependence structure between each subject's dynamic functional connectivity – represented by a time-indexed collection of positive definite covariance matrices – and high-dimensional data representing lifestyle, demographic, and psychometric measures. Specifically, we employ a

reformulation of canonical correlation analysis that enables efficient control of the complexity of the canonical directions using tangent space sieve approximations. Additionally, we enforce an interpretable group structure on the high-dimensional canonical directions via a sparsity-promoting penalty. The proposed method shows improved empirical performance over alternative approaches and comes with theoretical guarantees. Its application to data from the Human Connectome Project reveals a dominant mode of covariation between dynamic functional connectivity and lifestyle, demographic, and psychometric measures. This mode aligns with results from static connectivity studies but reveals a unique temporal non-stationary pattern that such studies fail to capture.

1.5 A correlation analysis approach to finding interpretable latent representations via conditional generative models

In Chapter 3, we extend the model proposed in Chapter 2 to automatically learn interpretable embeddings from the data, thereby estimating its underlying geometry. Here, the target view Y is $Y \in \mathcal{M}$ but where we do not know $\mathcal{M}$ in advance. To achieve this, we formulate a Partially Linear interpretable Canonical Correlation Analysis model (PLiCCA) and prove the existence of population solutions. The model jointly learns a nonlinear embedding $g : \mathbb{R}^q \to \mathbb{R}^d$ for the target view Y together with a linear embedding $T^\top$ for the auxiliary view X. This allows us to learn nonlinear embeddings for the target view that are interpretable thanks to their linear and sparse association with the auxiliary view. We establish formal connections between PLiCCA and conditional latent-variable models, specifically, conditional variational autoencoders and conditional normalizing flows. We show that these latent-variable models can be interpreted as relaxations of the PLiCCA problem, where difficult global constraints are replaced by tractable local ones. This perspective enables efficient solving of PLiCCA via 'proxy' problems derived from contemporary conditional generative models, providing an alternative to the models proposed in the first project when the underlying structure of the data is unknown.

Chapter 2

ASYMMETRIC CANONICAL CORRELATION ANALYSIS OF RIEMANNIAN AND HIGH-DIMENSIONAL DATA

2.1 Introduction

One of the primary goals of large-scale neuroimaging studies, such as the Human Connectome Project, ABCD, and the UK Biobank, is to understand the relationship between complex neuroimaging traits and non-imaging high-dimensional variables, including cognitive abilities, neurodegenerative conditions, mental health disorders, psychometric test scores, and other external factors (Zhu et al., 2023). In the context of functional connectivity studies, such complex imaging data are typically networks that are derived from fMRI data and are characterized by a single covariance matrix that captures the temporal correlation between the fMRI signals of different brain regions. For instance, Xia et al. (2018) study correlation patterns between functional connectivity and psychiatric symptoms. Other studies, such as Smith et al. (2015) and Liu et al. (2022b), investigate the relationship between functional connectivity and behavioral and demographic measures.

Traditional analyses often view brain functional networks as static. Yet, there is growing evidence that these networks are inherently dynamic and exhibit significant temporal fluctuations (Hutchison et al., 2013), which appear to be linked to various aspects of human behavior (Liégeois et al., 2019). Therefore, they are best represented by a time-indexed collection of covariances, that is, a Riemannian manifold-valued function where the manifold consists of the space of symmetric positive definite (SPD) matrices.

This work seeks to identify joint variation between these functional dynamic networks and multivariate variables, such as lifestyle, demographic, and psychometric measures. To this purpose, we develop a novel asymmetric canonical correlation analysis model that allows us

to explore the underlying relationships between two data views: Riemannian manifold-valued functional data and high-dimensional variables. We refer to this setting as *asymmetric* due to the different nature of the data views, which require different approaches to address their complexity. While our motivation stems from dynamic functional connectivity, the proposed method is general and can be applied to a variety of other settings.

Numerous models have been developed to model manifold-valued functional data, (see, e.g., Pigoli et al., 2014; Dai & Müller, 2018; Masarotto et al., 2019; Lin & Yao, 2019; Dubey & Müller, 2020, 2021; Zhang et al., 2020; Zhou & Müller, 2022; Ghodrati & Panaretos, 2022; Ghosal et al., 2023; Stöcker et al., 2023), which can be more broadly viewed as object data (Marron & Dryden, 2021) – a generalization of functional data (Ramsay & Silverman, 2015; Hsing & Eubank, 2015; Kokoszka & Reimherr, 2017). Regression models for manifold-valued data with low-dimensional predictors have been proposed in Petersen and Müller (2019), Zhao et al. (2021), and Zhou et al. (2023). See also Petersen et al. (2022) for a review. Nonetheless, models that facilitate the integration of manifold-valued functional data with high-dimensional variables have not been extensively explored.

Canonical correlation analysis (CCA) is one of the principal tools for data integration (Hotelling, 1936; Uurtio et al., 2018; Zhuang et al., 2020; Yang et al., 2021) and can be used to identify shared structure between two low-dimensional sets of variables by seeking linear combinations of these sets – with weights referred to as canonical vectors – that exhibit maximum correlation. CCA methods that go beyond low-dimensional data have largely focused on the symmetric setting, where both data views have the same structure or form. Extensions of CCA to high-dimensional data have been proposed, for instance, in Witten et al. (2009), Lin et al. (2013), Chen et al. (2013), Gaynanova et al. (2016), Gao et al. (2017), Yoon et al. (2020), and Wang and Zhou (2021a). The setting of functional data has been considered in He et al. (2010), Shin and Lee (2015), and Huang and Renaut (2015) and that of more complex imaging data in Cho et al. (2022) and Liu et al. (2021). CCA between data on Riemannian manifolds has been considered in Kim et al. (2014). Inferential aspects have been explored in Yang and Pan (2015), McKeague and Zhang (2022), and Kessler and Levina

(2023). Methods that estimate both shared and individual structure have been proposed in Lock et al. (2013), Feng et al. (2018), Carmichael (2020), Shu et al. (2020), and Yuan and Gaynanova (2022), and their connection to CCA has been studied in Murden et al. (2022).

Yet, despite the large body of literature on CCA and its extensions, existing approaches are not able to effectively estimate common structure between Riemannian manifold-valued functional data and high-dimensional multivariate variables, and more broadly, between imaging and high-dimensional data. To bridge this gap, we propose a model that leverages a regression-based characterization of CCA which allows us to incorporate appropriate notions of complexity for the functional and high-dimensional canonical directions. Specifically, our approach takes advantage of the inherent smoothness and geometric nature of the functional data, employing tangent space approximations based on a data-driven function basis computed using the Riemannian Functional Principal Components Analysis (RFPCA) framework (Dai & Müller, 2018; Lin & Yao, 2019; Shao et al., 2022). Moreover, it tackles the high dimensionality of the multivariate data by imposing sparsity. It therefore performs feature selection, resulting in models that are more interpretable and mitigate overfitting issues. In the motivating application, this will result in the identification of a small and interpretable set of multivariate variables linked to specific functional dynamic connectivity patterns. On the other hand, the tangent-space representation ensures that the estimated functional canonical directions remain constrained to the non-linear space to which the data belong, i.e., the space of SPD matrices.

The asymmetric setting considered in this work is of interest not just for its potential applications but also methodologically, as it has some distinct features that are not found in the purely sparse or functional settings. Specifically, we show that if the functional data can be efficiently represented using a finite subspace, the proposed method can consistently estimate the high-dimensional canonical vectors without requiring the direct estimation of the precision matrix of the high-dimensional data – a notoriously difficult problem and typically solvable only under specific structural assumptions (Cai et al., 2016). This feature renders the proposed methodology novel even in the simpler setting of classical functional

and high-dimensional data integration.

In addition to accommodating manifold-valued functional data and high-dimensional data, our proposed method has several other desirable properties in comparison to existing CCA models, which we highlight below:

1. It can estimate multiple canonical directions simultaneously, without requiring iterative deflation strategies and leveraging shared sparsity structure across canonical vectors.

2. It is computationally efficient, with its complexity essentially reducing to solving a regularized multivariate linear regression problem.

3. It does not require a consistent estimator for the precision matrix of the high-dimensional data.

4. The canonical vectors satisfy the correct orthogonality conditions, ensuring that the proposed approach is invariant to data rescaling, while simultaneously enforcing an interpretable sparsity structure on the high-dimensional canonical vectors.

5. When the number of observations is larger than the dimension of the high-dimensional data and no sparsity is imposed, our approach reduces to classical multivariate CCA.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the proposed asymmetric CCA model. In Section 2.3, we introduce the associated estimator and in Section 2.4, we explore its theoretical properties. In Section 2.5, we apply our method to data from the Human Connectome Project to study dynamic functional connectivity. Simulation studies, proofs, and more technical details are left to the supplementary materials.

2.2 Model

2.2.1 Elements of Riemannian geometry

Let $\mathcal{M}$ be an M-dimensional Riemannian manifold and let $T_x\mathcal{M}$ denote the tangent space at a point $x \in \mathcal{M}$ equipped with Riemannian metric $\langle \cdot, \cdot \rangle_x$. Moreover, for any $x \in \mathcal{M}$, denote the

exponential map by $\mathrm{Exp}_x : U \to \mathcal{M}$, where $U \subset T_x\mathcal{M}$ is an open set containing the origin that guarantees that this map is a bijection onto its range $\mathrm{Im}\left(\mathrm{Exp}_x\right)$. The logarithmic map at x, denoted by $\mathrm{Log}_x : \mathrm{Im}\left(\mathrm{Exp}_x\right) \to U$ is the inverse of the exponential map Exp_x. We denote by $d_{\mathcal{M}}(\cdot,\cdot)$ the Riemannian distance function on $\mathcal{M}$, which generalizes the Euclidean distance to manifolds. We refer to Lee (2012) and Lee (2018) for an introduction to the differential geometric concepts used in this work.

In our final application, $\mathcal{M}$ will represent the non-Euclidean manifold of SPD matrices equipped with the affine-invariant metric (see, e.g., Fletcher & Joshi, 2007; Pennec et al., 2019). This is particularly well-suited to studying covariance matrices thanks to its natural affine-invariant property: given two random vectors $X, Y \in \mathbb{R}^p$, let Σ_X and Σ_Y denote their covariance matrices; for any rotation matrix $R \in \mathbb{R}^{p\times p}$, the geodesic distance on the manifold is invariant to the rotation of X and Y by R, i.e. $d_{\mathcal{M}}\left(\Sigma_X, \Sigma_Y\right) = d_{\mathcal{M}}\left(\Sigma_{RX}, \Sigma_{RY}\right)$.

The affine-invariant metric at $P \in \mathcal{M}$ between $W, Z \in T_P\mathcal{M}$ is defined as $\langle W, Z\rangle_{\mathcal{M}} = \mathrm{tr}\left(P^{-1}WP^{-1}Z\right)$. Let exp and log denote the matrix exponential and logarithm, defined here on the sets of symmetric matrices and positive definite matrices, respectively, and let $\|\cdot\|_F$ denote the Frobenius norm of a matrix. Then, the affine-invariant Riemannian distance is defined as $d_{\mathcal{M}}(P,Q) = \|\log(P^{-1/2}QP^{-1/2})\|_F$. The logarithmic map $\mathrm{Log}_P(Q) = P^{1/2}\log\left(P^{-1/2}QP^{-1/2}\right)P^{1/2}$ will allow us to compute unconstrained tangent space representations of our data, i.e., symmetric matrices. Roughly speaking, the tangent space representations allow us to apply simple Euclidean mathematical operations without breaking the geometry of the space of SPD matrices and the exponential map $\mathrm{Exp}_P(W) = P^{1/2}\exp\left(P^{-1/2}WP^{-1/2}\right)P^{1/2}$ will allow us to map tangent space elements back to the manifold $\mathcal{M}$. In this case, the exponential and logarithmic maps Exp and Log are global bijections between $\mathcal{M}$ and the space of symmetric matrices. Alternative metrics that accommodate positive semi-definite covariances have been defined, for instance, in Dryden et al. (2009), Pigoli et al. (2014), and Masarotto et al. (2019).

Next, we present the mathematical tools necessary to model Riemannian-valued functions. Let $\mathcal{T}$ be a compact subset of $\mathbb{R}$ and let $\mu : \mathcal{T} \to \mathcal{M}$ be a sufficiently smooth curve

on $\mathcal{M}$. A vector field V along μ is a map from $\mathcal{T}$ to the tangent bundle $T\mathcal{M}$ such that $V(t) \in T_{\mu(t)}\mathcal{M}$ for all $t \in \mathcal{T}$. The collection of vector fields V along μ defines a vector space. Define $L^2(T\mu)$ to be the space of square integrable vector fields V along μ equipped with inner product $\langle\!\langle U, V \rangle\!\rangle_\mu := \int_{\mathcal{T}} \langle V(t), U(t) \rangle_{\mu(t)} \mathrm{d}t$ and induced norm defined by $\|\cdot\|_\mu^2 = \langle\!\langle \cdot, \cdot \rangle\!\rangle_\mu$, where U and V are both vector fields along μ. Then, $L^2(T\mu)$ is a separable Hilbert space (Lin & Yao, 2019).

For a curve μ and Riemannian-valued function $y : \mathcal{T} \to \mathcal{M}$, we denote as $\mathrm{Log}_\mu y$ the function $t \mapsto \mathrm{Log}_{\mu(t)} y(t)$. Similarly, for a vector field V along μ, we denote as $\mathrm{Exp}_\mu V$ the function $t \mapsto \mathrm{Exp}_{\mu(t)} V(t)$. In our setting, y will be random, and μ will represent the mean of y. Under appropriate assumptions, the vector field $\mathrm{Log}_\mu y$ along μ will be a random element of $L^2(T\mu)$, which intuitively represents a linearized and centered version of y. Indeed, if $\mathcal{M}$ is a Euclidean space $\mathcal{M} = \mathbb{R}^d$, then $\mathrm{Log}_{\mu(t)} y(t) = y(t) - \mu(t)$ for every $t \in \mathcal{T}$.

Later, we will need to compare vector fields along different curves μ and $\hat{\mu}$. To this purpose, following Lin and Yao (2019), we introduce the parallel transport operator. We denote the parallel transport operator on $\mathcal{M}$ along geodesics as $\mathcal{P}_{x,p} : T_x\mathcal{M} \to T_p\mathcal{M}$. A fundamental property of this operator is that it preserves inner products of tangent vectors, i.e., for any $u, v \in T_x\mathcal{M}$, $\langle u, v \rangle_x = \langle \mathcal{P}_{x,p} u, \mathcal{P}_{x,p} v \rangle_p$. We can then define parallel transport for vector fields U, V along curves $f, h : \mathcal{T} \to \mathcal{M}$. Specifically, given $U \in L^2(Tf)$ and $V \in L^2(Th)$, we define $\Gamma_{f,h} U \in L^2(Th)$ as the map $t \mapsto \mathcal{P}_{f(t),h(t)} U(t)$. Therefore, $\Gamma_{f,h}$ can be viewed as a map from $L^2(Tf)$ to $L^2(Th)$. Therefore, while U and V cannot be 'compared' directly since for every t, $U(t)$ and $V(t)$ may belong to different tangent spaces, we can compare $\Gamma_{f,h} U$ and V, since they are both elements of $L^2(Th)$. In particular, $\|\Gamma_{f,h} U - V\|_h$ quantitatively describes the difference between U and V. We refer to Proposition 2 of Lin and Yao (2019) for additional properties of the parallel transport operator.

2.2.2 Modeling Riemannian-valued data

Let (y, X) be a pair of random variables, where $X \in \mathbb{R}^p$ is a zero-mean random vector, with covariance $\Sigma_X \in \mathbb{R}^{p \times p}$, representing the high-dimensional multivariate variables, and the

process y is a Riemannian-valued random process with continuous sample paths. We assume that $\forall x \in \mathcal{M}, \forall t \in \mathcal{T}$ we have $\mathbb{E}\left[d^2_{\mathcal{M}}(y(t), x)\right] < \infty$.

Next, we define the Fréchet mean of the process y on $\mathcal{M}$ as

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

We assume that the Fréchet mean $\mu(t)$ exists and is unique for every $t \in \mathcal{T}$, and μ is a continuous function. For more details on the Fréchet mean, see Bhattacharya and Patrangenaru (2003). Following Lin and Yao (2019), we assume

$$\Pr\left\{\text{For all } t \in \mathcal{T} : y(t) \in \operatorname{Im}\left(\operatorname{Exp}_{\mu(t)}\right)\right\} = 1,$$

which ensures that $\operatorname{Log}_{\mu(t)} y(t)$ is defined almost surely for all $t \in \mathcal{T}$. This condition is superfluous for many common manifolds, for example whenever Exp_x is surjective onto $\mathcal{M}$ for all $x \in \mathcal{M}$. For instance, this holds on the manifold of SPD matrices equipped with the affine-invariant metric.

Let the tensor product $U \otimes V : L^2(T\mu) \to L^2(T\mu)$, between $U, V \in L^2(T\mu)$, be defined as $(U \otimes V)(W) = \langle\!\langle U, W \rangle\!\rangle_\mu V$ for all $W \in L^2(T\mu)$. If $\mathbb{E}\left[\|\operatorname{Log}_\mu y\|^2_\mu\right] < \infty$, then the covariance function $\mathcal{C}$ of $\operatorname{Log}_\mu y$ is defined as $\mathcal{C} = \mathbb{E}\left[\operatorname{Log}_\mu y \otimes \operatorname{Log}_\mu y\right]$ and is nonnegative and trace class. Therefore, it admits the eigendecomposition

$$\mathcal{C} = \sum_{j=1}^{\infty} \omega_j \phi_j \otimes \phi_j, \tag{2.1}$$

with ω_j a sequence of real numbers converging to 0, and $\phi_j \in L^2(T\mu)$ satisfying $\langle\!\langle \phi_j, \phi_k \rangle\!\rangle_\mu = \delta_{jk}$, where $\delta_{jk} = 1$ if $j = k$, and 0 otherwise. The functions $\{\phi_j\}$ are called the population loading functions, or population principal components, of $\operatorname{Log}_\mu y$. Moreover, with probability one, we have that the process $\operatorname{Log}_\mu y$ admits a Principal Component expansion

$$\operatorname{Log}_\mu y = \sum_{j=1}^{\infty} Y_j \phi_j,$$

where $Y_j = \langle\!\langle \phi_j, \operatorname{Log}_\mu y \rangle\!\rangle_\mu$ are pairwise uncorrelated random variables, and satisfy $\mathbb{E}[Y_j] = 0$ and $\operatorname{Var}(Y_j) = \omega_j$. The variables Y_j are called the population principal scores. For further details on the principal component basis and eigendecomposition of the $\mathcal{C}$, see Lemma A.2.1 in the supplementary materials.

2.2.3 Asymmetric Riemannian CCA

In this section, we introduce our proposed asymmetric CCA model, which can be naturally formalized by mirroring the multivariate and functional versions (He et al., 2010) of the problem. We define the first canonical direction pair (ψ_1, θ_1) as a solution, if one exists, to the following problem

$$\underset{\psi \in L^2(T\mu),\ \theta \in \mathbb{R}^p}{\text{maximize}} \ \text{Corr}^2 \left(\langle\!\langle \text{Log}_\mu y, \psi \rangle\!\rangle_\mu, \langle X, \theta \rangle \right), \tag{2.2}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in $\mathbb{R}^p$. Analogously, we can define the subsequent pairs (ψ_k, θ_k) to maximize the same objective function, with the condition that each pair is orthogonal to the previous ones, namely, $\langle\!\langle \psi_k, \mathcal{C}\psi_{k'} \rangle\!\rangle_\mu = \delta_{kk'}$ and $\theta_k^\top \Sigma_X \theta_{k'} = \delta_{kk'}$. When they exist, we refer to ψ_k as the kth canonical function, and to θ_k as the kth canonical vector. Given the canonical function $\psi_k \in L^2(T\mu)$, we can map it back to the original space via the exponential map. This procedure is illustrated in Figure 2.1.

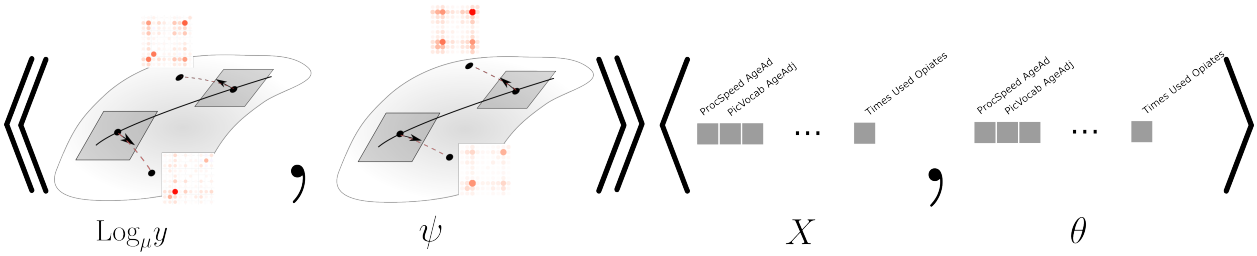


Figure 2.1: In this figure, we illustrate the process of projecting the Riemannian-valued functional data and the high-dimensional data to define maximally correlated variables. We leverage tools from differential geometry to compute linear tangent representations $\text{Log}_\mu y$ of the temporally-indexed Riemannian-valued data y, which are equipped with a notion of inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\mu$, that is, a projection operator. For the multivariate data, we use the conventional notion of projection, i.e., the Euclidean inner product. We therefore seek ψ and θ whose respective data projections define maximally correlated variables.

While equation (2.2) provides an intuitive formulation of the canonical correlation prob-

lem, it has been noted in Cupidon et al. (2008) that the maximum of this problem may not be attained by any $\psi \in L^2(T\mu)$, $\theta \in \mathbb{R}^p$. Despite this, the problem can still be reformulated with respect to the pair of canonical variables $(U, V) = \left(\langle\!\langle \mathrm{Log}_\mu y, \psi \rangle\!\rangle_\mu, \langle X, \theta \rangle\right)$, resulting in the following maximization problem:

$$\underset{U \in \overline{\mathcal{U}},\, V \in \overline{\mathcal{V}}}{\text{maximize}} \;\; \mathrm{Corr}^2\,(U, V)\,, \tag{2.3}$$

where $\mathcal{U} = \{\langle\!\langle \mathrm{Log}_\mu y, \psi \rangle\!\rangle_\mu : \; \psi \in L^2(T\mu)\}$, $\mathcal{V} = \{\langle X, \theta \rangle : \; \theta \in \mathbb{R}^p\}$, and $\overline{\mathcal{U}}$ and $\overline{\mathcal{V}}$ are appropriate closures of $\mathcal{U}$ and $\mathcal{V}$, respectively. This guarantees that an optimal canonical variable pair (U, V) does exist. We emphasize that they *cannot* necessarily be written in terms of the canonical directions, i.e., it does not necessarily hold that $U \in \overline{\mathcal{U}}$ can be written as $U = \langle\!\langle \mathrm{Log}_\mu y, \psi \rangle\!\rangle_\mu$ for some $\psi \in L^2(T\mu)$. To simplify the exposition, we defer the details of this formulation to Section A.2 of the supplementary materials (also see Remark 8). In Theorem 2.4.3, we show that despite the nonexistence of population canonical directions for functional data, the *canonical variables* corresponding to estimated canonical directions are consistent for U and V. To investigate the theoretical properties of the *canonical directions*, and in particular, to ensure that their population counterparts are well-defined, we make the following assumption.

Assumption 2.2.1. *There exists a complete orthonormal system $\{\zeta_i\}_{i=1}^\infty$ for $L^2(T\mu)$, and a set of indices $I \subset \{1, 2, \ldots\}$, with finite cardinality $|I| \equiv d^{(corr)} \leq p$, such that*

$$\begin{aligned} &\mathrm{Corr}\left(X_k, \langle\!\langle \mathrm{Log}_\mu y, \zeta_j \rangle\!\rangle_\mu\right) = 0, \qquad k = 1, \ldots p, \forall j \in I^c, \\ &\mathrm{Corr}\left(\langle\!\langle \mathrm{Log}_\mu y, \zeta_i \rangle\!\rangle_\mu, \langle\!\langle \mathrm{Log}_\mu y, \zeta_j \rangle\!\rangle_\mu\right) = 0, \qquad \forall i \in I, \forall j \in I^c, \end{aligned}$$

where I^c denotes the complement of I in $\{1, 2, \ldots\}$.

Intuitively, Assumption 2.2.1 implies that there are only a finite number of basis elements $\{\zeta_i\}_{i \in I}$ that capture the correlation between X and $\mathrm{Log}_\mu y$ through the scores $\left\{\langle\!\langle \mathrm{Log}_\mu y, \zeta_i \rangle\!\rangle_\mu\right\}_{i \in I}$. Crucially, through this assumption, we do not limit the dimensionality of the functional data. For more details on this assumption, see also Section A.2.3 of

the supplementary materials. Let $\{\zeta_j\}_{j=1}^{\infty}$ be an orthonormal system for $L^2(T\mu)$ satisfying Assumption 2.2.1, and reorder the $\{\zeta_j\}_{j=1}^{\infty}$ so that $I = \{1, \dots d^{(\text{corr})}\}$. Define $Y_j \equiv \langle\!\langle \text{Log}_\mu y, \zeta_j \rangle\!\rangle$ for $j = 1, \dots d^{(\text{corr})}$, so that the $\{Y_j\}$ are random variables with $\text{Var}(Y_j) < \infty$. Note that in practice, to define the orthonormal system $\{\zeta_j\}_{j=1}^{\infty}$, we could employ the principal components analysis in Section 2.2.2, or alternatively, we could design its basis functions to capture specific features of interest. Next, define $Y = (Y_1, \dots, Y_{d^{(\text{corr})}})$ and let Σ_Y be the $d^{(\text{corr})} \times d^{(\text{corr})}$ covariance of Y. Let $\|\cdot\|_2$ denote the Euclidean 2-norm of a vector in $\mathbb{R}^{d^{(\text{corr})}}$. Without loss of generality, we suppose that X and Y are mean 0. Under Assumption 2.2.1, the following theorem states that the canonical correlation problem in equation (2.3) is equivalent to solving a suitably formulated finite-dimensional regression problem.

Theorem 2.2.1. *Under Assumption 2.2.1, the CCA model in equation (2.2) admits at most $d^{(corr)}$ nontrivial canonical variable pairs $\{(U_k, V_k)\}$, and each pair (U_k, V_k) can be written in terms of the associated canonical directions: $U_k = \langle\!\langle \text{Log}_\mu y, \psi_k \rangle\!\rangle_\mu$ and $V_k = \langle X, \theta_k \rangle$ for some $\psi_k \in L^2(T\mu)$ and $\theta_k \in \mathbb{R}^p$. Additionally, suppose $\Sigma_X \in \mathbb{R}^{p \times p}$ and $\Sigma_Y \in \mathbb{R}^{d^{(corr)} \times d^{(corr)}}$ are invertible. Let B be the solution to the multivariate least-squares problem*

$$\underset{B \in \mathbb{R}^{p \times d^{(corr)}}}{\text{minimize}} \; \mathbb{E}\left[\|\Sigma_Y^{-1/2} Y - B^\top X\|_2^2 \right], \tag{2.4}$$

and let

$$B^\top \Sigma_X B = \tilde{H} D^2 \tilde{H}^\top \tag{2.5}$$

be an eigendecomposition of $B^\top \Sigma_X B$. Define

$$T = B\tilde{H} D^{-1} \in \mathbb{R}^{p \times d^{(corr)}}, \tag{2.6}$$

$$H = \Sigma_Y^{-1/2} \tilde{H} \in \mathbb{R}^{d^{(corr)} \times d^{(corr)}}. \tag{2.7}$$

Then, the kth column of H, η_k, characterizes the kth canonical function ψ_k through $\psi_k = \sum_{j=1}^{d^{(corr)}} \eta_{kj} \zeta_j$, and the kth column of T is the kth high-dimensional canonical vector θ_k. Moreover, the optimum values attained by the maximization problem in equation (2.3) are the diagonal entries of D^2, which we denote by $\gamma_1^2, \dots \gamma_{d^{(corr)}}^2$.

The proof of Theorem 2.2.1 can be found in Section A.2.6 of the supplementary materials. This suggests a novel methodology for deriving estimates of the canonical functions $\{\psi_k\}$ and the canonical vectors $\{\theta_k\}$. This entails defining a subspace, spanned by $\{\zeta_j\}_{j=1}^{d^{(\text{corr})}}$, onto which the tangent space representations of the functional data are projected. Subsequently, the canonical functions and vectors can be characterized by the equations (2.4)-(2.7), using empirical estimates in place of the theoretical population values. Note that in practice, we need to choose the dimension of this subspace by choosing a number of ζ_j to use, which we denote by d. In Theorem 2.4.3 we show that for consistency of the canonical directions, all that is required is choosing $d \geq d^{(\text{corr})}$; we do not need to choose d exactly equal to $d^{(\text{corr})}$.

Crucially, as opposed to other methods in the literature (see, e.g., Chen et al., 2013; Gao et al., 2017), the proposed model circumvents the direct estimation of Σ_X^{-1}, i.e., the precision matrix of the variable X, which is a notoriously difficult problem in high dimensions as it can be estimated only under restrictive structural assumptions. Our strategy yields interpretable results by enforcing sparsity directly on the canonical directions $\{\theta_k\}$ through an additional penalty term on the estimate of B. The complexity of the functional canonical direction is controlled by projecting the functional data on a finite-dimensional subspace. Such an approach leverages the smooth nature of the functional data (and its tangent space representation) — which is reflected in the eigenvalues of the covariance rapidly decaying to zero — suggesting that such a projection can serve as an efficient and interpretable approximation.

2.2.3.1 On the existence of canonical directions and connections with partial least-squares

In Theorem 2.4.3 we show that even without Assumption 2.2.1, given a new independent sample $(X_{\text{test}}, y_{\text{test}})$, if we use our canonical direction estimates $\hat{\psi}_k$ and $\hat{\theta}_k$ to form estimated scores $\left(\hat{U}_k, \hat{V}_k\right) \equiv \left(\langle\!\langle \text{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \rangle\!\rangle_{\hat{\mu}}, \langle \hat{\theta}_k, X_{\text{test}} \rangle\right)$, where $\hat{\mu}$ has also been estimated from the data, then $(\hat{U}, \hat{V})$ converge to the true scores (U, V) given a sufficiently large N. However, it should be noted that similar approximation results do not exist in general for the population canonical directions ψ_k and θ_k, which, as shown in Cupidon et al. (2008), may not even

be well defined. In other words, a maximizer of the population problem in equation (2.2) may not be attained in $L^2(T\mu)$. This highlights that it is crucial for practitioners to be cautious in their interpretation of results derived from a CCA model. While the estimated canonical variables can generally be regarded as estimates of the population counterparts, the canonical directions should only be interpreted as estimates of the population canonical directions under a d-dimensional subspace approximation of the process and not necessarily of the underlying infinite-dimensional process, which, as mentioned earlier, may not even be well-defined. Table 2.1 summarizes the conditions under which the quantities of interest are well-defined.

	$d^{(\mathrm{corr})} < \infty$	$d^{(\mathrm{corr})} = \infty$
Canonical directions (θ_k, ψ_k)	✓	×
Canonical variables (U_k, V_k)	✓	✓

Table 2.1: Existence of canonical directions and variables, depending on $d^{(\mathrm{corr})}$ in Assumption 2.2.1.

We also remark on how the situation changes when considering the partial least squares problem, as opposed to the canonical correlation problem. It is straightforward to show that by replacing equation (2.4) with

$$\underset{B \in \mathbb{R}^{p \times d}}{\text{minimize}} \; \mathbb{E}\left[\|Y - B^\top X\|_2^2\right], \tag{2.8}$$

namely, by omitting the standardization of the random coefficients in the basis expansion, the proposed reformulation defines a partial least squares model (see, e.g., Boulesteix and Strimmer (2007)). For this model, the population canonical directions are more broadly defined, and it can be shown that the d-dimensional approximation of the functional data results in asymptotically negligible errors for both the canonical variables and directions. Intuitively, by comparing equations (2.4) and (2.8), one can see that the functional data

enter the partial least squares model via the unnormalized coefficients Y of the first d basis functions, leading to negligible 'residual information' as d increases, due to the compactness of the covariance of the functional data. On the other hand, in the CCA model, the functional data are incorporated through the normalized coefficients $\Sigma_Y^{-1/2}Y$. This normalization step prevents the residual information from becoming negligible and may cause the canonical directions to diverge as d increases.

In this work, we focus on CCA due to its ability to detect components of y that are small in magnitude but correlated with X, as opposed to partial least squares models which are sensitive to the scale of the signal. Such a feature is particularly critical in neuroimaging applications (Wang et al., 2020).

2.3 Estimation

Suppose we are given N observations

$$(y_i, X_i), \quad i = 1, \ldots, N,$$

each being a realization of the pair (y, X). We propose the following estimation procedure, outlined in four steps.

Step A: RFPCA

We first compute the sample version of the Fréchet mean, defined as

$$\hat{\mu}(t) = \arg\min_{x \in \mathcal{M}} \frac{1}{N} \sum_{i=1}^{N} d^2_{\mathcal{M}}\left(y_i(t), x\right).$$

We then estimate the tangent space representations of the functional data observations using $\mathrm{Log}_{\hat{\mu}} y_i \in L^2\left(T\hat{\mu}\right)$. Next, we define an orthonormal basis for the tangent space representations using the RFPCA framework proposed in Dai et al. (2017), Lin and Yao (2019), and Shao et al. (2022) to estimate a data-driven basis $\{\hat{\phi}_j\}_{j=1}^{d}$. Note that d is not necessarily equal to $d^{(\mathrm{corr})}$ from Section 2.2.3. Specifically, we estimate the tangent-space covariance operator $\mathcal{C}$ using the sample covariance function $\hat{\mathcal{C}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Log}_{\hat{\mu}} y_i \otimes$

$\mathrm{Log}_{\hat{\mu}} y_i$. Each population loading function ϕ_j and associated eigenvalue ω_j can be estimated using the eigenfunction $\hat{\phi}_j$ and eigenvalue $\hat{\omega}_j$ of $\hat{\mathcal{C}}$. The empirical Principal Component expansion of $\{\mathrm{Log}_{\hat{\mu}} y_i\}$ is then given by

$$\mathrm{Log}_{\hat{\mu}} y_i = \sum_{j=1}^{d} \hat{Y}_{ij} \hat{\phi}_j,$$

where $\hat{Y}_{ij} = \langle\!\langle \hat{\phi}_j, \mathrm{Log}_{\hat{\mu}} y_i \rangle\!\rangle_{\hat{\mu}}$ are the PC scores. Here we assume that the rank of the Principal Component expansion d is such that $d < \min(p, N)$. For completeness, in Section A.7 of the supplementary materials, we provide a detailed description of the RFPCA algorithm, including a computationally efficient explicit basis construction for the space of SPD matrices equipped with the affine invariant metric.

Step B: Regularized regression

Next, we use the scores $\hat{Y}_{ij}$ to represent the manifold-valued functional data and estimate the canonical directions leveraging the characterization in Theorem 2.2.1. We let $\mathbb{X} \in \mathbb{R}^{N \times p}$ and $\hat{\mathbb{Y}} \in \mathbb{R}^{N \times d}$ denote the data matrices $(X_{ij})_{ij}$ and $(\hat{Y}_{ij})_{ij}$, respectively, where our notation emphasizes that the entries of $\hat{\mathbb{Y}}$ are estimates.

Define $\hat{\Sigma}_Y = \frac{1}{N} \hat{\mathbb{Y}}^T \hat{\mathbb{Y}}$ and $\hat{\Sigma}_X = \frac{1}{N} \mathbb{X}^T \mathbb{X}$. We estimate the matrix B in equation (2.4) using $\hat{B}$, which is derived by solving the following group lasso problem:

$$\hat{B} = \underset{B \in \mathbb{R}^{p \times d}}{\arg\min} \frac{2}{N} \left\| \hat{\mathbb{Y}} \hat{\Sigma}_Y^{-1/2} - \mathbb{X} B \right\|_F^2 + \lambda \left\| B \right\|_{\ell_1, \ell_2}, \tag{2.9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\|B\|_{\ell_1,\ell_2} = \sum_{i=1}^{p} \|b_i\|_2$ is a group lasso penalty. Here, b_i refers to the ith row of B. Note that the first term of the minimization problem in equation (2.9) is an empirical approximation of that in equation (2.4), where the data matrices $\mathbb{X}$ and $\hat{\mathbb{Y}}$ replace the random variables X and Y, respectively.

Step C: Eigenanalysis

Given $\hat{\Sigma}_X^{1/2} \hat{B}$, we then compute its right singular vectors $\hat{\tilde{H}} \in \mathbb{R}^{d \times d}$ and singular values matrix $\hat{D} \in \mathbb{R}^{d \times d}$, that is,

$$\hat{B}^T \hat{\Sigma}_X \hat{B} = \hat{\tilde{H}} \hat{D}^2 \hat{\tilde{H}}^T.$$

Step D: Estimates computation

We define

$$\hat{T} = \hat{B}\hat{\tilde{H}}\hat{D}^{-1}, \tag{2.10}$$

$$\hat{H} = \hat{\Sigma}_Y^{-1/2}\hat{\tilde{H}}, \tag{2.11}$$

where $\hat{T} \in \mathbb{R}^{p\times d}$ and $\hat{H} \in \mathbb{R}^{d\times d}$ are estimates of T and H, respectively. Then, $\hat{T} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_d\right]$ is a matrix whose columns $\hat{\theta}_k$ are the estimates of θ_k, and $\hat{H} = [\hat{\eta}_1, \ldots, \hat{\eta}_d]$ is a matrix whose columns $\hat{\eta}_k$ are the estimates of η_k. The estimated canonical functions are therefore given by $\hat{\psi}_k = \sum_{j=1}^{d} \hat{\eta}_{kj}\hat{\phi}_j$, for $k = 1, \ldots d$, resulting in the estimated canonical functions and vectors $(\hat{\psi}_k, \hat{\theta}_k)$.

Algorithm 1 Asymmetric Sparse-Functional CCA

Input: Pairs $(y_i, X_i)_{i=1,\ldots N}$ of manifold-valued functional data and high-dimensional data; number of principal components d chosen for the manifold-valued functional data.

1. Obtain $\hat{\phi}_j$, $\hat{\omega}_j$, for $j = 1, \ldots d$, and $\mathbb{Y}$ applying Intrinsic RFPCA to $(y_i)_{i=1,\ldots N}$.
2. Compute $\hat{\Sigma}_Y = \text{diag}\,(\hat{\omega}_j)$ and $\hat{\Sigma}_X = \frac{1}{N}\mathbb{X}^\top\mathbb{X}$.
3. Compute $\hat{B}$ solving the group lasso problem in equation (2.9) using the `glmnet` package (Friedman et al., 2010).
4. Compute $\hat{H} = [\hat{\eta}_1, \ldots, \hat{\eta}_d]$ and $\hat{T} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_d\right]$ in equations (2.10) and (2.11).
5. Compute the estimated canonical functions $\hat{\psi}_k = \sum_{j=1}^{d} \hat{\eta}_{kj}\hat{\phi}_j$ for $k = 1, \ldots d$.
6. Return $\{\hat{\theta}_k\}_{k=1}^{d}$, the estimated canonical vectors associated with X, and $\{\hat{\psi}\}_{k=1}^{d}$, the estimated canonical functions associated with y.

The sparsity-promoting regularization norm employed in equation (2.9) encourages entire rows of the matrix B to be set to zero. From the equation $\hat{T} = \hat{B}\hat{\tilde{H}}\hat{D}^{-1}$, it follows that the corresponding rows of $\hat{T}$ will also be zero. This yields canonical vectors $\{\theta_k\}$ with a group sparsity structure, meaning they share identical sparsity patterns.

The main steps of the estimation procedure are summarized in Algorithm 1, which we refer to as *asymmetric sparse-functional CCA*.

2.3.1 Selection of hyperparameters

In the RFPCA step of Algorithm 1, it is necessary to select the number of principal components d, which needs to be less than or equal to both N and p. Additionally, the regularization parameter λ needs to be chosen in the regularized regression step. We recommend using cross-validation to select these parameters. Specifically, for each choice of d, the optimal λ in the regression step can be selected via cross-validation, as implemented in the `glmnet` package. Then, we select the value of d that yields the largest out-of-sample (or cross-validated) canonical correlations. For instance, if the user is interested in finding the top k canonical directions, then we recommend using the sum of the first k estimates of out-of-sample canonical correlations.

If the user believes that the functional data has finite rank, a hypothesis testing approach can be used to obtain an upper bound on d, for example by adapting the method proposed in Charkaborty and Panaretos (2022).

2.3.2 Special instances

To demonstrate the versatility of our model, we present a few special cases. Although some of these settings are simpler than the motivating neuroimaging application, the proposed method still provides an innovative approach to analyzing such data.

- In situations where $y_i \in \mathcal{M}$, meaning our imaging data are manifold-valued observations without a temporal dimension, Algorithm 1 can be adapted by using tangent-space PCA (Marron & Dryden, 2021) rather than RFPCA, similar to the setting considered in Kim et al. (2014). This model is especially useful for studying static connectivity networks.

- When the imaging data take the form of classical functional data, that is $y_i(t) \in \mathcal{M} \subset \mathbb{R}$ for all $t \in \mathcal{T}$, one can apply Algorithm 1 by replacing RFPCA with classical FPCA (Ramsay & Silverman, 2015; Yao et al., 2005). In addition, when $y_i(t) \in \mathcal{M} \subset \mathbb{R}^d$, multivariate FPCA can be employed (Happ & Greven, 2018).

Algorithm 2 Asymmetric Sparse CCA

Input: Pairs $(Y_i, X_i)_{i=1,\dots N}$ of low- and high-dimensional data. Let $\mathbb{Y} = (Y_{ij})_{ij}$ and $\mathbb{X} = (X_{ij})_{ij}$.

1. Compute $\hat{\Sigma}_Y = \frac{1}{N}\mathbb{Y}^\top\mathbb{Y}$ and $\hat{\Sigma}_X = \frac{1}{N}\mathbb{X}^\top\mathbb{X}$.
2. Compute $\hat{B}$ solving the group lasso problem

$$\hat{B} = \underset{B\in\mathbb{R}^{p\times d}}{\arg\min} \frac{2}{N} \left\| \mathbb{Y}\hat{\Sigma}_Y^{-1/2} - \mathbb{X}B \right\|_F^2 + \lambda \|B\|_{\ell_1,\ell_2} \tag{2.12}$$

   using the `glmnet` package (Friedman et al., 2010).
3. Compute $\hat{H} = [\hat{\eta}_1, \dots, \hat{\eta}_d]$ and $\hat{T} = \left[\hat{\theta}_1, \dots, \hat{\theta}_d\right]$ in equations (2.10) and (2.11).
4. Return $\{\hat{\eta}_k\}_{k=1}^d$, the estimated canonical vectors associated with $\{Y_i\}_{i=1}^N$, and $\{\hat{\theta}_k\}_{k=1}^d$, the estimated canonical functions associated with $\{X_i\}_{i=1}^N$.

Central to the proposed methodology is a CCA model for pairs of observations (Y_i, X_i), where $Y_i \in \mathbb{R}^d$, $X_i \in \mathbb{R}^p$, $d \ll N$, and the covariance of Y_i is full-rank. In the imaging setting, we use a dimension reduction model to compute the low-dimensional component. However, this setting may also be of independent interest and plays a crucial role in the development of the theoretical results. Therefore, we outline the algorithm for this particular setting in Algorithm 2, and we refer to it as *asymmetric sparse CCA*.

In this special case, our approach is closely related to the Eigenvector-CCA model proposed in Wang and Zhou (2021a). Yet, notable differences exist between the two approaches. For example, we ensure that the estimated canonical vectors satisfy the correct orthogonality conditions $\hat{H}^\top\hat{\Sigma}_Y\hat{H} = I_d$ and $\hat{T}^\top\hat{\Sigma}_X\hat{T} = I_d$. Furthermore, our proposed model does not rely on the assumption that the data have been generated from a regression model.

2.4 Theory

Here we investigate the convergence properties of the proposed estimators. We first study the asymptotic properties of the asymmetric sparse CCA model outlined in Algorithm 2, which sets the stage for studying the asymptotic convergence properties of the asymmetric Sparse-Functional CCA model outlined in Algorithm 1.

2.4.1 Estimation error rates for Asymmetric Sparse CCA

In this section, we state error bounds for the asymmetric sparse CCA model outlined in Algorithm 2. We assume N observations $Y_i \in \mathbb{R}^d$ and $X_i \in \mathbb{R}^p$ are independent copies of the random variables Y and X, respectively. We denote with γ_k the kth canonical correlation attained in the population version of the problem and recall that $T = [\theta_1, \dots \theta_d] \in \mathbb{R}^{p \times d}$. Moreover, we denote with $K = \max\{i \in \{1, \dots d\} : \gamma_i > 0\}$ the number of nontrivial canonical vectors. To simplify the notation, we use the conventions $\gamma_{d+1}^2 = -\infty$ and $\gamma_0^2 = \infty$. We use $\operatorname{cond}(A) = \|A\|_2 \cdot \|A^{-1}\|_2$ to denote the condition number of an invertible matrix A, and $\|A\|_2$ to denote the operator norm of A, or equivalently, the square root of the largest eigenvalue of $A^\top A$. The norm $\|A\|_{2,\infty}$ denotes the maximum Euclidean norm of the rows of A, and $\|A\|_{\ell_1,\ell_2} = \sum_{i=1}^p \|a_i\|_2$, where a_i is the ith row of A. The notation $a \lesssim b$ indicates inequality up to an absolute constant, i.e., there exists an absolute constant $C > 0$ such that $a \leq Cb$. Next, we introduce the main assumptions.

Assumption 2.4.1. *The random variables X and Y are strict sub-Gaussian random vectors with invertible covariance matrices Σ_X and Σ_Y, respectively. Strict sub-Gaussian random vectors are introduced in Definition A.3.2 of the supplementary materials.*

Assumption 2.4.2. *It holds that $d \leq p$, $d\log(p) = o(N)$, $\operatorname{cond}(\Sigma_Y)^2 d = o(N)$, and $\gamma_1 > \dots > \gamma_K$ are bounded from below and are distinct.*

Assumption 2.4.3. *The norms $\|\Sigma_X\|_{2,\infty}, \|T\|_{\ell_1,\ell_2}$ are bounded from above and are larger than 1, $\|\Sigma_X^{-1}\|_2, \|\Sigma_Y^{-1}\|_2 \geq 1$, and $\hat{\eta}^\top \hat{\Sigma}_Y^{1/2} \Sigma_Y^{1/2} \eta \geq 0$ for $k = 1, \dots K$.*

The sub-Gaussian condition in Assumption 2.4.1 ensures that X and Y do not have heavy tails, allowing us to use standard concentration results for the estimation of Σ_X and Σ_Y. Strict sub-Gaussianity (Kereta & Klock, 2021) facilitates the proofs by allowing the sub-Gaussian norm of a random variable and its variance to be used interchangeably.

In Assumption 2.4.2, the condition that $d\log(p) = o(N)$ allows p to grow exponentially in N/d (i.e., $p \lesssim e^{N/d}$) while still retaining consistency of the estimator for the canonical vectors. The critical component of the condition $\operatorname{cond}(\Sigma_Y)^2 d = o(N)$ is that $d = o(N)$, which ensures that Σ_Y can be estimated at a sufficiently fast rate by its sample estimator $\hat{\Sigma}_Y$. The presence of $\operatorname{cond}(\Sigma_Y)^2$ allows us to show that $\|\hat{\Sigma}_Y\|_2 \lesssim \|\Sigma_Y\|_2$ and to ignore lower order terms of $\frac{d}{N}$, simplifying the theorem statement. We assume that the correlations $\gamma_1, \dots \gamma_K$ are distinct in order to estimate each canonical vector separately instead of estimating entire subspaces.

Assumption 2.4.3 is not essential, and mainly serves to simplify the statement of the theorem. Since the canonical vectors are defined only up to a sign, we use the condition $\hat{\eta}^\top \hat{\Sigma}_Y^{1/2} \Sigma_Y^{1/2} \eta \geq 0$ to account for the sign ambiguity of the CCA solutions, allowing us to compare the estimates of the canonical vectors with their population counterparts through the differences $\|\theta_k - \hat{\theta}_k\|_2$ and $\|\eta_k - \hat{\eta}_k\|_2$.

Theorem 2.4.1. *Suppose Assumptions 2.4.1-2.4.3 hold. Fix $\alpha \in (0,1)$, and for some absolute constant $C > 0$, define the regularization parameter in Algorithm 2 as $\lambda = C\sqrt{\frac{d}{N}\log(p\alpha^{-1})}$. Then, with probability $1-\alpha$, we have that, for $k = 1, \dots K$,*

$$\|\theta_k - \hat{\theta}_k\|_2^2 \lesssim \left(\frac{d}{N}\log\left(p\alpha^{-1}\right)\right)^{1/2} \frac{\gamma_1^2 \|\Sigma_X\|_{2,\infty} \|T\|_{\ell_1,\ell_2}^2}{\min_{j\neq k} |{\gamma_k}^2 - {\gamma_j}^2|^2} \frac{\|\Sigma_X^{-1}\|_2}{\gamma_k^2}, \tag{2.13}$$

$$\|\eta_k - \hat{\eta}_k\|_2^2 \lesssim \left(\frac{d}{N}\log\left(p\alpha^{-1}\right)\right)^{1/2} \frac{\gamma_1^2 \|\Sigma_X\|_{2,\infty} \|T\|_{\ell_1,\ell_2}^2}{\min_{j\neq k} |{\gamma_k}^2 - {\gamma_j}^2|^2} \|\Sigma_Y^{-1}\|_2, \tag{2.14}$$

where θ_k and η_k denote the high- and low-dimensional population canonical vectors, respectively.

The proof follows directly from Theorem A.3.2 in the supplementary materials. We refer to this bound as a "slow"-rate bound, as it makes fewer assumptions but results in slower con-

vergence rates relative to the sample size N. Specifically, we make no sparsity assumptions on the high-dimensional canonical vectors. In Theorem A.3.3 in the supplementary materials, we provide the "fast"-rate bound, where under more restrictive assumptions, the term $\left(\frac{d}{N}\log\left(p\alpha^{-1}\right)\right)^{1/2}$ is replaced by $\frac{d}{N}\log\left(p\alpha^{-1}\right)$, similar to what is observed in lasso regression problems (Hastie et al., 2015). The proof of Theorem 2.4.1 hinges on two key components: firstly, deterministic group lasso bounds for in-sample prediction error (Gaynanova, 2020), and secondly, the rates at which $\left\|\Sigma_{XY}-\hat{\Sigma}_{XY}\right\|_{2,\infty}$ and $\left\|B^{\top}(\Sigma_X-\hat{\Sigma}_X)B\right\|_2$ converge to zero under the sub-Gaussian assumptions for X and Y. Here, $\hat{\Sigma}_X$ and $\hat{\Sigma}_{XY}$ represent the sample covariance matrices. As an intermediate step in the proof, we show Theorem A.3.1 in the supplementary materials, which gives similar slow and fast rate bounds for the estimated canonical correlations $\hat{\gamma}_k$.

Under the stated assumptions the canonical vector estimates are consistent. Moreover, our rates of convergence depend on the dimension of the high-dimensional data, p, only through $\log(p)$. The bounds for the kth canonical directions depend on the nearest canonical correlation gaps, resembling those concerning the variance in the PCA literature.

We emphasize that our rates are dependent on Σ_X only through $\|\Sigma_X\|_{2,\infty}$, and not $\|\Sigma_X\|_2$. The norm $\|\Sigma_X\|_{2,\infty}$ can be much smaller than $\|\Sigma_X\|_2$, particularly when many of the X_j's are correlated with one another. This property highlights the robustness of the proposed methodology in the high-dimensional setting, where highly correlated covariates are commonplace.

We are able to establish our error bounds for each canonical vector θ_k, η_k, independently, and these bounds depend on each other only through the norms of the canonical vectors $\|T\|^2_{\ell_1,\ell_2}$, and through the neighboring canonical correlation gaps. It is also worth noting that the error associated with θ depends on Σ_X^{-1} but not Σ_Y^{-1}. Similarly, the error associated with η depends on Σ_Y^{-1} but not Σ_X^{-1}. Hence, Y can be poorly behaved without impacting the estimation of θ, and vice-versa.

2.4.2 *Estimation error rates for canonical directions from Asymmetric Sparse-Functional CCA*

In this section, we investigate the asymptotic properties of our proposed estimators $\hat{\psi}_k$ and $\hat{\theta}_k$, outlined in Algorithm 1, for the canonical functions ψ_k and canonical vectors θ_k. In this setting, the observations are pairs of Riemannian-valued functional data $y_i \in L^2(T\mu)$ and high-dimensional multivariate data $X_i \in \mathbb{R}^p$. Given the technical nature of many of the assumptions, we refer the reader to Assumptions A.4.1-A.4.4 in the supplementary materials for a complete list.

As in the multivariate case, we denote with γ_k the kth canonical correlation attained in the population version of the problem, and we denote with $K = \max\{i \in \{1, \dots d^{(\text{corr})}\} : \gamma_i > 0\}$ the number of nontrivial canonical vectors. We again use the conventions $\gamma_{K+1}^2 = -\infty$ and $\gamma_0^2 = \infty$. Recall that we denote by d the number of principal components we use in the estimation step and p the dimension of the multivariate data. We denote by $d^{(\text{corr})}$ the dimensionality with which the finite-correlation Assumption 2.2.1 holds, with $d^{(\text{corr})} \leq d \leq p$. We suppose that the canonical vectors $\{\theta_k\}$ are s-sparse with a group sparsity pattern. We let X_S denote the random vector where we omit covariates $\{X_j\}$ that do not contribute to the association structure with Y. For the high-dimensional terms to match the speed of convergence of the functional terms, we assume that $\Sigma_X^{1/2}$ satisfies the group restricted eigenvalue condition $\text{RE}(s, 3, d)$, introduced in Definition A.3.3 in the supplementary materials, with parameter $\kappa = \kappa(s, d, \Sigma_X^{1/2})$, which yields 'fast'-rate bounds.

We do not provide the 'slow'-rate bounds as in Theorem 2.4.1. The terms resulting from using Intrinsic RFPCA and estimating the Frechét mean μ converge at a rate of $1/N$, while the terms resulting from solving the multivariate CCA problem would converge at a rate of $1/\sqrt{N}$; considering the 'slow'-rate bound would amount to ignoring the contribution to the error from the functional estimation steps.

Theorem 2.4.2. *For some absolute constant $C > 0$, define the regularization parameter in Algorithm 1 as $\lambda = C\sqrt{\frac{d}{N}\log(p)}$. Then, under Assumptions A.4.1-A.4.4, for $k = 1, \dots K$, we*

have

$$\left\|\psi_k - \Gamma_{\hat{\mu},\mu}\hat{\psi}_k\right\|_\mu^2 = O_P\left(\frac{d^2 s\log(p)}{N}\frac{\left\|\psi_k\right\|_\mu^2 \kappa \left\|\Sigma_X\right\|_{2,\infty}}{\min_{j\neq k}\min\left\{\left|{\gamma_k}^2 - {\gamma_j}^2\right|, \left|\gamma_k - \gamma_j\right|\right\}^2}\right), \tag{2.15}$$

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\frac{ds\log(p)}{N}\frac{\left\|\Sigma_{X_S}^{-1}\right\|_2^{1/2} + \left(\frac{\gamma_1}{\gamma_k}\right)^2 \left\|\Sigma_X\right\|_{2,\infty}\left\|\Sigma_X^{-1}\right\|_2 \kappa^2}{\min_{j\neq k}\min\left\{\left|{\gamma_k}^2 - {\gamma_j}^2\right|, \left|\gamma_k - \gamma_j\right|\right\}^2}\right), \tag{2.16}$$

where we have omitted the terms $\mathbb{E}\left[\left\|\mathrm{Log}_\mu y\right\|_\mu^4\right]$ *and* $\mathrm{Var}\left(\left\langle\!\left\langle \phi_j, \mathrm{Log}_\mu y\right\rangle\!\right\rangle_\mu\right)$ *for* $j = 1, \ldots d$.

The theorem presented here is a special case of Theorems A.4.2 and A.4.3 in the supplementary materials. As in Theorem 2.4.1, the rate of convergence depends on p only through the term $\log(p)$ and on Σ_X only through $\left\|\Sigma_X\right\|_{2,\infty}$. Our rate also depends on the dimensionality of the reduced representation of the functional data, d, linearly and quadratically in the estimation of θ_k and ψ_k, respectively. The quadratic term d^2 is most likely not tight but arises from our choice to estimate each ϕ_j via $\hat{\phi}_j$, individually, rather than estimating subspaces. As in Theorem 2.4.1, the convergence rates depend on the neighboring correlation gaps.

It follows from Theorem 2.4.2 that if terms other than d, s, p, and N are treated as constants, then, if $d^2 s\log(p) = o(N)$, we have that $\hat{\psi}_k$ and $\hat{\theta}_k$ are consistent estimators for ψ_k and θ_k, respectively. Thus, for the proposed methodology, p is allowed to grow exponentially with respect to $\frac{N}{d^2 s}$ (i.e., $p \lesssim e^{\frac{N}{d^2 s}}$) and consistency is retained.

2.4.3 Estimation error rates for canonical variables from Asymmetric Sparse-Functional CCA

In this section, we investigate the asymptotic properties of the canonical variables. In general, without an assumption such as Assumption 2.2.1, the canonical directions ψ_k and θ_k do not necessarily exist. However, even in the absence of such an assumption (equivalently, when $d^{(\mathrm{corr})} = \infty$), using our proposed estimators $\hat{\psi}_k$ and $\hat{\theta}_k$ we still obtain a form of asymptotic consistency.

Recall that we denote by d the number of principal components used to represent the functional data, and p the dimension of the multivariate data. Given observations $(X_i, y_i)_{i=1}^N$, we use Algorithm 1 to obtain the estimates $\hat{\mu}$, $\left\{\hat{\phi}_j\right\}_{j=1}^d$, $\hat{H} = [\hat{\eta}_1, \ldots, \hat{\eta}_d]$, $\hat{T} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_d\right]$, and $\hat{\psi}_k = \sum_{j=1}^d \hat{\eta}_{kj}\hat{\phi}_j$ for $k = 1, \ldots d$.

We define the out-of-sample scores as follows. Let $(X_{\text{test}}, y_{\text{test}})$ be a new and independent data point drawn from the same distribution as the sample. We define

$$\left(\hat{U}_k, \hat{V}_k\right) \equiv \left(\langle\!\langle \operatorname{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \rangle\!\rangle_{\hat{\mu}}, \langle X_{\text{test}}, \hat{\theta}_k \rangle\right), \tag{2.17}$$

which represent the canonical scores obtained from the new data point using the canonical vectors estimated from the sample. Moreover, let (U_k, V_k) denote the solution to the infinite-dimensional population problem (2.3), defined with respect to y_{test} and X_{test}.

Following Theorem 10.2.3 in Hsing and Eubank (2015), we provide a probabilistic bound for $\mathbb{E}\left[\left(U_k - \hat{U}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]$ and $\mathbb{E}\left[\left(V_k - \hat{V}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]$ as the sample size N, the number of selected principal components d, and the dimension of the high-dimensional data p go to infinity. We choose this notion of error, where we condition on the sample, because it allows us to derive a result that is comparable to our canonical vector consistency results, while also integrating out the randomness of $(X_{\text{test}}, y_{\text{test}})$.

We make the same assumptions as in Section 2.4.2, with the important exception that we no longer make Assumption 2.2.1, and therefore allow the case $d^{(\text{corr})} = +\infty$. In other words, we allow for all principal scores of the functional data y to be correlated with the components of X. We denote with γ_k^* the kth canonical correlation attained in the infinite-dimensional population version of the problem (2.3), and we denote with $K = \max\{i \in \{1, \ldots p\} : \gamma_i^* > 0\}$ the number of nontrivial canonical vectors. We use the conventions ${\gamma_{K+1}^*}^2 = -\infty$ and ${\gamma_0^*}^2 = \infty$.

In Theorem 2.4.3, which we are about to present, the operator $\mathscr{C}_{12}$ appears, which is a cross-covariance operator containing information about the correlation between $\operatorname{Log}_\mu y$ and X. We defer its definition to equation (A.16) in A.2.4, due to its technical nature. We also employ its rank-d principal component approximation, which we denote as $\mathscr{C}_{12}^{(d)}$. The norm of the difference $\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|$ then represents how well our d-dimensional subspace

captures the true correlation structure between the infinite-dimensional functional data y and high-dimensional data X.

Theorem 2.4.3. *Let (U_k, V_k) be the solution pair to the problem in the infinite-dimensional population version of the problem (2.3) between y_{test} and X_{test}. Let $\left(\hat{U}_k, \hat{V}_k\right) \equiv \left(\langle\!\langle \mathrm{Log}_{\hat{\mu}} y_{test}, \hat{\psi}_k \rangle\!\rangle_{\hat{\mu}}, \langle X_{test}, \hat{\theta}_k \rangle\right)$, where $\hat{\theta}_k$, $\hat{\eta}_k$ and $\hat{\mu}$ have been estimated via Algorithm 1. For some absolute constant $C > 0$, define the regularization parameter in Algorithm 1 as $\lambda = C\sqrt{\frac{d}{N}\log(p)}$. Then, under Assumptions A.4.1-A.4.4, with the exception of Assumption 2.2.1, for $k = 1, \ldots K$, we have*

$$\max\left\{\mathbb{E}\left[\left(U_k - \hat{U}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right], \mathbb{E}\left[\left(V_k - \hat{V}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]\right\} \tag{2.18}$$

$$= O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2} - \gamma_j^{*2}\right|} + \frac{ds\log p}{N}\frac{\|\Sigma_X\|_{2,\infty}\,\kappa}{\min_{j\neq k}\min\left\{\left|\gamma_k^{*2} - \gamma_j^{*2}\right|, \left|\gamma_k^{*} - \gamma_j^{*}\right|\right\}^2}\right), \tag{2.19}$$

where we have omitted the terms $\mathbb{E}\left[\left\|\mathrm{Log}_\mu y\right\|_\mu^4\right]$ and $\mathrm{Var}\left(\langle\!\langle \phi_j, \mathrm{Log}_\mu y \rangle\!\rangle_\mu\right)$ for $j = 1, \ldots d$.

The theorem presented here is a special case of Theorem A.5.1 in the supplementary materials, whose proof can be found in Section A.5. Here, we recover the classical trade-off in selecting d: increasing d decreases the first term (a "bias" term arising from approximating the infinite-dimensional problem using a finite number of principal components), while increasing the second term (a "variance" term due to estimation error in solving the sample version of the finite-dimensional problem).

Comparing the variance term of these rates with the rates in Theorem 2.4.2 for the canonical directions, we observe one primary difference. Focusing on the canonical direction θ_k for simplicity, we see that the $\Sigma_X^{-1/2}$ factors have disappeared. These factors are necessarily present in the canonical vector rates because the θ_k are derived by taking unit-norm vectors and scaling them by $\Sigma_X^{-1/2}$. The canonical variables, on the other hand, are defined from the canonical directions as $\theta_k^\top X$, where the $\Sigma_X^{-1/2}$ hidden in the definition of θ_k "cancels out" with the $\Sigma_X^{1/2}$ in X, removing the factor $\Sigma_X^{-1/2}$ from the canonical variable rates.

2.5 Application to dynamic functional connectivity

2.5.1 Data and preprocessing

We analyze resting-state fMRI images from 1003 subjects in the Human Connectome Project dataset (Van Essen et al., 2012). Throughout the duration of these 15-minute fMRI scans, participants were at rest and not engaging in any specific activities. Details on the acquisition process can be found in Glasser et al. (2013) and Smith et al. (2013). The fMRI images have been pre-processed using the minimal pre-processing HCP pipeline (Glasser et al., 2013), including spatial artifact, distortion removal, and mapping onto a common reference template (Smith et al., 2013).

We define 360 spatially localized regions of interest (ROIs) using the multimodal parcellation proposed in Glasser et al. (2016). These 360 regions are further aggregated into 10 distinct functional systems following the definition in Power et al. (2011). These are the somatosensory/motor network (SMH), cingulo-opercular network (COP), auditory network (AUD), default mode network (DMN), visual network (VIS), frontoparietal network (FPT), salience network (SAL), ventral attention network (VAT), dorsal attention network (DAT), and a category for Other Regions (OTH), which includes areas that are not strictly classified within the aforementioned functional systems.

We partition the fMRI data into 20 time intervals of equal length. For each interval, we reduce the fMRI data to a 'functional fingerprint' representation that is a 10×10 SPD covariance that captures the temporal correlation between the fMRI signals of different functional systems within a specific time interval. These matrices are denoted as $y_i(t_j)$ where $i = 1, \ldots, N = 1003$ represents the subject and $j = 1, \ldots, 20$ denotes the time interval. We conducted sensitivity analysis by repeating the analysis with 10 and 40 time intervals of equal length. The results are virtually identical, as expected, since the time dynamics of the estimated mode of covariation, shown in Figure 2.3, does not seem to be constrained by the number of time points.

In addition, an extensive set of 150 subject traits of lifestyle, demographic, and psycho-

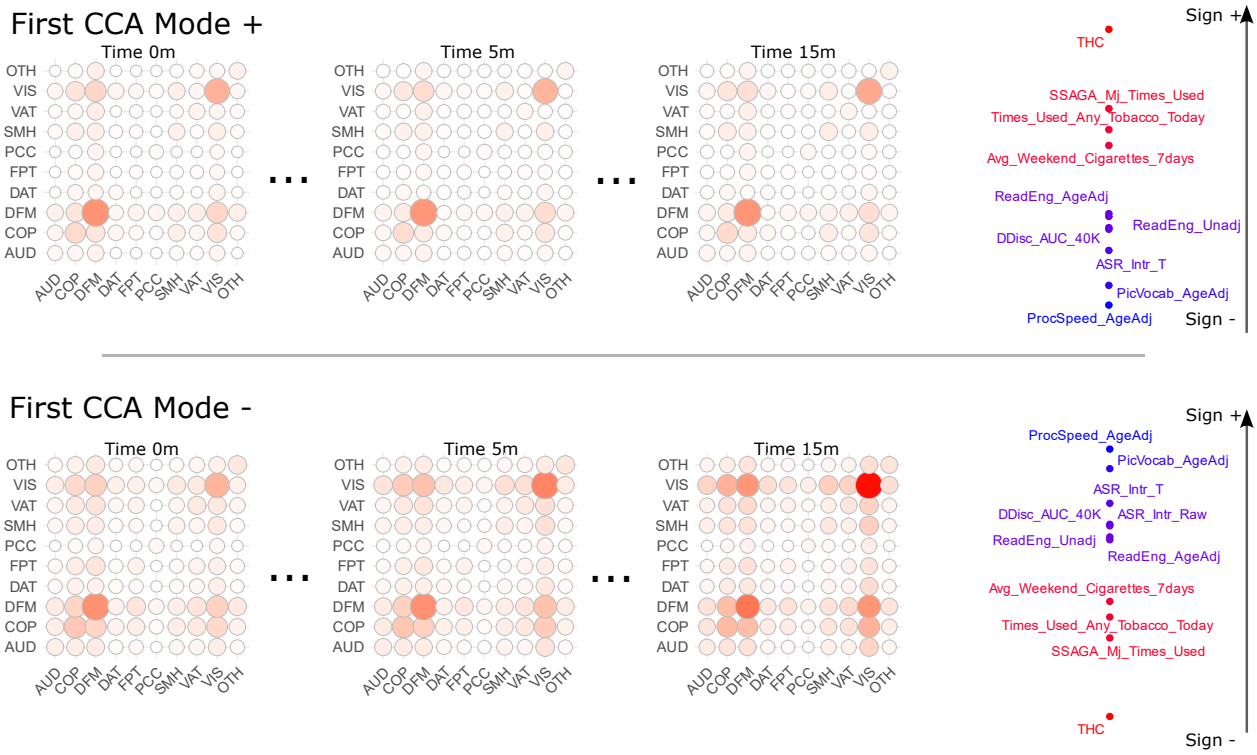


Figure 2.2: This figure illustrates the first mode of covariation between dynamic connectivity and behavioral measures. On the top panel, we show $\left(\mathrm{Exp}_{\hat{\mu}}\left(-c\hat{\psi}_1\right), -c\hat{\theta}_1\right)$, which we refer to as 'First CCA Mode +', on the bottom panel we show $\left(\mathrm{Exp}_{\hat{\mu}}\left(+c\hat{\psi}_1\right), +c\hat{\theta}_1\right)$, which we refer to as 'First CCA Mode -'. These represent two extremities of the spectrum identified by the first mode of covariation. Within each panel, we show the canonical function of SPD covariances $\mathrm{Exp}_{\hat{\mu}}\left(\pm\hat{\psi}_1\right)$ at three different times, and a subset of the selected entries of the canonical vector $\pm\hat{\theta}_1$. The depicted mode of covariation suggests that subjects with an increasing variance over time within the visual (VIS) and default mode (DFN) functional systems, as well as an increasing covariance between these systems, positively correlate with higher scores in 'ProcSpeed_AgeAdj' – assessing processing speed – and 'PicVocab_AgeAdj' – evaluating language/vocabulary comprehension and negatively correlate with using cannabis and opiates (variables THC, SSAGA_Mj_Use, and SSAGA_Times_Used_Opiates).

metric measures are also provided for the same cohort of 1003 subjects. We denote these by X_i, with $i = 1, \ldots, N = 1003$. To account for potential confounding factors, we regressed out of the 150 variables nine confounders identified in Smith et al. (2015), and the squares of the continuous ones, using multivariable linear regression.

2.5.2 Analysis

We apply Algorithm 1 to the pairs $(y_i(\cdot), X_i)$. Specifically, we model the SPD-valued functional data $\{y_i(\cdot)\}$ using the affine-invariant Riemannian metric. The choice of this metric is primarily driven by its ability to avoid the swelling effect (Dryden et al., 2009). The affine-invariant metric has been shown to be effective in prediction tasks (Barachant et al., 2013; Pervaiz et al., 2020).

The Fréchet mean $\hat{\mu}$ and tangent space representations $\{\mathrm{Log}_{\hat{\mu}}\, y_i\}$ are computed. See Section 2.3.2 for details. Both the hyperparameters λ and d, the number of PCs used to reduce the dimension of the SPD-valued functional data, are chosen by cross-validation. Specifically, for every candidate d, the parameter λ is chosen to minimize the cross-validated prediction error of the regression model in equation (2.9), while d is chosen by examining the scree plot of the cross-validated canonical correlations. We chose the smallest d for which the cross-validated correlations appear to level off, that is, $d = 12$. The outlined procedure results in a set of K estimated canonical directions $\left(\hat{\psi}_k, \hat{\theta}_k\right)_{k=1}^{K}$, where $\{\hat{\theta}_k\}$ are the canonical vectors associated with $\{X_i\}$, and $\{\hat{\psi}_k\}$ are the (tangent-space) representations of the canonical functions associated with $\{y_i\}$. After inspection of the cross-validated correlations and their associated variance, we decided to retain only the first pair of canonical directions.

2.5.3 Results and Discussion

In Figure 2.2, we display the first canonical direction $\left(\hat{\psi}_1, \hat{\theta}_1\right)$ by plotting

$$\left(\mathrm{Exp}_{\hat{\mu}}\left(-c\hat{\psi}_1\right), -c\hat{\theta}_1\right), \qquad \left(\mathrm{Exp}_{\hat{\mu}}\left(+c\hat{\psi}_1\right), +c\hat{\theta}_1\right),$$

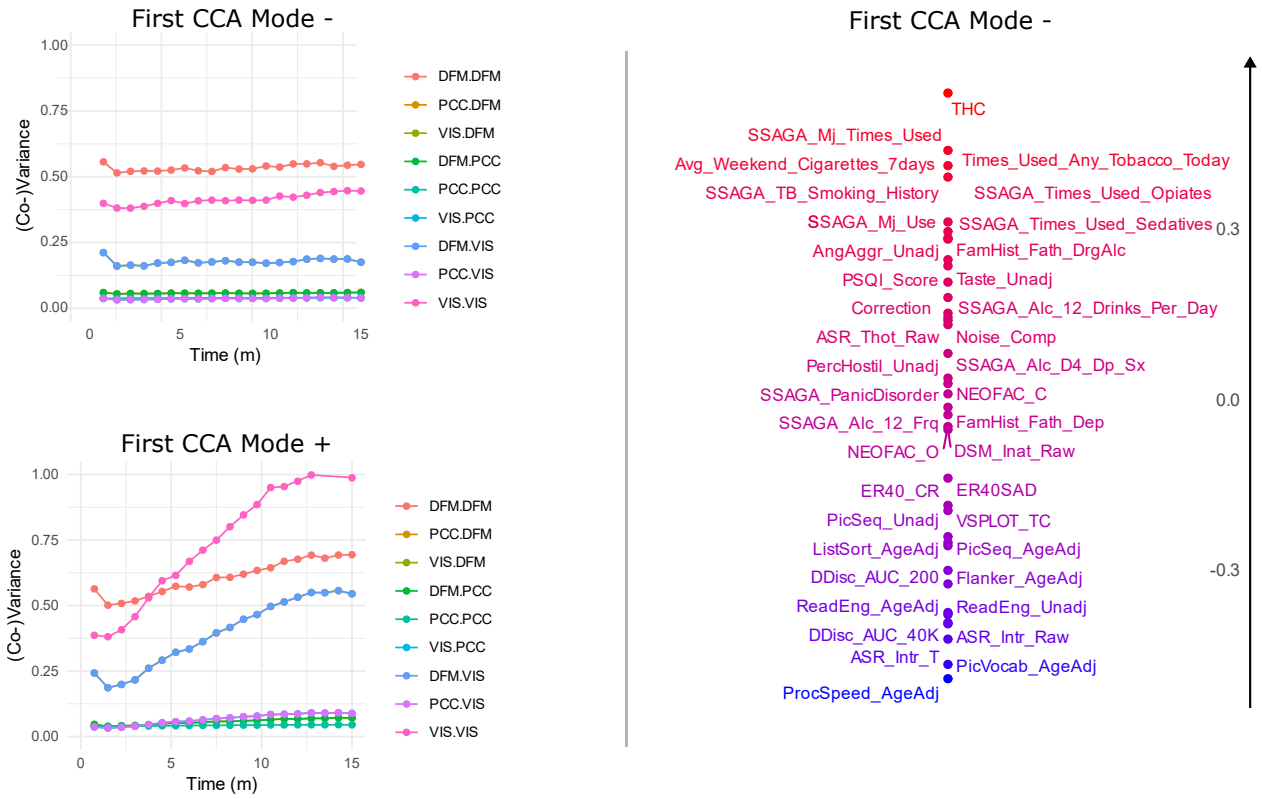


Figure 2.3: On the left panel, for both 'First CCA Mode -' and 'First CCA Mode +', we show the temporal dynamics of selected entries of the dynamic mode of connectivity shown in Figure‘2.2. Notably, some of these, e.g., the DFM-PCC covariance, remain stationary for both 'First CCA Mode +' and 'First CCA Mode -', while others, e.g., the DFM-VIS covariance, have markedly different patterns. On the right panel, we show a complete list of the 39 variables, of the canonical vector $\pm\hat{\theta}_1$, selected by the proposed model out of an initial set of 150, along with their relative importance.

for a fixed positive constant c. In the figure, we refer to $\left(\text{Exp}_{\hat{\mu}}\left(-c\hat{\psi}_1\right), -c\hat{\theta}_1\right)$ as 'First CCA Mode -' and to $\left(\text{Exp}_{\hat{\mu}}\left(+c\hat{\psi}_1\right), +c\hat{\theta}_1\right)$ as 'First CCA Mode +'. Intuitively, these represent the two extremities of the first mode of covariation between functional dynamic connectivity and lifestyle, demographic, and psychometric measures. The exponential map $\text{Exp}_{\hat{\mu}}(\cdot)$ allows us to map the canonical function back to the manifold of SPD-valued functions.

The estimated first pair of canonical directions appears to link subjects with increasing variance over time within the visual and default mode functional systems, and increasing co-variance over time between these functional systems, to 'positive' lifestyle, demographic, and psychometric measures, such as better 'ProcSpeed_AgeAdj' score, which tests the 'speed of processing' and better 'PicVocab_AgeAdj' score, which tests the ability to match an audio recording of a word to the most closely related picture. On the other hand, a more 'stationary' connectivity pattern is associated with more 'negative' lifestyle traits, such as a positive test for THC (THC), whether the subject has ever used cannabis (SSAGA_Mj_Use), and the number of times the subject has used opiates (SSAGA_Times_Used_Opiates). The cross-validated correlation of the identified mode of covariation is 0.075, whereas its in-sample correlation is 0.259. These values are relatively high compared to correlation analyses conducted in other studies using the same dataset (Lin et al., 2020).

The multivariate component of the identified mode of covariation resembles the one found between static functional connectivity and lifestyle, demographic, and psychometric variables in Smith et al. (2015). However, as illustrated in Figure 2.3, our analysis reveals the non-stationary nature of this mode, with the latter portion of the scan emerging as the most informative in terms of functional connectivity. It is during this phase that the differences between the extremities of the mode of covariation become more evident.

It is possible that a latent variable linked to both the identified dynamic connectivity and the behavioral components of the first mode of covariation is responsible for the observed correlation between them. This variable potentially reflects the subjects' experience, such as growing impatience or distractions, during the 15-minute resting-state MRI session where they were instructed not to engage in specific tasks. Indeed, it appears that the 'First CCA Mode -' subjects (who are more likely to test positive for THC and have used opiates) maintain a consistent 'wandering mind', whereas the 'First CCA +' subjects (who are likely to have better pattern completion skills and language/vocabulary comprehension) show a behavioral drift. This results in a progressive activation of the visual cortex and default mode network, and their cooperation, which might reflect a growing unease and consequent

search for external stimuli.

2.5.4 Comparison with other approaches

We also explore replacing our proposed asymmetric sparse CCA model with standard CCA, incorporating a ridge penalty added to the covariance of the behavioural data, and with the PMA-sparse CCA model proposed in Witten et al. (2009). Cross-validating standard CCA yielded an estimated out-of-sample correlation for the first mode of covariation of 0.051, while applying the PMA-sparse CCA model yielded a correlation of 0.027. However, standard CCA does not perform selection of the important lifestyle, demographic, and psychometric measures, and did not show a particularly high level of consistency, both in the weights and their sign assigned to the behavioural variables, compared to those obtained by our method. We believe this may be due to overfitting. On the other hand, the first mode of covariation computed using PMA-sparse CCA showed greater similarity to our mode of co-variation, although the out-of-sample correlation of the identified mode was lower than that of our proposed approach, indicating that the mode found might be suboptimal.

Further, we run an analogous analysis on static connectivity, that is, we compute the functional fingerprints using the whole time series (see also Section 2.3.2 for methodological details). The results of this analysis are qualitatively similar to those obtained from the dynamic connectivity analysis. Specifically, a very similar behavioural mode of variation is identified, and this is associated with a static connectivity mode of variation that resembles the one in Figure 2.2 at time 15m. However, completely removing the temporal component of the connectome does not provide insights into the possibility that the observed association reflects some subjects' growing unease in the scanner. This highlights the importance of dynamic connectivity studies.

2.6 Discussion and conclusions

In this chapter, we introduce a novel statistical model for identifying shared variation patterns between manifold-valued functional data and high-dimensional data. We refer to this

setting as asymmetric due to the differing nature of the data. The proposed asymmetric CCA approach is designed to control the complexity of the canonical directions associated with the functional data by using Riemannian FPCA. This facilitates the identification of a lower-dimensional, smooth subspace onto which these data can be projected. Moreover, our approach controls the complexity of the high-dimensional canonical directions, which lack spatial structure, through a sparsity-promoting penalty that leads to the selection of the important variables. As opposed to other methods in the literature, this is achieved without requiring the estimation of the precision matrix of the high-dimensional data, which is in general prohibitive.

We apply asymmetric CCA to explore the association structure between resting-state dynamic functional connectivity, represented as time-indexed covariance matrices, and high-dimensional behavioral, lifestyle, and demographic features. Our analysis reveals a non-stationary pattern in functional connectivity, indicating that the usual assumption of temporal stationarity may not hold, even in resting-state studies. While this work focuses on an application in dynamic connectivity, the proposed method can be easily adapted to accommodate different Riemannian structures and to employ different data representation models, paving the way for several future extensions.

Yet, our approach has some limitations. For example, the dimension reduction step is unsupervised, which may result in the loss of small signals that are highly correlated with the high-dimensional data. In future work, we hope to explore supervised extensions of our method to address this limitation. Moreover, our theoretical analysis does not directly generalize to the setting of sparsely observed Riemannian data, as the downstream analysis of the CCA estimators requires proving the convergence of the Intrinsic RFPCA estimators in expectation, rather than in probability. While we establish these results for the fully observed case, extending the proof to the sparsely observed setting is non-trivial.

Chapter 3

A CORRELATION ANALYSIS APPROACH TO SUPERVISED DISENTANGLEMENT

3.1 Introduction

Disentangled representation learning (Wang et al., 2024) aims to provide a low-dimensional representation of data in which distinct factors of variation are captured by separate latent coordinates. This has been recognized as a fundamental problem in interpretable machine learning (Rudin et al., 2022). While unsupervised disentanglement, where no auxiliary data is available, is actively studied (see, e.g., Balabin et al., 2024; Meo et al., 2024), it has been shown to be unidentifiable (Locatello et al., 2019). This challenge has motivated the development of supervised (Liu et al., 2022a; Wang et al., 2024) and weakly supervised approaches (Locatello et al., 2020; Shen et al., 2022; Shu et al., 2019), which construct a lower-dimensional representation of a 'target' data view using another 'auxiliary' data view to guide its construction. Indeed, latent low-dimensional representations are useful for their own sake, but they are most often used for downstream tasks where supervision may provide the inductive bias needed for generalization.

Broadly, supervised disentanglement methods follow one of two strategies: (i) a two-stage approach, first performing unsupervised disentanglement of the target view, and subsequently regressing the learned latent variables against an auxiliary interpretable view (Adel et al., 2018; Van et al., 2020)) or (ii) a joint approach, where both steps are carried out simultaneously (An & Jeon, 2024; Ding et al., 2020; Inecik et al., 2025; Lu et al., 2024; Nalisnick et al., 2019; Pati & Lerch, 2021; Zhao et al., 2019). We argue that canonical correlation analysis (CCA), rather than regression, offers a more principled approach to aligning the lower-dimensional representation of the target data view with the auxiliary data view.

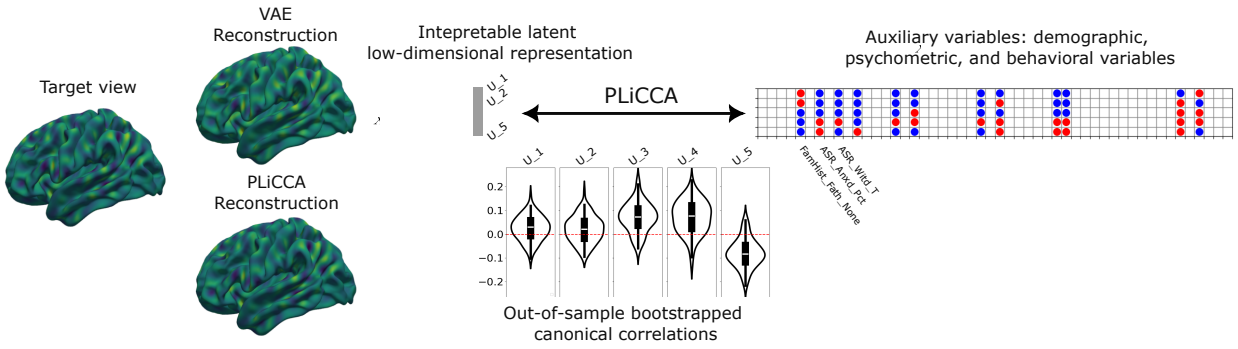


Figure 3.1: Results of the VAE formulation of PLiCCA applied to 700 subjects from the Human Connectome Project (Van Essen et al., 2013). Specifically, we consider cortical thickness as the main view and demographic, psychometric, and behavioral variables as auxiliary views. Both PLiCCA and standard VAEs provide expressive, low-dimensional latent embeddings that achieve satisfactory reconstruction errors. However, when applying an unsupervised VAE to learn the embedding map g, the canonical correlation model in Theorem 3.3.3 reveals *no* correlation between the constructed latent representations and the auxiliary variables. In contrast, PLiCCA yields latent variables that are linearly associated (from 300 validation samples) with sparse linear combinations of the auxiliary variables, thereby providing interpretable and scientifically meaningful representations. Out of the 150 auxiliary variables, the model selected 50. For simplicity, we show only the first 50 variables in the figure. Red circles indicate a positive effect, blue circles a negative effect, and no circle means the variable was not selected.

Our approach finds a latent representation of the target view that is *linearly correlated* with the auxiliary view. Imposing a (partially) linear structure on the latent space offers advantages for downstream tasks, such as improving interpretation of the latent space (Huben et al., 2023), enabling latent space interpolation (Bodin et al., 2025), facilitating conditional sampling (Härkönen et al., 2020; Jahanian et al., 2020), and supporting few-shot regression (Nitzan et al., 2022).

Our contributions are as follows:

1. We propose both joint and two-stage approaches to the supervised disentanglement problem via a novel partially linear invertible canonical correlation analysis (PLiCCA). PLiCCA constructs invertible, nonlinear latent representations of the target view that are maximally associated with sparse linear combinations of the auxiliary view, yielding embeddings that are ordered by interpretability.
2. We prove the existence of population solutions to PLiCCA, and provide a rigorous theoretical characterization of the problem as a nonlinear regression. Using this characterization, we show its connection with conditional latent variable models, in particular, conditional variational autoencoders and conditional normalizing flows. We demonstrate formally how both models can be viewed as relaxations of PLiCCA problems, replacing hard to enforce global constraints with local ones. This enables the efficient solving of PLiCCA via 'proxy' problems derived from contemporary conditional generative models.

3.2 Background and related work

CCA (Chapman & Wang, 2021; Guo & Wu, 2019; Hotelling, 1936; Yang et al., 2019) is a classical statistical method that, given two data views, aims to find latent variables defined as linear transformations of the variables in each view such that the resulting representations are maximally correlated. Hence, it provides an efficient basis (latent representations) for both data views, but unlike principal components analysis, each basis is informed by the other data view.

An important advantage of linear CCA over more complex methods is its simplicity, and

thus its interpretability (Gosiewska et al., 2021). To preserve this property and to ease estimation even for high-dimensional views, sparse CCA approaches have been proposed, which typically employ sparsity-inducing penalties to perform variable selection (see, e.g., Buenfil & Lila, 2024; Bykhovskaya & Gorin, 2023; Li et al., 2024, and references therein). The supervised disentanglement approach considered in this chapter, PLiCCA, builds on sparse CCA, but treats the two data views asymmetrically: one data view is modeled linearly, while the other data view is modeled nonlinearly.

Nonlinear CCA (Breiman & Friedman, 1985; Hannan, 1961; Lancaster, 1958; Michaeli et al., 2016) replaces linear transformations with nonlinear classes of functions, such as reproducing kernel Hilbert spaces (Akaho, 2006), and, more recently, neural networks (Andrew et al., 2013; Friedlander & Wolf, 2023). However, these approaches do not preserve the invertibility of the nonlinear mappings, hindering interpretability. To address this, Wang et al. (2015) proposed modifying the CCA objective function to additionally minimize a reconstruction error term using autoencoders. However, it takes a deterministic form and therefore does not leverage the advantages of variational autoencoders over standard autoencoders.

An approach more closely related to our own is that of Gundersen et al. (2019), who propose learning nonlinear latent representations of each view that maximize the likelihood of a linear probabilistic CCA model, although they do not explicitly maximize correlation. Zhang et al. (2023) also propose a related idea to maximize a classical CCA between the learned latent representations, but their method is more specialized to audio/visual applications. However, neither work provides a formal theoretical treatment of the underlying model.

There is a large body of work on CCA-inspired approaches to solving the multi-view data problem, a problem related to but distinct from supervised disentanglement (Aguila & Altmann, 2024; Guo et al., 2019). The general goal in multi-view learning is to find *shared* structure between two (or more) data views, often in the form of a learned shared subspace that embeds the views simultaneously. A large subset of these approaches relies on latent variable models (He et al., 2020; Karami & Schuurmans, 2021; Lyu et al., 2022; Qiu

et al., 2022; Senellart et al., 2023; Wang et al., 2016). Superficially, these approaches appear closely related to ours in that they combine CCA with latent variable models, but there are two major differences. First, these approaches primarily aim to solve the multi-view data problem, which is related to but distinct from the supervised disentanglement problem, as they learn shared embeddings of the views simultaneously. Second, they are typically inspired by CCA but lack a theoretically justified connection to the CCA formulation.

There is also a recent line of work relating conditional latent variable models to independent component analysis (ICA), enabling the application of ICA to disentanglement problems (see, e.g., Hyvärinen et al., 2023; Khemakhem et al., 2020; Zheng et al., 2025, and references therein). Our approach is related in that we also leverage the connection between conditional latent variable models and a component analysis problem. However, we advocate the use of CCA to provide more interpretable embeddings. The work of Basile et al. (2025) defines a notion of nonlinear correlation between manifolds and is thus tangentially related to ours.

3.3 Partially linear invertible canonical correlation analysis

3.3.1 Population Problem

In this section, we define the population problem of interest and establish that it is well-defined. We are interested in solving the *partially linear CCA* problem (Michaeli et al., 2016), but with the additional constraint that the nonlinear embedding is approximately invertible. We refer to the partially linear *invertible* CCA problem as PLiCCA. See Appendix Section B.2 for background on the non-invertible partially linear CCA problem.

We observe two random vectors $Y \in \mathcal{M} \subseteq \mathbb{R}^q$, where $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^q$, for which we seek a nonlinear invertible latent representation (i.e., the target view), and $X \in \mathbb{R}^p$ represents the high-dimensional auxiliary data. Without loss of generality, X and Y satisfy $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. Throughout the manuscript, we denote the covariance of a random vector Z as Σ_Z, and we call Z isotropic if $\Sigma_Z = I_d$.

We can formalize PLiCCA using a constraint set $\mathcal{C}$, which will formalize the notion of

invertibility of the latent representation and will be specified momentarily:

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d\\ g\in\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d}\mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{3.1}$$

where $T = [\theta_1, \ldots, \theta_d]$.

Intuitively, we can think of $U_i \equiv g_i(Y)$ and $V_i \equiv \theta_i^\top X$, as nonlinear and linear transformations of Y and X, respectively, which have maximal correlation between them while being pairwise uncorrelated, enforced through $\Sigma_U = \Sigma_V = I_d$. The variables (U_i, V_i) are referred to as *canonical variable pairs*, and we can order them so that the maximizing correlations $\gamma_i \equiv \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]$ are descending with i. The γ_i are referred to as the *canonical correlations*. The columns of T are referred to as the *canonical vectors* for X, and g is referred to as the *canonical embedding* for Y.

We remark that we have squared the correlations in contrast to the original partially linear CCA formulation. It can be shown that when g is linear, maximizing the squared correlations is equivalent to maximizing the sum of the correlations (see Appendix Section B.3). This minor modification of the objective allows us to draw closer connections to regression and latent variable models (see Theorem 3.3.3).

3.3.2 Supervised disentanglement via PLiCCA

Before specifying $\mathcal{C}$, we clarify how solving the PLiCCA problem provides an approach to the supervised disentanglement problem. The goal of *supervised* disentanglement is to find a low-dimensional, interpretable latent variable $U \in \mathbb{R}^d$ and a decoder $f : \mathbb{R}^d \to \mathbb{R}^q$ such that

$$Y = f(U) + \varepsilon, \tag{3.2}$$

where ε is small in magnitude and where the second data view X aids in the interpretability of U.

Suppose we have solved the PLiCCA problem with g invertible and inverse f. Then

$$Y = f(g(Y)) \tag{3.3}$$

$$= f(U). \tag{3.4}$$

This decomposition of Y has several desirable properties:

1. The variable U disentangles Y: $Y = f(U)$, with U satisfying $\Sigma_U = I_d$, so that the U_i are uncorrelated with one another.

2. If we estimate the canonical vectors T with a sparse structure, then the associated $V_i = \theta_i^\top X$ are easy to interpret as linear combinations of only a few selected X_i, while also being uncorrelated with one another.

3. Thus, the U_i are naturally ordered by interpretability, in the sense that the correlation between U_i and sparse linear combinations of X is nonincreasing. Latent traversals (see e.g. Song et al., 2023, and references therein) via f then allow us to see how changes in the selected X_i are associated with Y on its original scale, not just in its latent space.

3.3.3 Defining the invertibility constraint set $\mathcal{C}$

To formalize PLiCCA, we need to specify the set $\mathcal{C}$, which provides the correct notion of invertibility of the canonical embedding g. Given that g also needs to project the data onto a lower-dimensional space, ideally we would specify $\mathcal{C}$ as the set of diffeomorphisms from $\text{supp}(Y)$ to $\mathbb{R}^d$. However, this would require $\text{supp}(Y)$ to be contained in a manifold $\mathcal{M}$ with dimension $\dim(\mathcal{M}) = d$, which is unrealistic due to noise contamination in the absence of prior knowledge.

A more realistic approach is to impose the following autoencoder-type condition, which states that Y lies *approximately* on a d-dimensional manifold:

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

In the following theorem, we are able to establish the existence of a solution to our PLiCCA problem for a class of functions $\mathcal{C}$ with this property. We denote the constrained set of functions as $\mathcal{C}_{\text{VAE}}$, in anticipation of the methodology which employs variational autoencoders (VAEs).

We recall that a function $f : K \to \mathbb{R}^d$ is M-Lipschitz if it satisfies $\|f(x) - f(y)\|_2 \leq M \|x - y\|_2 \ \forall x, y \in K \subseteq \mathbb{R}^q$.

Theorem 3.3.1. *Suppose Σ_X is invertible and that $\text{supp}(Y) \subseteq K$, where K is a compact subset of $\mathbb{R}^q$. Fix constants $M, m > 0$, $\varepsilon > 0$. If $\mathcal{C}$ is chosen to be*

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

and $\mathcal{C}_{\text{VAE}} \cap \{g : \Sigma_{g(Y)} = I_d\} \neq \varnothing$, then there exists a solution (g, T) to the PLiCCA problem

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}_{\text{VAE}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2. \tag{3.5}$$

Remark 1. *For fixed ε, we can always find the smallest d such that $\mathcal{C}_{\text{VAE}}$ is non-empty. In fact, picking $d = q$ achieves a reconstruction error of 0 with the identity map.*

We emphasize that this is the first work to prove this, as many of the proposed methods only consider finite-sample versions of the problem. Establishing existence at the population level is important because it guarantees that the underlying problem is well-posed and that finite-sample methods are approximating a meaningful target.

3.3.4 Special case: a priori dimension reduction

Solving PLiCCA simplifies when Y lies on a *known* or *previously estimated* lower-dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^q$. In particular, if $\dim(\mathcal{M}) = d$ and $\mathcal{M}$ is a connected complete Riemannian manifold of non-positive sectional curvature, then thanks to the Cartan-Hadamard Theorem (Lee, 2018), there exists a global chart $\phi : \mathcal{M} \to \mathbb{R}^d$ for $\mathcal{M}$, i.e., $\mathcal{M}$ is diffeomorphic to $\mathbb{R}^d$.

Denoting $W \equiv \phi(Y) \in \mathbb{R}^d$ as the dimension-reduced representation of Y, we can reformulate the partially linear CCA problem in terms of W rather than Y, using functions $\tilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ in place of $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)} = \Sigma_{T^\top X} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\tilde{g}_i(W)\theta_i^\top X\right], \tag{3.6}$$

The equivalence follows from the invertibility of ϕ. To recover g by solving problem (3.1) over $\tilde{g}$, we simply set $g = \tilde{g} \circ \phi$.

Invertibility of g then follows from invertibility of $\tilde{g}$. In practice, this can be enforced by choosing $\tilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ from a restricted class of functions $\mathcal{C}$ that are invertible by design. This is in contrast to the previous setting where $g : \mathbb{R}^q \to \mathbb{R}^d$ mapped from high-to-low dimensional spaces and where strictly enforcing invertibility was not feasible. Here, we denote this constraint set $\mathcal{C}$ as $\mathcal{C}_{\text{NF}}$, in anticipation of one of the proposed approaches that employs normalizing flows.

We recall that a function $f : K \to \mathbb{R}^d$ is called bilipschitz with parameters (m, M) if it satisfies $m\|x - y\|_2 \le \|f(x) - f(y)\|_2 \le M\|x - y\|_2 \ \forall x, y \in K$. With $\mathcal{C} = \mathcal{C}_{\text{NF}}$, we are able to establish the existence of a solution to our PLiCCA problem, analogous to Theorem 3.3.1.

Theorem 3.3.2. *Suppose Σ_X is invertible, and that $\text{supp}(W) \subseteq K$, where K is a compact subset of $\mathbb{R}^d$. Fix constants $M, m > 0$. Then, if $\mathcal{C}$ is chosen to be*

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

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

$$\underset{\substack{\tilde{g}:\mathbb{R}^d \to \mathbb{R}^d,\ T \in \mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)} = \Sigma_{T^\top X} = I_d \\ \tilde{g} \in \mathcal{C}_{\text{NF}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\tilde{g}_i(W)\theta_i^\top X\right], \tag{3.7}$$

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

Remark 2. *In Theorem 3.3.2, as well as Theorem 3.3.1, we assume that the random vector of interest has compact support. This assumption simplifies the proof, allowing us to appeal to the simplest form of the Arzelà–Ascoli theorem (Rudin, 1976), but the statements still hold*

in the non-compact case if we add a vanishing at infinity-type assumption; see Theorem 5 of Krukowski (2018).

In cases where Y has a known manifold structure, for instance, positive definite matrices (Buenfil & Lila, 2024; Kim et al., 2014) or probability densities (Cho et al., 2022), (global) logarithmic maps can be used in place of ϕ. Hence, the model in equation (3.7) provides a natural approach to incorporating this geometry. In cases where the manifold structure is not known, a two-stage approach to the PLiCCA problem can be used. First, we learn an unsupervised latent representation of Y, $W \in \mathbb{R}^d$, via the approximation $Y = \psi(W) + \varepsilon$, or use pretrained models for this step. Second, we solve the nonlinear problem in equation (3.7).

3.3.5 Connection to regression models

We have now defined the population problem of interest. To establish the connection between PLiCCA and conditional latent variable models such as VAEs and normalizing flows, we first relate PLiCCA to a nonlinear regression. Below, $\mathcal{C}$ can refer to either $\mathcal{C}_{\text{VAE}}$ or $\mathcal{C}_{\text{NF}}$. We introduce the following notation: for $A, B \in \mathbb{R}^{d\times d}$ that are positive semidefinite, we write $A \preceq B$ to denote the Loewner ordering, meaning that $B - A$ is positive semidefinite. If $\mathcal{C}$ is a set of functions, then for $a > 0$ we denote $a\mathcal{C} \equiv \{af : f \in \mathcal{C}\}$.

Theorem 3.3.3. *Finding (g, T) that solve the PLiCCA problem,*

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d\\ g\in a^{-1/2}\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{3.8}$$

where the $\theta_i \in \mathbb{R}^p$ are the columns of T, is equivalent to finding (g', B) that solve the regression problem

$$\underset{\substack{g':\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ g'\in\mathcal{C},\Sigma_{g'(Y)}\succeq aI_d}}{\text{minimize}} \mathbb{E}\left[\|g'(Y) - B^\top X\|_2^2\right], \tag{3.9}$$

for any $a > 0$.

Furthermore, we have the following relationship between B, T, g' and g: if $\tilde{H}\Lambda^2\tilde{H}^\top$ is an eigendecomposition of

$$\Sigma_{g'(Y)}^{-1/2} B^\top \Sigma_X B \Sigma_{g'(Y)}^{-1/2}, \tag{3.10}$$

where $\tilde{H} \in \mathbb{R}^{d\times d}$ is orthogonal and Λ is diagonal, then letting $H = \Sigma_{g'(Y)}^{-1/2}\tilde{H}$, we have

$$T = BH\Lambda^{-1} \tag{3.11}$$

$$g(y) = H^\top g'(y). \tag{3.12}$$

While it is well known that CCA has a regression formulation (see Lyu et al. (2022), for example), here we additionally leverage the fact that one side of PLiCCA is linear, which results in an optimization over the unconstrained regression matrix B rather than the constrained canonical vectors T. Furthermore, this result shows that there is no need to enforce the strict, and typically computationally expensive constraint $\Sigma_{g(Y)} = I_d$ during optimization (Andrew et al., 2013; Friedlander & Wolf, 2023), since the whitening matrix H can correct for this after the regression. We further explore the relaxing of this constraint in the following section. Theorem 3.3.3 can also be understood as a generalization of those in Buenfil and Lila (2024) and Wang and Zhou (2021b), which apply only to the linear case.

3.4 Methodology

Next, we introduce conditional latent variable models—specifically, the conditional variational autoencoder (VAE) (Harvey et al., 2022; Khemakhem et al., 2020; Sohn et al., 2015) and conditional normalizing flow (NF) (Papamakarios et al., 2017; Winkler et al., 2019).

3.4.1 Conditional VAE

Let $Z \in \mathbb{R}^d$ be a latent random variable. A standard derivation (see Appendix Section B.4) shows that if we specify our model as

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right) \tag{3.13}$$

$$Y|Z, X \sim \mathcal{N}\left(f(Z), I_d\right), \tag{3.14}$$

and we model $q(z|y) \sim \mathcal{N}(g(y), I_d)$, then maximizing the evidence lower bound on the likelihood is equivalent to minimizing

$$\underset{g:\mathbb{R}^q \to \mathbb{R}^d, B \in \mathbb{R}^{p\times d}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right] \tag{3.15}$$

$$+\beta_{\text{VAE}} \mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right]. \tag{3.16}$$

The outermost expectation is taken with respect to the observed variables (X, Y), and we have added a tuning hyper-parameter β_{VAE} to the objective function. It is clear that the introduced conditional VAE objective already closely resembles the PLiCCA objective in equation (3.9) when choosing $\mathcal{C} = \mathcal{C}_{\text{VAE}}$. In the next section, we explore this connection more rigorously.

This model is a vanilla conditional VAE, which allows for a more thorough theoretical study. In practice, however, we use more flexible distributions, for example by allowing $\Sigma_{Y|Z}(z)$ and $\Sigma_{Z|Y}(y)$ to vary with their inputs. We also note that the encoder is chosen to be only a function of y, and not of the conditioning variable x, to meet our goals of disentangling the target view Y. In practice, we estimate the expectation of the above quantity using the sample $(X_i, Y_i)_{i=1}^N$.

Our methodology is a direct application of Theorem 3.3.3. However, instead of solving the problem in equation (3.9), we solve the simpler "proxy problem" of minimizing the conditional VAE objective (3.15) with respect to g and B. To aid interpretability, we augment the objective function with a sparsity-inducing group lasso penalty $\sum_{j=1}^p \|b_j\|_2$ (Yuan & Lin, 2006), where b_j denotes the jth row of B corresponding to covariate X_j. In practice, proximal gradient descent is applied, which is straightforward thanks to the closed-form solution of the proximal operator of $\sum_{j=1}^p \|b_j\|_2$ (see, for instance, Section 6.5.4 of Parikh (2014) for a derivation and Murray et al. (2019) for an implementation). After obtaining the estimates $\hat{g}$, $\hat{B}$, we estimate $\Sigma_{g(Y)}$ and Σ_X via their sample covariance estimates.

Then, as suggested by Theorem 3.3.3 (equations (3.10)-(3.12)), we take the eigendecomposition $\tilde{H}\hat{\Lambda}^2\tilde{H}^\top$ of $\hat{\Sigma}_{g(Y)}^{-1/2}\hat{B}^\top\hat{\Sigma}_X\hat{B}\hat{\Sigma}_{g(Y)}^{-1/2}$, where $\tilde{H} \in \mathbb{R}^{d\times d}$ is orthogonal and $\hat{\Lambda}$ is diagonal. Letting $\hat{H} = \hat{\Sigma}_{g(Y)}^{-1/2}\tilde{H}$, we finally obtain our estimates of the canonical vectors for X, $\hat{T} = \hat{B}\hat{H}\hat{\Lambda}^{-1}$,

as well as our correctly normalized canonical variables for Y, $\hat{g}(y) \equiv \hat{H}^\top \hat{g}(y)$.

Note that if the latent dimension of the VAE is misspecified, some latent variables $g_i(Y)$ will collapse to 0 (Bonheme & Grzes, 2023; Zheng et al., 2022). Hence, we inspect $\hat{\Sigma}_{g(Y)}$ for latent dimensions where $\hat{\mathrm{Var}}\left(g_i(Y)\right)$ has collapsed to 0, and we ignore these dimensions in $\hat{B}$, $\hat{g}$, and $\hat{\Sigma}_{g(Y)}$.

Our approach offers several advantages over existing methods: it avoids the need to estimate Σ_X^{-1}, and it does not require the enforcement of the global whitening constraint $\Sigma_{g(Y)} = I_d$ throughout optimization, enabling the use of standard batch gradient descent. The group sparsity imposed on B transfers directly to T since $T = B\left(H\Lambda^{-1}\right)$, retaining the variable selection done by the group lasso penalty, and if B is fixed at zero, the method reduces to a standard unsupervised VAE. Finally, because we compute covariance matrices from sample estimates, the resulting $\hat{T}$ and $\hat{g}$ automatically satisfy the correct orthogonality conditions $\hat{T}^\top \hat{\Sigma}_X \hat{T} = \hat{\Sigma}_{g(Y)} = I_d$.

3.4.2 Connection between PLiCCA and conditional VAE

Next, we elucidate the connection between the PLiCCA objective function and the 'proxy' conditional VAE objective. We begin with the natural regression problem we would be encouraged to solve by Theorem 3.3.3 for the $\mathcal{C} = \mathcal{C}_{\mathrm{VAE}}$ constraint set:

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ f:\mathbb{R}^d \to \mathbb{R}^q \\ B \in \mathbb{R}^{p \times d},\ \Sigma_{g(Y)} \succeq a I_d}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right] \tag{3.17}$$

$$+\mathbb{E}\left[\left\|Y - f(g(Y))\right\|_2^2\right], \tag{3.18}$$

We aim to show how the conditional VAE objective, (3.15), essentially reduces to the latter PLiCCA regression problem.

Clearly, the first terms in equation (3.15) and (3.17), respectively, are identical. Hence, we focus on the second term of the conditional VAE objective (3.16), $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\left\|Y - f(z)\right\|_2^2\right]\right]$. Using the form of the approximating posterior distribution $q(z|Y) \sim \mathcal{N}\left(g(Y), I_d\right)$, this term

becomes

$$\mathbb{E}\left[\mathbb{E}_{\varepsilon}\left[\|Y - f\left(g(Y) + \varepsilon\right)\|_2^2\right]\right], \tag{3.19}$$

where $\varepsilon \sim \mathcal{N}(0, I_d)$. This is clearly a noisy version of the reconstruction term (3.18) in the PLiCCA problem. More surprisingly, perhaps, is the fact that by adding this noise, the reconstruction term of the VAE (3.16) encourages the constraint $\Sigma_{g(Y)} \succeq aI_d$.

We first provide some intuition and then state our formal result. From equation (3.19), we see that if the decoder f wants to use the ith coordinate $g_i(Y)$ to reconstruct Y, then the variance of $g_i(Y)$ cannot be much smaller than 1, the variance of the noise ε_i. However, depending on the choice of d, the dimension of the latent space, the VAE may ignore certain latent dimensions by setting $g_i(y) = 0$ for all y and having the decoder f disregard its ith component. This phenomenon is closely related to the distinction between active and inactive latent dimensions in (conditional) VAEs (Bonheme & Grzes, 2023; Zheng et al., 2022). Therefore, the conditional VAE does not, and should not, *strictly* enforce the constraint $\Sigma_{g(Y)} \succeq aI_d$, since it should be free to represent Y with only as many latent dimensions as necessary. For this reason, in practice, we inspect and drop inactive latent dimensions when applying Theorem 3.3.3.

However, although the conditional VAE does not enforce $\Sigma_{g(Y)} \succeq aI_d$, it does enforce a weaker form of this constraint: a lower bound on $\mathrm{tr}(\Sigma_{g(Y)})$. In line with our prior intuition, this lower bound depends exactly on the quality of reconstruction by g and f and on the magnitude of the noise in $q(z|Y)$. Formally, we have the following result.

Theorem 3.4.1. *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,*

$$\mathrm{tr}\left(\Sigma_{g(Y)}\right) \geq \sigma^2_{\text{enc}} C(\delta), \tag{3.20}$$

where $C(\delta)$ is non-increasing in δ. We refer to Appendix Section B.4 for the form of the $C(\delta)$, which additionally depends on d and the distribution of Y.

Remark 3. *For* $\delta_0 = \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right]$*, we have* $C(\delta) = 0$*. This reflects the fact that if the target reconstruction error* δ *exceeds the variance of* Y*, no active dimensions* $g_i(Y)$ *are required: the decoder can simply be taken as the constant map* $f(z) \equiv \mathbb{E}[Y]$*.*

In solving the conditional VAE as a proxy for PLiCCA where the constraint $\Sigma_{g(Y)} \succeq aI_d$ is relaxed, a concern is that the regression term $\mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right]$ could drive g to the inactive solution 0. Theorem 3.4.1 says that if β_{VAE} is chosen sufficiently large, corresponding to a small δ, then this collapse can be avoided.

3.4.3 Conditional normalizing flow

The model introduced here corresponds to the two-stage population problem in equation (3.6). Specifically, for this problem we propose solving a "proxy" conditional normalizing flow problem. We assume an a priori dimension-reduced representation of Y, $W \in \mathbb{R}^d$, such that $Y = \psi(W) + \varepsilon$, with an encoder $\phi : \mathbb{R}^q \to \mathbb{R}^d$, trained, for instance, using an unsupervised VAE. Introducing a latent variable $Z \in \mathbb{R}^d$, a standard derivation (see Appendix Section B.5) shows that if we choose to specify the model as

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{3.21}$$

then maximizing the likelihood is equivalent to minimizing

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

where we have added the hyper-parameter β_{NF}, and $J_{\tilde{g}}(w)$ denotes the Jacobian matrix of a smooth function $\tilde{g}$ evaluated at w.

After estimating $\tilde{g}$ and B using the proxy NF problem, we estimate the remaining quantities of the model following Theorem 3.3 (equations (10)–(12)), as in the conditional VAE formulation. The only difference is that here the final encoder is defined as $g = \tilde{g} \circ \phi$ and the final decoder as $f = \psi \circ \tilde{g}^{-1}$. In the following, we take W, the latent representation of Y, to be an isotropic Gaussian. In fact, this choice is not critical and only serves to simplify the statements of the results.

3.4.4 Connection between PLiCCA and conditional NF

To elucidate the connection between PLiCCA and the conditional NF objective, we begin with the natural regression problem we would be encouraged to solve by Theorem 3.3.3 with $\mathcal{C} = \mathcal{C}_{\mathrm{NF}}$:

$$\underset{\substack{g:\mathbb{R}^d \to \mathbb{R}^d, g \in \mathcal{C}_{\mathrm{NF}} \\ B \in \mathbb{R}^{p \times d},\ \Sigma_{g(Y)} \succeq a I_d}}{\text{minimize}} \quad \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right]. \tag{3.22}$$

Next, we introduce a proxy problem, namely the conditional NF problem in its constrained form:

$$\underset{\substack{\tilde{g} \in \mathcal{C}_{\mathrm{NF}}, B \in \mathbb{R}^{p \times d} \\ \mathbb{E}[\ln|\det(J_{\tilde{g}}(W))|] \geq b}}{\text{minimize}} \quad \mathbb{E}\left[\|\tilde{g}(W) - B^\top X\|_2^2\right], \tag{3.23}$$

for some constant $b \in \mathbb{R}$. The NF problem we employ in practice corresponds to the Lagrangian form of (3.23).

The objective in both problems is identical. The constraint $\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b$ does not immediately appear related to the regression problem (3.22), but in fact, analogous to the reconstruction term of the conditional VAE, it is closely related to the constraint that $\Sigma_{g(Y)} \succeq a I_d$. We make this formal with the following theoretical results.

Theorem 3.4.2. *Fix $a > 0$. If $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, then we have*

$$\Sigma_{\tilde{g}(W)} \succeq a I_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a), \tag{3.24}$$

where $b(a) \equiv \frac{d}{2}\ln(a) - C$ where C is a constant that depends on the bilipschitz constants of $\mathcal{C}_{\mathrm{NF}}$, as well as the Hessian matrices of the coordinates $\tilde{g}_i$ of $\tilde{g}$. See Supplementary Section B.5 for the full expression of C.

Using Theorem 3.4.2, we can view the NF problem (3.23) as a relaxation of the regression problem (3.9). Moreover, since by Theorem 3.3.3 the regression and PLiCCA problems are equivalent, we have the following corollary.

Corollary 3.4.1. *Fix $a > 0$. The PLiCCA problem,*

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d\\ g\in a^{-1/2}\mathcal{C}_{\mathrm{NF}}}}{\text{maximize}} \sum_{i=1}^{d}\mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{3.25}$$

can be relaxed into an alternate NF problem via Theorem 3.4.2,

$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\mathrm{NF}},B\in\mathbb{R}^{p\times d}\\ \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right]\geq b(a)}}{\text{minimize}} \mathbb{E}\left[\left\|\tilde{g}(W)-B^\top X\right\|_2^2\right]. \tag{3.26}$$

Remark 4. *In the definition of $\mathcal{C}_{\mathrm{NF}}$, we could codify an upper bound on a matrix norm of the Hessian of the coordinates of $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$ as an explicit constraint, but we elect not to for simplicity of the presentation.*

In the upcoming Theorem 3.4.3, we establish that a "geometric" variant of the PLiCCA objective, which maximizes a geometric rather than an arithmetic mean of the canonical correlations, can be viewed as a relaxation of the NF problem. We first present the following auxiliary lemma, followed by the main theorem.

Lemma 3.4.1. *If $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$, then we have*

$$\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b \implies \det\left(\Sigma_{\tilde{g}(W)}\right) \geq c(b). \tag{3.27}$$

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

Theorem 3.4.3. *The NF problem,*

$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\mathrm{NF}},B\in\mathbb{R}^{p\times d}\\ \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right]\geq b}}{\text{minimize}} \mathbb{E}\left[\left\|\tilde{g}(W)-B^\top X\right\|_2^2\right], \tag{3.28}$$

can be relaxed into a 'geometric' PLiCCA problem via Lemma 3.4.1,

$$\sup_{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^\top X}=I_d\\ \tilde{g}\in c(b)^{-1/2}\mathcal{C}_{\mathrm{NF}}}} \left(\prod_{i=1}^{d} h\left(\rho_i\right)\right)^{1/d}, \tag{3.29}$$

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.

Remark 5. *We can perform a sanity check on the previous results by applying Theorem 3.4.2 and Lemma 3.4.1 in sequence. We have that* $\det\left(\Sigma_{\tilde{g}(W)}\right) \geq c(b(a)) = a^d e^{-2C}$. *Since* $\Sigma_{\tilde{g}}(W) \succeq aI_d$ *implies* $\det\left(\Sigma_{\tilde{g}}(W)\right) \geq a^d$, *we see that the constant* C *constitutes the only source of non-tightness in these inequalities.*

Together, Corollary 3.4.1 and Theorem 3.4.3 show that, from an optimization perspective, the NF problem is sandwiched between two variants of the PLiCCA problem: the standard PLiCCA problem and the geometric PLiCCA problem. This highlights the intermediate position of the NF problem, making its use as a proxy for PLiCCA well-grounded.

3.5 Demonstration on real data

We use 1000 subjects from the Human Connectome Project (HCP) (Van Essen et al., 2013) to demonstrate the utility of PLiCCA. For each subject i, cortical thickness derived from MRI is represented by a 32K-dimensional vector y_i, corresponding to measurements at 32K locations on the cortical surface (see Glasser et al., 2013 for preprocessing details). For each subject, we also have auxiliary variables, such as clinical and demographic features, represented by a 150-dimensional vector x_i. To incorporate spatial information, we expand the cortical thickness data on the basis of the Laplacian of a template cortical surface and retain 1000 coefficients. We apply the VAE variant of PLiCCA, with a two-hidden-layer perceptron encoder and decoder and a latent dimension of $d = 128$. Results are shown in Figure 3.1. We compare with a standard unsupervised VAE of the same architecture. Both methods achieve satisfactory reconstruction, but PLiCCA constructs latent representations that are linearly associated with the auxiliary clinical variables, providing more interpretable embeddings.

3.6 Discussion

In this work, we introduced PLiCCA, a correlation analysis approach to the supervised disentanglement problem, that may also be of independent interest for characterizing the

dependence structure between two data views. We provided theoretical results for our population model and showed a nontrivial connection to conditional latent variable models, enabling an efficient computation of its solution. Finally, we demonstrated the utility of our approach, and of supervised disentanglement more broadly, on real data from the Human Connectome Project.

The proposed approach lays the foundation for several theoretical and methodological extensions, such as establishing connections with flow matching conditional generative models and leveraging these models for supervised disentanglement. On the application side, an important direction is scaling the approach to larger biomedical imaging datasets.

BIBLIOGRAPHY

Adel, T., Ghahramani, Z., & Weller, A. (2018). Discovering interpretable representations for both deep generative and discriminative models. *International Conference on Machine Learning*, 50–59.

Aguila, A. L., & Altmann, A. (2024). A tutorial on multi-view autoencoders using the multi-view-ae library. *CoRR*.

Ahidar-Coutrix, A., Le Gouic, T., & Paris, Q. (2020). Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, *177*(1), 323–368.

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Alberti, G. S., Hertrich, J., Santacesaria, M., & Sciutto, S. (2024). Manifold learning by mixture models of vaes for inverse problems. *Journal of Machine Learning Research*, *25*(202), 1–35.

An, S., & Jeon, J.-J. (2024). Customization of latent space in semi-supervised variational autoencoder. *Pattern Recognition Letters*, *177*, 54–60.

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. *International conference on machine learning*, 1247–1255.

Balabin, N., Voronkova, D., Trofimov, I., Burnaev, E., & Barannikov, S. (2024). Disentanglement learning via topology. *International Conference on Machine Learning*, 2474–2504.

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, *112*, 172–178.

Basile, L., Acevedo, S., Bortolussi, L., Anselmi, F., & Rodriguez, A. (2025). Intrinsic dimension correlation: Uncovering nonlinear connections in multimodal representations. *The Thirteenth International Conference on Learning Representations*.

Bhatia, R. (2013). *Matrix Analysis* (Vol. 169). Springer Science & Business Media.

Bhattacharya, R., & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Annals of Statistics*, *31*(1), 1–29.

Bodin, E., Stere, A. I., Margineantu, D. D., Ek, C. H., & Moss, H. (2025). Linear combinations of latents in generative models: Subspaces and beyond. *The Thirteenth International Conference on Learning Representations*.

Bonheme, L., & Grzes, M. (2023). Be more active! understanding the differences between mean and sampled representations of variational autoencoders. *Journal of Machine Learning Research*, *24*(324), 1–30.

Boucheron, S., Lugosi, G., & Massart, P. (2013). A non asymptotic theory of independence.

Boulesteix, A.-L., & Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, *8*(1), 32–44.

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, *80*(391), 580–598.

Bright, P., Edelman, A., & Johnson, S. G. (2025). Matrix calculus (for machine learning and beyond). *arXiv preprint arXiv:2501.14787*.

Buenfil, J., & Lila, E. (2024, April). Asymmetric canonical correlation analysis of Riemannian and high-dimensional data.

Bykhovskaya, A., & Gorin, V. (2023). High-dimensional canonical correlation analysis. *arXiv preprint arXiv:2306.16393*.

Cai, T. T., Ren, Z., & Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, *10*(1), 1–59.

Cape, J., Tang, M., & Priebe, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, *47*(5), 2405–2439.

Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, *45*(1), 11–22.

Carmichael, I. (2020). Learning sparsity and block diagonal structure in multi-view mixture models.

Chadebec, C., & Allassonnière, S. (2022). A geometric perspective on variational autoencoders. *Advances in neural information processing systems*, *35*, 19618–19630.

Chapman, J., & Wang, H.-T. (2021). Cca-zoo: A collection of regularized, deep learning based, kernel, and probabilistic cca methods in a scikit-learn style framework. *Journal of Open Source Software*, *6*(68), 3823.

Charkaborty, A., & Panaretos, V. M. (2022). Testing for the rank of a covariance operator. *The Annals of Statistics*, *50*(6), 3510–3537.

Chen, M., Gao, C., Ren, Z., & Zhou, H. H. (2013). Sparse CCA via precision adjusted iterative thresholding.

Cheng, G., Ho, J., Salehian, H., & Vemuri, B. C. (2016). Recursive computation of the Fréchet mean on non-positively curved riemannian manifolds with applications. In *Riemannian computing in computer vision* (pp. 21–43). Springer.

Cho, M. H., Kurtek, S., & Bharath, K. (2022). Tangent functional canonical correlation analysis for densities and shapes, with applications to multimodal imaging data. *Journal of Multivariate Analysis*, *189*, 104870.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Cupidon, J., Eubank, R., Gilliam, D., & Ruymgaart, F. (2008). Some properties of canonical correlations and variates in infinite dimensions. *Journal of Multivariate Analysis*, *99*(6), 1083–1104.

Dai, X., & Müller, H. G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics*, *46*(6B), 3334–3361.

Dai, X., Müller, H. G., & Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, *104*(3), 545–560.

Ding, Z., Xu, Y., Xu, W., Parmar, G., Yang, Y., Welling, M., & Tu, Z. (2020). Guided variational autoencoder for disentanglement learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7920–7929.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real nvp. *International Conference on Learning Representations*.

Dryden, I. L., Koloydenko, A., & Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, *3*(3), 1102–1123.

Dubey, P., & Müller, H. G. (2020). Functional models for time-varying random objects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *82*(2), 275–327.

Dubey, P., & Müller, H. G. (2021). Modeling time-varying random objects and dynamic networks. *Journal of the American Statistical Association*.

Feng, Q., Jiang, M., Hannig, J., & Marron, J. S. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, *166*, 241–265.

Fletcher, P. T., & Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, *87*(2), 250–262.

Friedlander, T., & Wolf, L. (2023). Dynamically-scaled deep canonical correlation analysis. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 232–244.

Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gao, C., Ma, Z., & Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Annals of Statistics*, *45*(5), 2074–2101.

Gaynanova, I. (2020). Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *Bernoulli*, *26*(1), 286–322.

Gaynanova, I., Booth, J. G., & Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the p ≥ N setting. *Journal of the American Statistical Association*, *111*(514), 696–706.

Ghodrati, L., & Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, *109*(4), 957–974.

Ghosal, A., Meiring, W., & Petersen, A. (2023). Fréchet single index models for object response regression. *Electronic Journal of Statistics*, *17*(1).

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, *150*, 113556.

Gundersen, G., Dumitrascu, B., Ash, J. T., & Engelhardt, B. E. (2019). End-to-end training of deep probabilistic cca on paired biomedical observations. *Uncertainty in artificial intelligence*.

Guo, C., & Wu, D. (2019). Canonical correlation analysis (cca) based multi-view learning: An overview. *arXiv preprint arXiv:1907.01693*.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *Ieee Access*, *7*, 63373–63394.

Hannan, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, *2*(2), 229–242.

Happ, C., & Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, *113*(522), 649–659.

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, *33*, 9841–9850.

Harvey, W., Naderiparizi, S., & Wood, F. (2022). Conditional image generation by conditioning variational auto-encoders. *International Conference on Learning Representations*.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC press.

He, G., Müller, H.-G., Wang, J.-L., & Yang, W. (2010). Functional linear regression via canonical analysis. *Bernoulli*, *16*(3).

He, J., Pan, F., Zhuang, F., & He, Q. (2020). Cca-flow: Deep multi-view subspace learning with inverse autoregressive flow. *Asian Conference on Machine Learning*, 177–192.

Hörmander, L. (2015). *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Springer.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3-4), 321–377.

Hsing, T., & Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, With An Introduction to Linear Operators* (Vol. 997). John Wiley & Sons.

Huang, D., & Tropp, J. (2021). From poincaré inequalities to nonlinear matrix concentration. *Bernoulli*, *23*(3).

Huang, Q., & Renaut, R. (2015). Functional partial canonical correlation. *Bernoulli*, *21*(2).

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *The Twelfth International Conference on Learning Representations*.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., Handwerker,

D. A., Keilholz, S., Kiviniemi, V., Leopold, D. A., de Pasquale, F., Sporns, O., Walter, M., & Chang, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage*, *80*, 360–378.

Hyvärinen, A., Khemakhem, I., & Morioka, H. (2023). Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, *4*(10).

Inecik, K., Kara, A., Rose, A., Haniffa, M., & Theis, F. J. (2025). Tardis: Achieving robust and structured disentanglement of multiple covariates. *International Conference on Research in Computational Molecular Biology*, 285–289.

Jahanian, A., Chai, L., & Isola, P. (2020). On the" steerability" of generative adversarial networks. *International Conference on Learning Representations*.

Jirak, M., & Wahl, M. (2020). Perturbation bounds for eigenspaces under a relative gap condition. *Proceedings of the American Mathematical Society*, *148*(2), 479–494.

Kadelburg, Z., Dukic, D., Lukic, M., & Matic, I. (2005). Inequalities of karamata, schur and muirhead, and some applications. *The Teaching of Mathematics*, *8*(1), 31–45.

Karami, M., & Schuurmans, D. (2021). Deep probabilistic canonical correlation analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, 8055–8063.

Kendall, W. S., & Le, H. (2011). Limit theorems for empirical fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics*, *25*(3), 323–352.

Kereta, Ž., & Klock, T. (2021). Estimating covariance and precision matrices along subspaces. *Electronic Journal of Statistics*, *15*(1).

Kessler, D., & Levina, E. (2023). Computational inference for directions in canonical correlation analysis.

Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. *International conference on artificial intelligence and statistics*, 2207–2217.

Kim, H. J., Adluru, N., Bendlin, B. B., Johnson, S. C., Vemuri, B. C., & Singh, V. (2014). Canonical correlation analysis on Riemannian manifolds and its applications. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 251–267, Vol. 8690). Springer International Publishing.

Kokoszka, P., & Reimherr, M. (2017, September). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Krukowski, M. (2018). Natural proof of the characterization of relatively compact families in L^p- spaces on locally compact groups. *arXiv preprint arXiv:1801.01898*.

Krzyśko, M., & Waszak, Ł. (2013). Canonical correlation analysis for functional data. *Biometrical Letters*, *50*(2), 95–105.

Lancaster, H. O. (1958). The structure of bivariate distributions. *The Annals of Mathematical Statistics*, *29*(3), 719–736.

Lee, J. M. (2018). *Introduction to Riemannian Manifolds* (Vol. 2). Springer.

Lee, J. M. (2012). *Smooth Manifolds*. Springer New York.

Li, Y., Dey, S. S., & Xie, W. (2024). On sparse canonical correlation analysis. *Advances in Neural Information Processing Systems*, *37*, 10707–10734.

Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M. R., & Yeo, B. T. (2019). Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nature Communications*, *10*(1).

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, *14*(1), 245.

Lin, Y.-C., Baete, S. H., Wang, X., & Boada, F. E. (2020). Mapping brain–behavior networks using functional and structural connectome fingerprinting in the HCP dataset. *Brain and Behavior*, *10*(6), e01647.

Lin, Z., & Yao, F. (2019). Intrinsic Riemannian functional data analysis. *The Annals of Statistics*, *47*(6).

Liu, X., Sanchez, P., Thermos, S., O'Neil, A. Q., & Tsaftaris, S. A. (2022a). Learning disentangled representations in the imaging domain. *Medical Image Analysis*, *80*, 102516.

Liu, Z., Whitaker, K. J., Smith, S. M., & Nichols, T. E. (2022b). Improved interpretability of brain-behavior CCA with domain-driven dimension reduction. *Frontiers in Neuroscience*, *16*, 851827.

Liu, Z., Schulz, J., Taheri, M., Styner, M., Damon, J., Pizer, S., & Marron, J. S. (2021, September). Non-Euclidean analysis of joint variations in multi-object shapes.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. *international conference on machine learning*, 4114–4124.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. *International conference on machine learning*, 6348–6359.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, *7*(1), 523–542.

Lu, T., Marathe, A., & Martin, A. (2024). Supervising variational autoencoder latent representations with language. *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, 267–278.

Lyu, Q., Fu, X., Wang, W., & Lu, S. (2022). Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. *International Conference on Learning Representations*.

Marron, J., & Dryden, I. L. (2021). *Object Oriented Data Analysis* (1st ed.). Chapman and Hall/CRC.

Masarotto, V., Panaretos, V. M., & Zemel, Y. (2019). Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya: The Indian Journal of Statistics*, *81A*, 172–213.

McKeague, I. W., & Zhang, X. (2022). Significance testing for canonical correlation analysis in high dimensions. *Biometrika*, *109*(4), 1067–1083.

Mehta, R., & Harchaoui, Z. (2025). A generalization theory for zero-shot prediction. *arXiv preprint arXiv:2507.09128*.

Meilă, M., & Zhang, H. (2024). Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, *11*(1), 393–417.

Meo, C., Mahon, L., Goyal, A., & Dauwels, J. (2024). α-tcvae: On the relationship between disentanglement and diversity. *arXiv preprint arXiv:2411.00588*.

Michaeli, T., Wang, W., & Livescu, K. (2016, February). Nonparametric Canonical Correlation Analysis.

Murden, R. J., Zhang, Z., Guo, Y., & Risk, B. B. (2022). Interpretive JIVE: Connections with CCA and an application to brain connectivity. *Frontiers in Neuroscience*, *16*.

Murray, K., Kinnison, J., Nguyen, T. Q., Scheirer, W., & Chiang, D. (2019). Auto-sizing the transformer network: Improving speed, efficiency, and performance for low-resource machine translation. *Proceedings of the Third Workshop on Neural Generation and Translation*.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2019). Hybrid models with deep and invertible features. *International Conference on Machine Learning*, 4723–4732.

Nitzan, Y., Gal, R., Brenner, O., & Cohen-Or, D. (2022). Large: Latent-based regression through gan semantics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19239–19249.

Olkin, I., & Marshall, A. W. (2014). *Inequalities: Theory of majorization and its applications* (Vol. 143). Academic Press.

Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, *30*.

Parikh, N. (2014). Proximal Algorithms. *Foundations and Trends® in Optimization*, *1*(3), 127–239.

Pati, A., & Lerch, A. (2021). Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Computing and Applications*, *33*(9), 4429–4444.

Pennec, X. (2017). Hessian of the Riemannian squared distance. *Preprint*.

Pennec, X., Sommer, S., & Fletcher, T. (2019). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press.

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*, 116604.

Petersen, A., & Müller, H. G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, *47*(2), 691–719.

Petersen, A., Zhang, C., & Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, *21*, 159–178.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, *7*(15), 510.

Pigoli, D., Aston, J. A., Dryden, I. L., & Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, *101*(2), 409–422.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, *72*(4), 665–678.

Qiu, L., Chinchilli, V. M., & Lin, L. (2022). Variational interpretable deep canonical correlation analysis. *ICLR2022 Machine Learning for Drug Discovery*.

Ramsay, J. O., & Silverman, B. W. (2015). *Functional Data Analysis*. Springer-Verlag.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1–85.

Rudin, W. (1976). Principles of mathematical analysis. *3rd ed.*

Schötz, C. (2019). Convergence rates for the generalized Fréchet mean via the quadruple inequality.

Senellart, A., Chadebec, C., & Allassonnière, S. (2023). Improving multimodal joint variational autoencoders through normalizing flows and correlation analysis. *arXiv preprint arXiv:2305.11832*.

Serre, D. (2010). *Matrices: Theory and Applications*. Springer.

Shao, L., Lin, Z., & Yao, F. (2022). Intrinsic Riemannian functional data analysis for sparse longitudinal observations. *The Annals of Statistics*, *50*(3).

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., & Zhang, T. (2022). Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, *23*(241), 1–55.

Shin, H., & Lee, S. (2015). Canonical correlation analysis for irregularly and sparsely observed functional data. *Journal of Multivariate Analysis*, *134*, 1–18.

Shu, H., Wang, X., & Zhu, H. (2020). D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, *115*(529), 292–306.

Shu, R., Chen, Y., Kumar, A., Ermon, S., & Poole, B. (2019). Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*.

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, *18*(11), 1565–1567.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K. L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., ... Glasser, M. F. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage*, *80*, 144–168.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*.

Song, Y., Keller, A., Sebe, N., & Welling, M. (2023). Latent traversals in generative models as potential flows. *Proceedings of the 40th International Conference on Machine Learning*, 32288–32303.

Stimper, V., Liu, D., Campbell, A., Berenz, V., Ryll, L., Schölkopf, B., & Hernández-Lobato, J. M. (2023). Normflows: A pytorch package for normalizing flows. *The Journal of Open Source Software*, *8*(86), 5361.

Stöcker, A., Steyer, L., & Greven, S. (2023). Functional additive models on manifolds of planar shapes and forms. *Journal of Computational and Graphical Statistics*, *32*(4), 1600–1612.

Trefethen, L. N., & Bau, D. (2022). *Numerical Linear Algebra* (Vol. 181). Siam.

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., & Rousu, J. (2018). A tutorial on canonical correlation methods. *ACM Computing Surveys*, *50*(6), 1–33.

Van, T. P., Nguyen, T. M., Tran, N. N., Nguyen, H. V., Doan, L. B., Dao, H. Q., & Minh, T. T. (2020). Interpreting the latent space of generative adversarial networks using supervised learning. *2020 International Conference on Advanced Computing and Applications (ACOMP)*, 49–54.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: An overview. *Neuroimage*, *80*, 62–79.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science* (Vol. 47). Cambridge university press.

Wang, H.-T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S., & Bzdok, D. (2020). Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, *216*, 116745.

Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation learning. *International conference on machine learning*, 1083–1092.

Wang, W., Yan, X., Lee, H., & Livescu, K. (2016). Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*.

Wang, W., & Zhou, Y. H. (2021a). Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. *Journal of Multivariate Analysis*, *185*.

Wang, W., & Zhou, Y.-H. (2021b). Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. *Journal of Multivariate Analysis*, *185*, 104781.

Wang, X., Chen, H., Tang, S., Wu, Z., & Zhu, W. (2024). Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9677–9696.

Winkler, C., Worrall, D., Hoogeboom, E., & Welling, M. (2019). Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*(3), 515–534.

Xia, C. H., Ma, Z., Ciric, R., Gu, S., Betzel, R. F., Kaczkurkin, A. N., Calkins, M. E., Cook, P. A., García de la Garza, A., Vandekar, S. N., Cui, Z., Moore, T. M., Roalf, D. R., Ruparel, K., Wolf, D. H., Davatzikos, C., Gur, R. C., Gur, R. E., Shinohara, R. T., ... Satterthwaite, T. D. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, *9*(1).

Yang, X., Liu, W., Liu, W., & Tao, D. (2019). A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2349–2368.

Yang, X., Liu, W., Liu, W., & Tao, D. (2021). A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2349–2368.

Yang, Y., & Pan, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *The Annals of Statistics*, *43*(2), 467–500.

Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590.

Yoon, G., Carroll, R. J., & Gaynanova, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, *107*(3), 609–625.

Yu, Y., Wang, T., & Samworth, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, *102*(2), 315–323.

Yuan, D., & Gaynanova, I. (2022). Double-matched matrix decomposition for multi-view data. *Journal of Computational and Graphical Statistics*, *31*(4), 1114–1126.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67.

Zhang, J., Sun, W. W., & Li, L. (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *Journal of the American Statistical Association*, *115*(532), 2022–2036.

Zhang, J., Yu, Y., Tang, S., Wu, J., & Li, W. (2023). Variational autoencoder with cca for audio–visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(3s), 1–21.

Zhang, Y., & Chen, Y.-C. (2023). Linear convergence of the subspace constrained mean shift algorithm: From euclidean to directional data. *Information and Inference: A Journal of the IMA*, *12*(1), 210–311.

Zhao, Q., Adeli, E., Honnorat, N., Leng, T., & Pohl, K. M. (2019). Variational autoencoder for regression: Application to brain aging analysis. *International conference on medical image computing and computer-assisted intervention*, 823–831.

Zhao, Y., Wang, B., Mostofsky, S. H., Caffo, B. S., & Luo, X. (2021). Covariate assisted principal regression for covariance matrix outcomes. *Biostatistics*, *22*(3), 629–645.

Zheng, Y., He, T., Qiu, Y., & Wipf, D. P. (2022). Learning manifold dimensions with conditional variational autoencoders. *Advances in Neural Information Processing Systems*, *35*, 34709–34721.

Zheng, Y., Liu, Y., Yao, J., Hu, Y., & Zhang, K. (2025). Nonparametric factor analysis and beyond. *International Conference on Artificial Intelligence and Statistics*, 424–432.

Zhou, H., Yao, F., & Zhang, H. (2023). Functional linear regression for discretely observed data: From ideal to reality. *Biometrika*, *110*(2), 381–393.

Zhou, Y., & Müller, H.-G. (2022). Network regression with graph laplacians. *Journal of Machine Learning Research*, *23*(320), 1–41.

Zhu, H., Li, T., & Zhao, B. (2023). Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science*, *6*(1), 73–104.

Zhuang, X., Yang, Z., & Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, *41*(13), 3807–3833.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

Appendix A

APPENDIX TO CHAPTER 2

The appendix for this chapter is organized as follows. In Section A.1, we study the empirical performance of the proposed method by means of simulation studies. In Section A.2, we formalize the CCA problem for random elements of Hilbert spaces and prove Theorem 2.2.1. In Section A.3, we present intermediate results and associated proofs, and conclude with the proof of Theorem 2.4.1, our theoretical result on the asymptotic errors made by the canonical direction estimates of Asymmetric Sparse CCA. In Section A.4, we prove Theorem 2.4.2, our theoretical result on the asymptotic errors made by the canonical direction estimates of Asymmetric Sparse-Functional CCA. In Section A.5, we prove Theorem 2.4.3, our theoretical result on the asymptotic errors made by the canonical variable estimates of Asymmetric Sparse-Functional CCA, in the absence of a finite-dimensional correlation structure assumption. In Section A.6, we present several norm and matrix identities that are utilized throughout the appendix. Finally, in Section A.7, we provide additional details on the Intrinsic RFPCA algorithm (Lin & Yao, 2019), which is used in the proposed Algorithm 1. We also present an explicit basis construction for the space of symmetric positive definite matrices equipped with the affine invariant metric, providing a computational speed-up compared to the Gram-Schmidt procedure proposed in Lin and Yao (2019).

A.1 Simulations

We perform numerical experiments to investigate the finite sample performance of the proposed approach. First, we describe the data generation process. Then, we discuss the metrics utilized to evaluate the methods' performance. Lastly, we introduce the alternative approaches for comparison with our method and comment on the results.

A.1.1 Data generation

Recall that $y : \mathcal{T} \to \mathcal{M}$ is a random Riemannian process, $X \in \mathbb{R}^p$ is a high dimensional random vector, and $\mu : \mathcal{T} \to \mathcal{M}$ is a fixed smooth curve on $\mathcal{M}$ modeling the population mean of y. Here, we fix $p = 200$ and choose $\mathcal{M}$ to be the manifold of $m \times m$ SPD matrices, with $m = 3$. We let the time domain of y be $\mathcal{T} = [-1, 1]$. In the following, we aim to generate realizations of (y, X) according to a model that ensures that the K population canonical vectors and canonical functions are prespecified vectors $\{\theta_k\}_{k=1}^K \subset \mathbb{R}^p$ and functions $\{\psi_k\}_{k=1}^K \subset L^2(T\mu)$, respectively, with $K = 2$. We apply our proposed method and alternative approaches to this data to estimate the canonical vectors and functions, and then compare these estimates with the prespecified population quantities.

The procedure to generate the data is as follows. Take a random vector $Y \in \mathbb{R}^d$, a set of vectors $\{\eta_k\}_k \subset \mathbb{R}^d$, with $d = 3$, and an orthonormal basis $\{\phi_j\} \subset L^2(T\mu)$. Moreover, define $\operatorname{Log}_\mu y = \sum_{j=1}^d Y_j \phi_j$ and $y = \operatorname{Exp}_\mu \left(\operatorname{Log}_\mu y\right)$. It follows from Theorem 2.2.1 that if the multivariate data (Y, X) have population canonical vector (η_k, θ_k) then the functional/multivariate data (y, X) will have population canonical pairs (ψ_k, θ_k), with $\psi_k = \sum_{j=1}^d \eta_{kj} \phi_j$. Additionally, we impose a group-sparse structure on the canonical vectors $\{\theta_k\}$. To replicate a realistic setting, we add an extra mode of variation $W\phi_{d+1}$ to $\operatorname{Log}_\mu y$, with W a random variable that is independent of X and Y, and with $\operatorname{Var}(W) = 1/2$. This aims to contaminate the observations without affecting the canonical functions and vectors.

To generate the multivariate data (Y, X) given prespecified canonical pairs $\{(\eta_k, \theta_k)\}$, we use the model introduced in Chen et al. (2013). As is stated in Proposition 2.1 in Chen et al. (2013), this enables us to choose the canonical directions $\{(\eta_k, \theta_k)\}$ and correlations $\{\gamma_k\}$ for $k = 1, \ldots K$ freely, while retaining the flexibility to specify $\Sigma_X \in \mathbb{R}^{p \times p}$ and $\Sigma_Y \in \mathbb{R}^{d \times d}$. We define (Y, X) as

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{pmatrix}\right), \tag{A.1}$$

where $\mathcal{N}$ denotes the multivariate normal distribution and $\Sigma_{YX} = \Sigma_Y \left(\sum_{k=1}^K \gamma_k \eta_k \theta_k^T\right) \Sigma_X$. It

is easy to show that the population canonical vectors of (Y, X) are (η_k, θ_k) with correlations γ_k, for $k = 1, \dots K$. The set of canonical vectors $\{\eta_k\}$ is defined by generating K orthogonal random vectors, which are then normalized to satisfy the constraint $\eta_k^\top \Sigma_Y \eta_j = \delta_{kj}$. Similarly, the canonical vectors $\{\theta_k\}$ are randomly generated and constrained to satisfy the condition $\theta_k^\top \Sigma_X \theta_j = \delta_{kj}$. The group sparsity assumption is enforced by ensuring that only $k_1 = 20$ elements of each canonical vector (the same elements across all vectors) are non-zero. Additionally, the variables X_j corresponding to the non-zero components of θ_j have marginal covariance matrix $\Sigma_{X_S} = \text{diag}(\underbrace{2, \dots 2}_{10}, \underbrace{1, \dots 1}_{10}) \in \mathbb{R}^{k_1 \times k_1}$. The covariance Σ_X is then defined as

$$\Sigma_X = \begin{pmatrix} \Sigma_{X_S} & 0 \\ 0 & I_{p-k_1} \end{pmatrix}. \tag{A.2}$$

The covariance Σ_Y is set to be diagonal with diagonal values being $3, 2, 1$. The true canonical correlations are chosen to be $\gamma_1 = .95$ and $\gamma_2 = .6$.

We let the mean curve μ at each $t \in \mathcal{T}$ be a SPD matrix $\mu(t) \in \mathbb{R}^{m \times m}$. We set $\mu(0)$ by randomly generating its eigenvectors and setting the associated eigenvalues equal to $(1, 2, 3)$. The mean $\mu(t)$ at the other time-points $t \in \mathcal{T}$ is generated by applying a time-variant rotation to the eigenvectors of $\mu(0)$. We choose each principal component ϕ_j to take the form $\phi_j(t) = E(t)P(t)$, for $j = 1, \dots d$, where E is chosen at random from a set of orthogonal basis vectors for $L^2(T\mu)$, and P is chosen at random from a basis of orthogonal polynomials on $[-1, 1]$. This ensures that $\{\phi_j\}$ are orthogonal to one another as elements of $L^2(T\mu)$.

In our experiments, we generate N i.i.d. pairs (Y_i, X_i) from the multivariate CCA model in equation (A.1), for different choices of N. Next, we generate y_i via $\text{Exp}_\mu\left(\sum_{j=1}^d Y_{ij}\phi_j\right)$ and evaluate it at $L = 50$ locations $t_l \in [-1, 1]$, yielding $y_i(t_l)$ for $i = 1, \dots N$ and $l = 1, \dots L$. The observations

$$\left(\{y_i(t_l)\}_l, X_i\right)_i, \qquad i = 1, \dots, N,$$

are used to estimate the canonical vectors and functions and compare different approaches.

A.1.2 Metrics

We use the following metrics to compare the estimated accuracy of the models considered.

A. **Normalized Euclidean error for the canonical vector**

$$\left\| \theta_1 / \left\| \theta_1 \right\|_2 - \hat{\theta}_1 / \left\| \hat{\theta}_1 \right\|_2 \right\|_2$$

This is a natural metric for evaluating the estimation accuracy.

B. **F1-score for the canonical vector**

$$2 \cdot \frac{P \cdot R}{P + R},$$

where $P = \frac{TP}{TP+FP}$ is the precision, $R = \frac{TP}{TP+FN}$ is the recall, TP is the number of true positives, FP the number of false positives, FN the number of false negatives.

C. L^2 **Parallel transport error for the canonical function**

$$\left\| \Gamma_{\hat{\mu},\mu} \hat{\psi}_1 - \psi_1 \right\|_\mu$$

This metric allows us to use the $L^2(T\mu)$ norm to compare the estimates to the true population analog, by parallel transporting $\hat{\psi}_1 \in L^2(T\hat{\mu})$ and defining $\Gamma_{\hat{\mu},\mu}\hat{\psi}_1 \in L^2(T\mu)$.

D. **Tangent Correlation**

Using a large test set $\{\tilde{y}_i, \tilde{x}_i\}$, generated from the same distribution as the training data, we compute the sample correlation as follows:

$$\operatorname{Corr}\left(\left(\langle\!\langle \operatorname{Log}_\mu \tilde{y}_i, \Gamma_{\hat{\mu},\mu}\hat{\psi}_1 \rangle\!\rangle_\mu \right)_i, \left(\tilde{x}_i^\intercal \hat{\theta}_1 \right)_i \right).$$

We refer to this metric as the 'Tangent' correlation as it respects the manifold structure of the data.

E. **Euclidean Correlation**

Using a large test set $\{\tilde{y}_i, \tilde{x}_i\}$, generated from the same distribution as the training data, we compute the sample correlation as follows:

$$\text{Corr}\left(\left(\text{vec}\left(\tilde{y}_i\right)^\top \text{vec}\left(\hat{\psi}_1\right)\right)_i, \left(\tilde{x}_i^\top \hat{\theta}_1\right)_i\right).$$

We refer to this metric as the 'Euclidean' correlation as it ignores the manifold structure of the data.

A.1.3 Approaches for comparison

We compare 4 different approaches, detailed below.

1. **Proposed approach.** We apply Algorithm 1 without modifications. We use cross-validation to choose the regularization parameter λ in the group-lasso regression step as implemented by the `glmnet` package (Friedman et al., 2010).

2. **Sparse PCA-based approach:** Intrinsic RFPCA + sparse PCA + classical CCA. We use Intrinsic RFPCA, as in our approach, to reduce the dimensionality of the functional data. We use sparse PCA, using the `elasticnet` R package (Zou et al., 2006), to reduce the dimensionality of the multivariate data. Then, we use the estimated PCA scores as input for classical multivariate CCA. We provide sparse PCA with the exact number of principal components that are correlated with the functional data, i.e., $k_1 = 20$, and restrict the number of non-zero principal loadings per principal component to be 2.

3. **Sparse CCA-based approach:** Intrinsic RFPCA + sparse CCA. We again use Intrinsic RFPCA to reduce the dimension of the functional data. Next, we use the Penalized Matrix Analysis (PMA) approach to sparse CCA proposed in Witten et al. (2009) to compute canonical pairs between the PC scores from Intrinsic RFPCA and the high dimensional data. The PMA approach to sparse CCA assumes that the covariance matrices of the data are the identity matrices, giving it a slight disadvantage.

We choose the amount of penalization for θ_1 using the suggested permutation-type approach (Witten et al., 2009), and choose the penalization parameter for η_1 to induce virtually no penalization.

4. **Multivariate FPCA-based approach:** Multivariate FPCA + Asymmetric sparse CCA in Algorithm 2. This approach is analogous to the one proposed, except that the Intrinsic RFPCA step is replaced by multivariate FPCA (Happ & Greven, 2018). Therefore, it disregards the SPD manifold structure of the data. Specifically, it transforms each SPD matrix into a vector extracting the lower triangular part of the matrix. Then it applies multivariate FPCA to the resulting vector-valued functions.

We have chosen these alternative approaches in order to dissect specific components of the CCA problem. Specifically, approach 2 isolates the effect of selecting important features and identifying correlated components in two separate stages, and approach 3 isolates the effect of not taking advantage of the group sparsity structure in the canonical vectors and making restrictive assumptions on the covariance of the high-dimensional data. Approach 4 isolates the effect of treating manifold data as if it were Euclidean. Note that Approach 4 is technically solving a different canonical correlation problem than approaches 1-3 as it aims to maximize the Euclidean correlation rather than the tangent space correlation. For this reason, the underlying population canonical vectors and functions differ from those in the proposed model. Therefore, we only use metric E when evaluating the performance of approach 4.

Moreover, depending on the choice of Σ_X, either the PMA sparse CCA approach or the sparse PCA approach is at a disadvantage. Assuming Σ_X to be the identity matrix meets the assumptions of PMA sparse CCA but renders dimension reduction through sparse PCA less effective. Conversely, choosing Σ_X to not be the identity matrix benefits sparse PCA at the expense of the PMA sparse CCA approach.

A.1.4 Results and Discussion

In our experiments, we set $p = 200$ and vary N. For each value of N, we run 15 trials. We provide the Intrinsic RFPCA model with the true rank, indicating the number of functional principal components associated with the variable X, that is, $d = 3$.

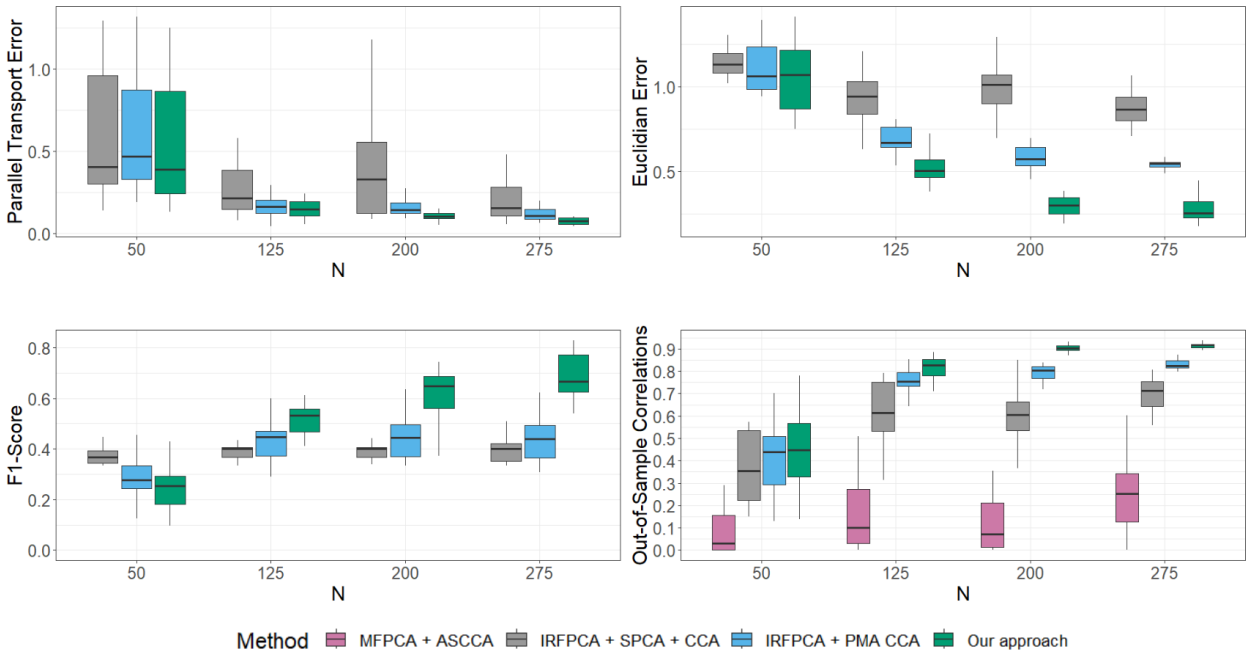


Figure A.1: (Top left): Performance evaluation using metric A, which measures the normalized Euclidean error in the first high-dimensional canonical vector, on approaches 1-3. (Top right): Performance evaluation using metric C, which is the parallel transport error in the first canonical function, on approaches 1-3. (Bottom left): Performance evaluation using metric B, the F1-score of the first estimated high dimensional canonical vector compared to the associated population vector, on approaches 1-3. (Bottom right): Performance evaluation using out-of-sample correlations. We use out-of-sample tangent correlation (metric D) for approaches 1-3, and out-of-sample Euclidean correlation (metric E) for approach 4.

In Figure A.1, we present the performance of approaches 1-3 measured using all defined metrics A-E. As previously mentioned, Approach 4 is assessed using only metric E due to its

differing underlying model. In the high-dimensional setting, where $N = 50$ and $p = 200$, all four approaches showed similar performance across all metrics. This setting likely identifies the detectability limits of CCA methods. However, when provided with more samples, our approach quickly outperforms the other approaches. Differences in performance were more notable in the estimation of the Euclidean error for the canonical vectors, F1-score, and out-of-sample correlation. This suggests that the most challenging aspect of the setting considered is estimating the canonical vectors as opposed to the canonical functions. This can be explained by the similar modeling strategies adopted for the functional data. Approach 4, while able to find correlated components in the data according to the Euclidean notion of correlation (E), it suffered from bias due to treating the functional data as Euclidean.

For approach 2, the differences in performance can be explained by its two-step strategy that involves first selecting the important features and reducing the dimension of the multivariate data, followed by identifying correlated components. Specifically, the sparse PCA step is based solely on the variance structure of X, and not on its correlation with the functional data. In our simulation, the variables of X correlated with Y have the same or smaller variance than those not correlated with Y. As a result, sparse PCA, which is unsupervised, struggles to tease them apart.

A.2 Canonical Correlation Analysis of Random Elements of Hilbert Spaces

In this section, we provide a more rigorous formalization of the CCA model. We mirror the development of Hsing and Eubank (2015) and Huang and Renaut (2015), but provide a less technical presentation by emphasizing the role of the canonical variables rather than the canonical vectors when formulating the general CCA problem. For an introduction to Hilbert space concepts and random elements taking values in Hilbert spaces, we refer to Hsing and Eubank (2015). For an introduction to classical CCA, we refer to Uurtio et al. (2018).

In Section A.2.1, we define the infinite-dimensional version of the CCA problem, and establish the existence of solutions in our asymmetric setting (Theorem A.2.1). In Section

A.2.2, we state preliminary definitions and results for the subsequent sections. In Section A.2.3, we state the necessary assumption (Assumption A.2.1) to reduce the infinite-dimensional CCA problem to a finite-dimensional CCA problem (Theorem A.2.2). In Section A.2.4, we study the difference between the canonical variable solutions of the finite- and infinite-dimensional CCA problems (Theorem A.2.3). In Section A.2.5, we prove the results of the section and additionally prove Theorem 2.2.1.

A.2.1 Problem Statement

Let χ_1 and χ_2 be measurable functions from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to separable Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively (See Section 7.2. of Hsing & Eubank, 2015). Here, $\mathbb{H}_1$ and $\mathbb{H}_2$ are arbitrary Hilbert spaces, but throughout the paper they correspond to $\mathbb{H}_1 = L^2(T\mu)$ and $\mathbb{H}_2 = \mathbb{R}^p$, and similarly χ_1 and χ_2 correspond to $\chi_1 = \mathrm{Log}_\mu y$ and $\chi_2 = X$. Hilbert space inner products are denoted by $\langle \cdot, \cdot \rangle$, with associated norms $\|\cdot\|$. The specific choice of the norm or inner product will be clear from the context. We assume that $\mathbb{E}[\|\chi_i\|^2] < \infty$ so that the mean and covariance of χ_i are well-defined for $i = 1, 2$. The mean element of χ_i is defined as $h_i \equiv \mathbb{E}[\chi_i] \in \mathbb{H}_i$, and for simplicity, we assume $h_i = 0$ for $i = 1, 2$.

A seemingly natural way to formalize the canonical correlation problem for the infinite-dimensional case, which is analogous to the finite-dimensional case, is

$$\underset{f \in \mathbb{H}_1,\, g \in \mathbb{H}_2}{\text{maximize}} \;\; \mathrm{Corr}^2\left(\langle \chi_1, f \rangle, \langle \chi_2, g \rangle\right) \tag{A.3}$$

where Corr is the usual correlation defined between two finite-variance real-valued random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. If they exist, the solution (f, g) would be the first canonical vector pair. Equivalently, we can write this problem in terms of the canonical variables U, V as

$$\underset{U \in \mathcal{U},\, V \in \mathcal{V}}{\text{maximize}} \;\; \mathrm{Corr}^2\left(U, V\right) \tag{A.4}$$

where $\mathcal{U} = \{\langle \chi_1, f \rangle : \; f \in \mathbb{H}_1\}$, $\mathcal{V} = \{\langle \chi_2, g \rangle : \; g \in \mathbb{H}_2\}$. However, the maximum of this problem may not be attained by any $U \in \mathcal{U}$, $V \in \mathcal{V}$ (Cupidon et al., 2008).

It turns out we can amend this by simply taking the closures of $\mathcal{U}$ and $\mathcal{V}$. Let $\mathbb{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ denote the Hilbert space of square-integrable random variables on Ω with inner product $\langle U, V \rangle = \text{Cov}(U, V)$ for $U, V \in \mathbb{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Note that U being square-integrable means that $\text{Var}(U) < \infty$. If we replace $\mathcal{U}$, $\mathcal{V}$ in the problem above with their closures as subsets of $\mathbb{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, denoted $\overline{\mathcal{U}}$, $\overline{\mathcal{V}}$, (for further discussion of $\overline{\mathcal{U}}$, $\overline{\mathcal{V}}$, see the discussion following Example 7.6.5 of Hsing and Eubank (2015)) then it can be shown that the maximum *will* be attained for some $U \in \overline{\mathcal{U}}$, $V \in \overline{\mathcal{V}}$, provided that a certain linear operator is assumed to be compact. In our setting, this compactness condition holds because we use $\mathbb{H}_2 = \mathbb{R}^p$, a finite-dimensional space. The result can be found in Theorem 10.1.2 of Hsing and Eubank (2015), which we restate in our context here.

The following result establishes the existence of solutions to the general CCA problem in our asymmetric setting where $\dim(\mathbb{H}_1) = \infty$ but $\dim(\mathbb{H}_2) = p < \infty$, and that there are at most p nontrivial solutions.

Theorem A.2.1. *If $\dim(\mathbb{H}_1) = \infty$ and $\dim(\mathbb{H}_2) = p < \infty$, then there exists $U_1 \in \overline{\mathcal{U}}$ and $V_1 \in \overline{\mathcal{V}}$ which are solutions to*

$$\sup_{U \in \overline{\mathcal{U}},\, V \in \overline{\mathcal{V}}} \text{Corr}^2(U, V),$$

with $\text{Var}(U_1) = \text{Var}(V_1) = 1$. For $k = 2, \ldots p$, there exists $U_k \in \overline{\mathcal{U}}$ and $V_k \in \overline{\mathcal{V}}$, which are solutions to

$$\sup_{\substack{U \in \overline{\mathcal{U}}: \text{Cov}(U, U_i) = 0, i = 1, \ldots, k-1 \\ V \in \overline{\mathcal{V}}: \text{Cov}(V, V_i) = 0, i = 1, \ldots, k-1}} \text{Corr}^2(U, V),$$

with $\text{Var}(U_k) = \text{Var}(V_k) = 1$. Moreover, for all $U \in \overline{\mathcal{U}}$ and $V \in \overline{\mathcal{V}}$ which are uncorrelated with $U_1, \ldots U_p$, $V_1, \ldots V_p$, respectively, the pair (U, V) is a trivial solution, that is,

$$\sup_{\substack{U \in \overline{\mathcal{U}}: \text{Cov}(U, U_i) = 0, i = 1, \ldots, k \\ V \in \overline{\mathcal{V}}: \text{Cov}(V, V_i) = 0, i = 1, \ldots, k}} \text{Corr}^2(U, V) = 0.$$

The pairs $\{(U_k, V_k)\}_{k=1}^p$ are called the canonical variable pairs. We refer to the problem of finding $\{(U_k, V_k)\}_{k=1}^p$ as the population canonical correlation problem.

Remark 6. *Intuitively, we must cast the problem in terms of the canonical variables* $U \in \overline{\mathcal{U}}$, $V \in \overline{\mathcal{V}}$ *rather than canonical vectors* $f \in \mathbb{H}_1$, $g \in \mathbb{H}_2$ *because* $\mathcal{U} = \{\langle \chi_1, f \rangle : f \in \mathbb{H}_1\}$ *and* $\mathcal{V} = \{\langle \chi_2, g \rangle : g \in \mathbb{H}_2\}$ *are not large enough for the supremum of the CCA problem to be attained. To emphasize this point, given an optimal* $U \in \overline{\mathcal{U}}$ *of the form* $U = \lim_{j\to\infty} \langle \chi_1, f_j \rangle$ *for some sequence* $(f_j)_j \in \mathbb{H}_1$, *then* U *can not necessarily be written as an inner product* $U = \langle f, \chi_1 \rangle$, *because* $(f_j)_j$ *may not even be a Cauchy sequence in* $\mathbb{H}_1$, *and thus not converge in* $\mathbb{H}_1$. *For further intuition on this property of the problem, see Remark 8.*

In the next section, we introduce an assumption that allows us to make the infinite-dimensional CCA problem finite-dimensional, and furthermore formulate the CCA problem in terms of the canonical vectors rather than the canonical variables. From now on, we assume $\dim(\mathbb{H}_2) = p < \infty$ as in Theorem A.2.1.

A.2.2 Background

We begin with preliminary definitions and properties of the random element χ_1. In particular, we introduce the covariance operator $\mathscr{K}_1$. The eigenvectors of $\mathscr{K}_1$, also known as the principal components of χ_1, are fundamental to our approach for two reasons: they provide a data-driven subspace for projecting χ_1 and their properties simplify our proofs.

Given that $\mathbb{E}[\|\chi_1\|^2] < \infty$ and χ_1 has mean 0, the covariance operator of χ_1 is well-defined as $\mathscr{K}_1 \equiv \mathbb{E}[\chi_1 \otimes \chi_1]$. Here, the tensor product $f \otimes g : \mathbb{H} \to \mathbb{H}$ between $f, g \in \mathbb{H}$ for a Hilbert space $\mathbb{H}$, is defined as $(f \otimes g)(h) = \langle f, h \rangle g$ for all $h \in \mathbb{H}$.

In the lemma below, we collect the properties of χ_1 and $\mathscr{K}_1$ used in what follows. An orthonormal sequence of elements $\{e_j\}_{j=1}^{\infty}$ of a Hilbert space $\mathbb{H}$ such that $\overline{\operatorname{span}\{e_j\}} = \mathbb{H}$ is referred to as a complete orthonormal system (CONS) for $\mathbb{H}$. For convenience, we state this result supposing that $\mathscr{K}_1$ has infinitely many eigenvalues, but an analogous result holds if $\mathscr{K}_1$ has only finitely many eigenvalues.

Lemma A.2.1. *Let* $\operatorname{Im}(\mathscr{K}_1)$ *denote the image of* $\mathscr{K}_1$ *and* $\overline{\operatorname{Im}(\mathscr{K}_1)} \subseteq \mathbb{H}_1$ *denote its closure in* $\mathbb{H}_1$. *Then, the following statements hold.*

1. *With probability 1,* $\chi_1 \in \overline{\mathrm{Im}(\mathscr{K}_1)}$*, and for any* $f \in \mathrm{Im}(\mathscr{K}_1)^\perp$*,* $\langle f, \chi_1 \rangle = 0$

2. $\mathscr{K}_1$ *has the eigendecomposition* $\mathscr{K}_1 = \sum_{j=1}^{\infty} \omega_j e_j \otimes e_j$*, where* $e_j \in \mathbb{H}_1$ *for* $j = 1, \ldots \infty$ *and* $\omega_1 \geq \omega_2 \geq \ldots > 0$*.* $\{e_j\}_{j=1}^{\infty}$ *forms a CONS of* $\overline{\mathrm{Im}(\mathscr{K}_1)}$*, and* $\omega_j \to 0$ *as* $j \to \infty$*. We refer to* $\{(e_j, \omega_j)\}_{j=1}^{\infty}$ *as the eigensystem of* $\mathscr{K}_1$*, with eigenvectors* e_j *and eigenvalues* ω_j*. The* e_j *are also referred to as principal components of* χ_1*.*

3. *With probability* 1, $\chi_1 = \sum_{j=1}^{\infty} \langle \chi_1, e_j \rangle e_j$*. We refer to the* $\{\langle \chi_1, e_j \rangle\}_{j=1}^{\infty}$ *as the principal scores, and they are uncorrelated random variables with* $\mathbb{E}[\langle \chi_1, e_j \rangle] = 0$ *and* $\mathrm{Var}(\langle \chi_1, e_j \rangle) = \omega_j$*.*

4. $\overline{\mathcal{U}} = \left\{ \sum_{j=1}^{\infty} a_j \langle \chi_1, e_j \rangle : \sum_{j=1}^{\infty} \omega_j a_j^2 < \infty \right\}$ *and* $\{\langle \chi_1, e_j \rangle / \omega_j^{1/2}\}_{j=1}^{\infty}$ *forms a CONS for* $\overline{\mathcal{U}}$*.*

Remark 7. *Item 1 elucidates the role of* $\overline{\mathrm{Im}(\mathscr{K}_1)}$ *as the subspace of* $\mathbb{H}_1$ *where* χ_1 *resides. Item 2 shows that the eigenvectors* $\{e_j\}$ *are a CONS for* $\overline{\mathrm{Im}(\mathscr{K}_1)}$*, which implies Item 3, that the principal scores* $\{\langle \chi_1, e_j \rangle\}$ *characterize* χ_1*. Item 4 establishes that the set of potential canonical variables,* $\overline{\mathcal{U}}$*, is equivalent to the set of linear combinations of the principal scores with finite variance.*

Remark 8. *Item 4 sheds light on why it is necessary to take the closure of* $\mathcal{U}$ *in order to guarantee a solution to the infinite-dimensional CCA problem in Theorem A.2.1, and further on why the canonical vectors are not defined in the infinite-dimensional CCA problem.*

From Items 1 and 2 it follows that

$$\mathcal{U} \equiv \{\langle f, \chi_1 \rangle : f \in \mathbb{H}_1\} = \left\{ \sum_{j=1}^{\infty} a_j \langle e_j, \chi_1 \rangle : \sum_{j=1}^{\infty} a_j^2 < \infty \right\}. \tag{A.5}$$

For an element $\langle f, \chi_1 \rangle \in \mathcal{U}$ *defined by* $f = \sum_{j=1}^{\infty} a_j e_j \in \mathbb{H}_1$*, since* $\mathrm{Var}(\langle e_j, \chi_1 \rangle) = \omega_j$*, we have that* $\mathrm{Var}(\langle f, \chi_1 \rangle) = \sum_{j=1}^{\infty} a_j^2 \omega_j$*. Since the* ω_j *decay to* 0 *as* j *approaches infinite, there are sequences* $\{a_j\}$ *that do not satisfy* $\sum_{j=1}^{\infty} a_j^2 < \infty$ *but do satisfy* $\sum_{j=1}^{\infty} a_j^2 \omega_j < \infty$*, in other words, there are finite variance linear combinations of the* $\langle e_j, \chi_1 \rangle$ *that are not contained in* $\mathcal{U}$*. These*

are the elements of $\overline{\mathcal{U}}$ *that are missing in* $\mathcal{U}$*:*

$$\mathcal{U} = \left\{ \sum_{j=1}^{\infty} a_j \langle e_j, \chi_1 \rangle : \sum_{j=1}^{\infty} a_j^2 < \infty \right\} \subset \overline{\mathcal{U}} = \left\{ \sum_{j=1}^{\infty} a_j \langle e_j, \chi_1 \rangle : \sum_{j=1}^{\infty} a_j^2 \omega_j < \infty \right\}. \tag{A.6}$$

If the canonical variable solution to the infinite-dimensional CCA problem (A.14) *belongs to* $\mathcal{U}$ *but not* $\overline{\mathcal{U}}$*, then it has no corresponding canonical vector.*

For an arbitrary CONS $\{e_j\}_{j=1}^{\infty}$ for $\overline{\mathrm{Im}\,(\mathscr{K}_1)}$, the associated scores $\{\langle e_j, \chi_1 \rangle\}$ may not be orthogonal in $\overline{\mathcal{U}}$, but as the following result shows, the associated scores still span $\overline{\mathcal{U}}$. This is the property that allows us to not rely on the principal component basis and instead use an arbitrary CONS for $\mathbb{H}_1$ in Assumption A.2.1.

Lemma A.2.2. *For any complete orthonormal system* $\{e_j\}_{j=1}^{\infty}$ *of* $\overline{\mathrm{Im}\,(\mathscr{K}_1)}$, $\overline{\mathcal{U}} = \overline{\mathrm{span}\,\{\langle e_j, \chi_1 \rangle\}}$. *Thus, any element of* $U \in \overline{\mathcal{U}}$ *can be written as* $\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j \rangle$ *where the* a_j *are such that* $\mathrm{Var}\,(U) < \infty$.

A.2.3 Reduction to a finite-dimensional problem

We can now introduce Assumption A.2.1 on the correlation structure between χ_1 and χ_2.

Assumption A.2.1. *There exists a complete orthonormal system* $\{e_j\}_{j=1}^{\infty}$ *for* $\mathbb{H}_1$ *and a set of indices* $I \subset \{1, 2, \ldots\}$*, with finite cardinality* $|I| = d$*, such that*

$$\mathrm{Corr}\,(V, \langle \chi_1, e_j \rangle) = 0, \qquad \forall V \in \overline{\mathcal{V}}, j \in I^c, \tag{A.7}$$

$$\mathrm{Corr}\,(\langle \chi_1, e_i \rangle, \langle \chi_1, e_j \rangle) = 0, \qquad \forall i \in I, \forall j \in I^c, \tag{A.8}$$

where I^c *denotes the complement of* I *in* $\mathbb{N} = \{1, 2, \ldots\}$.

Remark 9. *The complete orthonormal system* $\{e_j\}_j$ *is not required to be the principal component basis. In the case when* $\mathbb{H}_2 = \mathbb{R}^p$ *and* $\chi_2 = X$*, equation* (A.7) *can be rewritten as*

$$\mathrm{Corr}\,(X_i, \langle \chi_1, e_j \rangle) = 0, \qquad \forall i = 1, \ldots p, j \in I^c. \tag{A.9}$$

Intuitively, this assumption states that all elements ψ *of* $\mathbb{H}_1$ *whose projections* $\langle \chi_1, \psi \rangle$ *are correlated with* X *belong to a* d*-dimensional subspace.*

Remark 10. *Assumption A.2.1 is weaker than the assumption that χ_1 admits a finite-dimensional representation $\chi_1 = \sum_{j=1}^{d} \langle \chi_1, e_j \rangle e_j$, for a set of vectors $\{e_j\}_{j=1}^{d} \subset \mathbb{H}_1$. To see this, we first note that the elements $\{e_j\}_{j=1}^{d} \subset \mathbb{H}_1$ are orthonormal. Then, we complete $\{e_j\}_{j=1}^{d} \subset \mathbb{H}_1$ to form a CONS $\{e_j\}_{j=1}^{\infty}$ for $\mathbb{H}_1$, and take $I = \{1, \dots d\}$. Given that the elements $\{e_j\}$ are orthonormal, we have that $\langle \chi_1, e_j \rangle = 0$ for all $j \in I^c$, with probability 1. Hence, conditions (A.7) and (A.8) are satisfied.*

Making this assumption enables us to reduce the infinite-dimensional CCA problem to a finite-dimensional CCA problem; moreover, it allows us to formulate the CCA problem in terms of the population quantities of interest, the canonical vectors, rather than the canonical variables (Theorem A.2.2). Theorem A.2.2 can be viewed as a generalization of Theorem 1 of Krzyśko and Waszak (2013). There, it is assumed that χ_1 has a finite-dimensional representation, whereas here we make the weaker Assumption A.2.1.

Theorem A.2.2. *Reorder the complete orthonormal system $\{e_j\}_{j=1}^{\infty}$ for $\mathbb{H}_1$ in Assumption A.2.1 so that $I = \{1, \dots d\}$. Then, under Assumption A.2.1, the solution to the population canonical correlation problem in Theorem A.2.1 is found for a $U \in \overline{\mathcal{U}}_d = \left\{ \sum_{i=1}^{d} a_i \langle \chi_1, e_i \rangle : a_i \in \mathbb{R} \right\}$.*

Moreover, when $\mathbb{H}_2 = \mathbb{R}^p$ and $\chi_2 = X$, the problem is equivalent to the following multivariate (finite-dimensional) canonical correlation problem, where Y is the d-dimensional random vector such that $Y_j = \langle \chi_1, e_j \rangle$, the jth score associated with e_j, for $j = 1, \dots d$:

$$(a_1, b_1) = \underset{a \in \mathbb{R}^d, b \in \mathbb{R}^p, \mathrm{Var}(a^\top Y) = \mathrm{Var}(b^\top X) = 1}{\arg\max} \mathrm{Corr}^2 \left(a^\top Y, b^\top X \right), \tag{A.10}$$

$$(a_k, b_k) = \underset{\substack{a \in \mathbb{R}^d, b \in \mathbb{R}^p, \mathrm{Var}(a^\top Y) = \mathrm{Var}(b^\top X) = 1 \\ \mathrm{Cov}(a^\top Y, a_i^\top Y) = 0, i = 1, \dots, k \\ \mathrm{Cov}(b^\top X, b_i^\top X) = 0, i = 1, \dots, k}}{\arg\max} \mathrm{Corr}^2 \left(a^\top Y, b^\top X \right), \qquad k = 2, \dots min(p, d). \tag{A.11}$$

We call the pair $\left(\sum_{j=1}^{d} a_{kj} e_j, b_k \right)$ the kth canonical pair, since $a^\top Y = \langle \sum_{j=1}^{d} a_{kj} e_j, \chi_1 \rangle$, and $b^\top X = \langle b, \chi_2 \rangle$, where a_{kj} is the jth entry of a_k, for $k = 1, \dots \min(p, d)$.

This result is central to the proof of Theorem 2.2.1.

A.2.4 Error between finite- and infinite-dimensional problems

In this section, we analyze the error between the finite- and infinite-dimensional CCA problems. In particular, we quantify the error between the canonical variables obtained from solving these problems. We continue to write as (U_k, V_k) the canonical variable solutions to the infinite-dimensional CCA problem in Theorem A.2.1, while we denote as $(U_k^{(d)}, V_k^{(d)})$ the canonical variable solutions to the finite-dimensional CCA problem in Theorem A.2.2 where we have used a d-dimensional Assumption A.2.1. Thus, by definition, $U_k^{(d)} = \eta_k^\top Y$ and $V_k^{(d)} = \theta_k^\top X$.

The relationship between these problems is well-understood, and we refer to Chapter 10 of Hsing and Eubank (2015) for further background. We now introduce the notation and machinery necessary to write down the error between U_k and $U_k^{(d)}$, and V_k and $V_k^{(d)}$, provided in Theorem A.2.3. This result is used to show Theorem A.5.1.

For the remainder of this section, we suppose that the complete orthonormal system $\{e_j\}_{j=1}^\infty$ for $\mathbb{H}_1$ in Assumption A.2.1 are the principal components of χ_1 (extended from a CONS for $\overline{\mathrm{Im}\,(\mathscr{K}_1)}$).

A.2.4.1 Background

The idea central to deriving our bound is that there are spaces $\mathbb{G}_1 \subseteq \mathbb{H}_1$ and $\mathbb{G}_2 \subseteq \mathbb{H}_2$ that are congruent to $\overline{\mathcal{U}}$ and $\overline{\mathcal{V}}$, respectively, so that the infinite-dimensional CCA problem in Theorem A.2.1 can be written over $\mathbb{G}_1$ and $\mathbb{G}_2$ instead of $\overline{\mathcal{U}}$ and $\overline{\mathcal{V}}$, and subsequently in terms of an operator, $\mathscr{C}_{12}$, between $\mathbb{G}_1$ and $\mathbb{G}_2$. The infinite-dimensional CCA solution is then derived from the singular vector decomposition of $\mathscr{C}_{12}$, while the solution to the finite-dimensional CCA problem is derived from the singular vector decomposition of a principal component-approximation of $\mathscr{C}_{12}$, $\mathscr{C}_{12}^{(d)}$. The error between $\mathscr{C}_{12}$ and $\mathscr{C}_{12}^{(d)}$ quantifies the error between the canonical variable solutions for the two problems.

Define

$$\mathbb{G}_1 \equiv \mathrm{Im}\left(\mathscr{K}_1^{1/2}\right) = \left\{\sum_{j=1}^\infty \omega_j^{1/2} a_j e_j : \sum_{j=1}^\infty a_j^2 < \infty\right\}, \tag{A.12}$$

where (ω_j, e_j) are the eigenvector-eigenvalue pairs of $\mathscr{K}_1$. We define a new inner product on $\mathbb{G}_1 \subseteq \mathbb{H}_1$, which is not the one inherited from $\mathbb{H}_1$. For $f, g \in \mathbb{G}_1$ with representations $f = \sum_{j=1}^{\infty} \omega_j^{1/2} a_j e_j$ and $g = \sum_{j=1}^{\infty} \omega_j^{1/2} b_j e_j$, we define $\langle f, g \rangle_{\mathbb{G}_1} = \sum_{j=1}^{\infty} a_j b_j$. It is straightforward to verify that $\mathbb{G}_1$ with this inner product is a Hilbert space with a CONS $\{\tilde{e}_{1j}\}_{j=1}^{\infty}$ where $\tilde{e}_{1j} \equiv \omega_j^{1/2} e_j$.

We have the following congruency between $\overline{\mathcal{U}}$ and $\mathbb{G}_1$. We say that two Hilbert spaces, $\mathbb{H}_1$ and $\mathbb{H}_2$ with norms $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, are congruent if there exists a bijective function $\pi : \mathbb{H}_1 \to \mathbb{H}_2$ such that $\|f - g\|_1 = \|\pi(f) - \pi(g)\|_2$ for all $f, g \in \mathbb{H}_1$.

Lemma A.2.3. *$\mathbb{G}_1$ is congruent to $\overline{\mathcal{U}}$, with $\pi : \mathbb{G}_1 \to \overline{\mathcal{U}}$ defined as the map which takes $f \in \mathbb{G}_1$ with representation $f = \sum_{j=1}^{\infty} a_j \tilde{e}_{1j}$ to its image $\pi(f) \in \overline{\mathcal{U}}$:*

$$\pi(f) = \sum_{j=1}^{\infty} a_j \frac{\langle \chi_1, e_j \rangle}{\omega_j^{1/2}}. \tag{A.13}$$

The proof of this result follows directly from Theorem 2.4.16 of Hsing and Eubank (2015), recalling that $\left\{ \langle \chi_1, e_j \rangle / \omega_j^{1/2} \right\}_{j=1}^{\infty}$ form a CONS for $\overline{\mathcal{U}}$ by Item 4 of Lemma A.2.1.

We define $\mathbb{G}_2$ and $\pi_2 : \mathbb{G}_2 \to \overline{\mathcal{V}}$ analogously based on the covariance operator for χ_2 for clarity of our presentation, but note that since $\mathbb{H}_2$ is assumed to be finite-dimensional, $\mathbb{G}_2$ is also finite-dimensional, with $\dim(\mathbb{G}_2) = p$. We similarly define a CONS $\{\tilde{e}_{2j}\}_{j=1}^{p}$ for $\mathbb{G}_2$.

We can now write the problem over U, V, which reads as

$$\underset{\substack{U \in \overline{\mathcal{U}},\, V \in \overline{\mathcal{V}}, \\ \mathrm{Var}(U) = \mathrm{Var}(V) = 1}}{\text{maximize}} \quad \mathrm{Corr}^2(U, V), \tag{A.14}$$

in terms of elements of $\mathbb{G}_1$ and $\mathbb{G}_2$ instead:

$$\underset{\substack{f \in \mathbb{G}_1,\, g \in \mathbb{G}_2, \\ \mathrm{Var}(\pi_1(f)) = \mathrm{Var}(\pi_2(g)) = 1}}{\text{maximize}} \quad \mathrm{Corr}^2(\pi_1(f), \pi_2(g)). \tag{A.15}$$

We define an operator $\mathscr{C}_{12}$ between $\mathbb{G}_2$ and $\mathbb{G}_1$ such that the following property holds. For any $f \in \mathbb{G}_1$, $g \in \mathbb{G}_2$, we have

$$\mathrm{Cov}(\pi_1(f), \pi_2(g)) = \langle f, \mathscr{C}_{12} g \rangle_{\mathbb{G}_1}. \tag{A.16}$$

This operator exists, is bounded, and has $\|\mathscr{C}_{12}\| \leq 1$ by Theorem 10.1.1 of Hsing and Eubank (2015). We denote by $\mathscr{C}_{21} : \mathbb{G}_1 \to \mathbb{G}_2$ the adjoint of $\mathscr{C}_{12}$, which satisfies $\mathrm{Cov}\left(\pi_1(f), \pi_2(g)\right) = \langle \mathscr{C}_{21} f, g \rangle_{\mathbb{G}_2}$.

In our setting where $\mathbb{G}_2$ is finite-dimensional, $\mathscr{C}_{12}$ is immediately compact, so that we can solve (A.15) in terms of the singular vector decomposition of $\mathscr{C}_{12}$. In fact, taking the singular vector decomposition of $\mathscr{C}_{12}$ and applying equation (A.16) comprise the proof of Theorem A.2.1. We note that if $\mathbb{G}_2$ were infinite-dimensional, we would need to assume compactness of $\mathscr{C}_{12}$.

A.2.4.2 Equivalence of finite-dimensional canonical variables

Before stating our error bound, we need one more result on an equivalent definition of the canonical variables $U_k^{(d)}$ and $V_k^{(d)}$. We additionally introduce a principal component approximation of $\mathscr{C}_{12}$.

Since we have CONS $\{\tilde{e}_{1i}\}_{i=1}^{\infty}$ and $\{\tilde{e}_{2j}\}_{j=1}^{p}$ for $\mathbb{G}_1$ and $\mathbb{G}_2$, respectively, we can write $\mathscr{C}_{12}$ as

$$\mathscr{C}_{12} = \sum_{j=1}^{p} \sum_{i=1}^{\infty} \langle \tilde{e}_{2j}, \mathscr{C}_{12} \tilde{e}_{1i} \rangle_{\mathbb{G}_1}. \tag{A.17}$$

We then define its finite-dimensional approximation by the first d principal components of $\mathscr{K}_1$ as

$$\mathscr{C}_{12}^{(d)} = \sum_{j=1}^{p} \sum_{i=1}^{d} \langle \tilde{e}_{2j}, \mathscr{C}_{12} \tilde{e}_{1i} \rangle_{\mathbb{G}_1}. \tag{A.18}$$

We define $\mathscr{C}_{21}^{(d)}$ in an analogous way.

Lemma A.2.4. *Recall* $\left(U_k^{(d)}, V_k^{(d)}\right)$*, the canonical variables obtained by solving the finite-dimensional CCA problem in Theorem A.2.2 obtained by making a d-dimensional Assumption A.2.1:*

$$\left(U_k^{(d)}, V_k^{(d)}\right) = \underset{\substack{U \in \overline{\mathcal{U}}_d, V \in \overline{\mathcal{V}}, \\ \mathrm{Var}(U) = \mathrm{Var}(V) = 1}}{\arg\max} \mathrm{Corr}^2\left(U, V\right). \tag{A.19}$$

Let (f_k, g_k) *be the kth pair of right and left singular vectors of* $\mathscr{C}_{12}^{(d)}$*, respectively. Let*

$\left(\tilde{U}_k^{(d)}, \tilde{V}_k^{(d)}\right)$ *be defined by*

$$\tilde{U}_k^{(d)} = \pi_1\left(f_k\right), \tag{A.20}$$

$$\tilde{V}_k^{(d)} = \pi_2\left(g_k\right). \tag{A.21}$$

Then, $U_k^{(d)} = \tilde{U}_k^{(d)}$ *and* $\tilde{V}_k^{(d)} = V_k^{(d)}$.

Remark 11. *In this paper, we have defined* $\left(U_k^{(d)}, V_k^{(d)}\right)$ *directly as the solution to a finite-dimensional CCA problem. In Hsing and Eubank (2015), finite-dimensional canonical variables are studied and defined as* $\left(\tilde{U}_k^{(d)}, \tilde{V}_k^{(d)}\right)$. *This result shows that these two definitions are equivalent.*

A.2.4.3 Error bound

We can finally state the following result about the error in making a d-dimensional Assumption A.2.1. The proof follows directly from Theorem 5.2.2 and the proof of equation (10.19) in Theorem 10.2.3 in Hsing and Eubank (2015), the triangle inequality, and Lemma A.2.4.

Theorem A.2.3. *Let* (U_k, V_k) *be the infinite-dimensional canonical variables obtained by solving the problem in Theorem A.2.1. Let* $\left(U_k^{(d)}, V_k^{(d)}\right)$ *be the canonical variables obtained by solving the finite-dimensional CCA problem in Theorem A.2.2 obtained by making a d-dimensional Assumption A.2.1 with the principal components of* χ_1. *Then for* $k \leq d$, *as* $d \to \infty$, *we have*

$$\max\left\{\mathbb{E}\left[\left(U_k - U_k^{(d)}\right)^2\right], \mathbb{E}\left[\left(V_k - V_k^{(d)}\right)^2\right]\right\} \lesssim \frac{\gamma_1^{*2}\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2} - \gamma_j^{*2}\right|}, \tag{A.22}$$

where γ_i^* *denote the infinite-dimensional canonical correlations, i.e. the singular values of* $\mathscr{C}_{12}$.

This result is used in the proof of Theorem A.5.1.

A.2.5 Proofs

Proof of Lemma A.2.1:

The first item is part 3 of Theorem 7.2.5 of Hsing and Eubank (2015). The second and third items are Theorem 7.2.6 and Theorem 7.2.7 of Hsing and Eubank (2015), respectively. For the fourth item, equation (A.5) gives

$$\mathcal{U} \equiv \{\langle f, \chi_1\rangle : f \in \mathbb{H}_1\} = \left\{\sum_{j=1}^{\infty} a_j \langle e_j, \chi_1\rangle : \sum_{j=1}^{\infty} a_j^2 < \infty\right\}. \tag{A.23}$$

From this, it is clear that $\mathcal{U} \subseteq \left\{\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j\rangle : \sum_{j=1}^{\infty} \omega_j a_j^2 < \infty\right\}$, because $\sum_{j=1}^{\infty} a_j^2$ implies $\sum_{j=1}^{\infty} \omega_j a_j^2$. We also have that

$$\overline{\operatorname{span}\left(\{\langle \chi_1, e_j\rangle\}_{j=1}^{\infty}\right)} = \left\{\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j\rangle : \sum_{j=1}^{\infty} \omega_j a_j^2 < \infty\right\}, \tag{A.24}$$

from the definition of the norm in $\mathbb{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and because we can calculate the norm squared as $\operatorname{Var}\left(\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j\rangle\right) = \sum_{j=1}^{\infty} \omega_j a_j^2$ by the continuity of the inner product. In particular, this set is closed, so $\mathcal{U} \subseteq \left\{\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j\rangle : \sum_{j=1}^{\infty} \omega_j a_j^2 < \infty\right\}$ implies $\overline{\mathcal{U}} \subseteq \left\{\sum_{j=1}^{\infty} a_j \langle \chi_1, e_j\rangle : \sum_{j=1}^{\infty} \omega_j a_j^2 < \infty\right\}$. To show the reverse inclusion, it suffices to show that $\operatorname{span}\left(\{\langle \chi_1, e_j\rangle\}_{j=1}^{\infty}\right) \subseteq \mathcal{U}$, which is clear from equation (A.5).

That $\{\langle \chi_1, e_j\rangle / \omega_j^{1/2}\}_{j=1}^{\infty}$ forms a CONS for $\overline{\mathcal{U}}$ follows from $\overline{\mathcal{U}} = \overline{\operatorname{span}\left(\{\langle \chi_1, e_j\rangle\}_{j=1}^{\infty}\right)}$ as was just shown, and normalizing the already orthogonal $\langle \chi_1, e_j\rangle$ to have norm 1. □

Proof of Lemma A.2.2:

We begin by employing the fourth item of Lemma A.2.1, which states that $\{\langle \chi_1, e_j\rangle / \omega_j^{1/2}\}$ is a CONS for $\overline{\mathcal{U}}$, where the e_j are the eigenvectors of $\mathscr{K}_1$, and ω_j are the corresponding eigenvalues. Therefore, given an arbitrary CONS for $\overline{\operatorname{Im}(\mathscr{K}_1)}$, $\{f_j\}_{j=1,\ldots\infty}$, to complete the proof it suffices to show that $\overline{\operatorname{span}\{\langle f_j, \chi_1\rangle\}} = \overline{\operatorname{span}\{\langle e_j, \chi_1\rangle\}}$.

To show the $\subseteq$ direction, it suffices to show that $\langle f_k, \chi_1\rangle \in \overline{\operatorname{span}\{\langle e_j, \chi_1\rangle\}}$, for every k, by the definitions of closure and span of a set of vectors. Since the functions $\{e_j\}$ form a CONS for $\overline{\operatorname{Im}(\mathscr{K}_1)}$, there exists a sequence $(a_j)_{j=1}^{\infty}$ of scalars such that $f_k = \sum_{j=1}^{\infty} a_j e_j$.

Therefore, $\langle f_k, \chi_1 \rangle = \sum_{j=1}^{\infty} a_j \langle \chi_1, e_j \rangle$ by continuity of the inner product on $\mathbb{H}_1$, and we have $\langle f_k, \chi_1 \rangle \in \overline{\text{span}\{\langle e_j, \chi_1 \rangle\}}$.

To show the $\supseteq$ direction, we must show that $\langle e_k, \chi_1 \rangle \in \overline{\text{span}\{\langle f_j, \chi_1 \rangle\}}$ for every k, which follows by similar arguments. □

Proof of Theorem A.2.2:

We prove the statement for the first canonical pair; the proof for the remaining canonical pairs follows from a similar argument. Let (U, V) be the first canonical pair of the population canonical correlation problem in Theorem A.2.1. We consider the CONS $\{e_j\}_{j=1}^{\infty}$ for $\mathbb{H}_1$ from Assumption A.2.1, and reorder its elements so that $I = \{1, \dots d\}$. By Lemma A.2.2, we write $U \in \overline{\mathcal{U}}$ as $U = \sum_{j=1}^{\infty} a_j \langle \chi_1, e_j \rangle$. We will show that under Assumption A.2.1, we can find a $Q = \sum_{j=1}^{\infty} q_j \langle \chi_1, e_j \rangle \in \overline{\mathcal{U}}$, with $q_j = 0$ for $j > d$, that attains the same maximum value as U. Thus we will have $Q \in \overline{\mathcal{U}}_d \equiv \left\{ \sum_{i=1}^{d} a_i \langle \chi_1, e_i \rangle : a_i \in \mathbb{R} \right\}$, completing the proof of the first statement of the Theorem. For random variables with variance 1, such as U and V, we have that $\text{Cov}(U, V) = \text{Corr}(U, V)$. We use these interchangeably throughout the proof.

If the optimum value of $\text{Corr}^2(U, V)$ is 0, then we can select any Q with $q_j = 0$ for $j > d$ and $\text{Var}(Q) = 1$. Therefore, from now on, we focus on the case where $\text{Corr}^2(U, V) \neq 0$. Let

$$U = \sum_{j=1}^{d} a_j \langle \chi_1, e_j \rangle + \sum_{j=d+1}^{\infty} a_j \langle \chi_1, e_j \rangle \equiv W + Z. \tag{A.25}$$

Then, from continuity of the inner product and Assumption A.2.1, it follows that $\text{Cov}(Z, V) = 0$ and $\text{Cov}(W, Z) = 0$, from conditions (A.7) and (A.8) respectively. Before constructing Q, we note that the variance of W must be less than or equal to 1. To see this, we use

$$1 = \text{Var}(U) = \text{Var}(W + Z) = \text{Var}(W) + 2\,\text{Cov}(W, Z) + \text{Var}(Z) = \text{Var}(W) + \text{Var}(Z), \tag{A.26}$$

since $\text{Cov}(W, Z) = 0$. Then, $\text{Var}(W) \leq 1$ since both $\text{Var}(W)$ and $\text{Var}(Z)$ are positive and sum to 1.

Now, we construct a canonical variable Q with the desired property. The optimal value

of the CCA population problem in Theorem A.2.1 under assumption A.2.1 is

$$\begin{aligned}\mathrm{Corr}^2(U,V) &= \mathrm{Cov}^2(W+Z,V) && \text{(A.27)}\\ &= (\mathrm{Cov}(W,V)+\mathrm{Cov}(Z,V))^2 && \text{(A.28)}\\ &= \mathrm{Cov}^2(W,V) && \text{(A.29)}\\ &= \mathrm{Corr}^2(W,V)\,\mathrm{Var}(W). && \text{(A.30)}\end{aligned}$$

Having established $\mathrm{Var}(W) \le 1$, there are three cases, either $\mathrm{Var}(W) = 0$, $0 < \mathrm{Var}(W) < 1$, or $\mathrm{Var}(W) = 1$. In the case $\mathrm{Var}(W) = 1$, we take $Q = W$, and using equation (A.30), we see that the pair (W,V) attains the same maximum correlation as (U,V). This completes the proof as W is of the desired form. Now, we will show that the other two cases, $\mathrm{Var}(W) = 0$, $0 < \mathrm{Var}(W) < 1$, are not possible. $\mathrm{Var}(W) = 0$ cannot hold since, by equation (A.30), we would have $\mathrm{Cov}(U,V) = 0$, which we have already ruled out. Assume towards a contradiction that $0 < \mathrm{Var}(W) < 1$, let $c = \frac{1}{\mathrm{Var}(W)^{1/2}} > 1$, and define $Q = cW$. Then, we have that $\mathrm{Var}(Q) = c^2\,\mathrm{Var}(W) = 1$, and

$$\mathrm{Corr}^2(U,V) = \mathrm{Corr}^2(W,V)\,\mathrm{Var}(W) < \mathrm{Corr}^2(Q,V), \tag{A.31}$$

by equation (A.30), $\mathrm{Corr}^2(W,V) = \mathrm{Corr}^2(Q,V)$, and $\mathrm{Var}(W) < 1$. However, this is a contradiction as it would imply that the pair (Q,V) attains a larger value of the objective than (U,V). This completes the proof of the first statement.

Having established the existence of a solution of the stated form for Q, that we are able to reformulate the CCA problem in terms of the finite-dimensional vectors $\{a_k\}_k$ rather than $U \in \overline{\mathcal{U}}$ follows from the definition of $\overline{\mathcal{U}}_d$ and the bilinearity of the inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{H}_1$. In the case that $\mathbb{H}_2 = \mathbb{R}^p$ and $\chi_2 = X$, that we are able to reformulate the problem in terms of the finite-dimensional vectors $\{b_k\}_k$ rather than $V \in \overline{\mathcal{V}}$ is due to the following argument. We have $\mathcal{V} \equiv \{\langle\chi_2, g\rangle : g \in \mathbb{H}_2\} = \mathrm{span}\{\langle\chi_2, e_j\rangle, j = 1,\dots p\}$ (where the e_j here are the standard unit vectors for $\mathbb{R}^p$) is isomorphic to $\mathbb{R}^p$, which is complete. Thus, $\{\langle\chi_2, g\rangle : g \in \mathbb{H}_2\}$ is complete, so its completion in $\mathbb{L}^2(\Omega,\mathcal{F},\mathbb{P})$ is itself, i.e. $\overline{\mathcal{V}} = \mathcal{V}$. Therefore, $\overline{\mathcal{V}} = \mathrm{span}\{\langle\chi_2, e_j\rangle, j = 1,\dots p\} = \{g^\top X : g \in \mathbb{R}^p\}$, i.e. the set of linear combinations of $X_1,\dots X_p$.

The number of nontrivial canonical variables has changed from p in Theorem A.2.1 to $\min(p, d)$. This is because, in a finite-dimensional CCA problem concerning random vectors of dimensions p and d, the smaller of the two dimensions is the upper limit for the number of nontrivial canonical variables (Uurtio et al., 2018). This completes the proof. □

Proof of Lemma A.2.4:

Using the congruency of $\overline{\mathcal{U}}$ with $\mathbb{G}_1$ and $\overline{\mathcal{V}}$ with $\mathbb{G}_2$, the optimization problem in (A.19) can be rewritten as

$$\underset{\substack{f\in\text{span}(\tilde{e}_{11},\dots\tilde{e}_{1d}),g\in\mathbb{G}_2,\\ \text{Var}(\pi_1(f))=\text{Var}(\pi_2(g))=1}}{\text{maximize}} \text{Corr}^2\left(\pi_1(f),\pi_2(g)\right). \tag{A.32}$$

Using the definition of $\mathscr{C}_{12}$, we have

$$\text{Corr}^2\left(\pi_1(f),\pi_2(g)\right) = \langle \mathscr{C}_{21} f, g\rangle^2_{\mathbb{G}_2}. \tag{A.33}$$

Because $f \in \text{span}\left(\tilde{e}_{11},\dots\tilde{e}_{1d}\right)$, we obtain $\langle \mathscr{C}_{21} f, g\rangle_{\mathbb{G}_2} = \langle \mathscr{C}_{21}^{(d)} f, g\rangle_{\mathbb{G}_2} = \langle f, \mathscr{C}_{12}^{(d)} g\rangle_{\mathbb{G}_1}$. Therefore, the maximum of this problem is determined by the singular vectors of $\mathscr{C}_{12}^{(d)}$, and we see that $U_k^{(d)} = \pi_1\left(f_k\right)$ and $V_k^{(d)} = \pi_2\left(g_k\right)$. □

A.2.6 Proof of Theorem 2.2.1

Given that Assumption 2.2.1 is equivalent to Assumption A.2.1, by applying Theorem A.2.2, we readily derive the first part of the theorem. This establishes that there are at most $d^{(\text{corr})}$ nontrivial canonical variable pairs (U_k, V_k). Moreover, each pair (U_k, V_k) can be written in terms of the canonical directions $U_k = \langle\!\langle \text{Log}_\mu y, \psi_k\rangle\!\rangle_\mu$ and $V_k = X^\top\theta_k$, for some $\psi_k \in L^2(T\mu)$ and $\theta_k \in \mathbb{R}^p$. Additionally, $(\psi_k, \theta_k) = \left(\sum_{j=1}^d a_{kj}\phi_j, b_k\right)$, where the pairs (a_k, b_k) are defined in Theorem A.2.2 as the solution to a multivariate CCA problem, and the functions $\{\phi_j\}$ form the CONS for $L^2(T\mu)$, defined in Section 2.2.3.

It remains to be shown that the solutions (a_k, b_k) to the multivariate CCA problem can be characterized by the equations (2.4)-(2.7). We focus on the finite-dimensional optimization

problem, in equation (A.10), that defines the first canonical pair. This is equivalent to

$$\sup_{a_1^\top \Sigma_X a_1 = 1 = b_1^\top \Sigma_Y b_1} a_1^\top \Sigma_{XY} b_1.$$

Now using the assumption that Σ_X and Σ_Y are invertible, we make a change of variables $\tilde{a}_1 = \Sigma_X^{1/2} a_1$, $\tilde{b}_1 = \Sigma_Y^{1/2} b_1$ and obtain the equivalent problem

$$\sup_{\tilde{a}_1^\top \tilde{a}_1 = 1 = \tilde{b}_1^\top \tilde{b}_1} \tilde{a}_1^\top \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} \tilde{b}_1$$

Let $U\Gamma V^\top = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}$ be a singular value decomposition of $\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}$, where $U \in \mathbb{R}^{p \times d^{(\text{corr})}}$, $\Gamma \in \mathbb{R}^{d^{(\text{corr})} \times d^{(\text{corr})}}$, $V \in \mathbb{R}^{d^{(\text{corr})} \times d^{(\text{corr})}}$, $U^\top U = I_{d^{(\text{corr})}} = V^\top V$, and where Γ is a diagonal matrix with the diagonal elements $\gamma_1, \dots \gamma_{d^{(\text{corr})}}$, in descending order. Note that $p \geq d^{(\text{corr})}$. Then it follows from standard properties of the SVD that the first columns of U and V, denoted as u_1 and v_1 respectively, are the solutions to the above problem, i.e. $(\tilde{a}_1, \tilde{b}_1) = (u_1, v_1)$. Similarly, it can be shown that the kth columns of U and V, u_k and v_k respectively, are such that $(\tilde{a}_k, \tilde{b}_k) = (u_k, v_k)$, and that the optimal correlations are the singular values $\gamma_1, \dots \gamma_{d^{(\text{corr})}}$. Undoing the change of variables, it can be seen that the solutions to the original problems in equation (A.10) are the pairs formed by the kth columns of the matrices $\Sigma_X^{-1/2}U$ and $\Sigma_Y^{-1/2}V$. The associated squared correlations are the diagonal entries of Γ^2.

Now let B be the solution to the optimization problem

$$\underset{B \in \mathbb{R}^{p \times d^{(\text{corr})}}}{\text{minimize}} \ \mathbb{E}\left[\|\Sigma_Y^{-1/2} Y - B^\top X\|_2^2 \right]. \tag{A.34}$$

It is straightforward to show that $B = \Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2}$. Therefore, we have

$$\Sigma_X^{1/2} B = U\Gamma V^\top, \tag{A.35}$$

$$B^\top \Sigma_X B = V\Gamma^2 V^\top \tag{A.36}$$

and

$$B V \Gamma^{-1} = \Sigma_X^{-1/2} U. \tag{A.37}$$

Identifying $\tilde{H}$, H, and T in equations (2.4)-(2.7) with V, $\Sigma_Y^{-1/2}V$, and $\Sigma_X^{-1/2}U$, respectively, completes the proof.

A.3 Asymmetric Sparse CCA: Proof of Theorem 2.4.1

A.3.1 Notation

For a vector $x \in \mathbb{R}^p$ with entries $\{x_j\}$ we define its infinity norm $\|x\|_\infty = \max_j(|x_j|)$, its Euclidean norm $\|x\|_2 = \sqrt{\sum_{j=1}^p x_j^2}$, and its ℓ_1 norm $\|x\|_1 = \sum_{j=1}^p |x_j|$. For a matrix $A \in \mathbb{R}^{p\times d}$ with singular values $\sigma_1, \dots \sigma_d$, its operator norm is $\|A\|_2 = \max_i(|\sigma_i|)$. To denote the ith row of the matrix A, we use A_i, and for the entry in the ith row and jth column, we use a_{ij}. We define the matrix norms $\|A\|_F = \left(\sum_{i=1}^p \sum_{j=1}^d a_{ij}^2\right)^{1/2}$, $\|A\|_{\ell_1,\ell_2} = \sum_{i=1}^p \|A_i\|_2$, and $\|A\|_{\max} = \max_{(i,j)} |a_{i,j}|$.

Given the normed spaces $(\mathbb{R}^d, \|\cdot\|_\alpha)$ and $(\mathbb{R}^p, \|\cdot\|_\beta)$, and a matrix $A \in \mathbb{R}^{p\times d}$, we define the matrix norm induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ as

$$\|A\|_{\alpha,\beta} = \sup_{\|x\|_\alpha = 1} \|Ax\|_\beta. \tag{A.38}$$

For additional properties of the matrix norms used throughout the paper, we refer to Section A.6. We use the notation $x \lesssim y$ for $x, y \in \mathbb{R}$ to indicate that $x \le Cy$, with C some positive absolute constant.

A.3.2 Sub-Gaussian random vectors

Now we briefly define sub-Gaussian random vectors and state basic properties that we use in the proofs. We refer the reader to Vershynin (2018) for a more comprehensive introduction to sub-Gaussian random variables and vectors.

A random variable X is sub-Gaussian if, for some constant $C > 0$, it satisfies

$$\mathbb{P}\{|X| \ge t\} \le 2\exp\left(-t^2/C\right) \quad \text{for all } t \ge 0. \tag{A.39}$$

The sub-Gaussian norm of X is defined as

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp\left(X^2/t^2\right) \le 2\right\}. \tag{A.40}$$

A random vector $X \in \mathbb{R}^p$ is called sub-Gaussian if $\langle X, x \rangle$ is sub-Gaussian for all $x \in \mathbb{R}^p$. The sub-Gaussian norm of X is defined as

$$\|X\|_{\psi_2} = \sup_{x \in \mathbb{R}^p : \|x\|_2 = 1} \|\langle X, x \rangle\|_{\psi_2}. \tag{A.41}$$

From its definition, it is clear that $\|X_i\|_{\psi_2} \leq \|X\|_{\psi_2}$, where X_i is the ith element of X. To simplify our analysis, we will also assume that sub-Gaussian vectors X satisfy the variance-proxy condition defined below.

Definition A.3.1. *A sub-Gaussian random vector X satisfies the variance-proxy condition if there exists a constant K_X such that for any $x \in \mathbb{R}^p$, $\|\langle X, x \rangle\|_{\psi_2} \leq K_X \operatorname{Var}(\langle X, x \rangle)^{1/2}$.*

Intuitively, this condition implies that the sub-Gaussian norms of the one-dimensional marginals of X can be used as proxies for their standard deviations. Note that the reverse inequality $\operatorname{Var}(\langle X, x \rangle)^{1/2} \leq K \|\langle X, x \rangle\|_{\psi_2}$ for $K = \sqrt{2}$ is always satisfied when X has mean 0 (Proposition 2.5.2. (ii) of Vershynin (2018)). Moreover, for a Gaussian random vector X, this proxy assumption holds with $K_X = 1$. If X is a zero-mean sub-Gaussian random vector that satisfies the variance-proxy condition and has covariance matrix Σ_X, it follows from the definition above that $\|X\|_{\psi_2} \leq K_X \|\Sigma_X\|_2^{1/2}$. Additionally, it is straightforward to show that $\max_i(\|X_i\|_{\psi_2}) \leq K_X \|\Sigma_X\|_{2,\infty}^{1/2}$, where X_i is the ith entry of X. The proxy assumption allows us to compare sub-Gaussian norms of vectors to one another through their variances.

Throughout our proofs, we assume that the variance-proxy condition applies to the random vectors X, $B^\top X$, Y, $\Sigma_Y^{-1/2} Y$, and $\Sigma_Y^{-1} Y$. To simplify our assumptions, for the main theorems in this section, we conveniently assume that X and Y are strict sub-Gaussians, as defined in (Kereta & Klock, 2021):

Definition A.3.2. *A sub-Gaussian random vector X is called strict sub-Gaussian if there exists a constant K_X such that for any matrix $U \in \mathbb{R}^{k \times p}$, the following inequality is satisfied:*

$$\|UX\|_{\psi_2} \leq K_X \|\Sigma_{UX}\|_2^{1/2}. \tag{A.42}$$

A.3.3 Proof of Theorem 2.4.1

Recall that the matrices $\mathbb{X} \in \mathbb{R}^{N\times p}$ and $\mathbb{Y} \in \mathbb{R}^{N\times d}$ consist of N samples of the random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^d$, respectively. We assume that Σ_X and Σ_Y are invertible, and without loss of generality, we assume that X and Y have mean 0.

To estimate Σ_X and Σ_Y, we use their respective sample covariance estimates $\hat{\Sigma}_Y = \mathbb{Y}^\top\mathbb{Y}/N$ and $\hat{\Sigma}_X = \mathbb{X}^\top\mathbb{X}/N$. Define $B = \Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2}$, let $\hat{B}$ be the solution to the sample group lasso problem (2.9), and let λ be the associated penalization constant. In the setting of Theorem 2.2.1, if we define $\tilde{H}$ by the eigendecomposition $B^\top\Sigma_X B \equiv \tilde{H}D^2\tilde{H}^\top$, then by letting

$$T = B\tilde{H}D^{-1} \in \mathbb{R}^{p\times d}, \tag{A.43}$$

$$H = \Sigma_Y^{-1/2}\tilde{H} \in \mathbb{R}^{d\times d}, \tag{A.44}$$

it follows that the kth column of H, η_k, is the kth canonical vector associated with Y, and the kth column of T, θ_k, is the kth canonical vector associated with X. Moreover, the diagonal entries of D^2 are the squared population canonical correlations $\gamma_1 > \ldots > \gamma_d$, which we assume are distinct. This allows us to focus on estimating individual canonical vectors rather than subspaces spanned by canonical vectors sharing identical correlations.

We denote the columns of $\tilde{H}$ as $\tilde{\eta}_k$ and denote by $\{\hat{\theta}_k\}$ and $\{\hat{\eta}_k\}$ the estimates of the canonical vectors, and by $\{\hat{\gamma}_k\}$ the estimated canonical correlations, that is, the diagonal entries of $\hat{D}$. Note that by definition, the squared population correlations $\gamma_1^2 \ldots \gamma_d^2$ are the eigenvalues of $B^\top\Sigma_X B$ and the estimated squared correlations $\hat{\gamma}_1^2, \ldots \hat{\gamma}_d^2$ are the eigenvalues of $\hat{B}^T\hat{\Sigma}_X\hat{B}$. In the remainder of this section, we derive bounds on the estimation error for the canonical correlations, quantified by $|\gamma_k^2 - \hat{\gamma}_k^2|$, and the canonical vectors, quantified by $\|\eta_k - \hat{\eta}_k\|_2^2$ and $\left\|\theta_k - \hat{\theta}_k\right\|_2^2$.

A.3.3.1 Deterministic bounds

We begin by presenting our deterministic results. To establish fast-rate bounds, we use the Group restricted eigenvalue condition, analogously to the lasso regression problem (Hastie et al., 2015) and similar to Gaynanova (2020) in the context of penalized optimal scoring.

Definition A.3.3 (Group restricted eigenvalue condition)**.** *A matrix $Q \in \mathbb{R}^{q\times p}$ satisfies the Group restricted eigenvalue condition $RE(s,c,d)$ with parameter κ if for all sets $S \subset \{1,\dots p\}$ with $|S| \le s$, we have that, for all $A \in \mathbb{R}^{p\times d}$ such that $\|A_{\bar{S}}\|_{\ell_1,\ell_2} \le c\,\|A_S\|_{\ell_1,\ell_2}$,*

$$\|QA\|_F \geq \frac{\|A_S\|_F^2}{\kappa}. \tag{A.45}$$

Here, $|S|$ denotes the cardinality of S, and $\bar{S} = \{1,\dots p\}\backslash S$.

The following lemma establishes a deterministic bound for the 2-norm of the difference between the linear operators $B^\top\Sigma_X B$ and $\hat{B}^T\hat{\Sigma}_X\hat{B}$. In turn, this quantity will be used to bound the errors $|\gamma_k^2 - \hat{\gamma}_k^2|$, $\|\eta_k - \hat{\eta}_k\|_2^2$ and $\left\|\theta_k - \hat{\theta}_k\right\|_2^2$.

Lemma A.3.1. *The following inequality holds:*

$$\begin{aligned}\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \le \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\frac{1}{\sqrt{N}}\left\|\mathbb{X}(\hat{B}-B)\right\|_F + \left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2 + \gamma_1\left\|(\hat{B}-B)^\top\Sigma_X^{1/2}\right\|_2 \\ + \frac{1}{N}\left\|\mathbb{X}(\hat{B}-B)\right\|_F^2 + \left\|\hat{B}-B\right\|_{\ell_1,\ell_2}\left\|(\Sigma_X-\hat{\Sigma}_X)B\right\|_{2,\infty}.\end{aligned}$$

In the equation above, the first-order terms appear on the first line while the second-order terms appear on the second line of the equation. In this section, wherever possible, we will keep the convention.

Let $E = \mathbb{Y}\hat{\Sigma}_Y^{-1/2} - \mathbb{X}B$. Next, we derive 'slow'- and 'fast'-rate deterministic bounds.

Lemma A.3.2. *If $\lambda \ge \frac{2}{N}\left\|\mathbb{X}^\top E\right\|_{2,\infty}$, then the following slow-rate bound holds:*

$$\begin{aligned}\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\sqrt{\lambda}\,\|B\|_{\ell_1,\ell_2}^{1/2} + \left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2 \\ + \gamma_1\|B\|_{\ell_1,\ell_2}^{1/2}\left(\lambda\|B\|_{\ell_1,\ell_2} + \left\|\hat{\Sigma}_X-\Sigma_X\right\|_{\max}\right)^{1/2} \\ + \lambda\|B\|_{\ell_1,\ell_2} + \|B\|_{\ell_1,\ell_2}\left\|(\Sigma_X-\hat{\Sigma}_X)B\right\|_{2,\infty}.\end{aligned}$$

If, additionally, B has at most s non-zero rows, and $\frac{1}{\sqrt{N}}\mathbb{X}$ satisfies the Group restricted eigenvalue condition $RE(s,3,d)$ with parameter κ_X, then the following fast-rate bound holds:

$$\begin{aligned}\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\kappa_X^{1/2}s^{1/2}\lambda + \left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2 + \gamma_1\|\Sigma_X\|_2^{1/2}\kappa_X s^{1/2}\lambda \\ + \kappa_X s\lambda^2 + \left\|(\Sigma_X-\hat{\Sigma}_X)B\right\|_{2,\infty}\kappa_X s\lambda.\end{aligned}$$

Note that in the fast-rate bound $\sqrt{\lambda}$ and $\|B\|_{\ell_1,\ell_2}$ are replaced with λ and $\kappa_X s$, respectively. Next, we derive a bound for $\frac{1}{N}\|\mathbb{X}^\top E\|_{2,\infty}$.

Lemma A.3.3. *The following inequality holds:*

$$\frac{1}{N}\|\mathbb{X}^\top E\|_{2,\infty} \le \left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\Sigma_Y^{-1/2}\right\|_{2,\infty} + \left\|\Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2}-\Sigma_Y^{-1/2}\right)\right\|_{2,\infty} + \left\|\left(\Sigma_X-\hat{\Sigma}_X\right)B\right\|_{2,\infty} + \left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\left(\hat{\Sigma}_Y^{-1/2}-\Sigma_Y^{-1/2}\right)\right\|_{2,\infty}.$$

Denote the right-hand side of the equation in Lemma A.3.3 as λ_0. Given that $\frac{2}{N}\|\mathbb{X}^\top E\|_{2,\infty} \le 2\lambda_0$, choosing $\lambda \ge 2\lambda_0$ ensures that $\lambda \ge \frac{2}{N}\|\mathbb{X}^\top E\|_{2,\infty}$. Thus, we can replace the assumption $\lambda \ge \frac{2}{N}\|\mathbb{X}^\top E\|_{2,\infty}$ with the assumption $\lambda \ge 2\lambda_0$. Later, we will establish a high-probability bound for λ_0.

Due to the fact that $\left\|\left(\Sigma_X-\hat{\Sigma}_X\right)B\right\|_{2,\infty} \le \lambda_0$, we obtain the following simplification of Lemma A.3.2, where the fourth and fifth terms are combined.

Lemma A.3.4. *If $\lambda \ge 2\lambda_0$, then the following slow-rate bound holds:*

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B}-B^\top\Sigma_X B\right\|_2 \lesssim \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\sqrt{\lambda}\,\|B\|_{\ell_1,\ell_2}^{1/2} + \left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2 + \gamma_1\|B\|_{\ell_1,\ell_2}^{1/2}\left(\lambda\|B\|_{\ell_1,\ell_2}+\left\|\hat{\Sigma}_X-\Sigma_X\right\|_{\max}\right)^{1/2} + \lambda\|B\|_{\ell_1,\ell_2}.$$

If, additionally, B has at most s non-zero rows, and $\frac{1}{\sqrt{N}}\mathbb{X}$ satisfies the Group restricted eigenvalue condition $RE(s,3,d)$ with parameter κ_X, then the following fast-rate bound holds:

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B}-B^\top\Sigma_X B\right\|_2 \lesssim \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\kappa_X^{1/2}s^{1/2}\lambda + \left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2 + \gamma_1\kappa_X^{1/2}s^{1/2}\left(1+\kappa_X s\left\|\hat{\Sigma}_X-\Sigma_X\right\|_{\max}\right)^{1/2}\lambda + \kappa_X s\lambda^2.$$

Lemma A.3.4 shows that the rate of convergence will ultimately be determined by λ_0, $\left\|B^\top(\Sigma_X-\hat{\Sigma}_X)B\right\|_2$, and $\left\|\hat{\Sigma}_X-\Sigma_X\right\|_{\max}$.

A.3.3.2 Probabilistic bounds

From now on, we assume that X and Y are sub-Gaussian random vectors and that the variance-proxy condition in Definition A.3.1 holds for the random vectors X, $B^\top X$, Y, $\Sigma_Y^{-1/2}Y$, and $\Sigma_Y^{-1}Y$. We will repeatedly use the union bound and omit for simplicity the absolute constants arising from its applications.

First, we present an intermediary result that will be used to derive a probabilistic upper bound for λ_0.

Lemma A.3.5. *Let $X \in \mathbb{R}^p$ and $Z \in \mathbb{R}^d$ be zero-mean random vectors with covariance matrices Σ_X and Σ_Z and cross-covariance matrix Σ_{XZ}. Assume the entries of X and Z are sub-Gaussian random variables with norms $\|X_i\|_{\psi_2} = g_i$ and $\|Z_j\|_{\psi_2} = h_i$, for $i = 1, \dots p$ and $j = 1, \dots d$. Let $g = \max(g_i)$ and $h = \max(h_j)$. Let $\mathbb{X} \in \mathbb{R}^{N\times p}$ and $\mathbb{Z} \in \mathbb{R}^{N\times d}$ be data matrices such that the pairs of rows $\{(\mathbb{X}_i^\top, \mathbb{Z}_i^\top)\}$ are independent samples from the joint distribution (X, Z). If $d \le p$ and $\log(p) = o(N)$, then for any fixed $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$\left\|\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right\|_{2,\infty} \lesssim gh\sqrt{\frac{d}{N}\log(p\eta^{-1})}. \tag{A.46}$$

Remark 12. *In Lemma A.3.5, it is stated that for a fixed η, if $\lim(\log(p)/N) = 0$ as p and N go to infinity, then, eventually, the stated bound holds.*

Next, we derive probabilistic upper bounds for λ_0, bounding the terms in Lemma A.3.3.

Lemma A.3.6. *If $d \le p$ and $\log(p) = o(N)$, then for any fixed $\eta \in (0,1)$, with probability $1-\eta$,*

$$\left\|\left(\hat{\Sigma}_{XY} - \Sigma_{XY}\right)\Sigma_Y^{-1/2}\right\|_{2,\infty} \lesssim \max_i(\|X_i\|_{\psi_2})\sqrt{\frac{d}{N}\log(p\eta^{-1})} \tag{A.47}$$

and

$$\left\|\left(\Sigma_X - \hat{\Sigma}_X\right)B\right\|_{2,\infty} \lesssim \max_i(\|X_i\|_{\psi_2})\gamma_1\sqrt{\frac{d}{N}\log(p\eta^{-1})}. \tag{A.48}$$

Moreover, if $d = o(N)$, then

$$\left\|\Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_{2,\infty} \lesssim \|\Sigma_X\|_{2,\infty}^{1/2}\,\gamma_1\sqrt{\frac{d+\log(\eta^{-1})}{N}} \tag{A.49}$$

and

$$\left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\left(\hat{\Sigma}_{Y}^{-1/2}-\Sigma_{Y}^{-1/2}\right)\right\|_{2,\infty}\lesssim\left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\Sigma_{Y}^{-1/2}\right\|_{2,\infty}. \tag{A.50}$$

Remark 13. *As noted in Section A.3.1,* $\max_i(\|X_i\|_{\psi_2})\lesssim\|\Sigma_X\|_{2,\infty}^{1/2}$.

Using Lemmas A.3.3 and A.3.6, and Remark 13, it straightforward to derive the following result.

Lemma A.3.7. *If* $d\leq p$, $\log(p)=o(N)$, *and* $d=o(N)$, *then for any fixed* $\eta\in(0,1)$, *with probability* $1-\eta$,

$$\lambda_0\lesssim\|\Sigma_X\|_{2,\infty}^{1/2}\sqrt{\frac{d}{N}\log(p\eta^{-1})}. \tag{A.51}$$

Next, we establish bounds on the other terms appearing in A.3.4.

Lemma A.3.8. *If* $\log(p)=o(N)$, *then for any fixed* $\eta\in(0,1)$, *with probability* $1-\eta$,

$$\left\|\Sigma_X-\hat{\Sigma}_X\right\|_{\max}\lesssim\max(\|X_i\|_{\psi_2}^2)\sqrt{\frac{\log\left(p\eta^{-1}\right)}{N}}. \tag{A.52}$$

If $d=o(N)$, *then for any fixed* $\eta\in(0,1)$, *with probability* $1-\eta$,

$$\left\|B^{\intercal}(\Sigma_X-\hat{\Sigma}_X)B\right\|_2\lesssim\gamma_1^2\sqrt{\frac{d+\log\left(\eta^{-1}\right)}{N}}, \tag{A.53}$$

and

$$\frac{1}{\sqrt{N}}\left\|\mathbb{X}B\right\|_2\lesssim\gamma_1. \tag{A.54}$$

Before presenting our final bounds, we establish that the group-restricted eigenvalue condition holds for the design matrix $\frac{1}{\sqrt{N}}\mathbb{X}$, with high probability, assuming that the same condition holds for $\Sigma_X^{1/2}$.

Lemma A.3.9. *Suppose* $\Sigma_X^{1/2}$ *satisfies the group restricted eigenvalue condition* $RE(s,3,d)$ *with parameter* $\kappa=\kappa(s,d,\Sigma_X^{1/2})$. *If* $\max_{i=1,\dots p}(\|X_i\|_{\psi_2}^4)\kappa^2s^2\log(p)=o(N)$ *and* $s^2\log(p)=o(N)$, *then for any fixed* η, *with probability* $1-\eta$, $\frac{1}{\sqrt{N}}\mathbb{X}$ *satisfies the group restricted eigenvalue condition* $RE(s,3,d)$ *with parameter* κ_X, *where*

$$0<\kappa_X\leq 2\kappa. \tag{A.55}$$

Next, we state our probabilistic bound for $\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2$. The proof of the slow-rate bound follows straightforwardly from Lemmas A.3.4, A.3.7 and A.3.8. The proof of the fast-rate bound follows similarly from Lemmas A.3.4, A.3.7 and A.3.8, with the addition of Lemma A.3.9.

Theorem A.3.1. *Assume X and Y are sub-Gaussian random vectors and that X, $B^\top X$, $\Sigma_Y^{-1/2}Y$ satisfy the variance-proxy condition A.3.1. Moreover, assume that $d \le p$, $\log(p) = o(N)$, and $d = o(N)$. Fix $\eta \in (0,1)$, and for some absolute constant $C > 0$, let $\lambda = C\left\|\Sigma_X\right\|_{2,\infty}^{1/2}\sqrt{\frac{d}{N}\log(p\eta^{-1})}$. Then, with probability $1-\eta$, the following slow-rate bound holds:*

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim$$
$$\left(\frac{d}{N}\log(p\eta^{-1})\right)^{1/4}\left[\gamma_1^2 + \left\|\Sigma_X\right\|_{2,\infty}^{1/2}\left\|B\right\|_{\ell_1,\ell_2} + \gamma_1\left\|\Sigma_X\right\|_{2,\infty}^{1/4}\left\|B\right\|_{\ell_1,\ell_2}^{1/2}\left(1 + \left\|B\right\|_{\ell_1,\ell_2}^{1/2} + \left\|\Sigma_X\right\|_{2,\infty}^{1/4}\right)\right].$$

Under the additional assumption that $B \in \mathbb{R}^{p\times d}$ has at most s nonzero rows, $\Sigma_X^{1/2}$ satisfies the group restricted eigenvalue condition $RE(s,3,d)$ with parameter $\kappa = \kappa(s,d,\Sigma_X^{1/2})$, $s^2\log(p) = o(N)$, and $\left\|\Sigma_X\right\|_{2,\infty}^2\kappa^2 s^2\log(p) = o(N)$, then the following fast-rate bound holds:

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2}\left[\gamma_1\left(\gamma_1 + \kappa^{1/2}s^{1/2}\left\|\Sigma_X\right\|_{2,\infty}^{1/2}\right)\right]. \tag{A.56}$$

Corollary A.3.1. *In the setting of Theorem A.3.1, under the additional assumption that $\left\|\Sigma_X\right\|_{2,\infty}, \left\|B\right\|_{\ell_1,\ell_2} \ge 1$, then the expression of the slow-rate bound simplifies as follows:*

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/4}\left[\gamma_1\left\|\Sigma_X\right\|_{2,\infty}^{1/2}\left\|B\right\|_{\ell_1,\ell_2}\right]. \tag{A.57}$$

Under the additional assumption that $\left\|\Sigma_X\right\|_{2,\infty}, \kappa \ge 1$, then the expression of the fast-rate bound simplifies as follows:

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2}\left[\gamma_1\left\|\Sigma_X\right\|_{2,\infty}^{1/2}s^{1/2}\kappa^{1/2}\right]. \tag{A.58}$$

Remark 14. *The eigenvalues of the matrices $\hat{B}^T\hat{\Sigma}_X\hat{B}$ and $B^\top\Sigma_X B$ are $\{\hat{\gamma}_k^2\}$ and $\{\gamma_k^2\}$ respectively. Then by Weyl's inequality (Bhatia (2013) Corollary III.2.6.), because we have bounded the operator norm of the difference between these two matrices, we immediately obtain bounds on the estimation error of γ_k^2 by $\hat{\gamma}_k^2$ in Algorithm 2.*

To establish bounds for $\left\|\theta_k - \hat{\theta}_k\right\|_2$ and $\|\eta_k - \hat{\eta}_k\|_2$, we first introduce a few supporting lemmas.

Lemma A.3.10. *Under the slow-rate bound assumptions stated in Corollary A.3.1, for any fixed $\eta \in (0,1)$, with probability at least $1-\eta$, we have*

$$\left\|B - \hat{B}\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/4} \left\|\Sigma_X^{-1/2}\right\|_2 \|B\|_{\ell_1,\ell_2} \|\Sigma_X\|_{2,\infty}^{1/2}. \tag{A.59}$$

Under the fast-rate bound assumptions stated in Corollary A.3.1, for any fixed $\eta \in (0,1)$, with probability at least $1-\eta$, we have

$$\left\|B - \hat{B}\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2} \kappa s^{1/2} \|\Sigma_X\|_{2,\infty}^{1/2}. \tag{A.60}$$

Lemma A.3.11. *Suppose that $Y \in \mathbb{R}^d$ is a sub-Gaussian vector, $d = o(N)$ and that Y satisfies the variance-proxy condition. Then, for any fixed $\eta \in (0,1)$, with probability $1-\eta$, we have*

$$\left\|\Sigma_Y - \hat{\Sigma}_Y\right\|_2 \lesssim \|\Sigma_Y\|_2 \sqrt{\frac{d\log\left(\eta^{-1}\right)}{N}}. \tag{A.61}$$

Additionally, suppose that $\Sigma_Y^{-1}Y$ satisfies the variance-proxy condition, and $\|\Sigma_Y\|_2^2 d = o(N)$. Then for fixed $\eta \in (0,1)$, with probability $1-\eta$, we have

$$\left\|\Sigma_Y^{-1/2} - \hat{\Sigma}_Y^{-1/2}\right\|_2 \lesssim \left\|\Sigma_Y^{1/2}\right\|_2 \left\|\Sigma_Y^{-1/2}\right\|_2^2 \sqrt{\frac{d\log\left(\eta^{-1}\right)}{N}}. \tag{A.62}$$

Studying the theoretical properties of CCA through the lens of regression, using the matrix B, has been convenient thus far. However, for our final results, we bound $B = \Sigma_X^{-1/2}\tilde{T}D\tilde{H} = TD\tilde{H}$ in terms of quantities that are more directly related to the CCA problem. Using identity 12 in Section A.6, along with the standard properties of the 2-norm, and noting that $\tilde{T}$, $\tilde{H}$ are orthogonal matrices and D is a diagonal matrix with diagonal values no greater than 1, we observe that

$$\|B\|_{\ell_1,\ell_2} \le \|T\|_{\ell_1,\ell_2} \le \left\|\Sigma_X^{-1/2}\right\|_{\ell_1,\ell_2}, \tag{A.63}$$

$$\|B\|_2 \le \|T\|_2 \le \left\|\Sigma_X^{-1/2}\right\|_2. \tag{A.64}$$

Hence, in Corollary A.3.1, we can replace the assumption that $\|B\|_{\ell_1,\ell_2} \geq 1$ with the assumption that $\|T\|_{\ell_1,\ell_2} \geq 1$.

Next, we state our probabilistic bounds on the estimated canonical vectors. We denote with $K = \max\{i \in \{1,\dots d\} : \gamma_i > 0\}$ the number of nontrivial canonical vectors. Moreover, to simplify the notation, we use the conventions $\gamma_{K+1}^2 = -\infty$ and $\gamma_0^2 = \infty$.

Theorem A.3.2. *Under the slow-rate bound assumptions stated in Corollary A.3.1 and assuming that the canonical correlations $\gamma_1,\dots\gamma_K$ are bounded from below, Y and $\Sigma_Y^{-1}Y$ satisfy the variance-proxy condition, and $\hat{\eta}^\top \hat{\Sigma}_Y^{1/2}\Sigma_Y^{1/2}\eta \geq 0$, for $k = 1,\dots K$.*
If $d\log(p)\|\Sigma_X\|_{2,\infty}^2\|T\|_{\ell_1,\ell_2}^4 = o(N)$, then, for any fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/4} \frac{\gamma_1\|\Sigma_X\|_{2,\infty}^{1/2}\|T\|_{\ell_1,\ell_2}\left\|\Sigma_X^{-1/2}\right\|_2}{\gamma_k \min\left(\gamma_{k-1}^2-\gamma_k^2, \gamma_k^2-\gamma_{k+1}^2\right)}, \qquad k = 1,\dots K. \tag{A.65}$$

If $\|\Sigma_Y\|_2^2 d = o(N)$ and $\|\Sigma_Y\|_2^2\|\Sigma_Y^{-1}\|_2^2 d = o(N)$, then, for any fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/4} \frac{\gamma_1\|\Sigma_X\|_{2,\infty}^{1/2}\|T\|_{\ell_1,\ell_2}\left\|\Sigma_X^{-1/2}\right\|_2}{\min\left(\gamma_{k-1}^2-\gamma_k^2, \gamma_k^2-\gamma_{k+1}^2\right)}, \qquad k = 1,\dots K. \tag{A.66}$$

Theorem A.3.3. *Under the fast-rate bound assumptions stated in Corollary A.3.1 and assuming that the canonical correlations $\gamma_1,\dots\gamma_K$ are bounded from below, Y and $\Sigma_Y^{-1}Y$ satisfy the variance-proxy condition, and $\hat{\eta}_k^\top \hat{\Sigma}_Y^{1/2}\Sigma_Y^{1/2}\eta_k \geq 0$, for $k = 1,\dots K$.*
If $d\log(p)\|\Sigma_X\|_{2,\infty} s\kappa = o(N)$, then for any fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2} \frac{\gamma_1\|\Sigma_X\|_{2,\infty}^{1/2}\|T\|_2 s^{1/2}\kappa}{\gamma_k \min\left(\gamma_{k-1}^2-\gamma_k^2, \gamma_k^2-\gamma_{k+1}^2\right)}, \qquad k = 1,\dots K. \tag{A.67}$$

If $\|\Sigma_Y\|_2^2 d = o(N)$, then for any fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2} \left\|\Sigma_Y^{-1/2}\right\|_2 \max\left\{\frac{\gamma_1\|\Sigma_X\|_{2,\infty}^{1/2} s^{1/2}\kappa^{1/2}}{\min\left(\gamma_{k-1}^2-\gamma_k^2, \gamma_k^2-\gamma_{k+1}^2\right)}, \left\|\Sigma_Y^{1/2}\right\|_2\left\|\Sigma_Y^{-1/2}\right\|_2\right\}, \tag{A.68}$$

with $k = 1,\dots K$.

A.3.4 Proofs for the deterministic bounds in Section A.3.3.1 and for the probabilistic bounds in Section A.3.3.2

Proof of Lemma A.3.1:

The triangle inequality is used repeatedly without comment. By adding and subtracting $\hat{B}^\top \Sigma_X B$, we have

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \le \left\|\hat{B}^T(\hat{\Sigma}_X\hat{B} - \Sigma_X B)\right\|_2 + \left\|(\hat{B}-B)^\top\Sigma_X B\right\|_2. \tag{A.69}$$

Since

$$\hat{\Sigma}_X\hat{B} - \Sigma_X B = \hat{\Sigma}_X(\hat{B}-B) + (\hat{\Sigma}_X - \Sigma_X)B \tag{A.70}$$

by adding and subtracting $\hat{\Sigma}_X B$, we deduce that

$$\left\|\hat{B}^T\hat{\Sigma}_X\hat{B} - B^\top\Sigma_X B\right\|_2 \le \underbrace{\left\|\hat{B}^\top\hat{\Sigma}_X(\hat{B}-B)\right\|_2}_{\text{Term I}} + \underbrace{\left\|\hat{B}^\top(\hat{\Sigma}_X-\Sigma_X)B\right\|_2}_{\text{Term II}} + \underbrace{\left\|(\hat{B}-B)^\top\Sigma_X B\right\|_2}_{\text{Term III}}. \tag{A.71}$$

We bound each term individually. Recall that $\hat{\Sigma}_X = \frac{1}{N}\mathbb{X}^\top\mathbb{X}$.

Term I: We have

$$\left\|\hat{B}^\top\hat{\Sigma}_X(\hat{B}-B)\right\|_2 \le \frac{1}{\sqrt{N}}\left\|\mathbb{X}\hat{B}\right\|_2\left\|\mathbb{X}(\hat{B}-B)\right\|_F\frac{1}{\sqrt{N}} \tag{A.72}$$

using $\|AB\|_2 \le \|A\|_2\|B\|_2 \le \|A\|_2\|B\|_F$. Since

$$\left\|\mathbb{X}\hat{B}\right\|_2 \le \|\mathbb{X}B\|_2 + \left\|\mathbb{X}(\hat{B}-B)\right\|_2, \tag{A.73}$$

we have

$$\left\|\hat{B}^\top\hat{\Sigma}_X(\hat{B}-B)\right\|_2 \le \frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2\left\|\mathbb{X}(\hat{B}-B)\right\|_F\frac{1}{\sqrt{N}} + \frac{1}{N}\left\|\mathbb{X}(\hat{B}-B)\right\|_F^2. \tag{A.74}$$

Term II: We have

$$\left\|\hat{B}^\top(\hat{\Sigma}_X-\Sigma_X)B\right\|_2 \le \left\|B^\top(\hat{\Sigma}_X-\Sigma_X)B\right\|_2 + \left\|(\hat{B}-B)^\top(\hat{\Sigma}_X-\Sigma_X)B\right\|_2 \tag{A.75}$$

$$\le \left\|B^\top(\hat{\Sigma}_X-\Sigma_X)B\right\|_2 + \left\|\hat{B}-B\right\|_{\ell_1,\ell_2}\left\|(\hat{\Sigma}_X-\Sigma_X)B\right\|_{2,\infty} \tag{A.76}$$

using $\|A^\top B\|_2 \le \|A\|_{\ell_1,\ell_2} \|B\|_{2,\infty}$.

Term 3: We have

$$\left\|(\hat{B}-B)^\top \Sigma_X B\right\|_2 \le \left\|(\hat{B}-B)^\top \Sigma_X^{1/2}\right\|_2 \left\|\Sigma_X^{1/2} B\right\|_2 \le \gamma_1 \left\|(\hat{B}-B)^\top \Sigma_X^{1/2}\right\|_2 \tag{A.77}$$

since $\Sigma_X^{1/2} B = \tilde{T} D \tilde{H}$ where $\tilde{T}$ and $\tilde{H}$ are orthogonal and D is diagonal.

Combining these results we obtain the statement of the lemma. □

Proof of Lemma A.3.2:

For reference, Lemma A.3.1 gives the bound

$$\left\|\hat{B}^T \hat{\Sigma}_X \hat{B} - B^\top \Sigma_X B\right\|_2 \lesssim \frac{1}{\sqrt{N}} \|\mathbb{X} B\|_2 \sqrt{\lambda} \|B\|_{\ell_1,\ell_2}^{1/2} + \left\|B^\top (\Sigma_X - \hat{\Sigma}_X) B\right\|_2 \tag{A.78}$$

$$+ \gamma_1 \|B\|_{\ell_1,\ell_2}^{1/2} \left(\lambda \|B\|_{\ell_1,\ell_2} + \left\|\hat{\Sigma}_X - \Sigma_X\right\|_{\max}\right)^{1/2} \tag{A.79}$$

$$+ \lambda \|B\|_{\ell_1,\ell_2} + \|B\|_{\ell_1,\ell_2} \left\|(\Sigma_X - \hat{\Sigma}_X) B\right\|_{2,\infty}. \tag{A.80}$$

Proof of slow-rate bound:

Assuming that $\lambda \ge \frac{2}{N} \left\|\mathbb{X}^\top E\right\|_{2,\infty}$, then by Theorem 1 and Corollary 1 of Gaynanova (2020) we have

$$\frac{1}{N} \left\|\mathbb{X}(\hat{B}-B)\right\|_F^2 \lesssim \lambda \|B\|_{\ell_1,\ell_2}, \tag{A.81}$$

$$\left\|(\hat{B}-B)^\top \Sigma_X^{1/2}\right\|_F^2 \lesssim \|B\|_{\ell_1,\ell_2} \left(\lambda + \|B\|_{\ell_1,\ell_2} \left\|\hat{\Sigma}_X - \Sigma_X\right\|_{\max}\right), \tag{A.82}$$

and

$$\left\|\hat{B}\right\|_{\ell_1,\ell_2} \lesssim \|B\|_{\ell_1,\ell_2}. \tag{A.83}$$

This last equation implies that $\left\|\hat{B}-B\right\|_{\ell_1,\ell_2} \lesssim \|B\|_{\ell_1,\ell_2}$ by the triangle inequality. Applying these bounds to the terms in Lemma A.3.1 establishes the slow-rate bound.

Proof of fast-rate bound:

Assuming that $\lambda \ge \frac{2}{N} \left\|\mathbb{X}^\top E\right\|_{2,\infty}$, that B has at most s nonzero rows, and assuming the group restricted eigenvalue condition on $\frac{1}{\sqrt{N}}\mathbb{X}$, then by Theorem 2 and Corollary 2 of Gaynanova

(2020) we have

$$\frac{1}{N}\left\|\mathbb{X}(\hat{B}-B)\right\|_F^2 \lesssim \kappa_X s\lambda^2 \tag{A.84}$$

$$\left\|(\hat{B}-B)^\top\Sigma_X^{1/2}\right\|_F^2 \lesssim \kappa_X s\left(1+\kappa_X s\left\|\hat{\Sigma}_X-\Sigma_X\right\|_{\max}\right)\lambda^2 \tag{A.85}$$

and

$$\left\|\hat{B}-B\right\|_{\ell_1,\ell_2} \lesssim \kappa_X s\lambda. \tag{A.86}$$

Applying these bounds to the terms in Lemma A.3.1 establishes the fast-rate bound. □

Proof of Lemma A.3.3:

From the definition of E, $E = \mathbb{Y}\hat{\Sigma}_Y^{-1/2} - \mathbb{X}B$, we have

$$\frac{1}{N}\mathbb{X}^\top E = \frac{1}{N}\mathbb{X}^\top\mathbb{Y}\hat{\Sigma}_Y^{-1/2} - \frac{1}{N}\mathbb{X}^\top\mathbb{X}B \tag{A.87}$$

$$= \hat{\Sigma}_{XY}\hat{\Sigma}_Y^{-1/2} - \hat{\Sigma}_X B \tag{A.88}$$

$$= \left(\hat{\Sigma}_{XY}\hat{\Sigma}_Y^{-1/2} - \Sigma_{XY}\Sigma_Y^{-1/2}\right) + \left(\Sigma_{XY}\Sigma_Y^{-1/2} - \hat{\Sigma}_X B\right). \tag{A.89}$$

Now considering the first term in equation (A.89), by adding and subtracting $\Sigma_{XY}\hat{\Sigma}_Y^{-1/2}$, and subsequently adding and subtracting $\Sigma_Y^{-1/2}$ to $\hat{\Sigma}_Y^{-1/2}$, we have

$$\hat{\Sigma}_{XY}\hat{\Sigma}_Y^{-1/2} - \Sigma_{XY}\Sigma_Y^{-1/2} = \left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\hat{\Sigma}_Y^{-1/2} + \Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2}-\Sigma_Y^{-1/2}\right) \tag{A.90}$$

$$= \left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\left(\hat{\Sigma}_Y^{-1/2}-\Sigma_Y^{-1/2}\right) + \left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\Sigma_Y^{-1/2} + \Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2}-\Sigma_Y^{-1/2}\right). \tag{A.91}$$

For the second term, we have

$$\Sigma_{XY}\Sigma_Y^{-1/2} - \hat{\Sigma}_X B = \Sigma_X\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2} - \hat{\Sigma}_X B \tag{A.92}$$

$$= \left(\Sigma_X - \hat{\Sigma}_X\right)B. \tag{A.93}$$

Combining these equalities and using the triangle inequality completes the proof. □

Proof of Lemma A.3.5:

The proof is adapted from Lemma 7 of Gaynanova (2020) but considers cross-covariance

matrices rather than the covariance matrices. Let X_{ij} denote entry (i,j) of $\mathbb{X}$ and Z_{ij} denote entry (i,j) of $\mathbb{Z}$. Then

$$\frac{1}{N}\left(\mathbb{X}^\top\mathbb{Z}\right)_{kl} = \frac{1}{N}\sum_{i=1}^{N} X_{ik}Z_{il}, \qquad k = 1,\dots p,\ l = 1,\dots d. \tag{A.94}$$

Let Σ_{XZ} be the cross-covariance matrix of X and Z with (k,l) entry equal to $\sigma_{kl} = \mathbb{E}\left[X_{ik}Z_{il}\right]$. Then, $X_{ik}Z_{il} - \sigma_{kl}$ are each mean 0 subexponential random variables, since

$$\left\|X_{ik}Z_{il}\right\|_{\psi_1} \le \left\|X_{ik}\right\|_{\psi_2}\left\|Z_{il}\right\|_{\psi_2} = g_k h_l \le gh$$

by Lemma 2.7.7. of Vershynin (2018), and because

$$\left\|X_{ik}Z_{il} - \sigma_{kl}\right\|_{\psi_1} \lesssim \left\|X_{ik}Z_{il}\right\|_{\psi_1}$$

by Exercise 2.7.10. of Vershynin (2018). Thus, $\left(\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right)_{kl}$ is a sum of independently and identically distributed (i.i.d.) subexponential random variables, since each for fixed k and l, the $X_{ik}Z_{il} - \sigma_{kl}$ are mean 0 subexponential i.i.d. random variables over $i = 1,\dots N$.

By Corollary 2.8.3. of Vershynin (2018), (Berstein's inequality), for each k and l, and for every $t > 0$,

$$P\left(\left|\left(\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right)_{kl}\right| \ge t\right) \le 2\exp\left[-cN\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right], \tag{A.95}$$

where $K = Cgh$, for absolute constants c and C. Applying a union bound, we have

$$P\left(\left\|\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right\|_{\max} \ge t\right) \le 2dp\exp\left[-cN\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right]. \tag{A.96}$$

When $t \le K$, we have

$$P\left(\left\|\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right\|_{\max} \ge t\right) \le 2dp\exp\left[-cN\min\frac{t^2}{K^2}\right]. \tag{A.97}$$

since $t^2/K^2 \le t/K$ if and only if $t \le K$. Letting the right-hand side of equation (A.97) be denoted as η, we solve for t in terms of η to obtain

$$t = \sqrt{\log\left(2dp\eta^{-1}\right)\frac{K^2}{cN}}, \tag{A.98}$$

and so that, for $\eta \in (0,1)$, if $\sqrt{\log\left(2dp\eta^{-1}\right)\frac{K^2}{cN}} \leq Cgh$, then with probability at least $1-\eta$ we have

$$\left\|\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right\|_{\max} \lesssim gh\sqrt{\log\left(2dp\eta^{-1}\right)\frac{1}{N}}. \tag{A.99}$$

$\log\left(2dp\eta^{-1}\right) \leq \log(2) + 2\log\left(p\eta^{-1}\right)$ since $d \leq p$, and because we suppose $\log(p) = o(N)$, it follows that for fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\left\|\Sigma_{XZ} - \frac{1}{N}\mathbb{X}^\top\mathbb{Z}\right\|_{\max} \lesssim gh\sqrt{\log\left(p\eta^{-1}\right)\frac{1}{N}}. \tag{A.100}$$

Using $\|A\|_{2,\infty} \leq \sqrt{d}\,\|A\|_{\max}$ for any $A \in \mathbb{R}^{p\times d}$ completes the proof. □

Proof of Lemma A.3.6:

To establish

$$\left\|\left(\hat{\Sigma}_{XY} - \Sigma_{XY}\right)\Sigma_Y^{-1/2}\right\|_{2,\infty} \lesssim \max_i(\|X_i\|_{\psi_2})\sqrt{\frac{d}{N}\log(p\eta^{-1})}, \tag{A.101}$$

we can use Lemma A.3.5 where $X = X$ and $Z = \Sigma_Y^{-1/2}Y$, since $\Sigma_{XY}\Sigma_Y^{-1/2} = \Sigma_{X,\Sigma_Y^{-1/2}Y}$. Then, the sub-Gaussian norms are $g = \max_i(\|X_i\|_{\psi_2})$ and $h = \max_i\left(\left\|\left(\Sigma_Y^{-1/2}Y\right)_i\right\|_{\psi_2}\right)$, where $\left(\Sigma_Y^{-1/2}Y\right)_i$ is the ith entry of $\Sigma_Y^{-1/2}Y \in \mathbb{R}^d$. Using Definition A.3.1 for $\Sigma_Y^{-1/2}Y$, we have

$$h \leq K_{\Sigma_Y^{-1/2}Y}\left\|\Sigma_Y^{-1/2}\Sigma_Y\Sigma_Y^{-1/2}\right\|_2^{1/2} = K_{\Sigma_Y^{-1/2}Y}. \tag{A.102}$$

Treating $K_{\Sigma_Y^{-1/2}Y}$ as an absolute constant establishes equation (A.101).

Establishing

$$\left\|\left(\Sigma_X - \hat{\Sigma}_X\right)B\right\|_{2,\infty} \lesssim \max_i(\|X_i\|_{\psi_2})\gamma_1\sqrt{\frac{d}{N}\log(p\eta^{-1})} \tag{A.103}$$

follows an identical argument except that we let $Z = B^\top X$, so

$$h \leq K_{B^\top X}\,\|B^\top\Sigma_X B\|_2^{1/2}$$

in the final step. The identity $\|B^\top\Sigma_X B\|_2 = \gamma_1^2$ establishes equation (A.103).

To deduce that

$$\left\|\Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_{2,\infty} \lesssim \|\Sigma_X\|_{2,\infty}^{1/2}\,\gamma_1\sqrt{\frac{d+\log(\eta^{-1})}{N}}, \tag{A.104}$$

we begin by using $\|AB\|_{2,\infty} \leq \|A\|_{2,\infty} \|B\|_2$ and $\Sigma_Y^{-1/2}\Sigma_Y^{1/2} = I_d$ to obtain

$$\left\|\Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_{2,\infty} \leq \underbrace{\left\|\Sigma_{XY}\Sigma_Y^{-1/2}\right\|_{2,\infty}}_{\text{Term I}} \underbrace{\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2}_{\text{Term II}}. \tag{A.105}$$

Considering Term I, we have

$$\begin{aligned}\left\|\Sigma_{XY}\Sigma_Y^{-1/2}\right\|_{2,\infty} &= \left\|\Sigma_X^{1/2}\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\right\|_{2,\infty} &\text{(A.106)}\\ &\leq \left\|\Sigma_X^{1/2}\right\|_{2,\infty}\left\|\Sigma_X^{1/2}B\right\|_2 &\text{(A.107)}\\ &= \left\|\Sigma_X^{1/2}\right\|_{2,\infty}\gamma_1. &\text{(A.108)}\end{aligned}$$

In the below, we are able to bound Term II without incurring unnecessary factors of $\left\|\Sigma_Y^{-1/2}\right\|_2$ using the results of Kereta and Klock (2021) which pertain to precision matrix estimation along subspaces. The main idea is to bound $\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2$ in terms of $\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\right)\Sigma_Y^{1/2}\right\|_2$, to which the results of Kereta and Klock (2021) can be applied. That $\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2$ is not simply equal to $\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\right)\Sigma_Y^{1/2}\right\|_2$ is due to Σ_Y and $\hat{\Sigma}_Y$ not necessarily commuting with one another. We begin with

$$\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2 = \left\|\Sigma_Y^{1/2}\hat{\Sigma}_Y^{-1/2} - I\right\|_2. \tag{A.109}$$

Using identity 15 in Section A.6 and that both $\Sigma_Y^{1/2}$ and $\hat{\Sigma}_Y^{1/2}$ are positive definite along with their inverses,

$$\text{Term II} = \left\|\left(\Sigma_Y^{1/2}\hat{\Sigma}_Y^{-1}\Sigma_Y^{1/2}\right)^{1/2} - I\right\|_2 = \left\|\left(\Sigma_Y^{1/2}\hat{\Sigma}_Y^{-1}\Sigma_Y^{1/2}\right)^{1/2} - \left(\Sigma_Y^{1/2}\Sigma_Y^{-1}\Sigma_Y^{1/2}\right)^{1/2}\right\|_2. \tag{A.110}$$

Using $\|A^{1/2} - B^{1/2}\|_2 \leq \frac{1}{2}\max\left(\|A^{-1}\|_2, \|B^{-1}\|_2\right)^{1/2}\|A - B\|_2$ for positive definite matrices A and B, we deduce that

$$\begin{aligned}\text{Term II} &\leq \frac{1}{2}\max\left(\left\|\left(\Sigma_Y^{1/2}\hat{\Sigma}_Y^{-1}\Sigma_Y^{1/2}\right)^{-1}\right\|_2, \left\|I_d^{-1}\right\|_2\right)^{1/2}\left\|\Sigma_Y^{1/2}\hat{\Sigma}_Y^{-1}\Sigma_Y^{1/2} - \Sigma_Y^{1/2}\Sigma_Y^{-1}\Sigma_Y^{1/2}\right\|_2 &\text{(A.111)}\\ &= \underbrace{\frac{1}{2}\max\left(\left\|\Sigma_Y^{-1}\hat{\Sigma}_Y\right\|_2, 1\right)^{1/2}}_{\text{Term II.I}}\underbrace{\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\right)\Sigma_Y^{1/2}\right\|_2}_{\text{Term II.I}}. &\text{(A.112)}\end{aligned}$$

We bound $\left\|\Sigma_Y^{-1}\hat{\Sigma}_Y\right\|_2$ in Term II.I with

$$\left\|\Sigma_Y^{-1}\hat{\Sigma}_Y\right\|_2 \leq \left\|\Sigma_Y^{-1}\left(\hat{\Sigma}_Y - \Sigma_Y\right)\right\|_2 + \left\|\Sigma_Y^{-1}\Sigma_Y\right\|_2 \tag{A.113}$$

$$= \left\|\Sigma_Y^{-1/2}\left(\hat{\Sigma}_Y - \Sigma_Y\right)\Sigma_Y^{-1/2}\right\|_2 + 1. \tag{A.114}$$

We apply a result concerning covariance estimation along subspaces, Lemma 2 of Kereta and Klock (2021), to deduce that for fixed $\eta \in (0,1)$ with probability $1-\eta$,

$$\left\|\Sigma_Y^{-1/2}\left(\hat{\Sigma}_Y - \Sigma_Y\right)\Sigma_Y^{-1/2}\right\|_2 \lesssim \left\|\Sigma_Y^{-1/2}Y\right\|_{\psi_2}^2 \max\left(\sqrt{\frac{2d+\log(\eta^{-1})}{N}}, \frac{2d+\log(\eta^{-1})}{N}\right). \tag{A.115}$$

From Definition A.3.1 for $\Sigma_Y^{-1/2}Y$ combined with the assumption that $d = o(N)$, we have that in the limit this term is bounded by 1. Therefore, for fixed $\eta \in (0,1)$, with probability $1-\eta$, $\left\|\Sigma_Y^{-1}\hat{\Sigma}_Y\right\|_2 \lesssim 1$. From this, Term II.I $\lesssim 1$ as well.

Having bounded $\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2$ in terms of Term II.II $= \left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\right)\Sigma_Y^{1/2}\right\|_2$, we will bound the latter term. We use Theorem 10 of Kereta and Klock (2021) directly, implying that for fixed $\eta \in (0,1)$, if $N \gtrsim (d+\log(\eta^{-1})\left\|\Sigma_Y^{-1/2}Y\right\|_{\psi_2}^4$, then with probability $1-\eta$,

$$\left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\right)\Sigma_Y^{1/2}\right\|_2 \lesssim \left\|\Sigma_Y^{-1/2}Y\right\|_{\psi_2}^2 \sqrt{\frac{\operatorname{rank}(\Sigma_Y)+\log(\eta^{-1})}{N}}. \tag{A.116}$$

By Definition A.3.1 for $\Sigma_Y^{-1/2}Y$, $\left\|\Sigma_Y^{-1/2}Y\right\|_{\psi_2} \lesssim 1$, so that the assumption $d = o(N)$ ensures that for fixed η, $N \gtrsim (d+\log(\eta^{-1}))\left\|\Sigma_Y^{-1/2}Y\right\|_{\psi_2}^4$ eventually.

With our bounds for both Term II.I and Term II.II, we deduce that for fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\text{Term II} = \left\|\Sigma_Y^{1/2}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_2 \lesssim \sqrt{\frac{d+\log(\eta^{-1})}{N}}. \tag{A.117}$$

Now having bounded both Term I and Term II, we finally establish that for fixed $\eta \in (0,1)$, if $d = o(N)$, with probability $1-\eta$,

$$\left\|\Sigma_{XY}\left(\hat{\Sigma}_Y^{-1/2} - \Sigma_Y^{-1/2}\right)\right\|_{2,\infty} \lesssim \|\Sigma_X\|_{2,\infty}^{1/2}\,\gamma_1\sqrt{\frac{d+\log(\eta^{-1})}{N}}, \tag{A.118}$$

completing the proof of (A.104).

To show

$$\left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\left(\hat{\Sigma}_{Y}^{-1/2}-\Sigma_{Y}^{-1/2}\right)\right\|_{2,\infty} \lesssim \left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\Sigma_{Y}^{-1/2}\right\|_{2,\infty}, \tag{A.119}$$

we begin with

$$\left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\left(\hat{\Sigma}_{Y}^{-1/2}-\Sigma_{Y}^{-1/2}\right)\right\|_{2,\infty} \leq \left\|\left(\hat{\Sigma}_{XY}-\Sigma_{XY}\right)\Sigma_{Y}^{-1/2}\right\|_{2,\infty}\left\|\Sigma_{Y}^{1/2}\left(\hat{\Sigma}_{Y}^{-1/2}-\Sigma_{Y}^{-1/2}\right)\right\|_{2}, \tag{A.120}$$

using $\Sigma_Y^{-1/2}\Sigma_Y^{1/2} = I_d$ and $\|AB\|_{2,\infty} \leq \|A\|_{2,\infty}\|B\|_2$. From equation (A.117) we obtain that with probability $1-\eta$, the second factor in (A.120) is bounded by an absolute constant as $d = o(N)$, completing the proof. □

Proof of Lemma A.3.8:

That for fixed $\eta \in (0,1)$, if $\log(p) = o(N)$, then with probability $1-\eta$,

$$\left\|\Sigma_X - \hat{\Sigma}_X\right\|_{\max} \lesssim \max(\|X_i\|_{\psi_2}^2)\sqrt{\frac{\log\left(p\eta^{-1}\right)}{N}} \tag{A.121}$$

follows from Lemma 7 of Gaynanova (2020).

That for fixed $\eta \in (0,1)$, if $d = o(N)$, then with probability $1-\eta$,

$$\left\|B^\top(\Sigma_X - \hat{\Sigma}_X)B\right\|_2 \lesssim \gamma_1^2\sqrt{\frac{\operatorname{rank}\left(B^\top\Sigma_X\right)+\log\left(\eta^{-1}\right)}{N}} \tag{A.122}$$

follows from Lemma 2 of Kereta and Klock (2021) in addition to Definition A.3.1 applied to $B^\top X$, which implies that $\|B^\top X\|_{\psi_2} \leq K_{B^\top X}\|B^\top\Sigma_X B\|_2^{1/2}$. Using $\operatorname{rank}(AB) \leq \min\left(\operatorname{rank}(A), \operatorname{rank}(B)\right)$, we have

$$\operatorname{rank}\left(B^\top\Sigma_X\right) = \operatorname{rank}\left(\Sigma_X B\right) = \operatorname{rank}\left(\Sigma_{XY}\Sigma_Y^{-1/2}\right) \leq d, \tag{A.123}$$

which establishes the desired result.

To show that for fixed $\eta \in (0,1)$, if $d = o(N)$, then with probability $1-\eta$,

$$\frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2 \lesssim \gamma_1, \tag{A.124}$$

we begin with

$$\frac{1}{\sqrt{N}}\|\mathbb{X}B\|_2 = \frac{1}{\sqrt{N}}\|B^\top\mathbb{X}^\top\mathbb{X}B\|_2^{1/2} = \left\|B^\top\hat{\Sigma}_X B\right\|_2^{1/2}, \tag{A.125}$$

which holds since $\|A\|_2 = \|A^\top A\|_2^{1/2}$. Adding and subtracting $B^\top\Sigma_X B$ and using the triangle inequality, we obtain that for fixed $\eta \in (0,1)$, with probability $1-\eta$,

$$\left\|B^\top\hat{\Sigma}_X B\right\|_2 \le \|B^\top\Sigma_X B\|_2 + \left\|B^\top\left(\hat{\Sigma}_X - \Sigma_X\right)B\right\|_2 \lesssim \gamma_1^2. \tag{A.126}$$

In the last inequality, we have used $d = o(N)$ and equation (A.122) to deduce that $\left\|B^\top\left(\hat{\Sigma}_X - \Sigma_X\right)B\right\|_2$ becomes smaller that γ_1^2 eventually. This completes the proof. □

Proof of Lemma A.3.9:

By Lemma 6 of Gaynanova (2020), if suffices to show that under the condition $s^2\log(p) = o(N)$, that for fixed η, with probability $1-\eta$, we have $s\left\|\Sigma_X - \hat{\Sigma}_X\right\|_{\max} \le (32\kappa)^{-1}$. By the first item of Lemma A.3.8, we then have that for fixed η with probability $1-\eta$,

$$\kappa s\left\|\Sigma_X - \hat{\Sigma}_X\right\|_{\max} \lesssim \kappa s\max_i(\|X_i\|_{\psi_2}^2)\sqrt{\frac{\log\left(p\eta^{-1}\right)}{N}}. \tag{A.127}$$

It therefore suffices that $\max_i(\|X_i\|_{\psi_2}^4)\kappa^2 s^2\log(p) = o(N)$, since then with probability $1-\eta$, $\kappa s\left\|\Sigma_X - \hat{\Sigma}_X\right\|_{\max}$ is arbitrarily small. But this is the assumed condition, so the proof is complete. □

Proof of Lemma A.3.10:

Proof of slow-rate bound:

Under the slow-rate assumptions of Corollary A.3.1 and by Corollary 1 in Gaynanova (2020), for fixed η, with probability $1-\eta$, we have

$$\left\|B-\hat{B}\right\|_2^2 \lesssim \left\|\Sigma_X^{-1}\right\|_2\|B\|_{\ell_1,\ell_2}\left(\lambda + \|B\|_{\ell_1,\ell_2}\left\|\hat{\Sigma}_X - \Sigma_X\right\|_{\max}\right). \tag{A.128}$$

Then

$$\lambda \lesssim \|\Sigma_X\|_{2,\infty}^{1/2}\left(\frac{d}{N}\log\left(p\eta^{-1}\right)\right)^{1/2}, \tag{A.129}$$

and by Lemma A.3.8,

$$\left\|\Sigma_X - \hat{\Sigma}_X\right\|_{\max} \lesssim \|\Sigma_X\|_{2,\infty} \left(\frac{d}{N} \log\left(p\eta^{-1}\right)\right)^{1/2}. \tag{A.130}$$

The last two equations bound $\left\|B - \hat{B}\right\|_2$, and since $\|\Sigma_X\|_{2,\infty}, \|B\|_{\ell_1,\ell_2} \geq 1$, the proof of the slow-rate bound is complete.

Proof of fast-rate bound:

Under the slow-rate assumptions of Corollary A.3.1 and by Theorem 2 in Gaynanova (2020), for fixed η, with probability $1 - \eta$, we have

$$\left\|B - \hat{B}\right\|_2 \lesssim \kappa s^{1/2} \lambda. \tag{A.131}$$

Again using $\lambda \lesssim \|\Sigma_X\|_{2,\infty}^{1/2} \left(\frac{d}{N} \log\left(p\eta^{-1}\right)\right)^{1/2}$, the result is shown. □

Proof of Lemma A.3.11:

The proof of the first statement follows from Lemma 2 of Kereta and Klock (2021) and Definition A.3.1 applied to Y.

To show the bound on $\left\|\Sigma_Y^{-1/2} - \hat{\Sigma}_Y^{-1/2}\right\|_2$, we begin by using identity 13 in Section A.6 applied to Σ_Y^{-1} and $\hat{\Sigma}_Y^{-1}$. We have

$$\left\|\Sigma_Y^{-1/2} - \hat{\Sigma}_Y^{-1/2}\right\|_2 \leq \frac{1}{2} \max\left(\|\Sigma_Y\|_2, \left\|\hat{\Sigma}_Y\right\|_2\right)^{1/2} \left\|\Sigma_Y^{-1} - \hat{\Sigma}_Y^{-1}\right\|_2. \tag{A.132}$$

To bound $\left\|\Sigma_Y^{-1} - \hat{\Sigma}_Y^{-1}\right\|_2$, we use Corollary 11 of Kereta and Klock (2021): if

$$\left(\operatorname{rank}\left(\Sigma_Y\right) + \log\left(\eta^{-1}\right)\right) \left\|\Sigma_Y^{1/2} Y\right\|_{\psi_2}^4 \lesssim N$$

eventually, then with probability $1 - \eta$,

$$\left\|\Sigma_Y^{-1} - \hat{\Sigma}_Y^{-1}\right\|_2 \lesssim \left\|\Sigma_Y^{-1} Y\right\|_{\psi_2}^2 \sqrt{\frac{\operatorname{rank}\left(\Sigma_Y\right) + \log\left(\eta^{-1}\right)}{N}}. \tag{A.133}$$

The variance proxy condition on $\Sigma_Y^{1/2} Y$ (Definition A.3.1) implies that $\left\|\Sigma_Y^{-1/2} Y\right\|_{\psi_2}$ is bounded by an absolute constant. The condition $d = o(N)$ then ensures

$(\operatorname{rank}(\Sigma_Y) + \log(\eta^{-1}))\left\|\Sigma_Y^{1/2}Y\right\|_{\psi_2}^4 \lesssim N$ eventually. The variance proxy condition on $\Sigma_Y^{-1}Y$ implies $\|\Sigma_Y^{-1}Y\|_{\psi_2} \lesssim \left\|\Sigma_Y^{-1/2}\right\|_2$.

To bound $\left\|\hat{\Sigma}_Y\right\|_2$, we use the triangle inequality to obtain $\left\|\hat{\Sigma}_Y\right\|_2 \leq \left\|\hat{\Sigma}_Y - \Sigma_Y\right\|_2 + \|\Sigma_Y\|_2$. Then, using the first statement of the lemma and with the additional assumption that $\|\Sigma_Y\|_2^2 d = o(N)$, we have with probability $1-\eta$,

$$\left\|\hat{\Sigma}_Y\right\|_2 \lesssim \|\Sigma_Y\|_2 . \tag{A.134}$$

Combining these results together establishes the statement of the lemma and the proof is complete. □

Proof of Theorems A.3.2 and A.3.3:

Proof of bounds for θ:

By definition, we have that

$$\theta_k = B\tilde{\eta}_k\gamma_k^{-1} \tag{A.135}$$

$$\hat{\theta}_k = \hat{B}\hat{\tilde{\eta}}_k\hat{\gamma}_k^{-1}. \tag{A.136}$$

We bound $\left\|\theta_k - \hat{\theta}_k\right\|_2$ by bounding all three of $\left\|B - \hat{B}\right\|_2$, $\left\|\tilde{\eta}_k - \hat{\tilde{\eta}}_k\right\|_2$, and $\left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right|$. For ease of notation, we denote $\tilde{\eta}_k$ by v. We begin with

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \leq \left\|Bv\gamma_k^{-1} - Bv\hat{\gamma}_k^{-1}\right\|_2 + \left\|Bv\hat{\gamma}_k^{-1} - \hat{B}\hat{v}\hat{\gamma}_k^{-1}\right\|_2 \tag{A.137}$$

$$\leq \left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right| \|Bv\|_2 + \left|\hat{\gamma}_k^{-1}\right| \left\|Bv - \hat{B}\hat{v}\right\|_2 . \tag{A.138}$$

Examining $\left\|Bv - \hat{B}\hat{v}\right\|_2$ on the right-hand side of (A.138):

$$\left\|Bv - \hat{B}\hat{v}\right\|_2 \leq \|Bv - B\hat{v}\|_2 + \left\|B\hat{v} - \hat{B}\hat{v}\right\|_2 \tag{A.139}$$

$$\leq \|B\|_2 \|v - \hat{v}\|_2 + \left\|B - \hat{B}\right\|_2 \|\hat{v}\|_2 . \tag{A.140}$$

Using (A.138) and since $\|\hat{v}\|_2 = 1$, we have

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \leq \left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right| \|Bv\|_2 + \left|\hat{\gamma}_k^{-1}\right| \left(\|B\|_2 \|v - \hat{v}\|_2 + \left\|B - \hat{B}\right\|_2\right). \tag{A.141}$$

where we note that we now have bounded $\left\|\theta_k - \hat{\theta}_k\right\|_2$ in terms of $\left\|B - \hat{B}\right\|_2$, $\left\|\tilde{\eta}_k - \hat{\tilde{\eta}}_k\right\|_2$, and $\left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right|$. To bound $\left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right|$, we can use identity 16 from Section A.6, giving us that

$$\left|\gamma_k^{-1} - \hat{\gamma}_k^{-1}\right| \leq \min\left(\gamma_k, \hat{\gamma}_k\right)^{-3} \left|\gamma_k^2 - \hat{\gamma}_k^2\right|. \tag{A.142}$$

Bounding the two factors in equation (A.142) amounts to establishing that γ_k is close to $\hat{\gamma}_k$. For this, we apply Weyl's inequality (Bhatia (2013) Corollary III.2.6.) to obtain

$$\left|\gamma_k^2 - \hat{\gamma}_k^2\right| \leq \left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2. \tag{A.143}$$

since γ_k^2 is the kth eigenvalue of $B^\top \Sigma_X B$, and $\hat{\gamma}_k^2$ is the kth eigenvalue of $\hat{B}^\top \hat{\Sigma}_X \hat{B}$. We then deduce that

$$\gamma_k^2 - \left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2 \leq \hat{\gamma}_k^2. \tag{A.144}$$

From equation (A.144) we establish that $\frac{1}{2}\gamma_k^2 \leq \hat{\gamma}_k^2$ by using

$$\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2 \lesssim \frac{1}{2}\gamma_k^2, \tag{A.145}$$

which holds asymptotically in both the fast and slow rate cases under our assumption that the γ_k are bounded from below and from Corollary A.3.1. From $\frac{1}{2}\gamma_k^2 \leq \hat{\gamma}_k^2$ we also deduce

$$\min\left(\gamma_k, \hat{\gamma}_k\right)^{-2} \lesssim \frac{1}{\gamma_k^2}, \tag{A.146}$$

and additionally that

$$\frac{1}{\hat{\gamma}_k} \lesssim \frac{1}{\gamma_k}. \tag{A.147}$$

We now use our results thus far regarding γ_k and $\hat{\gamma}_k$ to obtain a simplified equation (A.141):

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \lesssim \frac{1}{\gamma_k^3} \left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2 \|Bv\|_2 + \frac{1}{\gamma_k}\left(\|B\|_2 \|v - \hat{v}\|_2 + \left\|B - \hat{B}\right\|_2\right) \tag{A.148}$$

To bound $\|v - \hat{v}\|_2$, we can use the Davis-Kahan theorem (Corollary 3 of Yu et al. (2015)). Assuming that $\tilde{\eta_k}^\top \hat{\tilde{\eta}}_k \geq 0$, then

$$\left\|\tilde{\eta}_k - \hat{\tilde{\eta}}_k\right\|_2 \leq \frac{2^{3/2}\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)}, \tag{A.149}$$

because $\tilde{\eta_k}$ is the kth eigenvector of $B^\top \Sigma_X B$, and $\hat{\tilde{\eta}}_k$ is the kth eigenvector of $\hat{B}^\top \hat{\Sigma}_X \hat{B}$. From this and equation (A.148), we obtain

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \lesssim \frac{1}{\gamma_k^2}\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2 \|\theta_k\|_2 + \frac{1}{\gamma_k}\left(\|B\|_2 \frac{\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)} + \left\|B - \hat{B}\right\|_2\right), \tag{A.150}$$

where we have also used the definition of θ_k, $\theta_k = Bv\gamma_k^{-1}$. Rearranging this expression, we have

$$\left\|\theta_k - \hat{\theta}_k\right\|_2 \lesssim \left(\frac{\|\theta_k\|_2}{\gamma_k^2} + \frac{\|B\|_2}{\gamma_k \min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)}\right)\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2 + \frac{\left\|B - \hat{B}\right\|_2}{\gamma_k}. \tag{A.151}$$

Now we use our bounds for $\left\|B^\top \Sigma_X B - \hat{B}^\top \hat{\Sigma}_X \hat{B}\right\|_2$ and $\left\|B - \hat{B}\right\|_2$ depending on if we are in the slow or fast rate case. In the slow rate case, we also use

$$\|\theta_k\|_2 = \left\|\Sigma_X^{-1/2}\tilde{\theta_k}\right\|_2 \leq \left\|\Sigma_X^{-1/2}\right\|_2 \left\|\tilde{\theta_k}\right\|_2 = \left\|\Sigma_X^{-1/2}\right\|_2, \tag{A.152}$$

where we have used the standard result of classical CCA that $\theta_k = \Sigma_X^{-1/2}\tilde{\theta_k}$ for some unit vector $\tilde{\theta_k}$ (Uurtio et al., 2018).

The proofs for both the fast and slow rate then follow directly from Lemma A.3.10, Theorem A.3.1, and rearranging of terms. This completes the proof of the bounds on $\left\|\theta_k - \hat{\theta}_k\right\|_2$.

Proof of bounds for η:

By definition, we have

$$\eta_k = \Sigma_Y^{-1/2}\tilde{\eta_k}, \tag{A.153}$$

$$\hat{\eta}_k = \hat{\Sigma}_Y^{-1/2}\hat{\tilde{\eta}}_k. \tag{A.154}$$

We bound $\|\eta_k - \hat{\eta}_k\|_2$ with the triangle inequality:

$$\|\eta_k - \hat{\eta}_k\|_2 \leq \left\|\Sigma_Y^{-1/2}\tilde{\eta_k} - \Sigma_Y^{-1/2}\hat{\tilde{\eta}}_k\right\|_2 + \left\|\Sigma_Y^{-1/2}\hat{\tilde{\eta}}_k - \hat{\Sigma}_Y^{-1/2}\hat{\tilde{\eta}}_k\right\|_2 \tag{A.155}$$

$$\leq \left\|\Sigma_Y^{-1/2}\right\|_2 \left\|\tilde{\eta_k} - \hat{\tilde{\eta}}_k\right\|_2 + \left\|\Sigma_Y^{-1/2} - \hat{\Sigma}_Y^{-1/2}\right\|_2 \left\|\hat{\tilde{\eta}}_k\right\|_2. \tag{A.156}$$

To simplify this expression, we can use $\left\|\hat{\tilde{\eta}}_k\right\|_2 = 1$, the Davis-Kahan Theorem for $\left\|\tilde{\eta_k} - \hat{\tilde{\eta}}_k\right\|_2$ as in the proof for the θ bounds, and the second statement of Lemma A.3.11. For clarity,

we state the assumptions required for these results: $d = o(N)$, $\|\Sigma_Y\|_2^2 d = o(N)$, the variance proxy condition (Definition A.3.1) for Y and $\Sigma_Y^{-1}Y$, and $\tilde{\eta_k}^\top \hat{\tilde{\eta}}_k \geq 0$ for $k = 1, \dots K$. Then, for $\eta \in (0,1)$, we have

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left\|\Sigma_Y^{-1/2}\right\|_2 \left[\frac{\left\|\hat{B}^T \hat{\Sigma}_X \hat{B} - B^\top \Sigma_X B\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)} + \|\Sigma_Y\|_2^{1/2} \left\|\Sigma_Y^{-1/2}\right\|_2 \sqrt{\frac{d \log\left(\eta^{-1}\right)}{N}} \right]. \quad \text{(A.157)}$$

Now we apply Corollary A.3.1 to equation (A.157) under the fast and slow rate assumptions. In the slow rate case, we have

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left\|\Sigma_Y^{-1/2}\right\|_2 \left[\frac{\gamma_1 \|\Sigma_X\|_{2,\infty}^{1/2} \|B\|_{\ell_1,\ell_2}}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)} \left(\frac{d}{N} \log\left(p\eta^{-1}\right)\right)^{1/4} \right.$$
$$\left. + \|\Sigma_Y\|_2^{1/2} \left\|\Sigma_Y^{-1/2}\right\|_2 \left(\frac{d \log\left(\eta^{-1}\right)}{N}\right)^{1/2} \right].$$

Factoring out $\left(\frac{d}{N}\right)^{1/4}$ we obtain

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left\|\Sigma_Y^{-1/2}\right\|_2 \left(\frac{d}{N}\right)^{1/4} \left[\frac{\gamma_1 \|\Sigma_X\|_{2,\infty}^{1/2} \|B\|_{\ell_1,\ell_2}}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)} \log\left(p\eta^{-1}\right)^{1/4} \right.$$
$$\left. + \|\Sigma_Y\|_2^{1/2} \left\|\Sigma_Y^{-1/2}\right\|_2 \log\left(\eta^{-1}\right)^{1/2} \left(\frac{d}{N}\right)^{1/4} \right].$$

In the bracketed expression we are able to combine the first and second terms, since we assume γ_1 is bounded from below, using the additional assumption that $\|\Sigma_Y\|_2^2 \|\Sigma_Y^{-1}\|_2^2 d = o(N)$, and because the other terms in the first term are greater than or equal to 1. Then, with probability $1 - \eta$,

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \frac{\gamma_1 \|\Sigma_X\|_{2,\infty}^{1/2} \|B\|_{\ell_1,\ell_2} \left\|\Sigma_Y^{-1/2}\right\|_2}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)} \left(\frac{d \log\left(p\eta^{-1}\right)}{N}\right)^{1/4}, \quad \text{(A.158)}$$

completing the proof of the slow-rate bound for η.

In the fast-rate case, under the fast-rate bound assumptions of Corollary A.3.1 and applying Corollary A.3.1 to equation (A.157), we establish

$$\|\eta_k - \hat{\eta}_k\|_2 \lesssim \left\|\Sigma_Y^{-1/2}\right\|_2 \max\left(\frac{\gamma_1 \|\Sigma_X\|_{2,\infty}^{1/2} s^{1/2} \kappa^{1/2}}{\min\left(\gamma_{k-1}^2 - \gamma_k^2, \gamma_k^2 - \gamma_{k+1}^2\right)}, \|\Sigma_Y\|_2^{1/2} \left\|\Sigma_Y^{-1}\right\|_2^{1/2} \right) \left(\frac{d \log\left(p\eta^{-1}\right)}{N}\right)^{1/2}, \quad \text{(A.159)}$$

completing the proof of the fast-rate bound for η. □

A.4 Asymmetric Sparse-Functional CCA: Proof of Theorem 2.4.2

In this section, we prove Theorem 2.4.2. Recall that we use Assumption 2.2.1 where the $\{\phi_j\}_{j=1}^{\infty}$ are the principal components of $\mathrm{Log}_\mu y$, and without loss of generality suppose that $I = \{1, \dots d^{(\mathrm{corr})}\}$. We suppose that d, the number of principal components we use in practice, satisfies $d \geq d^{(\mathrm{corr})}$. Define $Y_j \equiv \langle\!\langle \mathrm{Log}_\mu y, \phi_j \rangle\!\rangle$ for $j = 1, \dots d$, so that the $\{Y_j\}$ are random variables with $\mathrm{Var}(Y_j) \equiv \omega_j$. Let $Y = (Y_1, \dots, Y_d)^\top$. Then, by Theorem 2.2.1 the canonical pairs $\{(\psi_k, \theta_k)\}$ can be computed by solving a $d^{(\mathrm{corr})}$-dimensional multivariate CCA problem between X and $Y^{d^{(\mathrm{corr})}} \equiv (Y_1, \dots, Y_{d^{(\mathrm{corr})}})^\top$. It is straightforward to show that thanks to Assumption 2.2.1 and the properties of the singular value decomposition, we can equivalently solve a d-dimensional multivariate CCA problem between X and Y:

$$\underset{\mathrm{Var}(\eta_1^\top Y)=\mathrm{Var}(\theta_1^\top X)=1}{\mathrm{maximize}} \mathrm{Corr}^2\left(\eta_1^\top Y, \theta_1^\top X\right), \tag{A.160}$$

with subsequent canonical pairs defined analogously, where Y has replaced $Y^{d^{(\mathrm{corr})}}$. We denote with $K = \max\{i \in \{1, \dots d^{(\mathrm{corr})}\} : \gamma_i > 0\}$ the number of nontrivial canonical vectors in this problem, and to simplify the notation, we use the conventions $\gamma_{K+1}^2 = -\infty$ and $\gamma_0^2 = \infty$. Here, $\eta_k \in \mathbb{R}^d$ and $\theta_k \in \mathbb{R}^p$ for $k = 1, \dots K$. The canonical pairs are then given by $\left(\psi_k = \sum_{j=1}^d \phi_j \eta_{kj}, \theta_k\right)$, where $\eta_{kj} = 0$ for $j > d^{(\mathrm{corr})}$. Recall that $\mathcal{C} \equiv \mathbb{E}\left[\mathrm{Log}_\mu y \otimes \mathrm{Log}_\mu y\right]$ admits the expansion $\mathcal{C} = \sum_{j=1}^{\infty} \omega_j \phi_j \otimes \phi_j$, where $\{\phi_j\}$ are the eigenfunctions of $\mathcal{C}$ with associated eigenvalues $\{\omega_j\}$. We let $\gamma_1^2 \dots \gamma_K^2$ denote the squared canonical correlations attained by the pairs $(\psi_1, \theta_1), \dots (\psi_K, \theta_K)$.

We denote by $\hat{\psi}_k$ and $\hat{\theta}_k$ the canonical vectors estimated using our proposed Algorithm 1. In practice, we are given a sample of N independent pairs

$$(y_i, X_i), \quad i = 1, \dots, N, \tag{A.161}$$

where each pair (y_i, X_i) is an independent observation of the pair (y, X). Here, the functions $\{y_i\}$ are assumed to be fully observed on $\mathcal{T}$. We denote $\tau \equiv |\mathcal{T}|$, the length of the time interval of the functional data. We store the observations $\{X_i\}$ in a matrix $\mathbb{X} \in \mathbb{R}^{N \times p}$.

In Algorithm 1, we estimate μ using the sample Fréchet mean, denoted as $\hat{\mu}$, and estimate the eigenfunctions $\{\phi_j\}$ using $\hat{\phi}_j$, which are the eigenfunctions of the sample covariance function $\hat{\mathcal{C}} \equiv \frac{1}{N}\sum_{i=1}^{N} \operatorname{Log}_{\hat{\mu}} y_i \otimes \operatorname{Log}_{\hat{\mu}} y_i$. Hence, the functional data can be represented using the vector $Z \in \mathbb{R}^d$, where its jth element is

$$Z_j \equiv \langle\!\langle \operatorname{Log}_{\hat{\mu}} y_1, \hat{\phi}_j \rangle\!\rangle_{\hat{\mu}}. \tag{A.162}$$

We note that in the definition of Z, both $\hat{\mu}$ and $\hat{\phi}_j$ depend on y_1 since their estimation depends on the full sample $y_1, \ldots y_N$. We also note that the distribution of Z depends on the sample size N. In practice, we solve the following problem for the first pair of canonical variables:

$$\underset{\operatorname{Var}(a_1^\top Z)=\operatorname{Var}(b_1^\top X)=1}{\text{maximize}} \operatorname{Corr}^2\left(a_1^\top Z, b_1^\top X\right). \tag{A.163}$$

The subsequent canonical pairs can be defined analogously. We denote the solutions to these problems as $(a_1, b_1), \ldots (a_K, b_K)$. We let $\tilde{\gamma}_1^2 \ldots \tilde{\gamma}_K^2$ denote the squared canonical correlations attained by the pairs $(a_1, b_1), \ldots (a_K, b_K)$. We expect that, under appropriate assumptions, a_k and b_k will closely approximate η_k and θ_k, respectively, provided that $\hat{\mu}$ and $\{\hat{\phi}_j\}$ closely approximates μ and $\{\phi_j\}$, respectively.

We assume that the canonical vectors $\{\theta_k\}$ are group s-sparse, and that the associated vectors $\{b_k\}$ are also group s-sparse. Let the support $S \subseteq \{1, \ldots p\}$ represent the indices of non-zero elements of θ_k or b_k, with cardinality $|S| \leq 2s$. This sparsity condition allows us to simplify equations (A.160) and (A.163) by replacing X with $X_S \in \mathbb{R}^{|S|}$, the random vector consisting of only the entries $\{X_j : j \in S\}$. Moreover, θ_k and b_k can be replaced with $\theta_{k,S}$ and $b_{k,S}$, respectively.

We begin by deriving an error bound for the estimation of ψ_k using $\hat{\psi}_k$. Since the population quantity ψ_k belongs to $L^2(T\mu)$ and our estimate $\hat{\psi}_k$ belongs to $L^2(T\hat{\mu})$, we use the parallel transport operator $\Gamma_{\mu,\hat{\mu}}$ to define estimation error, as proposed in Lin and Yao (2019). For ease of notation, we denote $\Gamma_{f,g}U - V$ as $U \delta_\Gamma V \in L^2(Tg)$, for any vector fields $U \in L^2(Tf)$ and $V \in L^2(Tg)$.

A.4.1 Bounding the canonical function error

Next, we derive a bound for $\hat{\psi}_k \delta_\Gamma \psi_k \in L^2(T\mu)$. Recall that, by definition, $\psi_k = \sum_{j=1}^d \phi_j \eta_{kj}$ and $\hat{\psi}_k = \sum_{j=1}^d \hat{\phi}_j \hat{\eta}_{kj}$.

Lemma A.4.1. *The following inequality holds:*

$$\|\hat{\psi}_k \delta_\Gamma \psi_k\|_\mu^2 \lesssim \underbrace{\|\hat{\eta}_k - a_k\|_2^2}_{\text{Term I}} + \underbrace{\|a_k - \eta_k\|_2^2}_{\text{Term II}} + \underbrace{\left(\|\eta_k\|_\infty \sum_{j=1}^d \left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu\right)^2}_{\text{Term III}}. \tag{A.164}$$

Remark 15. *Term II, that is $\|a_k - \eta_k\|_2^2$, captures differences between the population CCA problem described in equations (A.160) and that in (A.163). The non-random nature of this term complicates the analysis, as it requires deriving bounds for the expectation rather than establishing probability bounds.*

The first term, $\|\hat{\eta}_k - a_k\|_2$, will be bounded using our multivariate CCA arguments. The second and third terms will be bounded in the following section.

A.4.1.1 Bounding Terms II and III of Lemma A.4.1

Assuming that Σ_Y and Σ_Z are invertible, and from the definitions of a_k and η_k as the solutions to the problems in equations (A.163) and (A.160) respectively, we have that

$$a_k = \Sigma_Z^{-1/2} \tilde{a}_k, \tag{A.165}$$

$$\eta_k = \Sigma_Y^{-1/2} \tilde{\eta}_k, \tag{A.166}$$

where $\tilde{a}_k$ is the kth eigenvector (unit vector) of $A^\top A$ and $\tilde{\eta}$ is the kth eigenvector (unit vector) of $C^\top C$, where we have

$$A = \Sigma_{X_S}^{-1/2} \Sigma_{X_S Z} \Sigma_Z^{-1/2}, \tag{A.167}$$

$$C = \Sigma_{X_S}^{-1/2} \Sigma_{X_S Y} \Sigma_Y^{-1/2}. \tag{A.168}$$

Applying inequality 17 in Section A.6, we have that

$$\|a_k - \eta_k\|_2^2 \lesssim \left\|\Sigma_Z^{-1/2} - \Sigma_Y^{-1/2}\right\|_2^2 + \left\|\Sigma_Y^{-1}\right\|_2 \|\tilde{a}_k - \tilde{\eta}_k\|_2^2. \tag{A.169}$$

Noting that $A^\top A$ has the same eigenvectors as $|A|$, where $|A| \equiv (A^\top A)^{1/2}$, we can apply the Davis-Kahan theorem (Corollary 3 of Yu et al. (2015)) and obtain

$$\|\tilde{a}_k - \tilde{\eta}_k\|_2 \lesssim \frac{\||A| - |C|\|_2}{\min(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1})}, \qquad k = 1, \dots K, \tag{A.170}$$

where we assume $\tilde{a}_k^\top \tilde{\eta}_k \geq 0$ and we use the conventions $\gamma_{K+1} = -\infty$, and $\gamma_0 = \infty$. Next, note that

$$\||A| - |C|\|_2 \leq \||A| - |C|\|_F \leq \sqrt{2}\,\|A - C\|_F\,, \tag{A.171}$$

where the second inequality is identity 19 from Section A.6. We can then bound $\|A - C\|_F^2$ in terms of $\left\|\Sigma_Z^{-1/2} - \Sigma_Y^{-1/2}\right\|_2^2$ and $\mathbb{E}\left[\|Y - Z\|_2^2\right]$ using the following lemma.

Lemma A.4.2. *Under the group s-sparsity assumptions on b_k and θ_k, we have that*

$$\|A - C\|_F^2 \leq 2s\left[\mathbb{E}\left[\|Z\|_2^2\right]\left\|\Sigma_Z^{-1/2} - \Sigma_Y^{-1/2}\right\|_2^2 + \left\|\Sigma_Y^{-1}\right\|_2 \mathbb{E}\left[\|Y - Z\|_2^2\right]\right]. \tag{A.172}$$

Next, we bound $\left\|\Sigma_Z^{-1/2} - \Sigma_Y^{-1/2}\right\|_2^2$ in terms of $\mathbb{E}\left[\|Y - Z\|_2^2\right]$ using the following lemma.

Lemma A.4.3. *It can be shown that*

$$\left\|\Sigma_Z^{-1/2} - \Sigma_Y^{-1/2}\right\|_2^2 \lesssim \mathbb{E}\left[\|Z - Y\|_2^2\right] \max\left(\left\|\Sigma_Y^{-1}\right\|_2, \left\|\Sigma_Z^{-1}\right\|_2\right)^3 \max\left(\mathbb{E}\left[\|Z\|_2^2\right], \mathbb{E}\left[\|Y\|_2^2\right]\right). \tag{A.173}$$

Additionally, we have that

$$\|\Sigma_Y - \Sigma_Z\|_2^2 \lesssim \max\left(\mathbb{E}\left[\|Z\|_2^2\right], \mathbb{E}\left[\|Y\|_2^2\right]\right) \mathbb{E}\left[\|Z - Y\|_2^2\right]. \tag{A.174}$$

Combining the results of this section we obtain the following bound on $\|a_k - \eta_k\|_2^2$ in terms of $\mathbb{E}\left[\|Z - Y\|_2^2\right]$.

Lemma A.4.4. *Under the group s-sparsity assumptions on b_k and θ_k, we have*

$$\begin{aligned}\|a_k - \eta_k\|_2^2 \lesssim\ & \mathbb{E}\left[\|Z - Y\|_2^2\right] \\ & \times \left[\left(1 + \frac{s\left\|\Sigma_Y^{-1}\right\|_2 \mathbb{E}\left[\|Z\|_2^2\right]}{\min(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1})^2}\right) \max\left(\left\|\Sigma_Y^{-1}\right\|_2, \left\|\Sigma_Z^{-1}\right\|_2\right)^3 \max\left(\mathbb{E}\left[\|Z\|_2^2\right], \mathbb{E}\left[\|Y\|_2^2\right]\right)\right. \\ & \left. + \frac{s\left\|\Sigma_Y^{-1}\right\|_2^2}{\min(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1})^2}\right]\end{aligned}$$

Additionally, if we assume that $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$, $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$, *and* $\|\Sigma_Y^{-1}\|_2, \mathbb{E}\left[\|Y\|_2^2\right] \geq 1$, *then the statement simplifies as follows:*

$$\|a_k - \eta_k\|_2^2 \lesssim \mathbb{E}\left[\|Z - Y\|_2^2\right] \frac{s\,\|\Sigma_Y^{-1}\|_2^4\, E\left[\|Y\|_2^2\right]^2}{\min\left(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1}\right)^2}. \tag{A.175}$$

Remark 16. *In subsequent discussions, we will detail the conditions necessary for the additional assumptions stated here to hold.*

Having established a bound for $\|a_k - \eta_k\|_2^2$ in terms of $\mathbb{E}\left[\|Z - Y\|_2^2\right]$, we now turn our attention to bounding $\mathbb{E}\left[\|Z - Y\|_2^2\right]$ using 'known' quantities. In the process, we will derive a bound for $\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu\right]$, which will enable us to derive a probabilistic bound. This, in turn, will be used to bound Term III in Lemma A.4.1.

To establish a bound for $\mathbb{E}\left[\|Z - Y\|_2^2\right]$, we first begin by introducing a lemma to bound $\|Z - Y\|_2^2$.

Lemma A.4.5. *It can be shown that*

$$\|Z - Y\|_2^2 \leq 2d \left\|\mathrm{Log}_{\hat{\mu}}\, y_i \delta_\Gamma \mathrm{Log}_\mu\, y_i\right\|_\mu^2 + 2 \sum_{j=1}^{d} \left\|\mathrm{Log}_\mu\, y_i\right\|_\mu^2 \left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2. \tag{A.176}$$

Next, we aim to establish a bound for $\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2$. Consider an operator $\hat{\mathcal{C}}$ on $L^2(T\hat{\mu})$. We use parallel transport to define $\Phi\hat{\mathcal{C}}$ as the operator on $L^2(T\mu)$ such that $\Phi\hat{\mathcal{C}}(V) = \Gamma_{\hat{\mu},\mu}\left(\hat{\mathcal{C}}\left(\Gamma_{\mu,\hat{\mu}} V\right)\right) \in L^2(T\mu)$, for every $V \in L^2(T\mu)$. We also define the operator $\hat{\mathcal{C}}_\mu \equiv \frac{1}{N}\sum_{i=1}^N \mathrm{Log}_\mu\, y_i \otimes \mathrm{Log}_\mu\, y_i$ on $L^2(T\mu)$. Moreover, we use $\|\cdot\|_{\mathrm{op}}$ to denote the operator norm on $L^2(T\mu)$.

Lemma A.4.6. *For any* $j \geq 1$, *we have that*

$$\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2 \lesssim \frac{\|\mathcal{C} - \hat{\mathcal{C}}\|_{\mathrm{op}}^2 + \|\hat{\mathcal{C}}_\mu - \Phi\hat{\mathcal{C}}\|_{\mathrm{op}}^2}{\min\left(\omega_{j-1} - \omega_j, \omega_j - \omega_{j+1}\right)^2}. \tag{A.177}$$

We introduce the following lemma to bound the expectation of the terms in the previous lemma.

Lemma A.4.7. *If* $\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right] < \infty$, *then*

$$\mathbb{E}\left[\|\mathcal{C} - \hat{\mathcal{C}}\|_{\mathrm{op}}^2\right] \leq \frac{1}{N}\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]. \tag{A.178}$$

Additionally, under the assumption that $\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^2\right]$, *we have*

$$\mathbb{E}\left[\|\hat{\mathcal{C}}_\mu - \Phi\hat{\mathcal{C}}\|_{\mathrm{op}}^2\right] \lesssim \left(\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_i\right\|_\mu^4\right]^{1/2} + \mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^2\right]\right)\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^2\right]. \tag{A.179}$$

The lemma above shows that, in order to bound $\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2\right]$, it is first necessary to bound $\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^2\right]$. To do this, we need the following more technical results. But first, we state some preliminary definitions. Let $T\mathcal{M}$ denote the tangent bundle of $\mathcal{M}$, and let ∇ denote the Riemannian connection on $\mathcal{M}$. The next result, which is a mean value theorem for the parallel transport operation, is used in the proof of Lemma A.4.9.

Lemma A.4.8. *For a smooth vector field* $U : \mathcal{M} \to T\mathcal{M}$, $x, y \in \mathcal{M}$, *with minimizing geodesic* $\gamma(t)$ *between* x *and* y *(so that* $\gamma(0) = x$ *and* $\gamma(d(x,y)) = y$*), we have that*

$$\|\mathcal{P}_{y,x}U(y) - U(x)\|_x \leq d(y,x) \sup_{c\in[0,d(x,y)]} \|\nabla_{\gamma'(c)}U(\gamma(c))\|_{\gamma(c)}. \tag{A.180}$$

Remark 17. *When the vector field* U *has bounded Hessian* H*, defined below, we can use this result to bound the parallel transport error by the geodesic distance* $d(x,y)$ *between the base points of the vector field.*

For $x \in \mathcal{M}$ and $t \in \mathcal{T}$, define $f_t(x) \equiv \frac{1}{2}d^2(x, y_1(t))$. Let H_t be the Riemannian Hessian of f_t, i.e. $H_t(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ such that for all $x \in \mathcal{M}$, $t \in \mathcal{T}$, and $v \in T_x\mathcal{M}$, $H_t(x)(v) = \nabla_v \,\mathrm{grad}\, f_t(x)$. For a mapping $A(x)$, which for each x is an operator $A(x) : T_x\mathcal{M} \to T_x\mathcal{M}$, we define the operator norm at x of A: $\|A\|_{\mathrm{op},x} \equiv \sup_{v\in T_x\mathcal{M}, \|v\|_x=1} \|A(x)(v)\|_x$. Recall that $\tau = |\mathcal{T}|$ is the length of the time interval of the functional data.

Lemma A.4.9. *Assume that*

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

2. $\sup_{t\in\mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.

3. $\sup_{t\in\mathcal{T}, x\in\mathcal{M}} \|H_t(x)\|_{\mathrm{op},x}^2 \lesssim 1$ *with probability* 1.

Then,

$$\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}}\, y_i \delta_\Gamma \,\mathrm{Log}_\mu\, y_i\right\|_\mu^2\right] \lesssim \frac{\tau}{N}. \tag{A.181}$$

The next lemma combines the results of Lemmas A.4.6, A.4.7, and A.4.9.

Lemma A.4.10. *Under the assumptions of Lemma A.4.9, and additionally assuming that* $\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}}\, y_i \delta_\Gamma \,\mathrm{Log}_\mu\, y_i\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}}\, y_i \delta_\Gamma \,\mathrm{Log}_\mu\, y_i\right\|_\mu^2\right]$, *and* $\mathbb{E}\left[\left\|\mathrm{Log}_\mu\, y_1\right\|_\mu^4\right] \geq 1$, *we have*

$$\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2\right] \lesssim \frac{1}{N} \frac{\tau \mathbb{E}\left[\left\|\mathrm{Log}_\mu\, y_1\right\|_\mu^4\right]}{\min\left(\omega_{j-1} - \omega_j, \omega_j - \omega_{j+1}\right)^2}. \tag{A.182}$$

From Lemma A.4.10 and Markov's inequality, we obtain the following inequality, which we can use to bound Term III in Lemma A.4.1.

Corollary A.4.1. *Under the assumptions of Lemma A.4.10, we have that*

$$\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2 = O_P\left(\frac{1}{N} \frac{\tau \mathbb{E}\left[\left\|\mathrm{Log}_\mu\, y_1\right\|_\mu^4\right]}{\min\left(\omega_{j-1} - \omega_j, \omega_j - \omega_{j+1}\right)^2}\right). \tag{A.183}$$

Now we can combine the results of Lemmas A.4.5, A.4.9 and A.4.10 to obtain a bound on $\mathbb{E}\left[\|Z - Y\|^2\right]$.

Lemma A.4.11. *Under the assumptions of Lemma A.4.10 and additionally assuming that* $\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2\right]$, *we have that*

$$\mathbb{E}\left[\|Z - Y\|_2^2\right] \lesssim \frac{\tau d}{N} \mathbb{E}\left[\left\|\mathrm{Log}_\mu\, y_1\right\|_\mu^4\right]^{3/2} \max_{j=1,\ldots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2. \tag{A.184}$$

The proof follows from applying the Cauchy-Schwarz inequality and collecting similar terms.

Now we can state the conditions under which the additional assumptions of Lemma A.4.4 will hold. From now on, we keep tracking the terms τ and $\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]$ in the error bounds, but we assume they are constant.

Lemma A.4.12. *We have that* $\|\Sigma_Y^{-1}\|_2 = \omega_d^{-1}$, $\|\Sigma_Y\|_2 = \omega_1$, *and* $\mathbb{E}\left[\|Y\|_2^2\right] = \sum_{j=1}^d \omega_j$. *Additionally, if* $\mathbb{E}\left[\|Y\|_2^2\right], \|\Sigma_Y^{-1}\|_2 \geq 1$, *and*

$$\frac{d}{N}\frac{\sum_{j=1}^d \omega_j}{\omega_d^2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2 = o(1), \tag{A.185}$$

then $\mathbb{E}\left[\|Z-Y\|_2^2\right] = o(1)$, $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$, $\|\Sigma_Z\|_2 \lesssim \|\Sigma_Y\|_2$ *and* $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$.

Now we can combine Lemmas A.4.4, A.4.11, and A.4.12 to obtain a final bound on $\|a_k - \eta_k\|_2^2$.

Theorem A.4.1. *Under the assumptions of Lemmas A.4.11 and A.4.12, we have that*

$$\|a_k - \eta_k\|_2^2 \lesssim \frac{\tau s d}{N}\max_{j=1,\dots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2\left(\frac{\sum_{j=1}^d \omega_j}{\omega_d^2}\right)^2 \frac{\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}}{\min\left(\gamma_{k-1}-\gamma_k, \gamma_k - \gamma_{k+1}\right)^2}. \tag{A.186}$$

Before we combine the bounds for the three components detailed in Lemma A.4.1, we show that $\tilde{\gamma}_k$ and γ_k are asymptotically equivalent, which allows us to simplify the expression of our final bounds. Recall that the correlations $\{\tilde{\gamma}_k\}$ are defined using equation (A.163) as the canonical correlations between X and Z, while the correlations $\{\gamma_k\}$ are the canonical correlations between X and Y. To this end, we first state the following bound, which follows directly from Lemmas A.4.2 and A.4.3. This will also be used later to derive a bound for $\left\|\theta_k - \hat{\theta}_k\right\|_2$.

Lemma A.4.13. *Under the assumptions of Lemma A.4.12, we have that*

$$\|A - C\|_F^2 \lesssim s\left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]^2 \mathbb{E}\left[\|Z-Y\|_2^2\right]. \tag{A.187}$$

We can prove the next lemma by using Lemmas A.4.11, A.4.12, and A.4.13.

Lemma A.4.14. *Under the assumptions of Lemmas A.4.11 and A.4.12, and further assuming that* τ *and* $\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]$ *are absolute constants, that the canonical correlations* $\{\gamma_k\}$ *and* $\{\tilde{\gamma}_k\}$ *are bounded from below and that*

$$\frac{ds}{N}\frac{\left(\sum_{j=1}^d \omega_j\right)^2}{\omega_d^3}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2 = o(1), \tag{A.188}$$

then γ_k *and* $\tilde{\gamma}_k$ *are asymptotically equivalent, i.e.,* $\gamma_k \lesssim \tilde{\gamma}_k$ *and* $\tilde{\gamma}_k \lesssim \gamma_k$*. Additionally,* $\gamma_k^2 \lesssim \tilde{\gamma}_k^2$ *and* $\tilde{\gamma}_k^2 \lesssim \gamma_k^2$*.*

Now we are now in a position to establish bounds for all three terms in Lemma A.4.1, using Theorem A.3.3 (applied to X and Z and using the bound for η), Theorem A.4.1, and Corollary A.4.1 for Term I, II, and III, respectively. This will yield our final bound for the canonical functions $\|\hat{\psi}_k\delta_\Gamma\psi_k\|_\mu^2$, which is presented in Section A.4.3.

A.4.2 Bounding high-dimensional canonical vector error

In this section, we derive a bound for the estimation error of the high-dimensional canonical vectors, $\left\|\theta_k-\hat{\theta}_k\right\|_2$. Having already derived a bound for the canonical functions, the proof is straightforward. We start with an application of the triangle inequality. For $k=1,\dots K$,

$$\left\|\theta_k-\hat{\theta}_k\right\|_2^2 \lesssim \|\theta_k-b_k\|_2^2+\left\|b_k-\hat{\theta}_k\right\|_2^2, \tag{A.189}$$

where b_k is the high dimensional canonical vector given by the solution to problem (A.163). Assuming s-group sparsity for θ_k and b_k, we represent the associated vectors with non-zero entries as $\theta_{k,S}$ and $b_{k,S}$. Recall that these are at most $2s$-dimensional. By definition, we then have $\|\theta_k-b_k\|_2=\|\theta_{k,S}-b_{k,S}\|_2$, hence

$$\left\|\theta_k-\hat{\theta}_k\right\|_2^2 \lesssim \|\theta_{k,S}-b_{k,S}\|_2^2+\left\|b_k-\hat{\theta}_k\right\|_2^2. \tag{A.190}$$

To bound the second term, we can use Theorem A.3.3, applied to the random vectors X and Z. To bound the second term, we can make the following argument which is similar to the

one made to bound $\|\eta_k - a_k\|_2$. Let

$$A = \Sigma_{X_S}^{-1/2} \Sigma_{X_S Z} \Sigma_Z^{-1/2}, \tag{A.191}$$

$$C = \Sigma_{X_S}^{-1/2} \Sigma_{X_S Y} \Sigma_Y^{-1/2}. \tag{A.192}$$

Then, from classical CCA (Uurtio et al., 2018), we have that $\theta_{k,S} = \Sigma_{X_S}^{-1/2} \tilde{\theta}_{k,S}$, where $\tilde{\theta}_{k,S}$ is the kth eigenvector of $CC^\top$, and $b_k = \Sigma_{X_S}^{-1/2} \tilde{b}_{k,S}$, where $\tilde{b}_{k,S}$ is the kth eigenvector of $AA^\top$. Therefore,

$$\|\theta_{k,S} - b_{k,S}\|_2 = \left\|\Sigma_{X_S}^{-1/2} \tilde{\theta}_{k,S} - \Sigma_{X_S}^{-1/2} \tilde{b}_{k,S}\right\|_2 \tag{A.193}$$

$$= \left\|\Sigma_{X_S}^{-1/2} \left(\tilde{\theta}_{k,S} - \tilde{b}_{k,S}\right)\right\|_2 \tag{A.194}$$

$$\leq \left\|\Sigma_{X_S}^{-1/2}\right\|_2 \left\|\tilde{\theta}_{k,S} - \tilde{b}_{k,S}\right\|_2. \tag{A.195}$$

Since $A^\top A$ has the same eigenvectors as $|A| = (AA^\top)^{1/2}$, we can then use the Davis-Kahan theorem (Yu et al., 2015) to derive the following bound. If $\tilde{\theta}_{k,S}^\top \tilde{b}_{k,S} \geq 0$, then

$$\left\|\tilde{\theta}_{k,S} - \tilde{b}_{k,S}\right\|_2 \lesssim \frac{\||A^\top| - |C^\top|\|_2}{\min\left(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1}\right)}. \tag{A.196}$$

The term $\||A^\top| - |C^\top|\|_2$ can be bounded as follows:

$$\||A^\top| - |C^\top|\|_2 \leq \||A^\top| - |C^\top|\|_F \leq \sqrt{2}\,\|A^\top - C^\top\|_F = \sqrt{2}\,\|A - C\|_F, \tag{A.197}$$

where the second inequality is identity 19 from Section A.6. Having established a bound for $\left\|\tilde{\theta}_{k,S} - \tilde{b}_{k,S}\right\|_2$ in terms of $\|A - C\|_F$, we can apply similar arguments to those used to bound $\|a_k - \eta_k\|_2$ in order to derive a bound for $\left\|\tilde{\theta}_{k,S} - \tilde{b}_{k,S}\right\|_2$ in terms of $\mathbb{E}\left[\|Z - Y\|_2^2\right]$. Combining the results of this section, using Lemmas A.4.11 and A.4.13, we establish the following result.

Lemma A.4.15. *Under the assumptions of Lemmas A.4.11 and A.4.12, and if* $\tilde{\theta}_{k,S}^\top \tilde{b}_{k,S} \geq 0$, *we have that*

$$\|\theta_{k,S} - b_{k,S}\|_2^2 \lesssim \frac{\left\|\Sigma_{X_S}^{-1/2}\right\|_2^2 s \left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]^2}{\min\left(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1}\right)^2} \frac{\tau d}{N} \mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]^{3/2} \max_{j=1,\ldots d} \left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2. \tag{A.198}$$

Next, by applying Theorem A.3.3, Lemma A.4.15, and equation (A.190), we can establish the final bound for the high-dimensional canonical vector error $\left\|\theta_k - \hat{\theta}_k\right\|_2^2$, which is presented in Section A.4.3.

A.4.3 Final rates

In this section, we present our final results. Recall that we denote as $K = \max\{i \in \{1, \dots d^{(\text{corr})}\} : \gamma_i > 0\}$ the number of nontrivial canonical vectors and we use the conventions $\gamma_{K+1}^2 = -\infty$ and $\gamma_0^2 = +\infty$. The random vector Z was defined in equation (A.162) and represents a 'sample' version of Y where ϕ_j and μ are replaced by their finite-sample estimates.

For clarity, we first provide a comprehensive list of our assumptions. For the definitions of the quantities that appear below, please see the beginning of Section A.4.

Assumption A.4.1 (Manifold Properties)**.**

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

2. *The curvature $\sup_{t \in \mathcal{T}, x \in \mathcal{M}} \|H_t(x)\|_{\text{op},x}^2$ is bounded with probability 1.*

Assumption A.4.2 (Distributional Assumptions)**.**

1. *The random vectors X and Z are strict sub-Gaussian random vectors (Definition A.3.2).*

2. *The covariance matrices Σ_X, Σ_Y, and Σ_Z are invertible.*

3. *The group s-sparsity assumption holds for $\{b_k\}$ and $\{\theta_k\}$.*

4. *The matrix $\Sigma_X^{1/2}$ satisfies the group restricted eigenvalue condition $RE(s, 3, d)$ (Definition A.3.3) with parameter $\kappa = \kappa(s, d, \Sigma_X^{1/2})$.*

5. *The functional data are such that* $\sup_{t\epsilon\mathcal{T}} \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right] < \infty$.

Assumption A.4.3 (Rate Assumptions)**.**

1. *There exists a complete orthonormal system for* $L^2(T\mu)$, $\{\phi_j\}_{j=1}^{\infty}$, *such that a* $d^{(corr)}$*-dimensional Assumption 2.2.1 holds with* $p \geq d \geq d^{(corr)}$, *where* d *is the chosen number of principal components in Algorithm 1;*

2. $\operatorname{cond}\left(\Sigma_Y\right)^2 d = o(N)$;

3. $\kappa^2 s^2 d \log(p) = o(N)$;

4. $ds\frac{\left(\sum_{j=1}^{d}\omega_j\right)^2}{\omega_d^3}\max_{j=1,\ldots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2 = o(N)$;

5. *The correlations* $\gamma_1, \ldots \gamma_K$ *are bounded from below and distinct from one another, as well as* $\tilde{\gamma}_1, \ldots \tilde{\gamma}_K$.

6. $\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2\right]$;

7. $\mathbb{E}\left[\left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu^2\right]$.

Assumption A.4.4 (Minor Assumptions)**.**

1. *The quantities* $\|\Sigma_X\|_{2,\infty}, \|T\|_{\ell_1,\ell_2}$ *are bounded from above and are* ≥ 1.

2. *The variables* τ *and* $\mathbb{E}\left[\left\|\operatorname{Log}_\mu y_1\right\|_\mu^4\right]$ *are constants.*

3. *The following quantities are larger than 1:* $\kappa, \omega_1, \omega_d^{-1}, \|\eta\|_\infty, \|\Sigma_X^{-1}\|_2, \|\Sigma_Z^{-1}\|_2$.

4. $a_k^\top \Sigma_Z^{1/2} \Sigma_Y^{1/2} \eta_k \geq 0$ *and* $\hat{\eta}_k^\top \hat{\Sigma_Z}^{1/2} \Sigma_Z^{1/2} a_k \geq 0$ *for* $k = 1, \ldots K$.

Next, we present our main results.

Theorem A.4.2 (Canonical Function Error Bound). *Under Assumptions A.4.1-A.4.4, we have*

$$\|\hat{\psi}_k \delta_\Gamma \psi_k\|_\mu^2 = O_P\left(\frac{d^2 s \log(p)}{N}\right) \frac{\tau}{\min_{j\neq k} \min\left\{|{\gamma_k}^2 - {\gamma_j}^2|, |\gamma_k - \gamma_j|\right\}^2} \tag{A.199}$$

$$\cdot \left(\frac{\sum_{j=1}^d \omega_j}{\omega_d^2}\right)^2 \max_{j=1,\ldots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2 \kappa \|\Sigma_X\|_{2,\infty} \|\eta\|_\infty^2 \mathbb{E}\left[\left\|\mathrm{Log}_\mu y\right\|_\mu^4\right]^{3/2}, \tag{A.200}$$

with $k = 1, \ldots K$.

Theorem A.4.3 (Canonical Vector Error Bound). *Under Assumptions A.4.1-A.4.4, and additionally assuming that* $\theta_{k,S}^\top \Sigma_{X_S} b_{k,S} \geq 0$, *we have*

$$\left\|\theta_k - \hat{\theta}_k\right\|_2^2 = O_P\left(\frac{ds\log(p)}{N}\right) \frac{\tau}{\min_{j\neq k} \min\left\{|{\gamma_k}^2 - {\gamma_j}^2|, |\gamma_k - \gamma_j|\right\}^2} \tag{A.201}$$

$$\cdot \left[\frac{\left(\sum_{j=1}^d \omega_j\right)^2}{\omega_d^3} \max_{j=1,\ldots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2 \mathbb{E}\left[\left\|\mathrm{Log}_\mu y\right\|_\mu^4\right]^{3/2} \left\|\Sigma_{X_S}^{-1/2}\right\|_2 + \left(\frac{\gamma_1}{\gamma_k}\right)^2 \|\Sigma_X\|_{2,\infty} \kappa^2 \left\|\Sigma_X^{-1}\right\|_2\right], \tag{A.202}$$

with $k = 1, \ldots K$.

Here we make a few remarks on the assumptions of Theorems A.4.2 and A.4.3. Some of the more technical assumptions arise from avoiding overly simplifying assumptions. For instance, with the exception of the curvature-related quantity $H_t(x)$, we do not assume that the random variables/functions are bounded. Furthermore, we avoid assuming Gaussianity of the random variables of interest and instead assume these are sub-Gaussians. As in Lin and Yao (2019), we do not assume that the Frechét mean μ is known and instead estimate it using its sample version $\hat{\mu}$; this choice introduces significant complexity, making it necessary to use the parallel transport operator to compare estimates and estimands, which are defined in different tangent spaces. These challenges are further compounded by the dimensionality reduction step that takes place before CCA, which requires that we derive bounds in *expectation* rather than in probability. Specifically, this requires showing that $\mathbb{E}\left[d\left(\hat{\mu}(t), \mu(t)\right)^2\right] \lesssim \frac{1}{N}$, i.e., equation (A.253), using the results of Schötz (2019) as opposed to those of Lin and Yao (2019). See also Remark 15.

Remarks on Assumption A.4.1:

Assumption A.4.1 is required to bound the term $\mathbb{E}\left[d\left(\hat{\mu}(t), \mu(t)\right)^2\right]$. See also Lemma A.4.9. Here, $H_t(x)$ is the Riemannian Hessian of the random function $d^2(x, y(t))$, and is related to curvature on the manifold (Pennec, 2017). Assuming this quantity is bounded allows us to bound the parallel transport distance in terms of the geodesic distance, as shown in Lemma A.4.8.

Remarks on Assumption A.4.2:

Items 1-4 of Assumption A.4.2 are used to facilitate the application of our multivariate CCA results in Section 2.4.1 to the sparse-functional setting considered here. Specifically, Items 3-4 are used in particular to get fast-rate bounds, which match the root-n estimation rate of the functional quantities. We note that in item 3, we do not require that θ_k and b_k have the same sparsity structure, but only that they are both s-sparse. Item 4 is a generalization of the standard restricted-eigenvalue condition in Lasso theory (Hastie et al., 2015) and is equivalent to the one proposed in Gaynanova (2020). Item 5 is a weak assumption about the boundedness of the variance of $y(t)$ on the manifold and along with Assumption A.4.1 is necessary to show that $\mathbb{E}\left[d\left(\hat{\mu}(t), \mu(t)\right)^2\right]$ is root-n consistent.

Remarks on Assumption A.4.3:

Item 1 of Assumption A.4.3, that the correlation with the high-dimensional data is captured in a finite-dimensional subspace of the functional data, is necessary to even define the canonical directions for functional data, as stated in the main manuscript. The condition that $d \leq p$ formalizes our asymmetrical treatment of the data, and allows for the CCA problem to become a sparse regression problem. $d \geq d^{(\text{corr})}$ is necessary to ensure we capture all of the components of $\text{Log}_\mu y$ which are correlated with X. Items 2 - 5 are mainly used to simplify the theorem statements by allowing us to bound norms of estimated quantities using the corresponding population quantities. In particular, Item 2 is used in conjunction with Lemma A.4.12 to show that $\left\|\hat{\Sigma}_Z\right\|_2 \lesssim \left\|\Sigma_Y\right\|_2$. Item 3 allows us to only make a group restricted eigenvalue assumption on $\Sigma_X^{1/2}$ rather than the data matrix $\frac{1}{\sqrt{N}}\mathbb{X}$ (see Lemma A.3.9). Item 4 states that the variances ω_j of the principal scores Y_j should not shrink too quickly as

N and d grow. Note that if d is assumed constant, the condition reduces to $s = o(N)$. This is used to establish the simplifying bounds given in Lemmas A.4.12 and A.4.14. Item 5 is used to replace $\hat{\gamma}_k$ and $\tilde{\gamma}_k$ with the population canonical correlations γ_k (see Lemma A.4.14 and the discussion in the proof of Theorem A.3.2). Items 6 and 7 are technical conditions. These mainly arise from the complexity of the setting considered here. The assumption $\mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^4\right]^{1/2} \lesssim \mathbb{E}\left[\left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2\right]$ could be replaced with a boundness assumption on $\left\|\mathrm{Log}_\mu y_i\right\|_\mu$, or alternatively, by adopting a sample splitting strategy to estimate $\hat{\mu}$, the principal components $\{\hat{\phi}_j\}$, and to carry out CCA. This can be seen from the second term of equation (A.176), where the absence of one of these assumptions requires that we use the Cauchy-Schwarz inequality, introducing fourth moments. Note that if we were to assume that $\hat{\mu} = \mu$, then $\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}} y_i \delta_\Gamma \mathrm{Log}_\mu y_i\right\|_\mu^4\right]^{1/2} = 0$, immediately satisfying the condition in item 7.

Remarks on Assumption A.4.4:

Items 1-3 are not critical and only serve the purpose of simplifying the theorem statements. Item 4 is introduced to account for the sign ambiguity of the CCA solutions.

A.4.4 Proofs for results in Section A.4

Proof of Lemma A.4.1:

We define $\tilde{\psi}_k = \sum_{j=1}^d \hat{\phi}_j a_{kj}$. For ease of notation, we drop the k in writing $\hat{\psi}_k$, ψ_k and $\tilde{\psi}_k$, η_k, etc. We have

$$\|\hat{\psi}\delta_\Gamma\psi\|_\mu^2 = \left\|\Gamma_{\hat{\mu},\mu}\hat{\psi} - \psi\right\|_\mu^2 \leq 2\left\|\Gamma_{\hat{\mu},\mu}\hat{\psi} - \Gamma_{\hat{\mu},\mu}\tilde{\psi}\right\|_\mu^2 + 2\left\|\Gamma_{\hat{\mu},\mu}\tilde{\psi} - \psi\right\|_\mu^2. \tag{A.203}$$

The first term in equation (A.203) is

$$2\left\|\Gamma_{\hat{\mu},\mu}\left(\hat{\psi} - \tilde{\psi}\right)\right\|_\mu^2 = \left\|\Gamma_{\mu,\hat{\mu}}\left(\Gamma_{\hat{\mu},\mu}\left(\hat{\psi} - \tilde{\psi}\right)\right)\right\|_{\hat{\mu}}^2 = 2\left\|\hat{\psi} - \tilde{\psi}\right\|_{\hat{\mu}}^2.$$

Define $\bar{\psi}$ as $\sum_{j=1}^d \hat{\phi}_j \eta_j$. Then the second term in equation (A.203) is

$$\leq 4\left\|\Gamma_{\hat{\mu},\mu}\tilde{\psi} - \Gamma_{\hat{\mu},\mu}\bar{\psi}\right\|_\mu^2 + 4\left\|\Gamma_{\hat{\mu},\mu}\bar{\psi} - \psi\right\|_\mu^2.$$

Therefore,

$$\|\hat{\psi}\delta_\Gamma\psi\|_\mu^2 \lesssim \left\|\hat{\psi}-\tilde{\psi}\right\|_{\hat{\mu}}^2 + \left\|\tilde{\psi}-\bar{\psi}\right\|_{\hat{\mu}}^2 + \left\|\bar{\psi}\delta_\Gamma\psi\right\|_\mu^2. \tag{A.204}$$

The first term in equation (A.204) is

$$\left\|\hat{\psi}-\tilde{\psi}\right\|_{\hat{\mu}}^2 = \left\|\sum_{j=1}^d \hat{\phi}_j\hat{\eta}_j - \sum_{j=1}^d a_j\hat{\phi}_j\right\|_{\hat{\mu}}^2 = \left\|\sum_{j=1}^d \hat{\phi}_j\left(\hat{\eta}_j - a_j\right)\right\|_{\hat{\mu}}^2 = \|\hat{\eta}-a\|_2^2, \tag{A.205}$$

where in the third equality we have used that the $\hat{\phi}_j$ are orthonormal in $L^2(T\hat{\mu})$. Similarly, the second term is

$$\left\|\tilde{\psi}-\bar{\psi}\right\|_{\hat{\mu}}^2 = \|a-\eta\|_2^2. \tag{A.206}$$

By the triangle inequality, the third term is

$$\left\|\bar{\psi}\delta_\Gamma\psi\right\|_\mu^2 \le \left(\|\eta\|_\infty \sum_{j=1}^d \left\|\hat{\phi}_j\delta_\Gamma\phi_j\right\|_\mu\right)^2, \tag{A.207}$$

completing the proof. □

Proof of Lemma A.4.2:

For ease of notation, we write X_S as X throughout the proof of the lemma. From the definition of A and C, we have

$$\|A-C\|_F = \left\|\Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2} - \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\right\|_F = \left\|\Sigma_X^{-1/2}\left(\mathbb{E}\left[XZ^\top\right]\Sigma_Z^{-1/2} - \mathbb{E}\left[XY^\top\right]\Sigma_Y^{-1/2}\right)\right\|_F. \tag{A.208}$$

From the linearity of expectation, this is

$$= \left\|\mathbb{E}\left[\Sigma_X^{-1/2}X\left(Z^\top\Sigma_Z^{-1/2} - Y^\top\Sigma_Y^{-1/2}\right)\right]\right\|_F = \left\|\mathbb{E}\left[\Sigma_X^{-1/2}X\left(\Sigma_Z^{-1/2}Z - \Sigma_Y^{-1/2}Y\right)^\top\right]\right\|_F. \tag{A.209}$$

By Theorem 2.6.7 of Hsing and Eubank (2015), using the Frobenious norm of an outer product, and the Cauchy-Schwarz inequality, we have

$$\|A-C\|_F \le \mathbb{E}\left[\left\|\Sigma_X^{-1/2}X\left(\Sigma_Z^{-1/2}Z - \Sigma_Y^{-1/2}Y\right)^\top\right\|_F\right] \tag{A.210}$$

$$= \mathbb{E}\left[\left\|\Sigma_X^{-1/2}X\right\|_2\left\|\Sigma_Z^{-1/2}Z - \Sigma_Y^{-1/2}Y\right\|_2\right] \tag{A.211}$$

$$\le \mathbb{E}\left[\left\|\Sigma_X^{-1/2}X\right\|_2^2\right]^{1/2}\mathbb{E}\left[\left\|\Sigma_Z^{-1/2}Z - \Sigma_Y^{-1/2}Y\right\|_2^2\right]^{1/2}. \tag{A.212}$$

We have that $\mathbb{E}\left[\left\|\Sigma_X^{-1/2}X\right\|_2^2\right] = \mathbb{E}\left[\mathrm{tr}\left(\Sigma_X^{-1}XX^\top\right)\right] = \mathrm{tr}\left(I_s\right) = s$. Using this along with identity 17 from Section A.6 to upper bound $\mathbb{E}\left[\left\|\Sigma_Z^{-1/2}Z - \Sigma_Y^{-1/2}Y\right\|_2^2\right]^{1/2}$ completes the proof.

Proof of Lemma A.4.3:

The first statement follows from the second statement, by identity 14 from Section A.6 and Lemma A.4.3. To show the second statement, we begin with

$$\|\Sigma_Y - \Sigma_Z\|_2 = \|\mathbb{E}\left[YY^\top\right] - \mathbb{E}\left[ZZ^\top\right]\|_2 = \|\mathbb{E}\left[YY^\top - ZZ^\top\right]\|_2, \tag{A.213}$$

and using Theorem 2.6.7 of Hsing and Eubank (2015), this is

$$\leq \mathbb{E}\left[\|YY^\top - ZZ^\top\|_2\right] = \mathbb{E}\left[\|YY^\top - ZY^\top + ZY^\top - ZZ^\top\|_2\right] = \mathbb{E}\left[\left\|Z\left(Z-Y\right)^\top + \left(Z-Y\right)Y^\top\right\|_2\right]. \tag{A.214}$$

From this, the triangle inequality, and using the two-norm of an outer product, we have

$$\|\Sigma_Y - \Sigma_Z\|_2 \leq \mathbb{E}\left[\left\|Z\left(Z-Y\right)^\top\right\|_2\right] + \mathbb{E}\left[\|\left(Z-Y\right)Y^\top\|_2\right] = \mathbb{E}\left[\|Z\|_2\|Z-Y\|_2\right] + \mathbb{E}\left[\|Z-Y\|_2\|Y\|_2\right] \tag{A.215}$$

By the Cauchy-Schwarz inequality, the right-hand side is

$$= \mathbb{E}\left[\left(\|Z\|_2 + \|Y\|_2\right)\|Z-Y\|_2\right] \leq \mathbb{E}\left[\left(\|Z\|_2 + \|Y\|_2\right)^2\right]^{1/2}\mathbb{E}\left[\|Z-Y\|_2^2\right]^{1/2} \tag{A.216}$$

The second statement of the lemma follows, and the proof is complete. □

Proof of Lemma A.4.5:

Define $W \in \mathbb{R}^d$ as the random vector with $W_j \equiv \langle\!\langle \mathrm{Log}_\mu y_1, \Gamma_{\hat{\mu},\mu}\hat{\phi}_j\rangle\!\rangle_\mu$ for $j = 1, \ldots d$. Then

$$\|Z-Y\|_2^2 \leq 2\|Z-W\|_2^2 + 2\|W-Y\|_2^2. \tag{A.217}$$

We have

$$Z_j - W_j = \langle\!\langle \mathrm{Log}_{\hat{\mu}} y_1, \hat{\phi}_j\rangle\!\rangle_{\hat{\mu}} - \langle\!\langle \mathrm{Log}_\mu y_1, \Gamma_{\hat{\mu},\mu}\hat{\phi}_j\rangle\!\rangle_\mu \tag{A.218}$$

$$= \langle\!\langle \mathrm{Log}_{\hat{\mu}} y_1, \hat{\phi}_j\rangle\!\rangle_{\hat{\mu}} - \langle\!\langle \Gamma_{\mu,\hat{\mu}}\,\mathrm{Log}_\mu y_1, \hat{\phi}_j\rangle\!\rangle_{\hat{\mu}} \tag{A.219}$$

$$= \langle\!\langle \mathrm{Log}_{\hat{\mu}} y_1 - \Gamma_{\mu,\hat{\mu}}\,\mathrm{Log}_\mu y_1, \hat{\phi}_j\rangle\!\rangle_{\hat{\mu}}, \tag{A.220}$$

and therefore, by the Cauchy-Schwarz inequality, and because the $\hat{\phi}_j$ are orthonormal along $\hat{\mu}$,

$$|Z_j - W_j| \leq \|\mathrm{Log}_{\hat{\mu}}\, y_1 - \Gamma_{\mu,\hat{\mu}} \mathrm{Log}_{\mu}\, y_1\|_{\hat{\mu}} \|\hat{\phi}_j\|_{\hat{\mu}} \tag{A.221}$$

$$= \left\|\mathrm{Log}_{\mu}\, y_1 \delta_\Gamma \mathrm{Log}_{\hat{\mu}}\, y_1\right\|_{\hat{\mu}} \tag{A.222}$$

$$= \left\|\mathrm{Log}_{\hat{\mu}}\, y_1 \delta_\Gamma \mathrm{Log}_{\mu}\, y_1\right\|_{\mu}, \tag{A.223}$$

where in the last equality we have used that $\|U\delta_\Gamma V\|_\mu = \|V\delta_\Gamma U\|_{\hat{\mu}}$ for $U \in L^2(T\hat{\mu})$ and $V \in L^2(T\mu)$. We also have

$$W_j - Y_j = \langle\!\langle \mathrm{Log}_{\mu}\, y_1, \Gamma_{\hat{\mu},\mu}\hat{\phi}_j \rangle\!\rangle_\mu - \langle\!\langle \mathrm{Log}_{\mu}\, y_1, \phi_j \rangle\!\rangle_\mu \tag{A.224}$$

$$= \langle\!\langle \mathrm{Log}_{\mu}\, y_1, \hat{\phi}_j \delta_\Gamma \phi_j \rangle\!\rangle_\mu, \tag{A.225}$$

so that again by the Cauchy-Schwarz inequality,

$$|W_j - Y_j| \leq \left\|\mathrm{Log}_{\mu}\, y_1\right\|_\mu \left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu. \tag{A.226}$$

Thus,

$$\|Z - Y\|_2^2 \leq 2\sum_{j=1}^{d} |Z_j - W_j|^2 + |W_j - Y_j|^2 \tag{A.227}$$

$$\leq 2\sum_{j=1}^{d} \left\|\mathrm{Log}_{\hat{\mu}}\, y_1 \delta_\Gamma \mathrm{Log}_{\mu}\, y_1\right\|_\mu^2 + \left\|\mathrm{Log}_{\mu}\, y_1\right\|_\mu^2 \left\|\hat{\phi}_j \delta_\Gamma \phi_j\right\|_\mu^2, \tag{A.228}$$

from which the statement of the lemma follows, and the proof is complete. □

Proof of Lemma A.4.6:

From Lin and Yao (2019) (page 3551), $\Gamma_{\hat{\mu},\mu}\hat{\phi}_j$ are the eigenvectors of $\Phi\hat{\mathcal{C}}$. By definition, the ϕ_j are the eigenvectors of $\mathcal{C}$. Thus, we can use the Davis-Kahan Theorem for Hilbert spaces (Jirak & Wahl, 2020) to obtain that

$$\Gamma_{\hat{\mu},\mu}\hat{\phi}_j - \phi_j \leq 2\sqrt{2}\max\left((\omega_{j-1} - \omega_j)^{-1}, (\omega_j - \omega_{j+1})^{-1}\right)\|\mathcal{C} - \Phi\hat{\mathcal{C}}\|_{\mathrm{op}}, \tag{A.229}$$

where $\omega_0 \equiv \infty$ and $\|\mathcal{C}\|_{\text{op}} \equiv \sup_{U \in L^2(T\mu), \|U\|_\mu = 1} \|\mathcal{C}U\|_\mu$. Adding and subtracting $\hat{\mathcal{C}}_\mu$ in $\|\mathcal{C} - \Phi\hat{\mathcal{C}}\|_{\text{op}}$ and using identity 18 from Section A.6, the proof is complete. □

Proof of Lemma A.4.7:

The first statement is equivalent to Lemma 5.2 of Cardot et al. (1999). To show the second statement, we begin with the definition of $\hat{\mathcal{C}}$ and Proposition 2 item 5 of Lin and Yao (2019), from which we deduce that

$$\Phi\hat{\mathcal{C}} = \frac{1}{N}\sum_{i=1}^{N}\left(\Gamma_{\hat{\mu},\mu}\operatorname{Log}_{\hat{\mu}} y_i\right) \otimes \left(\Gamma_{\hat{\mu},\mu}\operatorname{Log}_{\hat{\mu}} y_i\right). \tag{A.230}$$

This implies

$$\hat{\mathcal{C}}_\mu - \Phi\hat{\mathcal{C}} = \frac{1}{N}\sum_{i=1}^{N} a_i \otimes a_i - b_i \otimes b_i, \tag{A.231}$$

where we denote $a_i \equiv \operatorname{Log}_\mu y_i$ and $b_i \equiv \Gamma_{\hat{\mu},\mu}\operatorname{Log}_{\hat{\mu}} y_i$. It is straightforward to show that $a \otimes a - b \otimes b = a \otimes (a-b) + (a-b) \otimes (b-a) + (a-b) \otimes a$, where $a, b \in L^2(T\mu)$. From Theorem 3.4.7. of Hsing and Eubank (2015), $\|a \otimes b\|_{\text{op}} = \|a\|_\mu \|b\|_\mu$, and therefore,

$$\|a \otimes a - b \otimes b\|_{\text{op}} \leq 2\|a\|_\mu \|a-b\|_\mu + \|a-b\|_\mu^2. \tag{A.232}$$

Then,

$$\|\hat{\mathcal{C}}_\mu - \Phi\hat{\mathcal{C}}_{\text{op}}\| \leq \frac{1}{N}\sum_{i=1}^{N} 2\left\|\operatorname{Log}_\mu y_i\right\|_\mu \left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu + \left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu^2, \tag{A.233}$$

so that

$$\|\hat{\mathcal{C}}_\mu - \Phi\hat{\mathcal{C}}_{\text{op}}\|^2 \lesssim \left(\frac{1}{N}\sum_{i=1}^{N} 2\left\|\operatorname{Log}_\mu y_i\right\|_\mu \left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu\right)^2 + \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\operatorname{Log}_{\hat{\mu}} y_i \delta_\Gamma \operatorname{Log}_\mu y_i\right\|_\mu^2\right)^2. \tag{A.234}$$

It is straightforward to show that, for any i.i.d. random variables W_i with finite variance we have

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} W_i\right)^2\right] \leq \mathbb{E}\left[W_1^2\right], \tag{A.235}$$

where the W_i are not required to have mean 0. Taking the expectation in equation (A.234), applying this last result on each term, using the Cauchy-Schwarz inequality, and

subsequently using the assumption stated in the Lemma concerning $\left\|\mathrm{Log}_{\hat{\mu}}\, y_i \delta_\Gamma \,\mathrm{Log}_\mu\, y_i\right\|_\mu$, the proof is complete. □

Proof of Lemma A.4.8:

We start with a mean value theorem result for smooth functions f from $[0,t_1]$ into a normed vector space V, where $t_1 \in \mathbb{R}$. By Theorem 1.1.1. of Hörmander (2015), we have

$$\|f(t_1) - f(0)\| = |b-a| \sup_{c\in[0,t_1]} f'(c). \tag{A.236}$$

We set

$$f(t) = \Gamma_{\gamma(t),\gamma(0)} U\left(\gamma(t)\right) - U\left(\gamma(0)\right), \tag{A.237}$$

where U is a smooth vector field on $\mathcal{M}$, $U : \mathcal{M} \to T\mathcal{M}$, and $\gamma(t)$ is the minimizing geodesic between two points $x, y \in \mathcal{M}$ ($\gamma(0) = x, \gamma\left((x,y)\right) = y$. Letting $t_1 = d(x,y)$, then $\gamma : [0,t_1] \to \mathcal{M}$, and $f : [0,t_1] \to T_x\mathcal{M}$. Letting $\|W\|_x$ denote the norm of $W \in T_x\mathcal{M}$, and using $f(t)$ in equation (A.236), we have

$$\|\Gamma_{y,x} U(y) - U(x)\| \leq d(x,y) \sup_{c\in[0,d(x,y)]} \|f'(c)\|. \tag{A.238}$$

We can determine $f'(c)$:

$$f'(c) = \lim_{t\to 0^+} \frac{f(c+t) - f(c)}{t} \tag{A.239}$$

$$= \lim_{t\to 0^+} \frac{\Gamma_{\gamma(c+t),x} U\left(\gamma(c+t)\right) - U(x) - \Gamma_{\gamma(c),x} U\left(\gamma(c)\right) + U(x)}{t} \tag{A.240}$$

$$= \lim_{t\to 0^+} \frac{\Gamma_{\gamma(c+t),x} U\left(\gamma(c+t)\right) - \Gamma_{\gamma(c),x} U\left(\gamma(c)\right)}{t}. \tag{A.241}$$

Using that $\Gamma_{z,x} U(z) = \Gamma_{y,x}\left(\Gamma_{z,y} U(z)\right)$, where $x = x$, $y = \gamma(c)$, and $z = \gamma(c+t)$, we have

$$f'(c) = \lim_{t\to 0^+} \frac{\Gamma_{\gamma(c),x}\left[\Gamma_{\gamma(c+t),\gamma(c)} U\left(\gamma(c+t)\right) - U\left(\gamma(c)\right)\right]}{t} \tag{A.242}$$

$$= \Gamma_{\gamma(c),x}\left[\lim_{t\to 0^+} \frac{\Gamma_{\gamma(c+t),\gamma(c)} U\left(\gamma(c+t)\right) - U\left(\gamma(c)\right)}{t}\right] \tag{A.243}$$

$$= \Gamma_{\gamma(c),x}\left[\nabla_{\gamma'(c)} U\right]. \tag{A.244}$$

Since for any smooth vector field W, $\|\Gamma_{y,x}(W)\|_x = \|W\|_y$, the proof is complete. □

Proof of Lemma A.4.9:

We apply Lemma A.4.8 to the vector $V_t(p) \equiv \mathrm{Log}_p\, y_1(t)$ for fixed t. We have that (equation (25) of Kendall and Le (2011))

$$V_t(p) = \mathrm{grad}\left(-\frac{1}{2}d(p, y_1(t))^2\right) = \mathrm{grad}(f_t)(p). \tag{A.245}$$

Then, by the definition of the Riemannian Hessian H_t of f_t, for a smooth curve γ on $\mathcal{M}$ at time c,

$$\|\nabla_{\gamma'(c)} V_t(\gamma(c))\|_{\gamma(c)} = \|\nabla_{\gamma'(c)} \mathrm{grad}(f_t)(p)\|_{\gamma(c)} = \|H_t(\gamma(c))(\gamma'(c))\|_{\gamma(c)}. \tag{A.246}$$

Using Lemma A.4.8, we choose $x = \mu(t)$ and $y = \hat{\mu}(t)$, and we denote $\gamma_t(s)$ as the minimizing geodesic between $\mu(t)$ and $\hat{\mu}(t)$:

$$\|\Gamma_{\hat{\mu}(t),\mu(t)} \mathrm{Log}_{\hat{\mu}(t)}\, y_1(t) - \mathrm{Log}_{\mu(t)}\, y_1(t)\| \leq d(\hat{\mu}(t), \mu(t)) \sup_{c \in [0, d(\hat{\mu}(t),\mu(t))]} \|H_t(\gamma_t(c))(\gamma_t'(c))\|_{\gamma_t(c)} \tag{A.247}$$

$$\lesssim d(\hat{\mu}(t), \mu(t)) \tag{A.248}$$

for all $t \in \mathcal{T}$ with probability one, where in the last inequality we have used Assumption 3 in the Lemma statement along with the fact that, since $\gamma_t(s)$ is a minimizing geodesic, $\|\gamma'(s)\|_{\gamma(s)} = 1$ for all s. Then,

$$\mathbb{E}\left[\|\mathrm{Log}_{\hat{\mu}}\, y_1 \delta_\Gamma \mathrm{Log}_\mu\, y_1\|_\mu^2\right] = \mathbb{E}\left[\int_{\mathcal{T}} \|\Gamma_{\hat{\mu}(t),\mu(t)} \mathrm{Log}_{\hat{\mu}(t)}\, y_1(t) - \mathrm{Log}_{\mu(t)}\, y_1(t)\|^2 dt\right] \tag{A.249}$$

$$\lesssim \mathbb{E}\left[\int_{\mathcal{T}} d(\hat{\mu}(t), \mu(t))^2\, dt\right]. \tag{A.250}$$

By Tonelli's theorem,

$$\mathbb{E}\left[\int_{\mathcal{T}} d(\hat{\mu}(t), \mu(t))^2\, dt\right] = \int_{\mathcal{T}} \mathbb{E}\left[d(\hat{\mu}(t), \mu(t))^2\right] dt. \tag{A.251}$$

Next, we will apply Corollary 4 of Schötz (2019) to bound $\mathbb{E}\left[d(\hat{\mu}(t), \mu(t))^2\right]$. To do so, we need the following definitions related to the metric entropy of geodesic balls in $\mathcal{M}$. Let

$B_\delta(a) \equiv \{x \in \mathcal{M} : d(x,a) \leq \delta\}$ be the ball of radius r centered at a on $\mathcal{M}$, and $N(B,r) \equiv \min\left(k \in \mathbb{N} | \exists q_1, \dots q_k \in \mathcal{M} : B \subseteq \bigcup_{j=1}^k B_r(q_j)\right)$ be the covering number of a set B using radius r. Then, the entropy assumption in the statement of Corollary 4 of Schötz (2019), that there exists $0 < \beta < 1$ such that, for all $\delta, r > 0$, $\log\left(N\left(B_\delta\left(\mu(t)\right), r\right)\right)^{1/2} \lesssim \left(\frac{\delta}{r}\right)^\beta$, is satisfied (Ahidar-Coutrix et al. (2020) Example 2.3).

Now additionally using Assumption 1 made in the statement of the Lemma, we can apply Corollary 4 of Schötz (2019) with $\varepsilon = 1$ to obtain

$$\mathbb{E}\left[d\left(\hat{\mu}(t), \mu(t)\right)^2\right] \lesssim \mathbb{E}\left[d\left(y_1(t), y_2(t)\right)^3\right]^{2/3} \frac{1}{N} \tag{A.252}$$

for all $t \in \mathcal{T}$.

Using Assumption 2 made in the statement of the Lemma, we have

$$\mathbb{E}\left[d\left(\hat{\mu}(t), \mu(t)\right)^2\right] \lesssim \frac{1}{N}, \tag{A.253}$$

which we can combine with equation (A.250) to establish

$$\mathbb{E}\left[\left\|\mathrm{Log}_{\hat{\mu}}\, y_1 \delta_\Gamma \,\mathrm{Log}_\mu\, y_1\right\|_\mu^2\right] \lesssim \int_{\mathcal{T}} \frac{1}{N} dt = \frac{\tau}{N}, \tag{A.254}$$

completing the proof. □

Proof of Lemma A.4.12:

Note that $\mathbb{E}\left[\|Y\|_2^2\right] = \mathbb{E}\left[Y^\top Y\right] = \mathbb{E}\left[\mathrm{tr}\left(YY^\top\right)\right] = \mathrm{tr}\left(\mathbb{E}\left[YY^\top\right]\right) = \mathrm{tr}\left(\Sigma_Y\right)$. Then, the first two statements of the Lemma follow from observing that Σ_Y is diagonal with entries ω_j, the eigenvalues of the covariance operator of $\mathrm{Log}_\mu\, y$, by the results of Lemma A.2.1.

To show $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$, we begin with the triangle inequality:

$$\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Z - Y\|_2^2\right] + \mathbb{E}\left[\|Y\|_2^2\right]. \tag{A.255}$$

Using Lemma A.4.11 and $\mathbb{E}\left[\|Y\|_2^2\right] = \sum_{j=1}^d \omega_j$, we observe for $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$ to hold, it is necessary that $\frac{d}{N} \max_{j=1,\dots d} \left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2 \left(\sum_{j=1}^d\right)^{-1} \lesssim 1$. This follows from assumption (A.185) provided in the Lemma, because under the other assumptions provided in the Lemma, both

$\left(\sum_{j=1}^{d}\omega_j\right)/\omega_d^2$ and $\sum_{j=1}^{d}\omega_j$ are greater than or equal to 1. From this, we also deduce that $\mathbb{E}\left[\|Z-Y\|_2^2\right] = o(1)$.

To show that $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$, a more involved argument is required. We again begin with the triangle inequality:

$$\left\|\Sigma_Z^{-1}\right\|_2^2 \lesssim \left\|\Sigma_Z^{-1}-\Sigma_Y^{-1}\right\|_2^2 + \left\|\Sigma_Y^{-1}\right\|_2^2 \tag{A.256}$$

$$\leq \left\|\Sigma_Z^{-1}\right\|_2^2 \|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2 + \left\|\Sigma_Y^{-1}\right\|_2^2. \tag{A.257}$$

This implies

$$\left\|\Sigma_Z^{-1}\right\|_2^2 - \left\|\Sigma_Z^{-1}\right\|_2^2 \|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2 \lesssim \left\|\Sigma_Y^{-1}\right\|_2^2, \tag{A.258}$$

so that

$$\left\|\Sigma_Z^{-1}\right\|_2^2 \left(1 - \|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2\right) \leq \left\|\Sigma_Y^{-1}\right\|_2^2, \tag{A.259}$$

and therefore, if $\|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2 = o(1)$ held, then we would have $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$ and the proof would be complete. Thus, to show $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$, it suffices to show that $\frac{d}{N}\frac{\sum_{j=1}^{d}\omega_j}{\omega_d^2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2 = o(1)$ implies $\|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2 = o(1)$. From the second item of Lemma A.4.3 and from $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$ which has already been shown, we have

$$\|\Sigma_Z-\Sigma_Y\|_2^2 \leq \max\left(\mathbb{E}\left[\|Z\|_2^2\right], \mathbb{E}\left[\|Y\|_2^2\right]\right) \mathbb{E}\left[\|Z-Y\|_2^2\right] \tag{A.260}$$

$$\lesssim \mathbb{E}\left[\|Z-Y\|_2^2\right] \sum_{j=1}^{d}\omega_j. \tag{A.261}$$

Thus, for $\|\Sigma_Z-\Sigma_Y\|_2^2 \left\|\Sigma_Y^{-1}\right\|_2^2 = o(1)$ to hold, it suffices to show that $\frac{1}{\omega_d^2}\mathbb{E}\left[\|Z-Y\|_2^2\right]\sum_{j=1}^{d}\omega_j = o(1)$. From Lemma A.4.11, we have made exactly the assumption so that $\frac{1}{\omega_d^2}\mathbb{E}\left[\|Z-Y\|_2^2\right]\sum_{j=1}^{d}\omega_j = o(1)$ holds, completing the proof that $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$.

To show the final statement of the Lemma, that $\|\Sigma_Z\|_2 \lesssim \|\Sigma_Y\|_2$, we can make an argument similar to the one used to show $\|\Sigma_Z^{-1}\|_2 \lesssim \|\Sigma_Y^{-1}\|_2$. From this, it suffices to show that $\frac{d}{N}\frac{\sum_{j=1}^{d}\omega_j}{\omega_1^2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2 = o(1)$. Since $\omega_d \leq \omega_1$, $\frac{d}{N}\frac{\sum_{j=1}^{d}\omega_j}{\omega_1^2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2 = o(1)$ holds under our assumptions and the proof is complete. □

Proof of Lemma A.4.14:

To bound $|\gamma_k - \tilde{\gamma}_k|$, we can use Weyl's inequality (Bhatia (2013) Corollary III.2.6.), because γ_k and $\tilde{\gamma}_k$ are the k eigenvalues of the matrices $|C|$ and $|A|$ respectively (Uurtio et al., 2018):

$$|\gamma_k - \tilde{\gamma}_k| \leq \||A| - |C|\|_2 \leq \||A| - |C|\|_F \lesssim \|A - C\|_F \,, \tag{A.262}$$

where in the second inequality we have used identity 19 from Section A.6. Combining Lemmas A.4.11, A.4.12, and A.4.13, we observe that the stated assumption in the Lemma implies $|\gamma_k - \tilde{\gamma}_k| \lesssim \min\{\gamma_k, \tilde{\gamma}_k\}$, establishing the stated conclusion using the assumption that γ_k and $\tilde{\gamma}_k$ are bounded from below. That γ_k^2 and $\tilde{\gamma}_k^2$ are asymptotically equivalent follows from the same argument, in addition to the function $f(x) = x^2$ being Lipschitz continuous on the interval $[0, 1]$.

A.5 Asymmetric Sparse-Functional CCA: Proof of Theorem 2.4.3

In this section we prove Theorem 2.4.3, which shows that our sample estimates of the canonical variables, derived from our proposed estimates of the canonical vectors, grow asymptotically close to the population canonical variable solutions of the infinite-dimensional CCA problem in Theorem A.2.1. We denote by (U_k, V_k) the kth canonical variable solution pair to the infinite-dimensional problem in Theorem A.2.1. We denote by $\{\gamma_k^*\}_{k=1}^p$ the canonical correlations that are attained by these solution pairs. We denote with $K = \max\{i \in \{1, \dots p\} : \gamma_i^* > 0\}$ the number of nontrivial canonical vectors in this problem, and to simplify the notation, we use the convention ${\gamma_0^*}^2 = \infty$ and ${\gamma_{K+1}^*}^2 = -\infty$. We often refer to canonical variables as canonical scores or simply scores, which are distinct from the principal scores derived from the principal component expansion of χ_1. The difference will be clear from context, but in general, 'scores' with no preceding modifier refers to the canonical scores.

Since we are showing that two random variables are close, we need to specify the probability spaces and random variables of interest. In this section we suppose that we observe realizations of $(X_i, y_i)_{i=1}^N$ which is used in the Sparse-Functional Asymmetric CCA (Algo-

rithm 1). We assume these are i.i.d samples that all live on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. d denotes the number of principal components we choose to approximate $\mathrm{Log}_\mu y$ by. We obtain estimates of $\hat{\mu}$, $\left\{\hat{\phi}_j\right\}_{j=1}^d$, $\hat{B}$, $\{\hat{\omega}_j\}_{j=1}^d$, $\{\hat{\gamma}_j\}_{j=1}^d$, $\hat{\Sigma}_Y, \hat{\Sigma}_X$, $\hat{H} = [\hat{\eta}_1, \ldots, \hat{\eta}_d]$, $\hat{T} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_d\right]$, and $\hat{\psi}_k = \sum_{j=1}^d \hat{\eta}_{kj}\hat{\phi}_j$ for $k = 1, \ldots d$.

We will define the scores from the sample procedure as follows. Observing a new and independent data point $(X_{\text{test}}, y_{\text{test}})$ from the same distribution as the sample, which also lives on $(\Omega, \mathcal{F}, \mathbb{P})$, we define

$$\left(\hat{U}_k, \hat{V}_k\right) \equiv \left(\langle\!\langle \mathrm{Log}_{\hat{\mu}} y_{\text{test}}, \hat{\psi}_k \rangle\!\rangle_{\hat{\mu}}, \hat{\theta}_k^\top X_{\text{test}}\right), \tag{A.263}$$

which are the canonical scores we would obtain from the new data point and our sample canonical vector estimates. To be precise, we denote by (U_k, V_k) the kth canonical variable solution pair to the infinite-dimensional problem in Theorem A.2.1, between $\chi_1 = \mathrm{Log}_\mu y_{\text{test}}$ and $\chi_2 = X_{\text{test}}$, for $k = 1, \ldots K$. We would like to show that $\hat{U}_k$ becomes close to U_k and $\hat{V}_k$ becomes close to V_k in an asymptotic sense.

Throughout the proof, we make the same assumptions in Section A.4.3, with the important exception that, we no longer make a finite-dimensional correlation structure Assumption A.2.1 between X and y. Thus, we allow for all principal scores of the functional data y to potentially be correlated with the components of X.

The main idea of the proof of the bounds is similar to that of the canonical vectors, and indeed we rely on the results of the previous sections. We define several intermediate scores in order to show that the sample score and the infinite-dimensional score are close together.

The first intermediate scores come from the population d-dimensional canonical correlation problem described in Theorem A.2.2 and equation (A.19), which we denote as $\left(U_k^{(d)}, V_k^{(d)}\right)$. By Theorem A.2.2, these can equivalently be derived from the canonical vector

solutions given by the following CCA problem:

$$(\eta_1, \theta_1) = \underset{\eta \in \mathbb{R}^d, \theta \in \mathbb{R}^p, \mathrm{Var}(\eta^\top Y_{\mathrm{test}}) = \mathrm{Var}(\theta^\top X_{\mathrm{test}}) = 1}{\arg\max} \mathrm{Corr}^2 \left(\eta^\top Y_{\mathrm{test}}, \theta^\top X_{\mathrm{test}}\right), \tag{A.264}$$

$$(\eta_k, \theta_k) = \underset{\substack{\eta \in \mathbb{R}^d, \theta \in \mathbb{R}^p, \mathrm{Var}(\eta^\top Y_{\mathrm{test}}) = \mathrm{Var}(\theta^\top X_{\mathrm{test}}) = 1 \\ \mathrm{Cov}\left(\eta^\top Y_{\mathrm{test}}, \eta_i^\top Y_{\mathrm{test}}\right) = 0, i = 1, \ldots, k \\ \mathrm{Cov}\left(\theta^\top X_{\mathrm{test}}, \theta_i^\top X_{\mathrm{test}}\right) = 0, i = 1, \ldots, k}}{\arg\max} \mathrm{Corr}^2 \left(\eta^\top Y_{\mathrm{test}}, \theta^\top X_{\mathrm{test}}\right), \qquad k = 2, \ldots K \tag{A.265}$$

Here, Y_{test} is the d-dimensional random vector such that $(Y_{\mathrm{test}})_j = \langle\!\langle \mathrm{Log}_\mu y_{\mathrm{test}}, \phi_j \rangle\!\rangle_\mu$, the jth principal score associated with the population principal component ϕ_j of $\mathrm{Log}_\mu y_{\mathrm{test}}$, for $j = 1, \ldots d$. By definition, $\left(U_k^{(d)}, V_k^{(d)}\right) = \left(\eta_k^\top Y_{\mathrm{test}}, \theta_k^\top X_{\mathrm{test}}\right)$. We refer to $\{\gamma_k\}_{k=1}^K$ as the canonical correlations attained in this problem.

The second intermediate scores are derived from the population canonical correlation problem derived from the 'estimates' of Y that we make when using $\hat{\mu}$ and $\hat{\phi}_j$:

$$(a_1, b_1) = \underset{a \in \mathbb{R}^d, b \in \mathbb{R}^p, \mathrm{Var}(a^\top Z) = \mathrm{Var}(b^\top X) = 1}{\arg\max} \mathrm{Corr}^2 \left(a^\top Z, b^\top X\right), \tag{A.266}$$

$$(a_k, b_k) = \underset{\substack{a \in \mathbb{R}^d, b \in \mathbb{R}^p, \mathrm{Var}(a^\top Z) = \mathrm{Var}(b^\top X) = 1 \\ \mathrm{Cov}\left(a^\top Z, a_i^\top Z\right) = 0, i = 1, \ldots, k \\ \mathrm{Cov}\left(b^\top X, b_i^\top X\right) = 0, i = 1, \ldots, k}}{\arg\max} \mathrm{Corr}^2 \left(a^\top Z, b^\top X\right), \qquad k = 2, \ldots K. \tag{A.267}$$

Here, Z is defined as

$$(Z)_j \equiv \langle\!\langle \mathrm{Log}_{\hat{\mu}} y_1, \hat{\phi}_j \rangle\!\rangle_{\hat{\mu}}. \tag{A.268}$$

We let $\tilde{\gamma}_1^2 \ldots \tilde{\gamma}_K^2$ denote the squared canonical correlations attained by the pairs $(a_1, b_1), \ldots (a_K, b_K)$. We also define the random variable $Z_{\mathrm{test}} \in \mathbb{R}^d$ so that

$$(Z_{\mathrm{test}})_j \equiv \langle\!\langle \mathrm{Log}_{\hat{\mu}} y_{\mathrm{test}}, \hat{\phi}_j \rangle\!\rangle_{\hat{\mu}}. \tag{A.269}$$

Note that the distribution of Z_{test} is not equal to that of Z, since y_{test} is independent of $\hat{\mu}$ and the $\hat{\phi}_j$, while y_1 is not.

We define the secondary intermediate scores as $U_k^{Z,1} \equiv a_k^\top Y_{\mathrm{test}}$, $U_k^{Z,2} \equiv a_k^\top Z_{\mathrm{test}}$, and $V_k^Z \equiv b_k^\top X_{\mathrm{test}}$. The additional score for U_k arises due to the complexity of estimating $\hat{\mu}$ and $\hat{\phi}_j$, whereas the proof for the bound on V_k is slightly simpler.

Following Hsing and Eubank (2015) Theorem 10.2.3, our goal is to show a probabilistic bound on $\mathbb{E}\left[\left(U_k - \hat{U}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]$ and $\mathbb{E}\left[\left(V_k - \hat{V}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]$ as the sample size N, the number of principal components we select d, and the dimension of the high-dimensional data p go to infinite. We choose this notion of error, where we condition on the sample, because we can derive a result which is comparable to our results for the canonical vectors while also integrating out the randomness of $(X_{\text{test}}, y_{\text{test}})$.

We begin with the proof of the bound for V_k since it is slightly simpler than that of U_k. For convenience throughout the proofs, we drop the k in the notation of the scores and canonical vectors. For example, we write $U_k, U_k^{Z,1}, \eta_k$ as $U, U^{Z,1}, \eta$.

A.5.1 Proof of bound for high-dimensional score

The proof strategy is to decompose the quantity of interest $\left(V - \hat{V}\right)^2$ into three parts:

$$\left(V - \hat{V}\right)^2 \lesssim \left(V - V^{(d)}\right)^2 + \left(V^{(d)} - V^Z\right)^2 + \left(V^Z - \hat{V}\right)^2 \tag{A.270}$$

so that

$$\mathbb{E}\left[\left(V - \hat{V}\right)^2 \mid (X_i, y_i)_{i=1}^N\right] \lesssim \mathbb{E}\left[\left(V - V^{(d)}\right)^2 + \left(V^{(d)} - V^Z\right)^2 + \left(V^Z - \hat{V}\right)^2 \mid (X_i, y_i)_{i=1}^N\right] \tag{A.271}$$

$$\equiv \text{Term I} + \text{Term II} + \text{Term III} \tag{A.272}$$

Term I:

$$\text{Term I} = \mathbb{E}\left[\left(V - V^{(d)}\right)^2 \mid (X_i, y_i)_{i=1}^N\right] = \left(V - V^{(d)}\right)^2. \tag{A.273}$$

since V and $V^{(d)}$ are independent of the sample. Then, Theorem A.2.3 gives the probabilistic bound

$$\text{Term I} = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}{\min_{j \neq k}\left|\gamma_k^{*2} - \gamma_j^{*2}\right|}\right), \tag{A.274}$$

since bounds in expectation imply probabilistic bounds. See Section A.2.4 for definitions of the operator $\mathscr{C}_{12}$ and its principal component approximation $\mathscr{C}_{12}^{(d)}$.

Term II:

$$\text{Term II} = \mathbb{E}\left[\left(V^{(d)} - V^Z\right)^2 \mid (X_i, y_i)_{i=1}^N\right]. \tag{A.275}$$

Let the support $S \subseteq \{1, \dots p\}$ represent the indices of non-zero elements of θ or b, with cardinality $|S| \leq 2s$. Denote $X_{\text{test},S} \in \mathbb{R}^{|S|}$ as the random vector consisting of only the entries $\{(X_{\text{test}})_j : j \in S\}$. Similarly, we define θ_S and b_S from θ and b respectively, as well as $\Sigma_{X_{\text{test}},S}$ from $\Sigma_{X_{\text{test}}}$, following Section A.4.

By definition,

$$V^{(d)} - V^Z = \theta^\top X_{\text{test}} - b^\top X_{\text{test}} \tag{A.276}$$

$$= \theta_S^\top X_{\text{test},S} - b_S^\top X_{\text{test},S} \tag{A.277}$$

$$= (\theta_S - b_S)^\top X_{\text{test},S}. \tag{A.278}$$

Therefore,

$$\text{Term II} = \mathbb{E}\left[(\theta_S - b_S)^\top X_{\text{test},S} X_{\text{test},S}^\top (\theta_S - b_S) \mid (X_i, y_i)_{i=1}^N\right] \tag{A.279}$$

$$= (\theta_S - b_S)^\top \Sigma_{X,S} (\theta_S - b_S) \tag{A.280}$$

$$= \left\|\Sigma_{X,S}^{1/2} (\theta_S - b_S)\right\|_2^2. \tag{A.281}$$

We can bound this by appealing to equation (A.195) and Lemma A.4.15, so that we can remove the factor of $\left\|\Sigma_{X,S}^{-1/2}\right\|_2^2$. We obtain

$$\text{Term II} = O_P\left(\frac{s \left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]^2}{\min\left(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1}\right)^2} \frac{\tau d}{N} \mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2} \max_{j=1,\dots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2\right). \tag{A.282}$$

Term III:

$$\text{Term III} = \mathbb{E}\left[\left(V^Z - \hat{V}\right)^2 \mid (X_i, y_i)_{i=1}^N\right]. \tag{A.283}$$

By definition,

$$V^Z - \hat{V} = b^\top X_{\text{test}} - \hat{\theta}^\top X_{\text{test}} \tag{A.284}$$

$$= \left(b - \hat{\theta}\right)^\top X_{\text{test}}. \tag{A.285}$$

Similarly to Term II, we obtain

$$\text{Term III} = \left\|\Sigma_X^{1/2}\left(\hat{\theta} - b\right)\right\|_2^2. \tag{A.286}$$

We can appeal to equation (A.151) following the same argument used to bound $\theta - \hat{\theta}$ in the proof of Theorem A.3.3, but where we place a $\Sigma_X^{1/2}$ in front of the relevant terms throughout to obtain

$$\left\|\Sigma_X^{1/2}\left(\hat{\theta}-b\right)\right\|_2 \lesssim \left(\frac{\left\|\Sigma_X^{1/2}\theta\right\|_2}{\tilde{\gamma}^2} + \frac{\left\|\Sigma_X^{1/2}\tilde{B}\right\|_2}{\tilde{\gamma}\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2, \tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)}\right)\left\|\tilde{B}^\top\Sigma_X\tilde{B}-\hat{B}^\top\hat{\Sigma}_X\hat{B}\right\|_2 + \frac{\left\|\Sigma_X^{1/2}\left(\tilde{B}-\hat{B}\right)\right\|_2}{\tilde{\gamma}}. \tag{A.287}$$

Here, $\tilde{B} = \Sigma_X^{-1}\Sigma_{XZ}\Sigma_Z^{-1/2}$. Noting that

$$\Sigma_X^{1/2}\tilde{B} = \Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2} \tag{A.288}$$

so $\left\|\Sigma_X^{1/2}\tilde{B}\right\|_2 = \tilde{\gamma}_1$ and $\left\|\Sigma_X^{1/2}\theta\right\|_2 = 1$, we have

$$\text{Term III} \lesssim \frac{1}{\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2, \tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)^2}\left\|\tilde{B}^\top\Sigma_X\tilde{B}-\hat{B}^\top\hat{\Sigma}_X\hat{B}\right\|_2^2 + \tilde{\gamma}^{-2}\left\|\Sigma_X^{1/2}\left(\tilde{B}-\hat{B}\right)\right\|_2^2. \tag{A.289}$$

Now we bound the two relevant terms on the right-hand-side. By equation (A.85), Lemma A.3.8, Remark 13, and Lemma A.3.7, we have

$$\left\|\Sigma_X^{1/2}(\hat{B}-\tilde{B})\right\|_F^2 = O_P\left(\kappa_X s\left\|\Sigma_X\right\|_{2,\infty}\frac{d\log p}{N}\right). \tag{A.290}$$

By Corollary A.3.1,

$$\left\|\tilde{B}^\top\Sigma_X\tilde{B}-\hat{B}^\top\hat{\Sigma}_X\hat{B}\right\|_2^2 = O_P\left(\tilde{\gamma}_1^2\left\|\Sigma_X\right\|_{2,\infty}s\kappa\frac{d\log p}{N}\right). \tag{A.291}$$

Putting the last three equations together,

$$\text{Term III} = O_P\left(\frac{\tilde{\gamma}_1^2\left\|\Sigma_X\right\|_{2,\infty}s\kappa}{\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2, \tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)^2}\frac{d\log p}{N} + \tilde{\gamma}^{-2}\kappa_X s\left\|\Sigma_X\right\|_{2,\infty}\frac{d\log p}{N}\right) \tag{A.292}$$

$$= O_P\left(\frac{\left\|\Sigma_X\right\|_{2,\infty}s\kappa}{\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2, \tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)^2}\frac{d\log p}{N}\right). \tag{A.293}$$

Final bound for V_k:

Now we can combine our bounds for terms I, II, and III to obtain

$$\mathbb{E}\left[\left(V-\hat{V}\right)^2 \mid (X_i,y_i)_{i=1}^N\right] = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12}-\mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2}-\gamma_j^{*2}\right|}\right) \tag{A.294}$$

$$+ O_P\left(\frac{s\left\|\Sigma_Y^{-1}\right\|_2^3\mathbb{E}\left[\|Y\|_2^2\right]^2}{\min\left(\gamma_{k-1}-\gamma_k,\gamma_k-\gamma_{k+1}\right)^2}\frac{\tau d}{N}\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right) \tag{A.295}$$

$$+ O_P\left(\frac{\|\Sigma_X\|_{2,\infty}\, s\kappa}{\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2,\tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)^2}\frac{d\log p}{N}\right). \tag{A.296}$$

Simplifying this, we obtain

$$\mathbb{E}\left[\left(V_k-\hat{V}_k\right)^2 \mid (X_i,y_i)_{i=1}^N\right] = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12}-\mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2}-\gamma_j^{*2}\right|}\right) \tag{A.297}$$

$$+ O_P\left(\frac{\|\Sigma_X\|_{2,\infty}\,\tau s\kappa\omega_d^{-3}\left(\sum_{j=1}^d\omega_j\right)^2\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}}{\min_{j\neq k}\min\left\{\left|{\gamma_k}^2-{\gamma_j}^2\right|,\left|\gamma_k-\gamma_j\right|\right\}^2}\frac{d\log p}{N}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right). \tag{A.298}$$

A.5.2 Proof of bound for Riemannian-functional score

The proof strategy is to decompose the quantity of interest $\left(U-\hat{U}\right)^2$ into four parts:

$$\left(U-\hat{U}\right)^2 \lesssim \left(U-U^{(d)}\right)^2+\left(U^{(d)}-U^{Z,1}\right)^2+\left(U^{Z,1}-U^{Z,2}\right)^2+\left(U^{Z,2}-\hat{U}\right)^2 \tag{A.299}$$

so that

$$\mathbb{E}\left[\left(U-\hat{U}\right)^2 \mid (X_i,y_i)_{i=1}^N\right] \lesssim \mathbb{E}\left[\left(U-U^{(d)}\right)^2+\left(U^{(d)}-U^{Z,1}\right)^2+\left(U^{Z,1}-U^{Z,2}\right)^2+\left(U^{Z,2}-\hat{U}\right)^2 \mid (X_i,y_i)_{i=1}^N\right] \tag{A.300}$$

$$\equiv \text{Term I} + \text{Term II} + \text{Term III} + \text{Term IV} \tag{A.301}$$

Term I:

Following the same argument used in the proof for the bound on V, Theorem A.2.3 gives the probabilistic bound

$$\text{Term I} = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12}-\mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2}-\gamma_j^{*2}\right|}\right). \tag{A.302}$$

Term II:

$$\text{Term II} = \mathbb{E}\left[\left(U^{(d)} - U^{Z,1}\right)^2 \mid (X_i, y_i)_{i=1}^N\right]. \tag{A.303}$$

By definition,

$$U^{(d)} - U^Z = \eta^\top Y_{\text{test}} - a^\top Y_{\text{test}} \tag{A.304}$$

$$= (\eta - a)^\top Y_{\text{test}}, \tag{A.305}$$

so that

$$\text{Term II} = \mathbb{E}\left[(\eta - a)^\top Y_{\text{test}} Y_{\text{test}}^\top (\eta - a) \mid (X_i, y_i)_{i=1}^N\right] \tag{A.306}$$

$$= \left\|\Sigma_Y^{1/2}(\eta - a)\right\|_2^2. \tag{A.307}$$

$$= \left\|\tilde{\eta} - \Sigma_Y^{1/2} a\right\|_2^2 \tag{A.308}$$

$$= \left\|\tilde{\eta} - \Sigma_Y^{1/2}\Sigma_Z^{-1/2}\tilde{a}\right\|_2^2 \tag{A.309}$$

$$\lesssim \|\tilde{\eta} - \tilde{a}\|_2^2 + \left\|\left(I - \Sigma_Y^{1/2}\Sigma_Z^{-1/2}\right)\tilde{a}\right\|_2^2 \tag{A.310}$$

The first term on the right-hand side shares the same bound as Term II in the proof for the bound for V, using Lemma A.4.15 and equation (A.170):

$$\|\tilde{\eta} - \tilde{a}\|_2^2 = O_P\left(\frac{s\left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]^2}{\min\left(\gamma_{k-1} - \gamma_k, \gamma_k - \gamma_{k+1}\right)^2}\frac{\tau d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2\right). \tag{A.311}$$

To bound $\left\|\left(I - \Sigma_Y^{1/2}\Sigma_Z^{-1/2}\right)\tilde{a}\right\|_2^2$, we use $\|\tilde{a}\|_2 = 1$, Lemma A.4.3, and

$$\left\|\left(I - \Sigma_Y^{1/2}\Sigma_Z^{-1/2}\right)\tilde{a}\right\|_2^2 \leq \left\|\Sigma_Y^{1/2}\left(\Sigma_Y^{-1/2} - \Sigma_Z^{-1/2}\right)\right\|_2^2 \tag{A.312}$$

$$\lesssim \|\Sigma_Y\|_2 \mathbb{E}\left[\|Z - Y\|_2^2\right]\left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]. \tag{A.313}$$

By Lemma A.4.11, we then have

$$\left\|\left(I - \Sigma_Y^{1/2}\Sigma_Z^{-1/2}\right)\tilde{a}\right\|_2^2 \lesssim \|\Sigma_Y\|_2\left\|\Sigma_Y^{-1}\right\|_2^3 \mathbb{E}\left[\|Y\|_2^2\right]\frac{\tau d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j - \omega_{j+1}}\right)^2. \tag{A.314}$$

Combining these bounds and simplifying, we finally obtain

$$\text{Term II} = O_P\left(\frac{\omega_1\omega_d^{-3}\left(\sum_{j=1}^d \omega_j\right)^2}{\min\left(\gamma_{k-1}-\gamma_k, \gamma_k-\gamma_{k+1}\right)^2}\frac{\tau s d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right). \quad \text{(A.315)}$$

Term III:

$$\text{Term III} = \mathbb{E}\left[\left(U^{Z,1}-U^{Z,2}\right)^2 \mid (X_i, y_i)_{i=1}^N\right]. \quad \text{(A.316)}$$

By definition,

$$U^{Z,1}-U^{Z,2} = a^\top Y_{\text{test}} - a^\top Z_{\text{test}} \quad \text{(A.317)}$$

$$= a^\top\left(Y_{\text{test}}-Z_{\text{test}}\right), \quad \text{(A.318)}$$

so that by the Cauchy-Schwarz inequality and Lemma A.4.12,

$$\left(U^{Z,1}-U^{Z,2}\right)^2 \le \|a\|_2^2\,\|Y_{\text{test}}-Z_{\text{test}}\|_2^2 \quad \text{(A.319)}$$

$$\le \left\|\Sigma_Z^{-1/2}\right\|_2^2\,\|Y_{\text{test}}-Z_{\text{test}}\|_2^2 \quad \text{(A.320)}$$

$$\lesssim \left\|\Sigma_Y^{-1}\right\|_2\,\|Y_{\text{test}}-Z_{\text{test}}\|_2^2. \quad \text{(A.321)}$$

Despite the fact that here we have Z_{test}, we can appeal to Lemma A.4.11 to bound $\|Y_{\text{test}}-Z_{\text{test}}\|_2^2$, since replacing y_1 with y_{test} does not change the proof. Using Lemma A.4.11 and the law of total expectation we obtain

$$\text{Term III} \lesssim \left\|\Sigma_Y^{-1}\right\|_2\mathbb{E}\left[\|Y_{\text{test}}-Z_{\text{test}}\|_2^2 \mid (X_i, y_i)_{i=1}^N\right] \quad \text{(A.322)}$$

$$= O_P\left(\omega_d^{-1}\frac{\tau d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right). \quad \text{(A.323)}$$

Term IV:

$$\text{Term IV} = \mathbb{E}\left[\left(U^{Z,2}-\hat{U}\right)^2 \mid (X_i, y_i)_{i=1}^N\right]. \quad \text{(A.324)}$$

By definition,

$$U^{Z,2}-\hat{U} = a^\top Z_{\text{test}} - \hat{\eta}^\top Z_{\text{test}} \quad \text{(A.325)}$$

$$= (a-\hat{\eta})^\top Z_{\text{test}}, \quad \text{(A.326)}$$

so that similarly to Term III we obtain

$$\text{Term IV} \lesssim \|a-\hat{\eta}\|_2^2 \,\mathbb{E}\left[\|Z_{\text{test}}\|_2^2 \,|\, (X_i,y_i)_{i=1}^N\right]. \tag{A.327}$$

While $\mathbb{E}\left[\|Z_{\text{test}}\|_2^2\right] \neq \mathbb{E}\left[\|Z\|_2^2\right]$, we can use the same argument we used to show $\mathbb{E}\left[\|Z\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right]$ to show that $\mathbb{E}\left[\|Z_{\text{test}}\|_2^2\right] \lesssim \mathbb{E}\left[\|Y\|_2^2\right] = \sum_{j=1}^d \omega_j$ (Lemma A.4.12). Then by the law of total expectation and Theorem A.3.3 applied to Z and X, we have

$$\text{Term IV} = O_P\left(\frac{d\log(p)}{N}\left\|\Sigma_Z^{-1}\right\|_2 \max\left\{\frac{\tilde{\gamma}_1^2\left\|\Sigma_X\right\|_{2,\infty} s\kappa}{\min\left(\tilde{\gamma}_{k-1}^2-\tilde{\gamma}_k^2,\tilde{\gamma}_k^2-\tilde{\gamma}_{k+1}^2\right)^2}, \left\|\Sigma_Z\right\|_2\left\|\Sigma_Z^{-1}\right\|_2\right\}\right)\cdot O_P\left(\sum_{j=1}^d \omega_j\right) \tag{A.328}$$

$$= O_P\left(\frac{d\log(p)}{N}\left(\sum_{j=1}^d \omega_j\right)\omega_1\omega_d^{-2}\frac{\gamma_1^2\left\|\Sigma_X\right\|_{2,\infty} s\kappa}{\min\left(\gamma_{k-1}^2-\gamma_k^2,\gamma_k^2-\gamma_{k+1}^2\right)^2}\right). \tag{A.329}$$

Final bound for U_k:

Now we can combine our bounds for terms I, II, III and IV to obtain

$$\mathbb{E}\left[\left(U-\hat{U}\right)^2 \,|\, (X_i,y_i)_{i=1}^N\right] = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12}-\mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2}-\gamma_j^{*2}\right|}\right) \tag{A.330}$$

$$+ O_P\left(\frac{\omega_1\omega_d^{-3}\left(\sum_{j=1}^d\omega_j\right)^2}{\min\left(\gamma_{k-1}-\gamma_k,\gamma_k-\gamma_{k+1}\right)^2}\frac{\tau s d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right) \tag{A.331}$$

$$+ O_P\left(\omega_d^{-1}\frac{\tau d}{N}\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right) \tag{A.332}$$

$$+ O_P\left(\frac{d\log(p)}{N}\left(\sum_{j=1}^d\omega_j\right)\omega_1\omega_d^{-2}\frac{\gamma_1^2\left\|\Sigma_X\right\|_{2,\infty} s\kappa}{\min\left(\gamma_{k-1}^2-\gamma_k^2,\gamma_k^2-\gamma_{k+1}^2\right)^2}\right) \tag{A.333}$$

Simplifying this gives

$$\mathbb{E}\left[\left(U_k-\hat{U}_k\right)^2 \,|\, (X_i,y_i)_{i=1}^N\right] = O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12}-\mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2}-\gamma_j^{*2}\right|}\right) \tag{A.334}$$

$$+ O_P\left(\frac{\left\|\Sigma_X\right\|_{2,\infty}\tau s\kappa\omega_1\omega_d^{-3}\left(\sum_{j=1}^d\omega_j\right)^2\mathbb{E}\left[\left\|\text{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}}{\min_{j\neq k}\min\left\{\left|{\gamma_k}^2-{\gamma_j}^2\right|,\left|\gamma_k-\gamma_j\right|\right\}^2}\right) \tag{A.335}$$

$$\cdot O_P\left(\frac{d\log p}{N}\max_{j=1,\dots d}\left(\frac{1}{\omega_j-\omega_{j+1}}\right)^2\right). \tag{A.336}$$

We note that the second term on the right-hand side is nearly identical to the bound on $\mathbb{E}\left[\left(V-\hat{V}\right)^2 \mid (X_i, y_i)_{i=1}^N\right]$, except here we have an additional ω_1 term. Therefore, we may concisely write the final bound on the score errors for both U and V together using the bound for U, since we have already assumed for convenience that $\omega_1 \geq 1$.

Before stating the final bounds, we state a final lemma which allows us to use the infinite-dimensional canonical correlations.

Lemma A.5.1. *For every $k = 1, 2, \ldots K$, γ_k^* is asymptotically equivalent to γ_k as d goes to infinite.*

The proof follows directly from Theorem 4.2.8 of Hsing and Eubank (2015) applied to $\mathscr{C}_{12}\mathscr{C}_{21}$ and $\mathscr{C}_{12}^{(d)}\mathscr{C}_{21}^{(d)}$ which gives that

$$\sup_{k\geq 1}\left|\gamma_k^{*2} - \gamma_k^2\right| \leq \|\mathscr{C}_{12}\mathscr{C}_{21} - \mathscr{C}_{12}^{(d)}\mathscr{C}_{21}^{(d)}\|. \tag{A.337}$$

For bounded operators A and B with adjoints A^* and B^*, we have

$$AA^* - BB^* = (A-B)A^* - B(A^* - B^*), \tag{A.338}$$

so that

$$\|AA^* - BB^*\| \leq (\|A\| + \|B\|)\,\|A - B\|. \tag{A.339}$$

Applying this inequality with $A = \mathscr{C}_{12}$ and $B = \mathscr{C}_{12}^{(d)}$, the quantity on the right-hand side of equation (A.337) converges to 0 as d grows to infinity since

$$\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\| \leq \|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|_{\mathrm{HS}}, \tag{A.340}$$

which decays to 0 as d grows since $\mathscr{C}_{12}$ being finite-dimensional means is it necessarily Hilbert-Schmidt.

We finally state the error bound on the convergence of the estimated canonical variables to the population infinite-dimensional canonical variables.

Theorem A.5.1 (Convergence of canonical variables)**.** *Let (U_k, V_k) be the solution pair to the problem in Theorem A.2.1 between $\chi_1 = \mathrm{Log}_\mu y_{test}$ and $\chi_2 = X_{test}$, for $k = 1, \ldots K$. Let*

$\left(\hat{U}_k, \hat{V}_k\right) \equiv \left(\langle\!\langle \mathrm{Log}_{\hat{\mu}} y_{test}, \hat{\psi}_k \rangle\!\rangle_{\hat{\mu}}, \langle X_{test}, \hat{\theta}_k \rangle\right)$, *where* $\hat{\theta}_k$, $\hat{\eta}_k$ *and* $\hat{\mu}$ *have been estimated via Algorithm 1. Then, under Assumptions A.4.1-A.4.4, but without assuming a finite-dimensional correlation structure on* X *and* y *in Assumption A.4.3, we have*

$$\max\left\{\mathbb{E}\left[\left(U_k - \hat{U}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right], \mathbb{E}\left[\left(V_k - \hat{V}_k\right)^2 \mid (X_i, y_i)_{i=1}^N\right]\right\} \tag{A.341}$$

$$= O_P\left(\frac{\gamma_1^{*2}\|\mathscr{C}_{12} - \mathscr{C}_{12}^{(d)}\|^2}{\min_{j\neq k}\left|\gamma_k^{*2} - \gamma_j^{*2}\right|} + \frac{ds\log p}{N}\frac{\tau\left\|\Sigma_X\right\|_{2,\infty}\kappa\mathbb{E}\left[\left\|\mathrm{Log}_\mu y_1\right\|_\mu^4\right]^{3/2}}{\min_{j\neq k}\min\left\{\left|\gamma_k^{*2} - \gamma_j^{*2}\right|, \left|\gamma_k^* - \gamma_j^*\right|\right\}^2}\frac{\omega_1\left(\sum_{j=1}^d \omega_j\right)^2}{\omega_d^3 \min_{j=1,\ldots d}\left(\omega_j - \omega_{j+1}\right)^2}\right). \tag{A.342}$$

A.6 Additional identities and inequalities

In the proofs, we use several identities and inequalities involving matrices. For definitions of the various matrix operations used below we refer to Section A.3.1. In the following, A and B denote matrices for which the specified matrix multiplications are valid.

1. For $x \in \mathbb{R}^p$, $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

2. $\|x\|_2 \leq \sqrt{p}\,\|x\|_\infty$, $\|x\|_1 \leq \sqrt{p}\,\|x\|_2$, and $\|x\|_1 \leq p\,\|x\|_\infty$.

3. $\|A\|_{\max} \leq \|A\|_{2,\infty} \leq \|A\|_2 \leq \|A\|_F \leq \|A\|_{\ell_1,\ell_2}$

4. For induced norms, $\|AB\|_{\beta,\alpha} \leq \|A\|_{\gamma,\alpha}\|B\|_{\beta,\gamma}$ (Trefethen and Bau (2022) equation (3.14)). In particular, $\|AB\|_2 = \|AB\|_{2,2} \leq \|A\|_{\infty,2}\|B\|_{2,\infty}$.

5. For any matrix A, $\|A\|_2 = \|A^\intercal A\|_2^{1/2}$.

6. In addition to its definition as the largest singular value, $\|A\|_2$ is the norm induced by $\|\cdot\|_2$ and $\|\cdot\|_2$.

7. In addition to its definition as the norm induced by the $\|\cdot\|_2$ and $\|\cdot\|_\infty$ norms, $\|A\|_{2,\infty} = \max_i(\|A_i\|_2)$, where A_i is the ith row of A. (Cape et al. (2019) Proposition 6.1.)

8. $\|AB\|_{2,\infty} \leq \|A\|_{2,\infty} \|B\|_2$. (Cape et al. (2019) Proposition 6.5.)

9. For $B \in \mathbb{R}^{p\times d}$, $\|B\|_{2,\infty} \leq \sqrt{d}\,\|B\|_{\max}$.

10. For the induced norm $\|\cdot\|_{\infty,2}$, $\|B^\top\|_{\infty,2} \leq \|B\|_{\ell_1,\ell_2}$.

11. If $A \in \mathbb{R}^{d\times p}$, and $B \in \mathbb{R}^{p\times d}$, then $\|AB\|_2 \leq \|A^\top\|_{2,\infty} \|B\|_{\ell_1,\ell_2}$. Additionally, $\|AB\|_2 \leq \|A^\top\|_{\ell_1,\ell_2} \|B\|_{2,\infty}$.

12. $\|AB\|_{\ell_1,\ell_2} \leq \|A\|_{\ell_1,\ell_2} \|B\|_2$.

13. If $A, B \in \mathbb{R}^{d\times d}$ are positive definite, then $\|A^{1/2} - B^{1/2}\|_2 \leq \frac{1}{2}\max\left(\|A^{-1}\|_2, \|B^{-1}\|_2\right)^{1/2} \|A - B\|_2$.

14. If $A, B \in \mathbb{R}^{d\times d}$ are positive definite, then $\|A^{-1/2} - B^{-1/2}\|_2 \leq \frac{1}{2}\max\left(\|A^{-1}\|_2, \|B^{-1}\|_2\right)^{3/2} \|A - B\|_2$.

15. For any positive definite matrices A, B of the same size, the product AB is diagonalizable with positive eigenvalues. Additionally, AB has the same eigenvalues as $(AB^2A)^{1/2}$. In particular, $\|AB - I\|_2 = \|(AB^2A)^{1/2} - I\|_2$. Note that AB may not be symmetric, and therefore is not necessarily positive definite.

16. For positive real numbers x and y, $|x^{-1} - y^{-1}| \leq \min(x, y)^{-3} |x^2 - y^2|$.

17. For $A, B \in \mathbb{R}^{d\times d}$, $a, b \in \mathbb{R}^d$, $\|Aa - Bb\|_2^2 \lesssim \|A - B\|_2^2 \|a\|_2^2 + \|B\|_2^2 \|a - b\|_2^2$

18. For any norm $\|\cdot\|$, $\|a + b\|^2 \leq 3\left(\|a\|^2 + \|b\|^2\right)$.

19. For matrices A, B of the same dimensions, $\||A| - |B|\|_F \leq \sqrt{2}\,\|A - B\|_F$, where $|A| \equiv (A^\top A)^{1/2}$. (Bhatia (2013) equation VII.39)

Proof. We prove the non-standard or non-straightforward results.

Proof of 10:

To show $\|B^\top\|_{\infty,2} \leq \|B\|_{\ell_1,\ell_2}$, we use the definition:

$$\|B^\top\|_{\infty,2} \equiv \sup_{\|x\|_\infty=1} \|B^\top x\|_2 . \tag{A.343}$$

Without loss of generality, let B be an element of $\mathbb{R}^{p\times d}$. $\|B^\top x\|_2 = \|\sum_{i=1}^p b_i x_i\|_2$ where b_i is the ith row of B, and x_i is the ith entry of $x \in \mathbb{R}^p$. For $x \in \mathbb{R}^p$ with $\|x\|_\infty = 1$, we have

$$\left\|\sum_{i=1}^p b_i x_i\right\|_2 \leq \sum_{i=1}^p |x_i| \, \|b_i\|_2 \leq \sum_{i=1}^p \|b_i\|_2 = \|B\|_{\ell_1,\ell_2} , \tag{A.344}$$

using the triangle inequality and because $\|x\|_\infty = 1$. This completes the proof.

Proof of 11:

To show $\|AB\|_2 \leq \|A^\top\|_{2,\infty} \|B\|_{\ell_1,\ell_2}$, we begin by using item 4 to obtain

$$\|AB\|_2 = \|B^\top A^\top\|_2 \leq \|B^\top\|_{\infty,2} \|A^\top\|_{2,\infty}. \tag{A.345}$$

Using item 10 we have

$$\|B^\top\|_{\infty,2} \|A^\top\|_{2,\infty} \leq \|B\|_{\ell_1,\ell_2} \|A^\top\|_{2,\infty}, \tag{A.346}$$

completing the proof of the first statement. To show the second statement, we proceed similarly but apply item 4 to $\|AB\|_2$ instead of $\|B^\top A^\top\|_2$.

Proof of 12:

To show $\|AB\|_{\ell_1,\ell_2} \leq \|A\|_{\ell_1,\ell_2} \|B\|_2$, we begin with the definition. Without loss of generality, $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{p\times r}$. We have

$$\|AB\|_{\ell_1,\ell_2} = \sum_{i=1}^p \|(AB)_i\|_2 = \sum_{i=1}^p \|A_i^\top B\|_2 \leq \sum_{i=1}^p \|a_i\|_2 \|B\|_2 = \|A\|_{\ell_1,\ell_2} \|B\|_2 , \tag{A.347}$$

where $(C)_i$ denotes the ith row of a matrix C and we have used item 6.

Proof of 13:

The statement follows from equation X.46 of Bhatia (2013) by choosing $r = 1/2$ and since the 2-norm is unitarily invariant.

Proof of 14:

We begin with the equality $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, which holds for any square invertible matrices A and B of the same size. This implies

$$\left\|A^{-1} - B^{-1}\right\|_2 \leq \left\|A^{-1}\right\|_2 \|A - B\|_2 \left\|B^{-1}\right\|_2. \tag{A.348}$$

We apply this to the matrices $A^{1/2}$ and $B^{1/2}$ to obtain

$$\left\|A^{-1/2} - B^{-1/2}\right\|_2 \leq \left\|A^{-1/2}\right\|_2 \left\|A^{1/2} - B^{1/2}\right\|_2 \left\|B^{-1/2}\right\|_2. \tag{A.349}$$

Using item 13 on $\|A^{1/2} - B^{1/2}\|_2$ in the last inequality, we deduce that

$$\left\|A^{-1/2} - B^{-1/2}\right\|_2 \leq \frac{1}{2}\max\left(\left\|A^{-1}\right\|_2, \left\|B^{-1}\right\|_2\right)^{1/2} \|A - B\|_2 \left\|A^{-1/2}\right\|_2 \left\|B^{-1/2}\right\|_2 \tag{A.350}$$

$$= \frac{1}{2}\left\|A^{-1/2}\right\|_2 \left\|B^{-1/2}\right\|_2 \max\left(\left\|A^{-1/2}\right\|_2, \left\|B^{-1/2}\right\|_2\right) \|A - B\|_2 \tag{A.351}$$

One of $\|A^{-1/2}\|_2$ and $\|B^{-1/2}\|_2$ is larger, and in either case, the statement to be proven holds.

Proof of 15:

The first claim, that AB is diagonalizable with positive eigenvalues, follows from Proposition 6.1 of Serre (2010). To show that AB has the same eigenvalues as $(AB^2A)^{1/2}$, let $C = AB$, so that $(AB^2A)^{1/2} = (CC^\top)^{1/2}$. Letting $U\Sigma V^\top = C$ be a singular value decomposition of C, we have

$$(CC^\top)^{1/2} = (U\Sigma V^\top V\Sigma U^\top)^{1/2} = \left(U\Sigma^2 U^\top\right)^{1/2} = U\Sigma U^\top, \tag{A.352}$$

where in the last line we have used that C has positive singular values from the first claim. The last claim follows from the previous claim, since for diagonalizable matrix C where $C = WDW^{-1}$ with D diagonal, we have

$$C - I = WDW^{-1} - WIW^{-1} = W(D - I)W^{-1}, \tag{A.353}$$

so that $C - I$ has eigenvalues equal to those of C minus 1.

Proof of 16:

Letting $f(x) = x^{-1/2}$, the mean value theorem implies

$$|f(x) - f(y)| \leq \min(x, y)^{-3/2} |x - y|. \tag{A.354}$$

Plugging in x^2 for x and y^2 for y, we obtain the result.

Proof of 17:

The statement follows from adding and subtracting Ba and the inequality$\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$.

Proof of 18:

From the triangle inequality, we have

$$\|a+b\|_2^2 \leq \left(\|a\|_2 + \|b\|_2\right)^2 = \|a\|_2^2 + 2\|a\|_2\|b\|_2 + \|b\|_2^2 \tag{A.355}$$

One of $\|a\|_2$ or $\|b\|_2$ is larger than the other, and in either case, we deduce that $\|a\|_2^2 + 2\|a\|_2\|b\|_2 + \|b\|_2 \leq 3\left(\|a\|_2^2 + \|b\|_2^2\right)$, completing the proof. □

A.7 Intrinsic RFPCA algorithm

For completeness, in this section, we outline the main steps of the Intrinsic RFPCA algorithm. Let M be the dimension of $\mathcal{M}$, d the number of principal components to compute, N the number of observations, and L the number of time-steps.

1. Estimate $\mu : \mathcal{T} \to \mathcal{M}$, the functional Frechét mean of y, by computing the Frechét mean of $\{y_i(t_l)\}_i$ separately for each point in $\{t_l\}_{l=1,\dots,L}$ where the functional data are observed.

2. Compute the linear representations $\mathrm{Log}_{\hat{\mu}} y_i \in L^2(T\hat{\mu})$, for $i = 1, \dots N$. In practice, this can be done by computing $\mathrm{Log}_{\hat{\mu}(t_l)} y_i(t_l) \in T_{\hat{\mu}(t_l)}\mathcal{M}$ for every t_l by using the Log map on $\mathcal{M}$.

3. Let $E(x)$ be an orthonormal frame for the tangent space centered at $x \in \mathcal{M}$ (see Section A.7.1 for an example). An orthonormal frame is a collection of tangent bases $E(x) = \{E_1(x), \dots E_M(x)\}$ for $T_x\mathcal{M}$, with $x \in \mathcal{M}$, which varies smoothly with x, i.e. for each $k = 1, \dots M$, $E_k(x)$ is a smooth map from $\mathcal{M}$ to $T\mathcal{M}$. For any fixed $x \in \mathcal{M}$, the functions $\{E_k(x)\}$ are orthonormal with respect to the inner product on $T_x\mathcal{M}$, that is, $\langle \cdot, \cdot \rangle_x$.

Compute the functional 'coefficients' $\hat{Z}_i : \mathcal{T} \to \mathbb{R}^M$ of the expansion of $\mathrm{Log}_{\hat{\mu}} y_i : \mathcal{T} \to T\mathcal{M}$ relative to E, for each $i = 1, \dots N$. In practice, $\hat{Z}_i(t_l) \in \mathbb{R}^M$ is computed separately for every $l = 1, \dots L$. The kth entry of $\hat{Z}_i(t_l) \in \mathbb{R}^M$ is computed as $\langle \mathrm{Log}_{\hat{\mu}(t_l)} y_i(t_l), E_k(\hat{\mu}(t_l)) \rangle_{\hat{\mu}(t_l)}$ for $k = 1, \dots M$. The resulting $\{\hat{Z}_i\}$ are estimates of the realizations of a real vector-valued random process $Z : \mathcal{T} \to \mathbb{R}^M$, with kth component $Z_k : \mathcal{T} \to \mathbb{R}$ given by $Z_k(t) = \langle \mathrm{Log}_{\mu(t)} y(t), E_k(\mu(t)) \rangle_{\mu(t)}$.

The process Z is a real vector-valued process with the same principal scores as the process $\mathrm{Log}_{\mu} y$, and the jth principal component of Z, $\pi_j : \mathcal{T} \to \mathbb{R}^M$, is related to the jth principal component of $\mathrm{Log}_{\mu} y$, $\phi_j \in L^2(T\mu)$, via $\phi_j = \sum_{k=1}^{M} \pi_{jk} E_k$. Here, π_{jk} denotes the kth entry of π_j for $k = 1, \dots M$.

4. Apply Multivariate Functional Principal Component Analysis (MFPCA) (Happ & Greven, 2018) to the functions $\{\hat{Z}_i\}$ to estimate d principal component functions $\hat{\pi}_j$ of the functional coefficients.

   Estimate the principal component functions $\{\phi_j\}$ of the Log representations of the functional data as $\hat{\phi}_j = \sum_{k=1}^{M} \hat{\pi}_{jk} E_k(\hat{\mu})$, for $j = 1, \dots d$. Then, estimate the associated scores as $\hat{Y}_{ij} = \frac{1}{L} \sum_{l=1}^{L} Z_i(t_l)^\top \hat{\pi}_j(t_l)$, for $j = 1, \dots d$ and $i = 1, \dots N$. The score estimates do not depend on the choice of the orthonormal frame E (Proposition 5, item 2 of Lin and Yao (2019)). Compute variance estimates $\hat{\omega}_j = \mathrm{Var}(\hat{Y}_{ij})$.

5. Return the estimated scores $\{\hat{Y}_{ij}\}$, variances $\{\hat{\omega}_j\}$ and principal component functions $\{\hat{\phi}_j\}$.

A.7.1 An orthonormal frame for the manifold of SPD matrices

Here we provide an explicit construction of an orthonormal frame in the setting where $\mathcal{M}$ is the manifold of $\mathbb{R}^{m \times m}$ symmetric positive matrices equipped with the affine invariant metric, as in our application setting. The maps Log and Exp maps are defined as $\mathrm{Log}_F(G) =$

$F^{1/2}\log\left(F^{-1/2}GF^{-1/2}\right)F^{1/2}$ and $\mathrm{Exp}_F(W) = F^{1/2}\exp\left(F^{-1/2}WF^{-1/2}\right)F^{1/2}$. Moreover, the inner product at $F \in \mathcal{M}$ between $W, Z \in T_F\mathcal{M}$ is defined as $\langle W, Z\rangle_{\mathcal{M}} = \mathrm{tr}\left(F^{-1}WF^{-1}Z\right)$. In this setting, we can use the result from Section 3.3.3.3. of Pennec et al. (2019), which provides an explicit construction of an orthonormal frame $E(F)$ for the tangent bundle, evaluated at an arbitrary $F \in \mathcal{M}$. This can be defined as

$$E_{ij}(F) = \begin{cases} \left(F^{1/2}e_i\right)\left(F^{1/2}e_i\right)^\top & (1 \le i = j \le m) \\ \frac{1}{\sqrt{2}}\left(\left(F^{1/2}e_i\right)\left(F^{1/2}e_j\right)^\top + \left(F^{1/2}e_j\right)\left(F^{1/2}e_i\right)^\top\right) & (1 \le i < j \le m), \end{cases} \tag{A.356}$$

where e_i denotes the ith standard unit vector in $\mathbb{R}^m$, and $F \in \mathbb{R}^{m\times m}$. For a fixed $F \in \mathcal{M}$, there are $M = m(m+1)/2$ unique $E_{ij}(F)$, where M is the dimension of $\mathcal{M}$. We also note that the iterative algorithm proposed in Cheng et al. (2016) can be used to estimate the Fréchet mean μ, and is detailed in equations (13) and (14) of Cheng et al. (2016).

Appendix B

APPENDIX TO CHAPTER 3

B.1 Simulations

In this section, we provide a detailed analysis of the proposed methods on a synthetic data set. We take inspiration from the rings and discs example of Chadebec and Allassonnière (2022) and generate a dataset of 250 greyscale images. The images are 80×80 pixels, with independent Gaussian noise with 0 mean and variance 0.005 added to each pixel's greyscale value. This is the 'target' data view, referred to as Y. The rings and discs are parameterized by r_1 ="radius of the hole" and r_2 ="width of the ring" parameters. For a disc, r_1 is equal to 0, since there is no hole at its center, and r_2 is the radius of the disc. For the rings, we set r_2 equal to a constant $r_2 = 0.2$. The dataset is comprised of 125 images of discs, with r_2 sampled uniformly between 0.2 and 0.5, and 125 images of rings, with r_1 sampled uniformly between 0.1 and 0.5. The second, auxiliary, data view is then $X = (0.5 \cdot r_1 + 0.5 \cdot r_2, 0.5 \cdot r_1 - 0.5 \cdot r_2, 0, 0, \cdots 0) + \varepsilon_X \in \mathbb{R}^p$ with $p = 10$, where ε_X is independent Gaussian noise with variance 0.005. The objective of this simulation is to learn latent representations for Y, while leveraging the dataset X to guide these representations. From the data generative model adopted, it is clear that in this setting X contains information relevant to defining interpretable latent representations for Y. However, its sparse and noisy form makes such a task non-trivial.

For all methods, hyperparameters are chosen through K-fold cross-validation on a computing cluster, with K equal to 4 or 5. We run each method until the the in-sample loss has effectively stopped decreasing. Each method uses a $d = 2$ dimensional latent space, and all VAE approaches use the same encoder and decoder architectures, 4 layer perceptrons with ReLu nonlinearities. Notably, the last layer of the decoder is fed into a sigmoid, reflecting

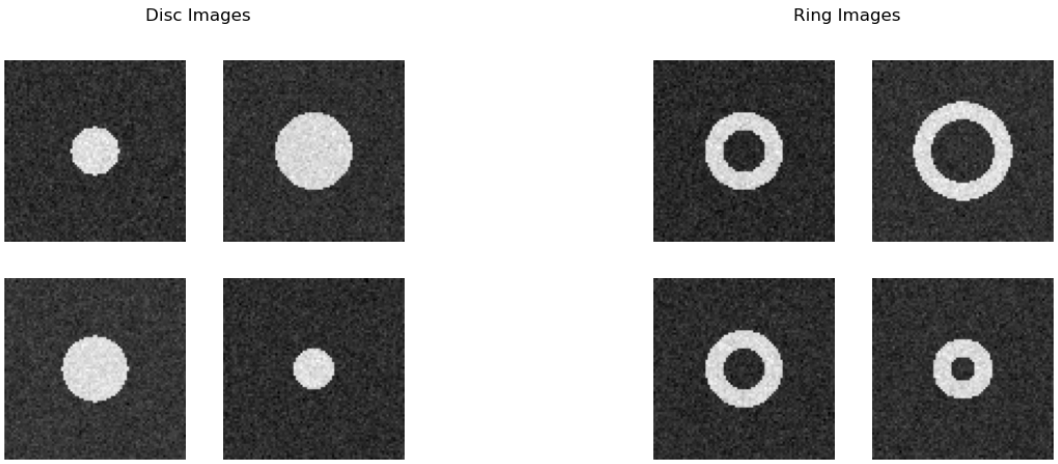
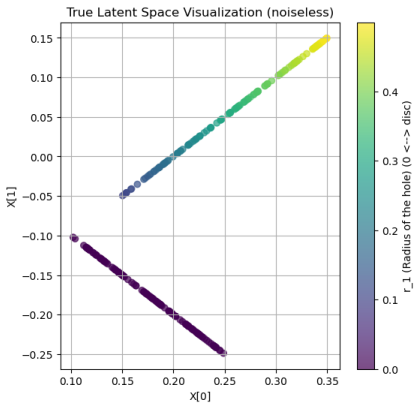
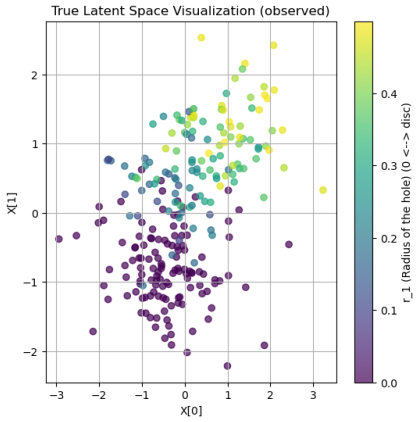


Figure B.1: Example of discs and rings generated using the proposed data generation process.



(a) The noiseless latent space, $0.5 \cdot r_1 + 0.5 \cdot r_2, 0.5 \cdot r_1 - 0.5 \cdot r_2$, colored by r_1.



(b) The first two coordinates latent space provided as the second data view, i.e. (X_1, X_2), standardized, and colored by r_1.

Figure B.2: The true latent space of the image dataset, hidden and observed. The observed latent space is found by adding Gaussian noise to the noiseless latent space.

the structure of the underlying (binary) image data.

We compared three different approaches. The first is to train a vanilla unsupervised VAE on the image dataset, and subsequently apply the sparse asymmetric CCA of Buenfil and Lila (2024), which is a linear sparse CCA approach also encouraging group sparsity on the estimated T, between the latent representation learned by the VAE and X. In training the VAE + sparse CCA model, we use the sum of the first two canonical correlations as the cross-validation metric. The model is then retrained on the full training set.

The second approach is the proposed conditional VAE, applied to X and Y jointly. In training the conditional VAE, we performed cross-validation again over the sum of first two canonical correlations. The model is then retrained on the full training set.

The third approach is the proposed conditional normalizing flow, applied to X and Y jointly. To reflect a realistic scenario, where we have access to an unsupervised pretrained model, we train an unsupervised vanilla VAE on the image data with hyperparameters selected through cross-validation based on reconstruction error. Then, we train a conditional normalizing flow based on Dinh et al. (2017). We construct the NF model by modifying the real NVP model provided by the normflows package of Stimper et al. (2023). Then, we select the hyperparameters of the conditional NF using cross-validation over the sum of the first two canonical correlations. The model is then retrained on the full training set.

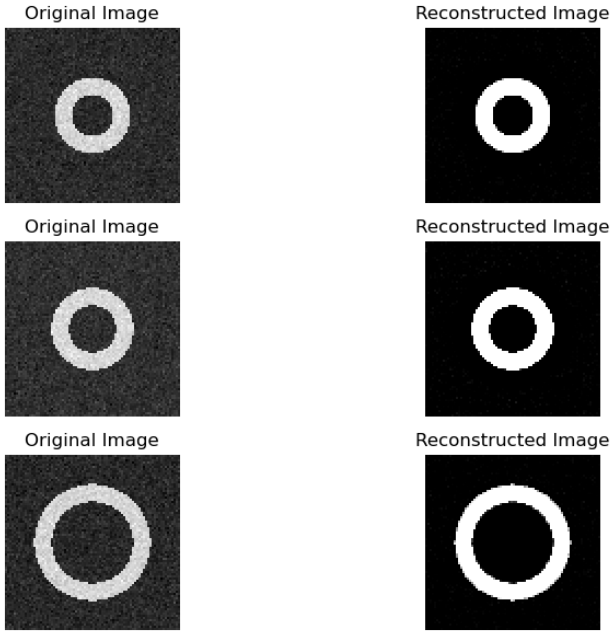
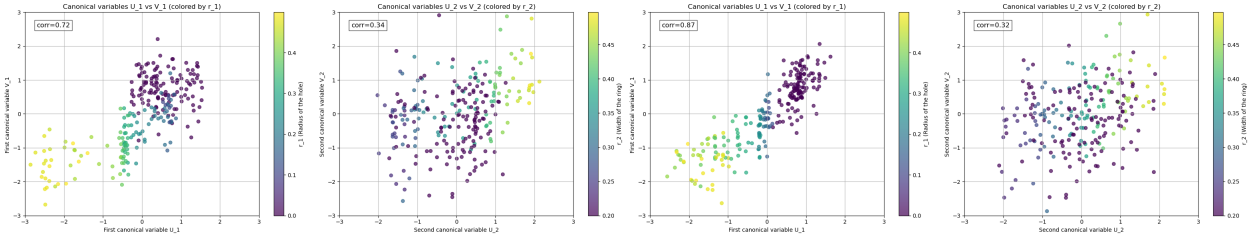


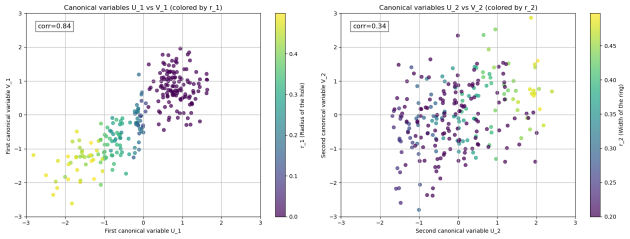
Figure B.3: Typical in-sample reconstructions of rings and discs, which were consistent across the methods compared.

We compare the methods based on their out-of-sample performance on a much larger dataset of 2500 images that follows the same distribution as the training data. The out-of-sample metrics we use are the image reconstruction error, the sum of first two canonical correlations, and the $F1$ score between the sparsity pattern of $\hat{T}$, the estimated high-dimensional canonical vectors, and the true sparsity pattern $((1, 1, 0, 0, 0, 0, 0, 0, 0, 0))$.



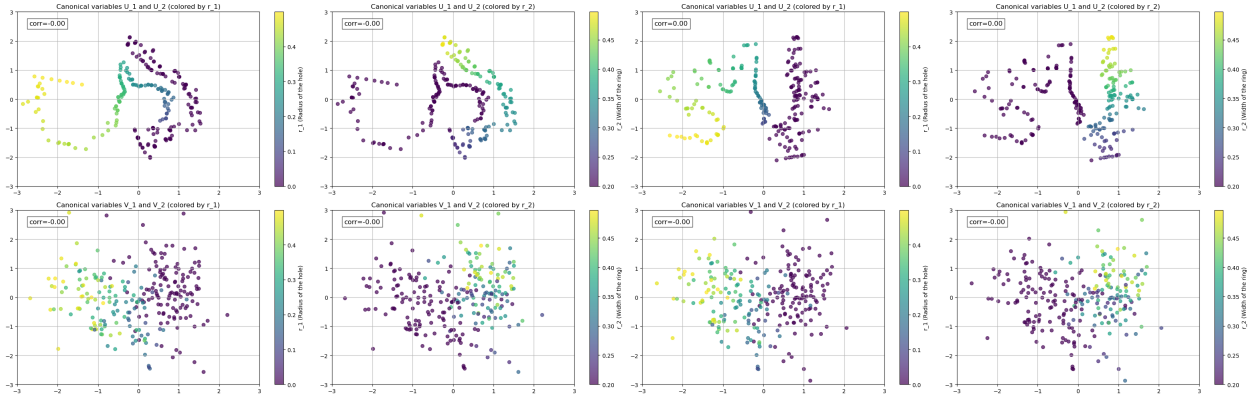
(a) VAE + sparse CCA

(b) Conditional VAE



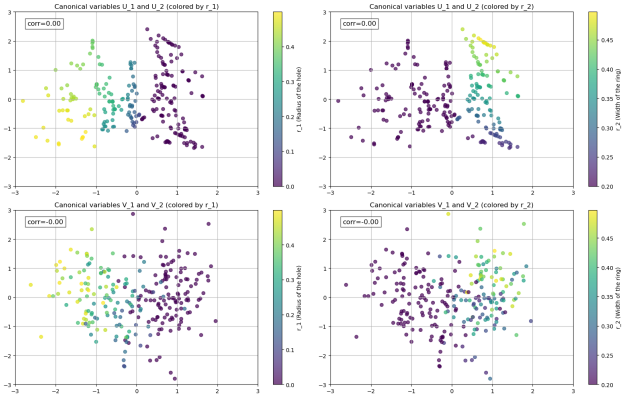
(c) Conditional NF

Figure B.4: The in-sample canonical variables U_1 versus V_1 and U_2 versus V_2, where $U = \hat{H}^\top \hat{g}(Y)$, while $V = \hat{T}^\top X$, for each approach. Here, we plot U_1 versus V_1 colored by r_1, and U_2 versus V_2 colored by r_2, to demonstrate how the different latent dimensions captured different notions of interpretability. In this synthetic example, the latent representation learned from the unsupervised VAE happened to be quite linear, leading to perhaps uncharacteristically high correlations for the VAE + sparse CCA approach, in contrast with our connectome application.



(a) VAE + sparse CCA

(b) Conditional VAE



(c) Conditional NF

Figure B.5: The in-sample canonical variables U and V, for each approach, where $U = \hat{H}^\top \hat{g}(Y)$, while $V = \hat{T}^\top X$. In this synthetic example, the conditional latent variable models are able to capture more global linear structure than the linear sparse CCA approach. However, the inherent nonlinearity in the latent representations, due to the limited encoder and decoder structures, makes linearizing the latent space more challenging.

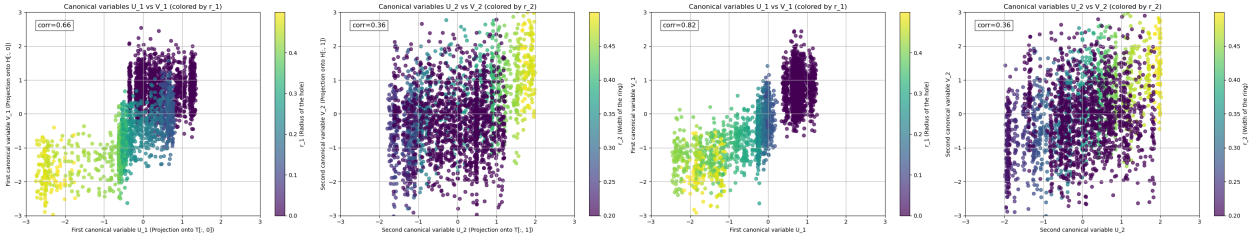
In Figures B.4-B.6 we show both in-sample and out-of-sample comparisons. In Table B.1 we show the out-of-sample metrics computed for each approach. As the validation set is very large, the entries in the table are very close to the population quantities being estimated.

Table B.1: Out-of-sample metrics.

Approach/out-of-sample metric	Sum of Correlations	Reconstruction error	F1 score for T
VAE + sparse CCA	1.02	0.0071	0.66
Conditional VAE	1.19	0.0070	**0.80**
Conditional NF	**1.20**	**0.0069**	0.5

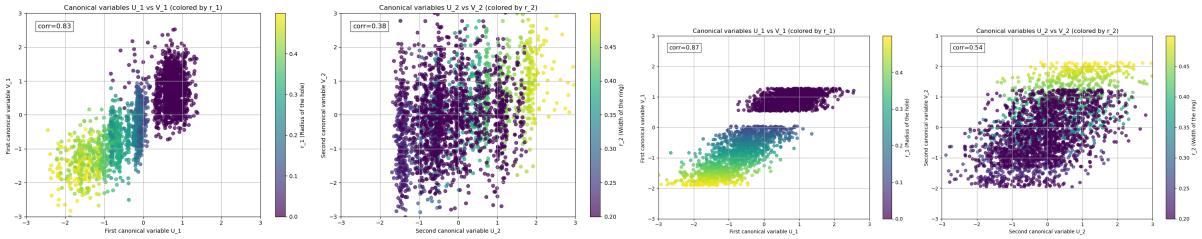
We make several remarks on the results. The conditional latent variable models we proposed clearly outperform the VAE + sparse CCA approach. The VAE + sparse CCA approach still does well, and we attribute this to the simplicity of the example, where a vanilla VAE is already disentangling both r_1 and r_2 effectively. However, in more complex settings, as in our real-world application, guiding the latent representations jointly during training turned out to be essential.

Despite the fact that the conditional latent VAE was cross-validated over correlation, not reconstruction error, its reconstruction error is still comparable to both other methods which used unsupervised VAEs, which do not have to balance minimizing multiple competing objectives. We attribute the relatively high $F1$ score of the VAE + sparse CCA approach to the fact that the sparse CCA approach performs its own cross-validation over its sparsity parameter. In this setting, with a high signal-to-noise ratio, the advantage the conditional models have, performing variable selection jointly while learning the latent representations, is less critical than in more challenging settings.



(a) VAE + sparse CCA

(b) Conditional VAE



(c) Conditional NF

(d) Ideal correlations

Figure B.6: The out-of-sample canonical variables U and V, for each approach, where $U = \hat{H}^\top \hat{g}(Y_{\text{val}})$, while $V = \hat{T}^\top X_{\text{val}}$. We also include the ideal correlation analysis between X and $0.5 \cdot r_1 + 0.5 \cdot r_2, 0.5 \cdot r_1 - 0.5 \cdot r_2$, which each approach tries to approximate.

B.2 Non-invertible nonlinear and partially linear CCA

In this section we provide background on both the (non-invertible) nonlinear and partially linear CCA problems. We are given two observed random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, which without loss of generality we suppose each have mean 0. We define the function classes of interest, the square integrable functions with respect to Y and X:

$$L_Y^2(\mathbb{R}^q, \mathbb{R}^d) \equiv \left\{ g : \mathbb{R}^q \to \mathbb{R}^d : \mathbb{E}\left[g(Y)\right] = 0 \text{ and } \forall i = 1, \ldots d,\ \mathbb{E}\left[g_i(Y)^2\right] < \infty \right\},$$
$$L_X^2(\mathbb{R}^p, \mathbb{R}^d) \equiv \left\{ f : \mathbb{R}^p \to \mathbb{R}^d : \mathbb{E}\left[f(Y)\right] = 0 \text{ and } \forall i = 1, \ldots d,\ \mathbb{E}\left[f_i(Y)^2\right] < \infty \right\}.$$

Then, the nonlinear, or nonparametric CCA problem, is defined as

$$\underset{\substack{g \in L_Y^2(\mathbb{R}^q, \mathbb{R}^d),\ f \in L_X^2(\mathbb{R}^p, \mathbb{R}^d) \\ \Sigma_{g(Y)} = \Sigma_{f(X)} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) f_i(X)\right]. \tag{B.1}$$

It has been shown (Breiman & Friedman, 1985; Michaeli et al., 2016) that the nonparametric CCA problem is equivalent to finding the singular value decomposition (SVD) of an operator S_{12} between $L^2_X(\mathbb{R}^p, \mathbb{R})$ and $L^2_Y(\mathbb{R}^p, \mathbb{R})$, sometimes referred to as the conditional mean operator (Mehta & Harchaoui, 2025)). This operator is in general not compact, and thus the existence of a SVD (and thus a solution to the problem) depends on the dependence structure between X and Y.

The *partially linear canonical correlation analysis* (PLCCA) problem, coined by Michaeli et al. (2016), is defined as follows:

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

where $\theta_i \in \mathbb{R}^p$ is the ith column of $T \in \mathbb{R}^{p \times d}$. We remark that we have altered the formulation in Michaeli et al. (2016) by imposing the constraint that $\mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)] = 0$; this only has the effect of eliminating the first trivial canonical variable solution $f_1(x) = g_1(y) = 1$, and removing the first trivial singular value of S_{12} (see Michaeli et al. (2016) for more details).

PLCCA is a special case of the more general nonparametric CCA problem, and as such, it is also equivalent to an SVD problem. However, in PLCCA, due to the additional constraint that the functions in L^2_X must be linear, L^2_X becomes a finite-dimensional space, so that the operator S_{12} is immediately compact without any assumptions on the dependence structure between X and Y. Thus, the PLCCA problem is fundamentally simpler and of independent interest, and we do not view it as only a special case of the more general problem.

Thanks to the compactness of this operator, the solution to the PLCCA problem can be written down in closed form. Let $\hat{X} \equiv \mathbb{E}[X|Y]$. Then, $T = \Sigma_X^{-1/2}\tilde{T}$, where $\tilde{T} \in \mathbb{R}^{p \times d}$ contains the first d columns of the SVD of $\Sigma_X^{-1/2}\Sigma_{\hat{X}}\Sigma_X^{-1/2}$, and $g(y) = \Sigma_{T^\top \hat{X}}^{-1/2} T^\top \mathbb{E}[X|Y = y]$. We note that in practice, we cannot necessarily use these formulas to compute the solutions, as the conditional expectation $\mathbb{E}[X|Y]$ is not easy to evaluate.

B.3 Supporting results and proofs for Section 3.3

B.3.1 Proof of Theorem 3.3.1

The following lemma is used in the proof of Lemma B.3.2 as well as Theorem 3.3.3. We will also show a more general version of this lemma, Lemma B.5.5, which is used in the proof of Theorem 3.4.3.

Lemma B.3.1. *For a random vector $Z \in \mathbb{R}^d$ with $\Sigma_Z = I_d$, and random vector $X \in \mathbb{R}^p$ with Σ_X invertible with $p \geq d$, with $\mathbb{E}[X] = 0$, $\mathbb{E}[Z] = 0$, we have*

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \left\|\Sigma_Z^{-1/2} \Sigma_{ZX} \Sigma_X^{-1/2}\right\|_F^2, \tag{B.3}$$

where the $\theta_i \in \mathbb{R}^p$ are the columns of T, and the $\eta_i \in \mathbb{R}^d$ are the columns of H. This is equivalent to the fact that, in classical CCA, if we optimize over the sum of squares of the correlations, rather than the usual sum of the correlations, then the solution does not change.

Proof of Lemma B.3.1

We begin from

$$\sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \sum_{i=1}^{d} \mathbb{E}\left[\theta_i^\top X Z^\top \eta_i\right]^2 \tag{B.4}$$

$$= \sum_{i=1}^{d} \left(\theta_i^\top \Sigma_{XZ} \eta_i\right)^2 \tag{B.5}$$

Changing variables from T to $\tilde{T} = \Sigma_X^{1/2} T$, whose columns we denote by $\tilde{\theta}$, we have

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \left(\theta_i^\top \Sigma_{XZ} \eta_i\right)^2 \tag{B.6}$$

$$= \sup_{\substack{\tilde{T} \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \sum_{i=1}^{d} \left(\tilde{\theta}_i^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} \eta_i\right)^2 \tag{B.7}$$

$$\leq \sup_{\substack{\tilde{T} \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \left\|\tilde{T}^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} H\right\|_F^2 \tag{B.8}$$

where in the last inequality, we have used that the diagonal entries of $\tilde{T}^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} H$ are the $\tilde{\theta}_i^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} \eta_i$, so that there is equality when this matrix is diagonal.

Taking the singular value decomposition of $\Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} = U \Lambda V^\top$, where $U \in \mathbb{R}^{p \times d}$, $\Lambda \in \mathbb{R}^{d \times d}$, $V \in \mathbb{R}^{d \times d}$ with $U^\top U = V^\top V = I_d$ and Λ diagonal, we have that

$$\sup_{\substack{T \in \mathbb{R}^{p \times d}, H \in \mathbb{R}^{d \times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 \leq \sup_{\substack{\tilde{T} \in \mathbb{R}^{p \times d}, H \in \mathbb{R}^{d \times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \left\| \tilde{T}^\top U \Lambda V^\top H \right\|_F^2 \tag{B.9}$$

$$\leq \|\Lambda\|_F^2 \tag{B.10}$$

$$= \left\| \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} \right\|_F^2 \tag{B.11}$$

where in the second inequality we have used the fact that multiplication by orthogonal matrices can only make the Frobenius norm smaller. We can attain this upper bound by choosing $\tilde{T} = U$ and $H = V$, which also makes the matrix $\tilde{T}^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} H$ diagonal. Thus, both inequalities are equalities, and the proof is complete. □

The following result is used in proof of several theorems.

Lemma B.3.2. *The nonlinear solution g maximizes the PLCCA problem*

$$\underset{\substack{g: \mathbb{R}^q \to \mathbb{R}^d, \ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}_{\text{VAE}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) \theta_i^\top X\right]^2. \tag{B.12}$$

if and only if it minimizes the nonlinear regression problem

$$\underset{\substack{g: \mathbb{R}^q \to \mathbb{R}^d, \ B \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = I_d, \ g \in \mathcal{C}_{\text{VAE}}}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right]. \tag{B.13}$$

Proof of Lemma B.3.2

We start from

$$\underset{\substack{g: \mathbb{R}^q \to \mathbb{R}^d, \ B \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = I_d, \ g \in \mathcal{C}_{\text{VAE}}}}{\text{minimize}} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right]. \tag{B.14}$$

and suppress the constraints in the notation for convenience. We have

$$\inf_{g,B} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] = \inf_{g,B} \mathbb{E}\left[g(Y)^\top g(Y) - 2g(Y)^\top B^\top X + X^\top B B^\top X\right] \tag{B.15}$$

$$= \inf_{g,B} \mathbb{E}\left[\operatorname{tr}\left(g(Y)^\top g(Y) - 2g(Y)^\top B^\top X + X^\top B B^\top X\right)\right] \tag{B.16}$$

$$= \inf_{g,B} \operatorname{tr}\left(\mathbb{E}\left[g(Y)g(Y)^\top - 2B^\top X g(Y)^\top + B^\top X X^\top B\right]\right) \tag{B.17}$$

$$= \inf_{g,B} \operatorname{tr}\left(\Sigma_{g(Y)} - 2B^\top \Sigma_{Xg(Y)} + B^\top \Sigma_X B\right) \tag{B.18}$$

$$= \inf_{g,B} \operatorname{tr}\left(\Sigma_{g(Y)}\right) - 2\operatorname{tr}\left(B^\top \Sigma_{Xg(Y)}\right) + \operatorname{tr}\left(B^\top \Sigma_X B\right). \tag{B.19}$$

Since this is a least squares problem, we can plug in the optimal B for fixed g, which is simply $B = \Sigma_X^{-1}\Sigma_{Xg(Y)}$:

$$\inf_{g,B} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] = \inf_g \inf_B \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] \tag{B.20}$$

$$= \inf_g \inf_B \operatorname{tr}\left(\Sigma_{g(Y)}\right) - 2\operatorname{tr}\left(B^\top \Sigma_{Xg(Y)}\right) + \operatorname{tr}\left(B^\top \Sigma_X B\right) \tag{B.21}$$

$$= \inf_g \left[\operatorname{tr}\left(\Sigma_{g(Y)}\right) - 2\operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right) + \operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_X\Sigma_X^{-1}\Sigma_{Xg(Y)}\right)\right] \tag{B.22}$$

$$= \inf_g \left[\operatorname{tr}\left(\Sigma_{g(Y)}\right) - \operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right)\right], \tag{B.23}$$

and using the fact that $\Sigma_{g(Y)} = I_d$, we obtain

$$\inf_{g,B} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] = \inf_g \left[d - \left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2\right]. \tag{B.24}$$

From classical CCA we recognize the matrix $\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}$ as the matrix whose singular value decomposition provides the solutions to the canonical correlation problem between $g(Y)$ and X. In particular, we can use the Lemma B.3.1 to obtain

$$\left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2 = \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^d \mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2, \tag{B.25}$$

where the $\theta_i \in \mathbb{R}^p$ are the columns of T and the $\eta_i \in \mathbb{R}^d$ are the columns of H. Combining

this with equation (B.24), we obtain

$$\inf_{g,B} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] = \inf_{g}\left[d - \left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2\right] \tag{B.26}$$

$$= \inf_{g}\left[d - \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^{d}\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right] \tag{B.27}$$

$$= d - \sup_{g}\sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}}\left[\sum_{i=1}^{d}\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right] \tag{B.28}$$

$$= d - \sup_{g}\sup_{\substack{T\in\mathbb{R}^{p\times d},\\ \Sigma_{T^\top X}=I_d}}\left[\sum_{i=1}^{d}\mathbb{E}\left[g(Y)\theta_i^\top X\right]^2\right], \tag{B.29}$$

where in the last step we have absorbed the optimization over H into g. This completes the proof. □

Proof of Theorem 3.3.1

Uniform convergence of a sequence of $\mathbb{R}^d$ valued functions on K refers to convergence in the supremum norm $\|g\|_\infty \equiv \sup_{y\in K}\|g(y)\|_2$. From Lemma B.3.2, the maximization problem of interest is equivalent to

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=I_d,\ g\in\mathcal{C}_{\mathrm{VAE}}}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right]. \tag{B.30}$$

From this, we can see that given an optimal g, the optimal B is the least-squares solution $B \equiv \Sigma_X^{-1}\Sigma_{Xg(Y)}$, and plugging this B back in (using equation (B.23)), we obtain an optimization problem over only g,

$$\underset{g\in\mathcal{C}_{\mathrm{VAE}},\Sigma_{g(Y)}=I_d}{\text{minimize}} J(g), \tag{B.31}$$

where

$$J(g) \equiv d - \mathrm{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right). \tag{B.32}$$

Therefore, finding an optimal pair (g,T) reduces to finding $g \in \mathcal{C}_{\mathrm{VAE}}$ which attains the value $\inf_{g\in\mathcal{C}_{\mathrm{VAE}},\Sigma_{g(Y)}=I_d} J(g)$, since by Theorem 3.3.3, an optimal T can be derived from B. This

infimum is not $-\infty$ provided that $\mathcal{C}_{\text{VAE}} \cap \{g : \Sigma_{g(Y)} = I_d\} \neq \varnothing$, so that the feasible set is not empty.

We begin with a sequence $\{g_n\} \subset \mathcal{C}_{\text{VAE}}$ for which $J(g_n)$ converges to $\inf_{g \in \mathcal{C}_{\text{VAE}}} J(g)$, and set out to show that $\{g_n\}$ has a subsequence that converges uniformly to a $g^* \in \mathcal{C}_{\text{VAE}}$ with $\Sigma_{g^*(Y)} = I_d$. Having shown this, once we show $J(g)$ is continuous in g with respect to the supremum norm, the proof will be complete, since continuity implies that the limit of the subsequence satisfies $J(g^*) = \inf_{g \in \mathcal{C}_{\text{VAE}}, \Sigma_{g(Y)} = I_d} J(g)$ with $g^* \in \mathcal{C}_{\text{VAE}}$ and $\Sigma_{g^*(Y)} = I_d$.

We show continuity of $J(g)$ with respect to the supremum norm. We need to show that if $\{g_n\} \subset \mathcal{C}_{\text{VAE}}$ and $g \in \mathcal{C}_{\text{VAE}}$ are such that $\lim_{n \to \infty} \sup_{y \in K} \|g_n(y) - g(y)\|_2 = 0$, then $\Sigma_{g_n(Y)X}$ converges to $\Sigma_{g(Y)X}$ in the Frobenious norm, since the trace function is continuous with respect to the Frobenious norm and because the composition of continuous functions is continuous. We have

$$\begin{aligned}
\left\|\Sigma_{g_n(Y)X} - \Sigma_{g(Y)X}\right\|_F &= \left\|\mathbb{E}\left[\left(g_n(Y) - g(Y)\right) X^\top\right]\right\|_F && \text{(B.33)} \\
&\leq \mathbb{E}\left[\left\|\left(g_n(Y) - g(Y)\right) X^\top\right\|_F\right] && \text{(B.34)} \\
&= \mathbb{E}\left[\left\|g_n(Y) - g(Y)\right\|_2 \left\|X\right\|_2\right] && \text{(B.35)} \\
&\leq \left\|g_n - g\right\|_\infty \mathbb{E}\left[\left\|X\right\|_2\right]. && \text{(B.36)}
\end{aligned}$$

Therefore, $J(g)$ is continuous.

Now we show that $\{g_n\}$ has a convergent subsequence. Since g is M-Lipschitz, $\|g(y)\|_2$ is contained within a closed interval $I \subset \mathbb{R}$ with length $\operatorname{diam}(K)M$. Because $\mathbb{E}[g(Y)] = 0$, I necessarily contains 0, so that $\mathcal{C}_{\text{VAE}}$ is uniformly bounded.

Since the Lipschitz assumption gives equicontinuity of $\mathcal{C}_{\text{VAE}}$, uniform boundedness on the compact set K allows us to use the Arzelà–Ascoli Theorem (Rudin (1976) Theorem 7.25) on each coordinate of $\{g_n\}$ to construct d subsequences, each containing the last, so that the final subsequence admits a subsequence converging uniformly to some continuous g^*.

Corresponding to each g_n in this sequence is an $f_n : \mathbb{R}^d \to \mathbb{R}^q$ which satisfies $\mathbb{E}[f_n(g_n(Y))] = 0$, f_n is m-Lipschitz, and $\mathbb{E}\|Y - f_n(g_n(Y))\|_2^2 \leq \varepsilon$. Similarly to $\{g_n\}$, the condition that $\mathbb{E}[f_n(g_n(Y))] = 0$ along with f_n being m-Lipschitz implies that the $\{f_n\}$ are

uniformly bounded. Since the $\{g_n\}$ are uniformly bounded, their range can be restricted to a closed ball of $\mathbb{R}^d$, which is in particular compact. Therefore, without loss of generality we can restrict the domain of each f_n to this compact set K_1, and again use the Arzelà–Ascoli Theorem to construct a final subsequence $\{f_n\}$ which uniformly converges to some continuous f^* on K_1, with a corresponding subsequence $\{g_n\}$ retaining its previous properties, which we do not relabel for notational convenience.

It remains to be shown that g^* satisfies $\Sigma_{g(Y)} = I_d$, and that g^* belongs to $\mathcal{C}_{\text{VAE}}$. Namely, we must show that $\mathbb{E}[g^*(Y)] = 0$, g^* is M-Lipschitz, and that f^* satisfies the decoder condition for g^*: that $\mathbb{E}[f^*(g^*(Y))] = 0$, f^* is m-Lipschitz, and that $\mathbb{E}\|Y - f^*(g^*(Y))\|_2^2 \le \varepsilon$.

That g^* is M-Lipschitz follows directly from the uniform convergence of $\{g_n\}$, and similarly, f^* is m-Lipschitz. Since $\mathcal{C}_{\text{VAE}}$ is uniformly bounded, we can apply the dominated convergence theorem on each coordinate of $\{g_n\}$, as well as $\{g_n g_n^\top\}$ to show that

$$\lim_{n\to\infty} \mathbb{E}[g_n(Y)] = \mathbb{E}[g^*(Y)] \text{ and} \tag{B.37}$$

$$\lim_{n\to\infty} \mathbb{E}[g_n(Y)g_n(Y)^\top] = \mathbb{E}[g^*(Y)g^*(Y)^\top]. \tag{B.38}$$

Therefore, g^* satisfies $\mathbb{E}[g^*(Y)] = 0$ and $\Sigma_{g^*(Y)} = I_d$. Finally, we have

$$\|f_n(g_n(Y)) - f^*(g^*(Y))\|_2 \le \|f_n(g_n(Y)) - f_n(g^*(Y))\|_2 + \|f_n(g^*(Y)) - f^*(g^*(Y))\|_2 \tag{B.39}$$

$$\le m\,\|g_n(Y) - g^*(Y)\|_2 + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty \tag{B.40}$$

$$\le m\,\|g_n - g^*\|_\infty + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty, \tag{B.41}$$

Therefore,

$$\mathbb{E}\left[\|(Y - f_n(g_n(Y))) - (Y - f^*(g^*(Y)))\|_2^2\right]^{1/2} \le m\,\|g_n - g^*\|_\infty + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty, \tag{B.42}$$

which converges to 0 by the uniform convergence of $\{f_n\}$ and $\{g_n\}$. Letting $\|Z\|_{L^2} \equiv \mathbb{E}\left[\|Z\|_2^2\right]^{1/2}$ denote the L^2 norm of a random vector $Z \in \mathbb{R}^q$, the reverse triangle inequality implies that

$$|\|Y - f_n(g_n(Y))\|_{L^2} - \|Y - f^*(g^*(Y))\|_{L^2}| \le \|Y - f_n(g_n(Y))) - (Y - f^*(g^*(Y)))\|_{L^2}, \tag{B.43}$$

so that

$$\lim_{n\to\infty} \mathbb{E}\left[\|Y - f_n(g_n(Y))\|_2^2\right]^{1/2} = \mathbb{E}\left[\|Y - f^*(g^*(Y))\|_2^2\right]^{1/2}, \tag{B.44}$$

and squaring, we obtain

$$\mathbb{E}\left[\|Y - f^*(g^*(Y))\|_2^2\right] = \lim_{n\to\infty} \mathbb{E}\left[\|Y - f_n(g_n(Y))\|_2^2\right] \le \varepsilon. \tag{B.45}$$

Similarly,

$$\mathbb{E}\left[f^*(g^*(Y))\right] = \lim_{n\to\infty} \mathbb{E}\left[f_n(g_n(Y))\right] = 0. \tag{B.46}$$

This completes the proof. □

B.3.2 Proof of Theorem 3.3.3

The proof is analogous to the proof of Lemma B.3.2. We start with

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\mathrm{VAE}}}}{\text{minimize}} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right]. \tag{B.47}$$

From equation (B.23), for the optimal B for a fixed g,

$$\begin{aligned}
\mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] &= \operatorname{tr}\left(\Sigma_{g(Y)}\right) - \operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right) && \text{(B.48)}\\
&= \operatorname{tr}\left(\Sigma_{g(Y)}\right) - \operatorname{tr}\left(\Sigma_{g(Y)}\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}\right) && \text{(B.49)}\\
&= \operatorname{tr}\left(\Sigma_{g(Y)}\left[I_d - \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}\right]\right). && \text{(B.50)}
\end{aligned}$$

We denote $D \equiv I_d - \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}$ for notational ease. We note that $\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2} = AA^\top$ where $A = \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}$. From classical CCA, the singular values of A are the canonical correlations between $g(Y)$ and X, so they are all between 0 and 1. Therefore, D is positive semi-definite, with eigenvalues equal to $1 - \gamma_i^2$, where γ_i is the ith canonical correlation between $g(Y)$ and X. Thus,

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\mathrm{VAE}}}} \mathbb{E}\left[\|B^\top X - g(Y)\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\mathrm{VAE}}}} \operatorname{tr}\left(\Sigma_{g(Y)}D\right) \tag{B.51}$$

It is straightforward to show that as a function of $\Sigma_{g(Y)}$, that subject to the constraint that $\Sigma_{g(Y)} \succeq cI_d$ this problem can be minimized by $\Sigma_{g(Y)} = cI_d$, thanks to the fact that D does not depend on $\Sigma_{g(Y)}$. Therefore,

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\text{VAE}}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\text{VAE}}}} c\sum_{i=1}^{d}(1-\gamma_i^2) \tag{B.52}$$

$$= c\left(d - \sup_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\text{VAE}}}} \sum_{i=1}^{d}\gamma_i^2\right), \tag{B.53}$$

Now using Lemma B.3.1, we have

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\text{VAE}}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = c\left(d - \sup_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}_{\text{VAE}}}} \sup_{\substack{T\in\mathbb{R}^{p\times d},H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^{d}\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right). \tag{B.54}$$

Writing this last supremum over $h \equiv H^\top g$ completes the proof of the first part of the statement. Additionally, we can see that to obtain h from g (by finding H) as well as find T, we take the optimal g from the regression problem and solve a linear CCA between g and X (appealing to Lemma B.3.1 that maximizing the sum of squares of the correlations is the same as maximizing the sum of the correlations in linear CCA).

Fixing the optimal (g, B) for the regression problem, where g may not satisfy $\Sigma_{g(Y)} = I_d$, from classical CCA we know that we can find (T, H) that solves the CCA problem between $g(Y)$ and X via the SVD of the following matrix:

$$\Sigma_X^{1/2}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2} = \tilde{T}\Lambda\tilde{H}^\top, \tag{B.55}$$

where the solutions are

$$T = \Sigma_X^{-1/2}\tilde{T} \tag{B.56}$$

$$H = \Sigma_{g(Y)}^{-1/2}\tilde{H}. \tag{B.57}$$

On the other hand, from its definition, it is straightforward to show that $B = \Sigma_X^{-1}\Sigma_{Xg(Y)}$. Therefore, we have

$$\Sigma_{g(Y)}^{-1/2} B^\top \Sigma_X B \Sigma_{g(Y)}^{-1/2} = \tilde{H}\Lambda\tilde{H}^\top, \tag{B.58}$$

and

$$T = BH\Lambda^{-1}. \tag{B.59}$$

which follow from matrix algebra. This completes the proof. □

B.4 Supporting results and proofs for conditional VAEs

B.4.1 Derivation of conditional VAE objective

Introducing the latent variable $Z \in \mathbb{R}^d$, the conditional density $p(y|x)$ can be written as

$$p(y|x) = \frac{p(x,y)}{p(x)}\frac{p(x,y,z)}{p(x,y,z)} \tag{B.60}$$

$$= \frac{p(y,z|x)}{p(z|x,y)}\frac{q(z|y)}{q(z|y)}. \tag{B.61}$$

The log-likelihood is

$$\ln p(y|x) = \ln\left(\frac{p(y,z|x)}{q(z|y)}\right) + \ln\left(\frac{q(z|y)}{p(z|x,y)}\right) \tag{B.62}$$

$$= \mathbb{E}_{q(z|y)}\left[\ln\left(\frac{p(y,z|x)}{q(z|y)}\right)\right] + D_{KL}\left(q(z|y), p(z|y,x)\right). \tag{B.63}$$

Since the KL divergence is nonnegative, we obtain the ELBO lower bound:

$$\text{ELBO} = \mathbb{E}_{q(z|y)}\left[\ln\left(\frac{p(y,z|x)}{q(z|y)}\right)\right]. \tag{B.64}$$

Supposing $p(y,z|x) = p(y|z)p(z|x)$, we can write this as

$$\text{ELBO} = \mathbb{E}_{q(z|y)}\left[\ln p(y|z)\right] - D_{KL}\left(q(z|y), p(z|x)\right). \tag{B.65}$$

If we choose to specify the model as

$$Y|Z,X \sim \mathcal{N}\left(f(Z), I_d\right) \tag{B.66}$$

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{B.67}$$

and we model $q(z|y) \sim \mathcal{N}(g(Y), I_d)$, then to maximize the ELBO we can equivalently minimize

$$\frac{1}{2}\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right] + \frac{1}{2}\|g(Y) - B^\top X\|_2^2, \tag{B.68}$$

where we have used the expression for the KL divergence between two Gaussian distributions with the same covariance, and where we have plugged in the population quantities X and Y in place of x and y. In practice, X and Y are observed, and we estimate the expectation of the above quantity with respect to (X, Y):

$$\underset{g:\mathbb{R}^q \to \mathbb{R}^d,\ B \in \mathbb{R}^{p\times d}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right] + \mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right], \tag{B.69}$$

where the outermost expectation is with respect to (X, Y), the observed quantities.

B.4.2 Proof of Theorem 3.4.1

We introduce tools from information theory that will be used in the upcoming proofs (see Chapters 2 and 8 of Cover (1999)).

The *differential entropy*, or just entropy for short, of a random vector Y with probability density function $f_Y(y)$, is denoted by

$$h(Y) \equiv -\mathbb{E}\left[\log\left(p(Y)\right)\right]. \tag{B.70}$$

when it exists and is finite.

The *mutual information* between random vectors X,Y with density functions f_X and f_Y and joint density $f_{X,Y}$ is defined as

$$I(X;Y) \equiv \mathbb{E}\left[\log\left(\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right)\right]. \tag{B.71}$$

when it exists and is finite.

The *conditional entropy* between X and Y, where $f_{X|Y}(x|y)$ denotes the conditional density of X given Y, as

$$h(Y|X) \equiv \mathbb{E}\left[\log\left(f_{X|Y}(X|Y)\right)\right] \tag{B.72}$$

when it exists and is finite.

We collect the following results from Cover (1999), in particular Theorems 8.4.1, 8.6.5, 2.8.1, and equation (2.39).

Lemma B.4.1. *1. For a Gaussian random vector $Y \in \mathbb{R}^d$ with covariance Σ_Y, we have the following expression for its differential entropy:*

$$h(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det(\Sigma_Y). \tag{B.73}$$

2. Differential entropy is maximized for Gaussian distributions: for any random vector $Z \in \mathbb{R}^d$ with covariance Σ_Z whose entropy exists, we have

$$h(Z) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det(\Sigma_Z). \tag{B.74}$$

3. Data processing inequality: If $X \to Y \to Z$ is a Markov chain, then $I(X;Y) \geq I(X;Z)$.

4. $I(X;Y) = h(X) - h(X|Y)$.

We state an additional auxiliary inequality, which will be used in the proofs of Theorems 3.4.1 and 3.4.3.

Lemma B.4.2 (GM-AM Eigenvalue Inequality)**.** *For positive semi-definite $A \in \mathbb{R}^{d\times d}$, we have*

$$\frac{\mathrm{tr}(A)}{d} \geq \det(A)^{1/d}. \tag{B.75}$$

We have equality when A is a multiple of the identity matrix I_d.

Proof of Lemma B.4.2

This is the well-known GM-AM (geometric mean-arithemetric mean) inequality applied to the eigenvalues of a positive semi-definite matrix. We give a proof for completeness Let λ_i be the eigenvalues of A. Then Jensen's inequality applied to log gives

$$\log\left(\frac{\mathrm{tr}(A)}{d}\right) = \log\left(\frac{\sum_{i=1}^d \lambda_i}{d}\right) \geq \frac{\sum_{i=1}^d \log(\lambda_i)}{d} = \log\left(\prod_{i=1}^d \lambda_i^{1/d}\right) = \log\left(\det(A)^{1/d}\right). \tag{B.76}$$

Taking the exponential of each side completes the proof. Equality when A is a multiple of the identity matrix I_d can be seen by evaluating both sides. □

Proof of Theorem 3.4.1

Following Theorem 10.2.1 of Cover (1999), we define the rate distortion function $R_Y(D)$ of Y under the squared error as

$$R_Y(D) \equiv \inf_{P_{\hat{Y}|Y}:\mathbb{E}\left[\|Y-\hat{Y}\|_2^2 \le D\right]} I(Y;\hat{Y}) \tag{B.77}$$

where $P_{\hat{Y}|Y}$ denotes the set of conditional distributions over a random vector $\hat{Y} \in \mathbb{R}^q$ given $Y \in \mathbb{R}^q$. We note that for $D \ge \mathbb{E}\left[\|Y\|_2^2\right]$ we have $R_Y(D) = 0$, since then $\hat{Y} = 0$ satisfies the given constraint that $\mathbb{E}\left[\left\|Y - \hat{Y}\right\|_2^2 \le D\right]$ and $I(Y;0) = 0$. The infimum is then 0 since mutual information is always nonnegative.

From the distribution $q(Z|Y)$, $Z = g(Y) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, I_d\sigma_{\text{enc}}^2)$ and ε is independent from Y. Let $\hat{Y} \equiv f(Z)$, i.e. the noisy reconstruction of Y. Then from the definition of $R_Y(\delta)$,

$$R_Y(\delta) \le I(Y;\hat{Y}). \tag{B.78}$$

We have that

$$I(Y;\hat{Y}) \le I(Y;Z) \tag{B.79}$$

by the data processing inequality (Lemma B.4.1) since $Y \to Z \to \hat{Y}$ is a Markov chain. Then,

$$\begin{aligned} I(Y;Z) &= I(Z;Y) && \text{(B.80)}\\ &= h(Z) - h(Z|Y) && \text{(B.81)}\\ &= h(Z) - h(Z|g(Y)) && \text{(B.82)} \end{aligned}$$

where we have used the symmetry of mutual information, item 4 of Lemma B.4.1, and the fact that Z only depends on Y through $g(Y)$.

Since $Z = g(Y) + \varepsilon$, we have $h(Z|g(Y)) = h(\varepsilon)$. Because ε is Gaussian, Lemma B.4.1 implies that

$$h(Z|g(Y)) = \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log\left(\sigma_{\text{enc}}^2\right). \tag{B.83}$$

Noting that mutual information is invariant to addition by constants, we can suppose that Z is mean 0 without loss of generality, so that $\Sigma_Z = \Sigma_{g(Y)} + \Sigma_\varepsilon = \Sigma_{g(Y)} + \sigma_{\text{enc}}^2 I_d$. Since differential entropy is maximized for Gaussian distributions, Lemma B.4.1 applied to Z implies that

$$h(Z) \le \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_{g(Y)} + \sigma_{\text{enc}}^2 I_d\right). \tag{B.84}$$

Putting these inequalities together, we have shown that

$$\begin{aligned} R_Y(\delta) &\le \frac{1}{2}\log\det\left(\Sigma_{g(Y)} + \sigma_{\text{enc}}^2 I_d\right) - \frac{d}{2}\log\left(\sigma_{\text{enc}}^2\right) && \text{(B.85)}\\ &= \frac{1}{2}\log\det\left(\frac{1}{\sigma_{\text{enc}}^2}\Sigma_{g(Y)} + I_d\right) + \frac{d}{2}\log\left(\sigma_{\text{enc}}^2\right) - \frac{d}{2}\log\left(\sigma_{\text{enc}}^2\right) && \text{(B.86)}\\ &= \frac{1}{2}\log\det\left(\frac{1}{\sigma_{\text{enc}}^2}\Sigma_{g(Y)} + I_d\right). && \text{(B.87)} \end{aligned}$$

Applying the GM-AM inequality Lemma B.4.2, this is upper bounded:

$$\begin{aligned} \frac{1}{2}\log\det\left(\frac{1}{\sigma_{\text{enc}}^2}\Sigma_{g(Y)} + I_d\right) &\le \frac{1}{2}\log\left(\frac{\operatorname{tr}\left(\frac{1}{\sigma_{\text{enc}}^2}\Sigma_{g(Y)} + I_d\right)}{d}\right)^d && \text{(B.88)}\\ &= \frac{d}{2}\log\left(\operatorname{tr}\left(\frac{1}{d\sigma_{\text{enc}}^2}\Sigma_{g(Y)}\right) + 1\right). && \text{(B.89)} \end{aligned}$$

Combining this with equation (B.87) and rearranging for $\Sigma_{g(Y)}$, we finally have

$$\operatorname{tr}\left(\Sigma_{g(Y)}\right) \ge \sigma_{\text{enc}}^2 d\left(e^{\frac{2}{d}R_Y(\delta)} - 1\right), \tag{B.90}$$

so that $C(\delta)$ from the statement of the proof is $C(\delta) \equiv d\left(e^{\frac{2}{d}R_Y(\delta)} - 1\right)$. Since for $D \ge \mathbb{E}\left[\|Y\|_2^2\right]$ we have $R_Y(D) = 0$, we can check that this translates into $C(D) = 0$ as well. This completes the proof. □

Using equation (B.87) directly, we also have the tighter bound related to the determinant of $\Sigma_{g(Y)}$:

$$\det\left(\frac{1}{\sigma_{\text{enc}}^2}\Sigma_{g(Y)} + I_d\right) \ge e^{2R_Y(\delta)}. \tag{B.91}$$

This reflects the same intuition as the trace bound: the eigenvalues of $\Sigma_{g(Y)}$ being large depends on the reconstruction error δ being small and the encoder variance σ^2_{enc} being large.

B.5 Supporting results and proofs for conditional NFs

B.5.1 Derivation of conditional NF objective

In this case, we already have access to a dimension-reduced Y, which we call W, so that $Y = \psi(W)$.

We introduce the model

$$W = \tilde{f}(Z) \tag{B.92}$$

$$Z = B^\top X + \varepsilon_2 \tag{B.93}$$

where $\tilde{f}$ is injective. We would like to derive the form of the conditional distribution $p(w|x)$.

We begin with the joint distribution, and use the standard transformation of variables formula to obtain

$$p_{W,X}(w,x) = p_{Z,X}\left(\tilde{f}(w),x\right)\left|\det\left(J_{\tilde{f}}(w)\right)\right|, \tag{B.94}$$

where $J_{\tilde{g}}(w)$ denotes the Jacobian matrix of a smooth function $\tilde{g}$ evaluated at w. Then, denoting $\tilde{g} \equiv \tilde{f}^{-1}$, the conditional density is

$$p_{W|X}(w|x) = p_{Z|X}\left(\tilde{g}(w)|x\right)\left|\det\left(J_{\tilde{g}}(w)\right)\right|, \tag{B.95}$$

so that the log of the conditional likelihood is

$$\ln p_{W|X}(w|x) = \ln p_{Z|X}\left(\tilde{g}(w)|x\right) + \ln\left|\det\left(J_{\tilde{g}}(w)\right)\right|. \tag{B.96}$$

Using the model

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{B.97}$$

and writing the maximization of the log-likelihood in its population form, we obtain

$$\underset{\tilde{g},B}{\text{maximize}}\ \mathbb{E}\left[-\frac{1}{2}\left\|\tilde{g}(W) - B^\top X\right\|_2^2 + \ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right]. \tag{B.98}$$

This is equivalent to minimizing

$$\underset{\tilde{g}\in\mathcal{C}_{\mathrm{NF}},B\in\mathbb{R}^{p\times d}}{\text{minimize}}\ \mathbb{E}\left[\frac{1}{2}\left\|\tilde{g}(W)-B^{\top}X\right\|_2^2-\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right], \tag{B.99}$$

where we have now specified the constraint set for feasible $\tilde{g}$.

B.5.2 Proof of Theorem 3.4.2

Lemma B.5.1 (Gaussian Poincaré Inequality)**.** *For Z Gaussian and isotropic, and $f\in\mathcal{C}_{\mathrm{NF}}$, we have*

$$\Sigma_{f(Z)}\preceq\mathbb{E}\left[J_f(Z)J_f(Z)^{\top}\right] \tag{B.100}$$

Proof of Lemma B.5.1

We provide the proof for completeness; for the idea of the proof and more general results, see Theorem 2.4 of Huang and Tropp (2021). The standard Gaussian Poincaré inequality says that (see Theorem 3.20 of Boucheron et al. (2013)), for differentiable $g:\mathbb{R}^d\to\mathbb{R}$,

$$\mathrm{Var}\left(g(Z)\right)\leq\mathbb{E}\left[\left\|\nabla g(Z)\right\|_2^2\right]. \tag{B.101}$$

To show the desired result, we simply apply this inequality to the function $u^{\top}f(y):\mathbb{R}^d\to\mathbb{R}$ for $u\in\mathbb{R}^d$ to obtain

$$\mathrm{Var}\left(u^{\top}f(Z)\right)\leq\mathbb{E}\left[\left\|\nabla\left(u^{\top}f\right)(Z)\right\|_2^2\right]. \tag{B.102}$$

The left hand side is

$$\begin{aligned}\mathrm{Var}\left(u^{\top}f(Z)\right)&=\mathbb{E}\left[\left(u^{\top}f(Z)\right)^2\right] &\text{(B.103)}\\ &=\mathbb{E}\left[u^{\top}f(Z)^{\top}f(Z)u\right] &\text{(B.104)}\\ &=u^{\top}\Sigma_{f(Z)}u, &\text{(B.105)}\end{aligned}$$

while the right hand side is

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla\left(u^{\top}f\right)(Z)\right\|_2^2\right]&=\mathbb{E}\left[\left\|J_f(Z)^{\top}u\right\|_2^2\right] &\text{(B.106)}\\ &=\mathbb{E}\left[u^{\top}J_f(Z)J_f(Z)^{\top}u\right] &\text{(B.107)}\\ &=u^{\top}\mathbb{E}\left[J_f(Z)J_f(Z)^{\top}\right]u. &\text{(B.108)}\end{aligned}$$

Therefore, for all $u \in \mathbb{R}^d$, we have

$$u^\top \Sigma_{f(Z)} u \leq u^\top \mathbb{E}\left[J_f(Z) J_f(Z)^\top\right] u, \tag{B.109}$$

completing the proof. □

Lemma B.5.2. *For $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$ with bilipschitz parameters (m, M), we have*

$$\frac{1}{2}\ln\det\left(\mathbb{E}\left[J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top\right]\right) - \frac{1}{4m^4}\mathbb{E}\left[\left\|J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top - \mathbb{E}\left[J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top\right]\right\|_F^2\right] \leq \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(Z)\right)\right|\right] \tag{B.110}$$

Proof of Lemma B.5.2

The idea of the proof is that we can lower bound the expectation of $\ln\det(J_{\tilde{g}}(Z))$ in terms of $\mathbb{E}\left[J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top\right]$ and the variance of $J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top$. Intuitively, this is like upper bounding $\mathbb{E}[f(V)]$ for convex $f : \mathbb{R} \to \mathbb{R}$ and $V \in \mathbb{R}$ a random variable in terms of $\mathbb{E}[V]$ and the variance of V; this upper bound will depend on the curvature of f, i.e. its second derivative. Here, V takes the role of $J_{\tilde{g}}(Z) J_{\tilde{g}}(Z)^\top$, and we are flipping the signs to obtain the concave version of this simpler example.

For fixed symmetric $H \in \mathbb{R}^{d \times d}$, let

$$f(X) \equiv \frac{1}{2}\ln\det\left(X\right), \tag{B.111}$$

$$X_t \equiv X + tH, \tag{B.112}$$

$$h(t) \equiv f(X_t), \tag{B.113}$$

for $t \in [0, 1]$. We will use the second order Taylor expansion with integral remainder,

$$h(1) = h(0) + h'(0) + \int_0^1 (1-t) h''(t) dt, \tag{B.114}$$

and with a lower bound on $h''(t) \geq l$ on $[0, 1]$ and integrating $(1-t)$, we have

$$h(1) \geq h(0) + h'(0) + \frac{1}{2} l. \tag{B.115}$$

Let $A \equiv J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\intercal$ for notational convenience. Choosing $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$,

$$h(1) = f(X_1) \tag{B.116}$$
$$= \frac{1}{2}\ln\det\left(\mathbb{E}[A] + (A - \mathbb{E}[A])\right) \tag{B.117}$$
$$= \frac{1}{2}\ln\det\left(J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\intercal\right) \tag{B.118}$$
$$= \ln\det|J_{\tilde{g}}(Z)|\,. \tag{B.119}$$

Similarly,

$$h(0) = \frac{1}{2}\ln\det\left(\mathbb{E}[A]\right) \tag{B.120}$$
$$= \frac{1}{2}\ln\det\left(\mathbb{E}[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\intercal]\right), \tag{B.121}$$

so that, taking the expectation, we have

$$\mathbb{E}[\ln\det|J_{\tilde{g}}(Z)|] \geq \frac{1}{2}\ln\det\left(\mathbb{E}[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\intercal]\right) + \mathbb{E}[h'(0)] + \frac{1}{2}\mathbb{E}[l]\,. \tag{B.122}$$

The remainder of the proof is finding the expressions of $h'(0)$ and $h''(t) \geq l$, and subsequently taking their expectations.

Beginning with $h'(t)$, this is just the directional derivative of f in the direction H, $f'(X_t)[H]$ (see Section 2.2.1 of Bright et al. (2025)). From Appendix section A.4.1 of Boyd and Vandenberghe (2004), $f'(X)[H] = \frac{1}{2}\mathrm{tr}\left(X^{-1}H\right)$, so that

$$h'(t) = \frac{1}{2}\mathrm{tr}\left(X_t^{-1}H\right), \tag{B.123}$$

and, for $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$, $h'(0) = \frac{1}{2}\mathrm{tr}\left(\mathbb{E}[A]^{-1}(A - \mathbb{E}[A])\right)$, so that

$$\mathbb{E}[h'(0)] = 0. \tag{B.124}$$

Now we evaluate $h''(t)$. Call $h'(t) \equiv g(X_t)$, so that $g(X) = \frac{1}{2}\mathrm{tr}\left(X^{-1}H\right)$. We again identify $h''(t)$ as a directional derivative, in this case of g in the direction H:

$$h''(t) = g'(X_t)[H]. \tag{B.125}$$

A short computation shows that, using equation (124) of Petersen, Pedersen, et al. (2008),

$$g'(X)[H] = -\frac{1}{2}\mathrm{tr}\left(X^{-1}HX^{-1}H\right), \tag{B.126}$$

so for $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$, we have

$$h''(t) = -\frac{1}{2}\mathrm{tr}\left(X_t^{-1}(A - \mathbb{E}[A])X_t^{-1}(A - \mathbb{E}[A])\right). \tag{B.127}$$

To obtain a lower bound on $h''(t)$ for $t \in [0,1]$, we note that $X_t = \mathbb{E}[A] + t(A - \mathbb{E}[A]) = tA + (1-t)\mathbb{E}[A]$, a convex combination of A and $\mathbb{E}[A]$, always satisfies $X_t \succeq m^2 I_d$, since $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$ guarantees that the smallest singular value of $J_{\tilde{g}}$ is always greater than m, so that the smallest singular value of $A = J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top$ is always larger than m^2. Thus, $X_t^{-1/2} \preceq \frac{1}{m} I_d$, and we have,

$$-h''(t) = \frac{1}{2}\left\|X_t^{-1/2}\left(A - \mathbb{E}[A]\right)X_t^{-1/2}\right\|_F^2 \tag{B.128}$$

$$\leq \frac{1}{2}\left\|X_t^{-1/2}\right\|_2^4 \|A - \mathbb{E}[A]\|_F^2 \tag{B.129}$$

$$\leq \frac{1}{2m^4}\|A - \mathbb{E}[A]\|_F^2. \tag{B.130}$$

where we have used the property of the Frobenious norm that $\|AC\|_F \leq \|A\|_2 \|C\|_F$. From this, we see that we can take $l = -\frac{1}{2m^4}\|A - \mathbb{E}[A]\|_F^2$, completing the proof. □

Proof of Theorem 3.4.2

Now, we will relax the constraint that $\Sigma_{g(Y)} \succeq aI_d$ for $g \in \mathcal{C}_{\mathrm{NF}}$. We start with Lemma B.5.1, which says that

$$\Sigma_{g(Y)} \preceq \mathbb{E}[J_g(Y)J_g(Y)^\top]. \tag{B.131}$$

This along with $\Sigma_{g(Y)} \succeq aI_d$ implies

$$aI_d \preceq \mathbb{E}[J_g(Y)J_g(Y)^\top]. \tag{B.132}$$

From Lemma B.5.2, we have

$$\frac{1}{2}\ln\det\left(\mathbb{E}[J_g(Y)J_g(Y)^\top]\right) - \frac{1}{4m^4}\mathbb{E}\left[\|J_g(Y)J_g(Y)^\top - \mathbb{E}[J_g(Y)J_g(Y)^\top]\|_F^2\right] \leq \mathbb{E}\left[\ln\left|\det\left(J_g(Y)\right)\right|\right]. \tag{B.133}$$

Using $aI_d \preceq \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]$, we can lower bound the first term in equation (B.133):

$$\frac{d}{2}\ln(a) \leq \frac{1}{2}\ln\det\left(\mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right). \tag{B.134}$$

Thus far we have shown that

$$\Sigma_{g(Y)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_g(Y)\right)\right|\right] \geq \frac{d}{2}\ln(a) - \frac{1}{4m^4}\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right], \tag{B.135}$$

where we can think of the second term in equation (B.133) as the 'variance' of $J_g(Y)J_g(Y)^\top$. We now derive an upper bound for $\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right]$. We denote the matrix function $J_g(y)J_g(y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right] \equiv F(y) \in \mathbb{R}^{d\times d}$. We apply the one-dimensional Gaussian Poincare inequality (equation (B.101)) to each entry of $F(y)$:

$$\operatorname{Var}\left(F_{ij}(Y)\right) \leq \mathbb{E}\left[\left\|\nabla F_{ij}(Y)\right\|_2^2\right] \tag{B.136}$$

$$= \mathbb{E}\left[\sum_{k=1}^{d}\left(\frac{\partial F_{ij}(Y)}{\partial y_k}\right)^2\right]. \tag{B.137}$$

Therefore,

$$\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right] = \mathbb{E}\left[\left\|F(Y)\right\|_F^2\right] \tag{B.138}$$

$$= \sum_{i,j=1}^{d}\operatorname{Var}\left(F_{ij}(Y)\right) \tag{B.139}$$

$$\leq \sum_{i,j,k}^{d}\mathbb{E}\left[\left(\frac{\partial F_{ij}(Y)}{\partial y_k}\right)^2\right] \tag{B.140}$$

$$= \sum_{k=1}^{d}\mathbb{E}\left[\left\|\frac{\partial F(Y)}{\partial y_k}\right\|_F^2\right] \tag{B.141}$$

where $\frac{\partial F(Y)}{\partial y_k}$ denotes the matrix with entry (i,j) equal to the $\frac{\partial F_{ij}(Y)}{\partial y_k}$. Using equations (37) and (44) of Petersen, Pedersen, et al. (2008), we have

$$\frac{\partial F(y)}{\partial y_k} = \frac{\partial J_g(y)J_g(y)^\top}{\partial y_k} \tag{B.142}$$

$$= \frac{\partial J_g(y)}{\partial y_k}J_g(y)^\top + J_g(y)\frac{\partial J_g(y)}{\partial y_k}^\top, \tag{B.143}$$

so that by the triangle inequality, the fact that $\|A\|_F = \|A^\top\|_F$, and the fact that $\|AB\|_F \leq \|A\|_2 \|B\|_F$,

$$\left\| \frac{\partial F(y)}{\partial y_k} \right\|_F \leq 2 \left\| J_g(y) \frac{\partial J_g(y)}{\partial y_k}^\top \right\|_F \tag{B.144}$$

$$\leq 2 \|J_g(y)\|_2 \left\| \frac{\partial J_g(y)}{\partial y_k} \right\|_F . \tag{B.145}$$

Putting these inequalities together, we have

$$\mathbb{E}\left[\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\|_F^2 \right] \leq \mathbb{E}\left[4 \sum_{k=1}^d \|J_g(Y)\|_2^2 \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 \right] \tag{B.146}$$

$$= \mathbb{E}\left[4 \|J_g(Y)\|_2^2 \sum_{k=1}^d \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 \right]. \tag{B.147}$$

Since $g \in \mathcal{C}_{\mathrm{NF}}$, the bound $\|J_g(Y)\|_2^2 \leq M^2$ holds. We can rewrite

$$\sum_{k=1}^d \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 = \sum_{i=1}^d \|H_{g_i}(Y)\|_F^2 , \tag{B.148}$$

where $H_{g_i}(y)$ denotes the Hessian matrix of coordinate g_i evaluated at $y \in \mathbb{R}^d$. Putting these final inequalities together, we have now shown that

$$\mathbb{E}\left[\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\|_F^2 \right] \leq 4M^2 \sum_{i=1}^d \mathbb{E}\left[\|H_{g_i}(Y)\|_F^2 \right]. \tag{B.149}$$

Denoting our lower bound on the right hand side of equation (B.135) as $b(a)$, we finally have

$$b(a) \equiv \frac{d}{2} \ln(a) - \frac{M^2}{m^4} \sum_{i=1}^d \mathbb{E}\left[\|H_{g_i}(Y)\|_F^2 \right], \tag{B.150}$$

so that $C \equiv \frac{M^2}{m^4} \sum_{i=1}^d \mathbb{E}\left[\|H_{g_i}(Y)\|_F^2 \right]$ in the statement of the Theorem. This completes the proof. □

B.5.3 Proof of Theorem 3.4.1

Theorem 3.4.1 follows immediately from the following Lemma B.5.3, since then

$$b \leq \mathbb{E}\left[\ln \left| \det \left(J_{\tilde{g}}(W) \right) \right| \right] \tag{B.151}$$

$$\leq \frac{1}{2} \log \det \left(\Sigma_{\tilde{g}(W)} \right), \tag{B.152}$$

and therefore,

$$b \leq \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \implies e^{2b} \equiv C \leq \det\left(\Sigma_{\tilde{g}(W)}\right). \tag{B.153}$$

Lemma B.5.3. *For $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$ and isotropic Gaussian $W \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}\left[\log\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \leq \frac{1}{2}\log\det\left(\Sigma_{\tilde{g}(W)}\right). \tag{B.154}$$

Proof of Lemma B.5.3

We denote $\tilde{g}$ by g and W by Z for notational ease. We begin with by using[1], for $g \in \mathcal{C}_{\mathrm{VAE}}$,

$$h(g(Y)) = h(Y) + \mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right]. \tag{B.155}$$

Having assumed that Y is an isotropic multivariate Gaussian, we can use Lemma B.4.1 for the expression of its differential entropy $h(Y) \equiv -\mathbb{E}\left[\log\left(p(Y)\right)\right]$:

$$h(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Y\right) = \frac{d}{2}\log(2\pi e). \tag{B.156}$$

Lemma B.4.1 also says that differential entropy is maximized for Gaussian distributions: for any random vector Z with covariance Σ_Z whose entropy exists, we have

$$h(Z) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Z\right). \tag{B.157}$$

Combining these facts together, picking $Z = g(Y)$, we have

$$\begin{aligned} h(g(Y)) &= h(Y) + \mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] && \text{(B.158)}\\ &= \frac{d}{2}\log(2\pi e) + \mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] && \text{(B.159)}\\ &\leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_{g(Y)}\right). && \text{(B.160)} \end{aligned}$$

This implies that

$$\mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] \leq \frac{1}{2}\log\det\left(\Sigma_{g(Y)}\right), \tag{B.161}$$

completing the proof. □

[1] See https://statproofbook.github.io/P/dent-noninv

B.5.4 Proof of Theorem 3.4.3

We provide some machinery needed for the proof of Lemma B.5.5. It comes from Olkin and Marshall (2014).

Given two vectors of real numbers of length d, $x, y \in \mathbb{R}^d$, we say that x majorizes y if

1. $x_1 \geq x_2 \geq \ldots \geq x_d$ and $y_1 \geq y_2 \geq \ldots \geq y_d$;
2. $x_1 + x_2 + \ldots + x_k \geq y_1 + y_2 + \ldots + y_k$ for $k = 1, \ldots d$;
3. $x_1 + x_2 + \ldots + x_d = y_1 + y_2 + \ldots + y_d$.

When the first two conditions hold but in the absence of the third, we say that x weakly submajorizes y. The result we need is known as Tomic's inequality (Theorem 4.B.2 of Olkin and Marshall (2014)).

Lemma B.5.4 (Tomic's inequality)**.** *Suppose $x, y \in \mathbb{R}^d$ are both vectors with entries contained in the interval $(\alpha, \beta) \subseteq \mathbb{R}$. Then, if $f : (\alpha, \beta) \to \mathbb{R}$ is convex, increasing, and x weakly submajorizes y, then*

$$\sum_{i=1}^{d} f(y_i) \leq \sum_{i=1}^{d} f(x_i). \tag{B.162}$$

A more well-known version of this theorem is referred to as Karamata's inequality, where f is not assumed to be increasing, but where we require that x majorizes y. In this case, the conclusion still holds (Kadelburg et al., 2005).

In the following lemma, we generalize Lemma B.3.1.

Lemma B.5.5. *Let $f : \mathbb{R} \to \mathbb{R}$ be an even function with $f(0) = 0$, $f(x) \geq 0$ $\forall x$, and such that $f(\sqrt{x})$ is convex and increasing on $\mathbb{R}_{\geq 0}$. Then, given a random vector $Z \in \mathbb{R}^d$ with invertible Σ_Z, and a random vector $X \in \mathbb{R}^p$ with Σ_X invertible with $p \geq d$, with $\mathbb{E}[X] = 0$, $\mathbb{E}[Z] = 0$, we have*

$$\sup_{\substack{T \in \mathbb{R}^{p \times d}, H \in \mathbb{R}^{d \times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} f\left(\mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]\right) = \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{B.163}$$

where the $\theta_i \in \mathbb{R}^p$ are the columns of T, and the $\eta_i \in \mathbb{R}^d$ are the columns of H, and the γ_i are the singular values of $\Sigma_Z^{-1/2}\Sigma_{ZX}\Sigma_X^{-1/2}$. In addition, the supremum is attained by the (T, H) that solve the classical CCA problem between X and Z.

This is equivalent to saying that in classical CCA, for a large class of convex functions, maximizing the sum of the function applied to the correlations rather than the usual sum of the correlations does not change the solution.

Remark 18. *Taking $f(x) = x^2$, we recover Lemma B.3.1.*

Proof of Lemma B.5.5

We begin by using a similar argument to the proof of Lemma B.3.1. We have, using a change of variables $T = \Sigma_X^{-1/2}\tilde{T}$, $H = \Sigma_Z^{-1/2}\tilde{H}$,

$$\sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d f\left(\mathbb{E}\left[\eta_i^\top Z\theta_i^\top X\right]\right) = \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d \left(\theta_i^\top\Sigma_{XZ}\eta_i\right)^2 \tag{B.164}$$

$$= \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d f\left(\tilde{\theta}_i^\top\Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_i\right) \tag{B.165}$$

$$\leq \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\tilde{\theta}_i^\top\Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_j\right) \tag{B.166}$$

where in the last inequality we have used that $f(x) \geq 0$. Then,

$$\sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\tilde{\theta}_i^\top\Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_j\right) = \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\left(\tilde{T}^\top U\Lambda V^\top\tilde{H}\right)_{ij}\right), \tag{B.167}$$

where $U\Lambda V^\top$ is the SVD of $\Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}$, where $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal, and $\Lambda \in \mathbb{R}^{p\times d}$ is diagonal. Since U and V are orthogonal, we can perform another change of variables to $L = U^\top\tilde{T}$ and $R = V^\top\tilde{H}$:

$$\sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\left(\tilde{T}^\top U\Lambda V^\top\tilde{H}\right)_{ij}\right) = \sup_{\substack{L\in\mathbb{R}^{p\times d}, R\in\mathbb{R}^{d\times d},\\ L^\top L=R^\top R=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\left(L^\top\Lambda R\right)_{ij}\right) \tag{B.168}$$

$$= \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}} \sup_{\substack{R\in\mathbb{R}^{d\times d},\\ R^\top R=I_d}} \sum_{i=1}^d\sum_{j=1}^d f\left(\left(K^\top R\right)_{ij}\right), \tag{B.169}$$

where in the last step we let $K \equiv \Lambda^\top L \in \mathbb{R}^{d\times d}$ and split the supremum into two suprema. Denote by $t_i \equiv \left((k_i^\top r_j)^2\right)_{j=1}^d \in \mathbb{R}^d$. Then, $(K^\top R)_{ij} = k_i^\top r_j = \sqrt{(k_i^\top r_j)^2} = \sqrt{t_{ij}}$, where k_i denotes the ith column of K, and r_j denotes the jth column of R.

Since R is orthogonal, its columns form a basis of $\mathbb{R}^d$, and the t_i contain the squared values of the representation of k_i in the basis of the columns of R. In particular, $\|k_i\|_2^2 = \sum_{j=1}^d t_i$, and the vector $w_i = \left(\|k_i\|_2^2, 0, 0 \dots 0\right) \in \mathbb{R}^d$ majorizes t. We can then apply Tomic's inequality, Lemma B.5.4, to t and w, applied to $f(\sqrt{x})$ which was assumed to be convex and increasing. We obtain, for $i = 1, \dots d$,

$$\sum_{j=1}^d f\left((K^\top R)_{ij}\right) = \sum_{j=1}^d f\left(\sqrt{t_{ij}}\right) \tag{B.170}$$

$$\leq \sum_{j=1}^d f\left(\sqrt{w_{ij}}\right) \tag{B.171}$$

$$= f\left(\sqrt{\|k_i\|_2^2}\right) \tag{B.172}$$

where we have used that $f(0) = 0$ in the final equality. Therefore,

$$\sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L = I_d}} \sup_{\substack{R\in\mathbb{R}^{d\times d},\\ R^\top R = I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left((K^\top R)_{ij}\right) \leq \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L = I_d}} \sum_{i=1}^d f\left(\sqrt{\|k_i\|_2^2}\right) \tag{B.173}$$

$$= \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L = I_d}} \sum_{i=1}^d f\left(\sqrt{l_i^\top \Lambda\Lambda^\top l_i}\right) \tag{B.174}$$

since $k_i = \Lambda^\top l_i$ where l_i is the ith column of L. Letting $b(L) \equiv (l_i^\top \Lambda\Lambda^\top l_i)_{i=1}^d$, then $b(L)$ is the diagonal of the matrix $B(L) = L^\top \Lambda\Lambda^\top L \in \mathbb{R}^{d\times d}$, where our notation emphasizes that b and B are functions of $L \in \mathbb{R}^{p\times d}$. We let $\lambda_{1:d}(\Lambda\Lambda^\top) \equiv (\gamma_i^2)_{i=1}^d$ denote the first d eigenvalues of $\Lambda\Lambda^\top$.

We denote the descending eigenvalues of any symmetric matrix $A \in \mathbb{R}^{d\times d}$ by $\lambda(A) \in \mathbb{R}^d$, and by $\lambda_i(A)$ the ith entry of $\lambda(A)$. By Schur's Theorem, Theorem 4.3.45 of Horn and Johnson (2012), for every L, $\lambda(B(L))$ majorizes $b(L)$. Corollary 4.3.39 of Horn and Johnson (2012) applied to $\Lambda\Lambda^\top$ immediately implies that $\lambda_{1:d}(\Lambda\Lambda^\top)$ weakly submajorizes $\lambda(B(L))$ for

every orthogonal $L \in \mathbb{R}^{p\times d}$. Two final applications of Tomic's inequality to $f(\sqrt{x})$ give that

$$\sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}} \sum_{i=1}^{d} f\left(\sqrt{l_i^\top \Lambda\Lambda^\top l_i}\right) \leq \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}} \sum_{i=1}^{d} f\left(\sqrt{\lambda_i(B(L))}\right) \tag{B.175}$$

$$\leq \sum_{i=1}^{d} f\left(\sqrt{\lambda_i\left(\Lambda\Lambda^\top\right)}\right) \tag{B.176}$$

$$= \sum_{i=1}^{d} f\left(\sqrt{\gamma_i^2}\right) \tag{B.177}$$

$$= \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{B.178}$$

and we note that we have equality when $L \in \mathbb{R}^{p\times d}$ has 1s on its diagonal and 0s elsewhere.

We have now established that

$$\sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^{d} f\left(\mathbb{E}\left[\eta_i^\top Z\theta_i^\top X\right]\right) \leq \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{B.179}$$

and tracing back the inequalities and changes of variables, we have equality when we choose L to have 1s on its diagonal with 0s elsewhere and R to be I_d, corresponding to choices of $\tilde{T} = U$ and $\tilde{H} = V$. Equality in equation (B.166) follows from the fact that these choices of $\tilde{T}$ and $\tilde{H}$ make $\tilde{T}^\top U\Lambda V^\top \tilde{H}$ diagonal. This corresponds to $T = \Sigma_X^{-1/2}\tilde{T}$ and $H = \Sigma_Z^{-1/2}\tilde{H}$, which are exactly the solutions to the classical CCA problem between X and Z, and the proof is complete. □

Proof of Theorem 3.4.3

We denote $\tilde{g}$ by g and W by Y for ease of notation. We begin with

$$\underset{\substack{g\in\mathcal{C}_{\mathrm{NF}}, B\in\mathbb{R}^{p\times d}\\ \mathbb{E}[\ln|\det(J_g(Y))|]\geq b}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right], \tag{B.180}$$

the equivalent formulation of the normalizing flow problem. From Theorem 3.4.1, we know that this problem is a constrained problem relative to

$$\underset{\substack{g\in\mathcal{C}_{\mathrm{NF}}, B\in\mathbb{R}^{p\times d}\\ \det\left(\Sigma_{g(Y)}\right)\geq c}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right], \tag{B.181}$$

where $c = e^{2b}$ Now, we show that this relaxed problem is equivalent to a partially linear CCA problem.

Following the proof of Theorem 3.3.3 up until equation (B.51), we have

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \det\left(\Sigma_{g(Y)}\right)\geq c,\ g\in\mathcal{C}_{\mathrm{NF}}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \det\left(\Sigma_{g(Y)}\right)\geq c,\ g\in\mathcal{C}_{\mathrm{NF}}}} \operatorname{tr}\left(\Sigma_{g(Y)}D\right), \tag{B.182}$$

where we have denoted $D \equiv I_d - \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}$ for notational ease. We note that $\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2} = AA^\top$ where $A = \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}$. From classical CCA, the singular values of A are the canonical correlations between $g(Y)$ and X, so they are all between 0 and 1. Therefore, D is positive semi-definite, with eigenvalues equal to $1-\gamma_i^2$, where γ_i is the ith canonical correlation between $g(Y)$ and X.

Applying the GM-AM inequality Lemma B.4.2, we have

$$\begin{aligned}
\operatorname{tr}\left(\Sigma_{g(Y)}D\right) &= \operatorname{tr}\left(\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}\right) && \text{(B.183)}\\
&\geq d\det\left(\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}\right)^{1/d} && \text{(B.184)}\\
&= d\det\left(\Sigma_{g(Y)}\right)^{1/d}\det\left(D\right)^{1/d}, && \text{(B.185)}
\end{aligned}$$

with equality when $\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}$ is a multiple of I_d. Therefore,

$$\begin{aligned}
\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] &\geq \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} d\det\left(\Sigma_{g(Y)}\right)^{1/d}\det\left(D\right)^{1/d} && \text{(B.186)}\\
&\geq \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} dc^{1/d}\det\left(D\right)^{1/d}. && \text{(B.187)}
\end{aligned}$$

Picking $\Sigma_{g(Y)} = \left(c\det\left(D\right)\right)^{1/d}D^{-1}$, $\Sigma_{g(Y)}$ satisfies $\det\left(\Sigma_{g(Y)}\right) = c$ and $\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}$ is a multiple of I_d. Therefore the inequalities become equalities and we have

$$\begin{aligned}
\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] &= \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} dc^{1/d}\det\left(D\right)^{1/d} && \text{(B.188)}\\
&= dc^{1/d}\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \left(\prod_{i=1}^d\left(1-\gamma_i^2\right)\right)^{1/d}. && \text{(B.189)}
\end{aligned}$$

Therefore, minimizing the relaxed NF problem (B.181) is equivalent to maximizing

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d \\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sum_{i=1}^{d} \log\left(\frac{1}{1-\gamma_i^2}\right). \tag{B.190}$$

Applying Lemma B.5.5, with $Z = g(Y)$, observing that $f(x) = \log\left(\frac{1}{1-x^2}\right)$ satisfies the conditions of the Lemma, namely evenness, that $f(x) \geq 0$, and that $f\left(\sqrt{x}\right) = \log\left(\frac{1}{1-x}\right)$ is convex and increasing on $\mathbb{R}_{\geq 0}$, we have

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d \\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sum_{i=1}^{d} \log\left(\frac{1}{1-\gamma_i^2}\right) = \sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d \\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d}, \\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^{d} \log\left(\frac{1}{1-\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2}\right), \tag{B.191}$$

where the $\theta_i \in \mathbb{R}^p$ are the columns of T, and the $\eta_i \in \mathbb{R}^d$ are the columns of H. This is equivalent to

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d \\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d}, \\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \left(\prod_{i=1}^{d} \frac{1}{1-\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2}\right)^{1/d}, \tag{B.192}$$

and writing this problem over $h(y) \equiv H^\top g(y)$ completes the proof. □