A Correlation Analysis Approach to Finding Interpretable Latent Representations via Conditional Generative Models

James Buenfil
Department of Statistics
University of Washington

Eardi Lila
Department of Biostatistics
University of Washington

Abstract

Supervised disentanglement, that is, learning interpretable nonlinear latent representations of a target data view informed by an auxiliary data view, is a central challenge in interpretable machine learning. We formulate this problem as a partially linear invertible canonical correlation analysis (PLiCCA). Specifically, given two data views, (i) complex data lying near a potentially high-dimensional manifold, and (ii) auxiliary high-dimensional multivariate data, PLiCCA learns latent variables for the complex view that are maximally correlated with sparse linear combinations of the auxiliary variables. In contrast to regression-based approaches to supervised disentanglement, the proposed method yields a latent embedding whose coordinates are explicitly ordered by their interpretability with respect to the auxiliary variables. We formalize the population PLiCCA problem and establish existence results. We then show a close theoretical connection between PLiCCA and conditional latent variable models, in particular conditional variational autoencoders and conditional normalizing flows, which enables practical estimation. We demonstrate our approach on brain imaging data, where PLiCCA is used to learn embeddings informed by auxiliary demographic, psychometric, and behavioral variables.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Disentangled representation learning (Wang et al., 2024; Moran and Aragam, 2026) aims to learn low-dimensional representations in which distinct factors of variation are encoded in separate latent coordinates. It is widely regarded as a fundamental problem in interpretable machine learning (Rudin et al., 2022). Although unsupervised disentanglement, where no auxiliary data are available, has been studied extensively (see, e.g., Meo et al., 2024; Balabin et al., 2024), it has been shown to be unidentifiable in full generality (Locatello et al., 2019). This has motivated the development of supervised (Liu et al., 2022; Wang et al., 2024) and weakly supervised approaches (Shu et al., 2019; Locatello et al., 2020; Shen et al., 2022), which leverage an auxiliary data view to guide the learning of a low-dimensional representation of a target data view. Such representations are valuable in their own right, but they are most often used in downstream tasks, where supervision can provide the inductive bias needed for generalization.

Broadly, supervised disentanglement methods follow one of two strategies: (i) a two-stage approach, which first performs unsupervised disentanglement of the target view and subsequently regresses the learned latent variables on an auxiliary interpretable view (Adel et al., 2018; Van et al., 2020) or (ii) a joint approach, in which both steps are carried out simultaneously (Nalisnick et al., 2019; Zhao et al., 2019; Ding et al., 2020; Pati and Lerch, 2021; Lu et al., 2024; Inecik et al., 2025; An and Jeon, 2024). We pursue both approaches but argue that canonical correlation analysis (CCA), rather than regression, provides a more principled way to align a low-dimensional representation of the target view with the auxiliary view.

To formalize this idea, we consider two random vectors: the target data view $Y \in \mathcal{M} \subseteq \mathbb{R}^q$, where $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^q$, for which we seek a nonlinear invertible latent representation, and the high-dimensional auxiliary data view $X \in \mathbb{R}^p$. We

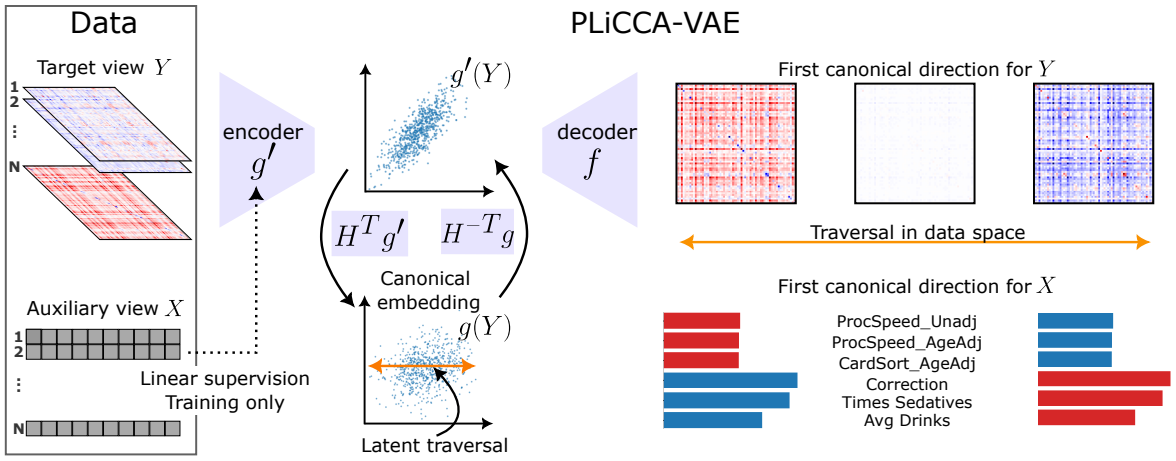

Figure 1: Diagram of PLiCCA-VAE, with the notation of Theorem 3.3, and $\mathcal{C} = \mathcal{C}_{\text{VAE}}$, applied to 700 subjects from the Human Connectome Project (Van Essen et al., 2013). Functional connectivity is treated as the target view, while demographic, psychometric, and behavioral variables form the auxiliary view. PLiCCA learns nonlinear latent variables of the target view that are linearly associated with the auxiliary variables. To illustrate the interpretability of the latent representation of Y, we display the trajectory induced by the first canonical direction in the latent space of Y (latent traversal), mapped back to the data space through the decoder, alongside the linear trajectory induced by the corresponding canonical direction of X. Only the six auxiliary variables with the largest weights are shown. Positive weights are shown in red and negative weights are shown in blue. The results appear to recover a positive–negative mode consistent with that identified in Smith et al. (2015).

aim to learn a supervised disentangled representation, that is, a low-dimensional, interpretable latent variable $U \in \mathbb{R}^d$, with $d \ll q$, and a decoder-encoder pair $f : \mathbb{R}^d \to \mathbb{R}^q$ and $g : \mathbb{R}^q \to \mathbb{R}^d$, such that

$$Y = f(U) + \varepsilon, \tag{1}$$

where ε denotes residual error, $U = g(Y)$, and the auxiliary data view X informs the construction and interpretation of U.

Our approach learns a latent representation U of the target view Y that is *linearly correlated* with the auxiliary view X. Imposing a (partially) linear structure on the latent space offers several 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 (Jahanian et al., 2020; Härkönen et al., 2020), and supporting few-shot regression (Nitzan et al., 2022).

Contributions.
(i) We propose novel joint and two-stage approaches to supervised disentanglement via partially linear invertible canonical correlation analysis (PLiCCA). PLiCCA learns invertible, nonlinear latent representations of the target view that are maximally associated with sparse linear combinations of the auxiliary view, yielding embeddings whose coordinates are ordered by their association with interpretable auxiliary variables.

(ii) We prove the existence of population solutions to PLiCCA and provide a rigorous theoretical characterization of the problem as a nonlinear regression. Building on this characterization, we establish a connection to conditional latent variable models, in particular, conditional variational autoencoders and conditional normalizing flows. We show formally that both models can be viewed as relaxations of PLiCCA, in which hard-to-enforce global constraints are replaced by local ones. This connection enables efficient estimation of PLiCCA via proxy conditional generative models.

2 BACKGROUND AND RELATED WORK

CCA (Hotelling, 1936; Guo and Wu, 2019; Yang et al., 2019; Chapman and Wang, 2021) is a classical statistical method that, given two data views, learns linear transformations of each view to produce latent variables whose resulting representations are maximally correlated. Hence, it provides bases (latent representations) for both data views, but unlike principal components analysis, each basis is informed by the other data view.

A key advantage of linear CCA over more complex methods is its simplicity and 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., Li et al., 2024; Bykhovskaya and Gorin, 2023; Buenfil and Lila, 2024, and references therein). The supervised disentanglement approach considered in this paper, 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 (Lancaster, 1958; Breiman and Friedman, 1985; Hannan, 1961; Michaeli et al., 2016) replaces linear transformations with nonlinear classes of functions, such as reproducing kernel Hilbert spaces (Lai and Fyfe, 2000; Akaho, 2006), and, more recently, neural networks (Andrew et al., 2013; Friedlander and Wolf, 2023). However, these approaches do not guarantee the invertibility of the nonlinear mappings, hindering interpretability. To mitigate this issue, Wang et al. (2015a) introduced DCCAE, which augments the CCA objective with a reconstruction term based on autoencoders. However, DCCAE is deterministic and therefore does not leverage the advantages of variational autoencoders over standard autoencoders.

We defer to Section C for an extended literature review, including works relating CCA and latent variable models to the multi-view data problem (Guo et al., 2019), as well as on applications of independent component analysis to supervised disentanglement (Hyvärinen et al., 2023).

There is also a large body of work on *unsupervised* disentanglement (see, e.g., Meo et al., 2024; Balabin et al., 2024; Mathieu et al., 2016; Kim and Mnih, 2018; Schmidhuber, 1992). Although our contributions focus on supervised disentanglement, our formulation remains compatible with many existing notions of unsupervised disentanglement. In particular, a common approach to disentanglement in latent variable models, including variational autoencoders (VAEs), is to augment the objective function with a disentanglement-inducing term (see Table 1 of Wang et al. (2024) for a succinct summary), and our use of conditional VAEs is compatible with this strategy.

3 PLiCCA

3.1 Population problem

In this section, we define the population problem of interest and show that it is well posed. Specifically, we study the *partially linear CCA* problem (Michaeli et al., 2016) under the additional constraint that the nonlinear embedding is approximately invertible. We refer to the partially linear *invertible* CCA problem as PLiCCA. See Section D for background on the partially linear CCA without the invertibility constraint.

Recall that we consider two random vectors: the target view $Y \in \mathcal{M} \subseteq \mathbb{R}^q$, where $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^q$, for which we seek a nonlinear invertible latent representation, and $X \in \mathbb{R}^p$, the high-dimensional auxiliary data. Without loss of generality, we assume 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 $\Sigma_Z \equiv \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top]$, and we say that Z is isotropic if $\Sigma_Z = I_d$, where I_d denotes the $d \times d$ identity matrix.

We define PLiCCA as the following constrained optimization problem, where the constraint set $\mathcal{C}$ encodes the invertibility of the latent representation:

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

where $T = [\theta_1, \ldots, \theta_d]$ and $g(Y) = (g_1(Y), \ldots, g_d(Y))^\top$.

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, as enforced by $\Sigma_U = \Sigma_V = I_d$. The variables (U_i, V_i) are referred to as *canonical variable pairs*, and can be ordered so that the maximizing correlations $\gamma_i \equiv \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]$ are decreasing in i. The variables γ_i are referred to as the *canonical correlations*. The columns of T are referred to as the *canonical vectors* for X, and $g(Y)$ is referred to as the *canonical embedding* for Y.

We note that, unlike the original partially linear CCA formulation, we maximize squared correlations. It can be shown that when g is linear, maximizing the squared correlations is equivalent to maximizing the sum of the correlations (see Lemma E.1). This minor modification of the objective allows us to draw closer connections to regression and latent variable models (see Theorem 3.3).

3.2 Supervised disentanglement via PLiCCA

Before defining our notion of invertibility through $\mathcal{C}$, we explain how solving PLiCCA provides an approach to supervised disentanglement. Suppose g is a solution to the PLiCCA problem and is 'invertible', with inverse f. We have

$$Y \approx f(g(Y)) \tag{3}$$
$$= f(U), \tag{4}$$

where $U = g(Y)$. Then, PLiCCA provides a solution to the supervised disentanglement problem for (Y, X) and enjoys the following desirable properties:

(i) Ordered disentanglement of Y. The invertibility of g guarantees that U disentangles Y, since $Y = f(U)$ and $\Sigma_U = I_d$, so the coordinates of U are uncorrelated with one another. Moreover, the coordinates of U can be naturally ordered by interpretability, where interpretability is determined by correlation with the auxiliary variables, in the sense that they can be arranged so that the correlations γ_i between U_i and V_i are decreasing. In contrast to other CCA-based approaches, such as deep CCA, the invertibility of the embedding g enables us to understand how changes to the latent space translate into changes in the target view Y.

For instance, to interpret the ith latent variable U_i, we consider the one-dimensional latent subspace corresponding to the ith coordinate. Mapping this subspace through the inverse map f induces a nonlinear trajectory in the space of the target view (e.g., neuroimaging data), revealing how variation along this coordinate is expressed in the observed data. Similar constructions arise in other VAE-based approaches through so-called latent traversals (see, e.g., Song et al., 2023, and references therein).

(ii) Interpretability through sparse linear signatures. The trajectory corresponding to U_i in the auxiliary view is the one-dimensional linear subspace spanned by θ_i. If the canonical vectors θ_i are estimated to be sparse, then the associated variables $V_i = \theta_i^\top X$ are readily interpretable as linear combinations of only a few selected variables in X, remaining mutually uncorrelated via $\Sigma_V = I_d$. In this way, each learned latent coordinate U_i is paired with a sparse linear "signature" in the auxiliary view. The partially linear formulation then allows U_i to be interpreted through its association with the corresponding sparse linear combination V_i. The signs and sparsity pattern of the coefficients in θ_i further facilitate the interpretation of the latent representations U_i. An illustration in the context of our data analysis is provided in Figure 1.

3.3 Definition of the invertibility constraint set $\mathcal{C}$

To define PLiCCA, we need to specify the set $\mathcal{C}$, which provides an appropriate notion of invertibility of the canonical embedding g. Since g must also project the data into a lower-dimensional space, the natural idealization would be to define $\mathcal{C}$ as the set of diffeomorphisms from $\mathrm{supp}(Y)$ to $\mathbb{R}^d$. However, this would require $\mathrm{supp}(Y)$ to be contained in a manifold $\mathcal{M}$ with dimension $\dim(\mathcal{M}) = d$, an assumption that is often unrealistic, particularly in the presence of noise.

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 establish the existence of a solution to our PLiCCA problem for a class of functions $\mathcal{C}$ satisfying this property. We denote this constrained set of functions as $\mathcal{C}_{\mathrm{VAE}}$, in anticipation of the VAE-based methodology developed later.

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

Theorem 3.1 *Suppose Σ_X is invertible and that $\mathrm{supp}(Y) \subseteq K$, where K is a compact subset of $\mathbb{R}^q$. Fix constants $M, m, \varepsilon, d > 0$. If $\mathcal{C}$ is chosen as*

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

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

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}_{\mathrm{VAE}}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2. \tag{5}$$

Remark 1 *The assumption that $\mathcal{C}_{\mathrm{VAE}}$ is non-empty only serves to ensure that the pair (ε, d) is compatible, i.e., that Y can be compressed to dimension d with reconstruction error at most ε. For fixed ε, one can always find a d such that $\mathcal{C}_{\mathrm{VAE}}$ is non-empty. Indeed, taking $d = q$ achieves a reconstruction error of 0 via the identity map.*

We emphasize that, to our knowledge, this is the first work to establish such an existence result, as many related approaches consider only finite-sample formulations of the problem.

3.4 Two-stage approach

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 $\phi : \mathcal{M} \to \mathbb{R}^d$, i.e., $\mathcal{M}$ is diffeomorphic to $\mathbb{R}^d$. Letting $W \equiv \phi(Y) \in \mathbb{R}^d$ denote the dimension-reduced representation of Y, assuming that Σ_W is invertible, we can reformulate partially linear CCA in terms of W

rather than Y, using the function $\tilde{g}: \mathbb{R}^d \to \mathbb{R}^d$ in place of $g: \mathbb{R}^q \to \mathbb{R}^d$:

$$\underset{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^\top X}=I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[\tilde{g}_i(W)\theta_i^\top X\right]^2, \tag{6}$$

The equivalence of such formulations follows from the invertibility of ϕ. A solution g to problem (2) is then obtained by setting $g = \tilde{g} \circ \phi$.

Invertibility of g then follows from invertibility of $\tilde{g}$. Note that a stricter notion of invertibility can now be adopted, since $\tilde{g}$ maps between spaces of the same dimension. Here, we denote this constraint set $\mathcal{C}$ as $\mathcal{C}_{\text{NF}}$, in anticipation of our proposed two-stage approach based on normalizing flows.

We recall that a function $f: K \to \mathbb{R}^d$ is called bi-Lipschitz with parameters (m, M) if it satisfies $m\|x-y\|_2 \le \|f(x)-f(y)\|_2 \le M\|x-y\|_2$ for all $x, y \in K$. With $\mathcal{C} = \mathcal{C}_{\text{NF}}$, we establish an existence result for the PLiCCA problem analogous to Theorem 3.1.

Theorem 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 $M, m > 0$. If $\mathcal{C}$ is chosen as*

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

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

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

where, as before, $\theta_i \in \mathbb{R}^p$ denotes the ith column of $T \in \mathbb{R}^{p\times d}$.

Remark 2 *In Theorems 3.2 and 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 by imposing a vanishing-at-infinity-type assumption; see Theorem 5 of Krukowski (2018).*

When Y has a known manifold structure, such as positive definite matrices (Kim et al., 2014; Buenfil and Lila, 2024) or probability densities (Cho et al., 2022), a (global) logarithmic map may be used in place of ϕ. Hence, the model in equation (7) provides a natural approach to incorporate the underlying geometry. When the manifold structure is not known, a two-stage approach to PLiCCA may be used instead. First, one learns an unsupervised latent representation $W \in \mathbb{R}^d$ of Y through an approximation of the form $Y = \psi(W) + \varepsilon$, or uses a pretrained model for this step. Next, one solves the nonlinear problem in equation (7).

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. We adopt 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 define $a\mathcal{C} \equiv \{af : f \in \mathcal{C}\}$. In Theorem 3.3, $\mathcal{C}$ can refer to either $\mathcal{C}_{\text{VAE}}$ or $\mathcal{C}_{\text{NF}}$.

Theorem 3.3 *Finding a pair (g, T) that solves the PLiCCA problem*

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d \\ g\in a^{-1/2}\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{8}$$

where $\theta_i \in \mathbb{R}^p$ denotes the ith column of T, is equivalent to finding a pair (g', B) that solves the regression problem

$$\underset{\substack{g':\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d} \\ g'\in\mathcal{C},\Sigma_{g'(Y)}\succeq aI_d}}{\text{minimize}} \mathbb{E}\left[\left\|g'(Y) - B^\top X\right\|_2^2\right], \tag{9}$$

for any $a > 0$.

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

$$\Sigma_{g'(Y)}^{-1/2} B^\top \Sigma_X B \Sigma_{g'(Y)}^{-1/2}, \tag{10}$$

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

$$T = BH\Lambda^{-1} \tag{11}$$

$$g(y) = H^\top g'(y). \tag{12}$$

Theorem 3.3 can also be understood as a generalization of results in Wang and Zhou (2021); Buenfil and Lila (2024), which apply only to the linear case. 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 and Wolf, 2023), since the whitening matrix H can correct for this after the regression. Intuitively, the equality constraint $\Sigma_{g(Y)} = I_d$ has been

replaced by the lower bound $\Sigma_{g'(Y)} \succeq aI_d$, which prevents the collapse of g' to 0 when solving the regression problem. We further explore relaxing this constraint in the following section. The scaling constant a is introduced for convenience in later results, particularly those in Section 4.4.

4 METHODOLOGY

Next, we introduce the two conditional latent variable models underlying conditional VAEs (Sohn et al., 2015; Harvey et al., 2022; Khemakhem et al., 2020) and conditional normalizing flows (NFs) (Papamakarios et al., 2017; Winkler et al., 2019). We use these models to develop two practical implementations of PLiCCA: PLiCCA-VAE and PLiCCA-NF.

4.1 PLiCCA-VAE

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

$$Y|Z, X \sim \mathcal{N}\left(f(Z), I_q\right) \tag{13}$$

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{14}$$

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

$$\underset{g:\mathbb{R}^q \to \mathbb{R}^d, B \in \mathbb{R}^{p\times d}}{\text{minimize}} \; \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right] \tag{15}$$

$$+\lambda_{\text{VAE}} \mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right]. \tag{16}$$

Here, the outer expectation is taken with respect to the observed variables (X, Y), and we have added a tuning hyperparameter λ_{VAE} to the objective function, as is common in practice, to allow more flexible balancing of the two terms. This formulation may be viewed as an implementation of a conditional VAE.

From the PLiCCA characterization in Theorem 3.3, it is clear that the conditional VAE objective introduced resembles the PLiCCA regression objective in (9) when $\mathcal{C} = \mathcal{C}_{\text{VAE}}$ and such constraint is enforced through a Lagrangian relaxation. In Section 4.2, we make this connection precise. This observation motivates our proposed methodology: rather than solving (9) directly, we solve the simpler proxy problem in equations (15)–(16), and then use the resulting (g, B) in place of the analogous quantities in equation (9) to compute the canonical directions via equations (10)–(12). We refer to this approach as PLiCCA-VAE.

Note that the conditional VAE introduced here is the vanilla version, which allows for a more thorough theoretical analysis. In practice, however, more flexible distributions for $q(z|y)$ can be used, for example by allowing $\Sigma_{Y|Z,X}(z)$ and $\Sigma_{Z|Y}(y)$ to vary with their inputs. We also note that the encoder g is taken to depend only on Y, and not on the conditioning variable X. This is intentional, since once the supervised representation model has been learned, the representation can be computed for new observations of the target view without requiring the associated auxiliary-view data.

In practice, we estimate the expectations in our proxy problem in equations (15)–(16) using the sample $(X_k, Y_k)_{k=1}^N$. To aid interpretability, we augment the objective function with a sparsity-inducing group lasso penalty $\lambda_s \sum_{j=1}^p \|b_j\|_2$ (Yuan and Lin, 2006), where b_j denotes the jth row of B, corresponding to covariate X_j. The regularization strength is controlled by the parameter λ_s. We optimize the resulting objective using proximal gradient descent, which is convenient because the proximal operator of $\sum_{j=1}^p \|b_j\|_2$ has a closed-form solution (see, for instance, Section 6.5.4 of Parikh et al. (2014) for a derivation and Murray et al. (2019) for an implementation). After obtaining estimates $\hat{g}$, $\hat{B}$, we estimate $\Sigma_{g(Y)}$ and Σ_X via their sample covariances.

Then, as suggested by Theorem 3.3, we compute the canonical directions for X and Y by solving equations (10)-(12), using $\hat{g}$, $\hat{B}$ and empirical covariance matrices in place of their population counterparts. The full PLiCCA-VAE algorithm is given in Algorithm 1.

Note that if the latent dimension of the VAE is misspecified, some latent variables $g_i(Y)$ may collapse to 0 (Zheng et al., 2022; Bonheme and Grzes, 2023). Hence, we recommend inspecting $\hat{\Sigma}_{g(Y)}$ for latent coordinates i such that $\hat{\text{Var}}\left(g_i(Y)\right)$ is approximately 0, and remove the corresponding variables from $\hat{B}$, $\hat{g}$, 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 carries over directly to T since $T = B\left(H\Lambda^{-1}\right)$, retaining the variable selection induced by the group lasso penalty. If B is zero, i.e., there is no association between Y and X, the method reduces to a standard unsupervised VAE. Finally, the resulting $\hat{T}$ and $\hat{g}$ automatically satisfy the correct disentanglement orthogonality conditions.

4.2 Connection between PLiCCA and conditional VAEs

Next, we elucidate the connection between PLiCCA's objective function and the proxy conditional VAE ob-

Algorithm 1 PLiCCA-VAE

Input: Pairs $(Y_k, X_k)_{k=1,\dots,N}$, comprising *target* and *auxiliary* data views; latent space dimension d.

1. Solve the sample version of the conditional VAE (15)–(16), augmented with the penalty $\lambda_s \sum_{j=1}^{p} \|b_j\|_2$.
   Obtain the estimated encoder $\hat{g}$, decoder $\hat{f}$, and sparse regression matrix $\hat{B}$.
2. Estimate $\Sigma_{g(Y)}$ and Σ_X via their sample covariance matrices $\hat{\Sigma}_{g(Y)}$ and $\hat{\Sigma}_X$.
3. Inspect $\hat{\Sigma}_{g(Y)}$ for latent coordinates i such that $\hat{\mathrm{Var}}(g_i(Y)) \approx 0$, and discard them from $\hat{B}$, $\hat{g}$, and $\hat{\Sigma}_{g(Y)}$.
4. Compute the eigendecomposition $\tilde{H}\hat{\Lambda}^2\tilde{H}^\top$ of $\hat{\Sigma}_{g(Y)}^{-1/2}\hat{B}^\top\hat{\Sigma}_X\hat{B}\hat{\Sigma}_{g(Y)}^{-1/2}$.
5. Compute $\hat{H} = \hat{\Sigma}_{g(Y)}^{-1/2}\tilde{H}$.
6. Compute the estimated canonical vectors for X as $\hat{T} = \hat{B}\hat{H}\hat{\Lambda}^{-1}$.
7. With a slight abuse of notation, replace $\hat{g}(y)$ by its normalized version: $\hat{g}(y) \equiv \hat{H}^\top\hat{g}(y)$.

jective. We begin with the natural regression problem (9) that Theorem 3.3 suggests solving under the constraint set $\mathcal{C} = \mathcal{C}_{\mathrm{VAE}}$:

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ f:\mathbb{R}^d\to\mathbb{R}^q \\ B\in\mathbb{R}^{p\times d},\ \Sigma_{g(Y)}\succeq aI_d}}{\text{minimize}} \mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right] \tag{17}$$

$$+\,\mathbb{E}\left[\|Y - f(g(Y))\|_2^2\right], \tag{18}$$

We aim to show how the conditional VAE objective, (15)–(16), can be viewed as a relaxation of this PLiCCA regression problem.

The regression terms in equations (15) and (17), respectively, are identical. Hence, we focus on the second term of the conditional VAE objective, $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right]$. Using a form of the approximating posterior distribution $q(z|Y) = \mathcal{N}(g(Y), I_d)$, this term can be rewritten as

$$\mathbb{E}\left[\mathbb{E}_{\varepsilon}\left[\|Y - f(g(Y) + \varepsilon)\|_2^2\right]\right], \tag{19}$$

where $\varepsilon \sim \mathcal{N}(0, I_d)$. Thus, the VAE reconstruction term is a noisy version of the PLiCCA reconstruction term (18). More interestingly, the injected noise also encourages the constraint $\Sigma_{g(Y)} \succeq aI_d$.

We first provide some intuition and then state our formal result. From equation (19), if the decoder f relies on 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 term ε_i, for the signal to be distinguishable from noise. 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 (Zheng et al., 2022; Bonheme and Grzes, 2023). Of course, if $g_i(y) = 0$ for any i then $\Sigma_{g(Y)}$ will have a 0 eigenvalue. Therefore, the conditional VAE does not, and should not, *strictly* enforce the constraint $\Sigma_{g(Y)} \succeq aI_d$, since it should remain free to represent Y using only as many latent dimensions as needed. For this reason, in practice, we inspect and drop inactive latent dimensions when applying Theorem 3.3.

Although the conditional VAE does not enforce $\Sigma_{g(Y)} \succeq aI_d$, it does enforce a weaker 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 achieved by g and f and on the magnitude of the noise in $q(z|Y)$. Formally, we have the following result.

Theorem 4.1 *Fix positive constants δ and σ^2_{enc}. Suppose g and f are such that the reconstruction error $\mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right] \le \delta$, and suppose that we model $q(z|y) = \mathcal{N}\left(g(y), \sigma^2_{\mathrm{enc}} I_d\right)$. Then,*

$$\mathrm{tr}\left(\Sigma_{g(Y)}\right) \ge \sigma^2_{\mathrm{enc}} C(\delta), \tag{20}$$

where $C(\delta)$ is decreasing in δ and is defined as $C(\delta) \equiv d\left(e^{\frac{2}{d}R_Y(\delta)} - 1\right)$. We refer to equation (107), the rate-distortion function of Y, for the form of $R_Y(\delta)$, which depends on the distribution of Y.

Remark 3 *For $\delta_1 = \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right]$, we have $R_Y(\delta_1) = 0$. Thus, the lower bound $C(\delta)$ in (20) reduces to 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 $f(z) \equiv \mathbb{E}[Y]$. For $\delta_0 = 0$, i.e., perfect reconstruction, we have $R_Y(\delta_0) = H(Y)$, the entropy of Y.*

When using the conditional VAE as a proxy for PLiCCA, with the constraint $\Sigma_{g(Y)} \succeq aI_d$ relaxed, a concern is that the regression term $\mathbb{E}\left[\|g(Y) - B^\top X\|_2^2\right]$ could drive g toward the zero solution. Theorem 4.1 says that if λ_{VAE} is chosen sufficiently large, corresponding to a sufficiently small δ, then this collapse can be avoided. Although this paper focuses on conditional VAEs, the result also holds for unconditional VAEs; this formalizes the intuition of adding noise to any autoencoder's latent space to prevent its latent representation from becoming arbitrarily close to 0, thereby stabilizing training.

4.2.1 Moving beyond isotropic noise

In Theorem 4.1, we model the posterior uncertainty as $\sigma_{\text{enc}}^2 I_d$ instead of the more general form $\Sigma_{Z|Y}(Y)$, which depends on the data Y. We adopt this simplification because it yields an intuitive interpretation: the lower bound depends only on the magnitude of the encoder error, captured by the single parameter σ_{enc}^2, and the reconstruction error δ. However, this result generalizes to the case of a data-dependent $\Sigma_{Z|Y}(Y)$. Due to space constraints, we defer the details of this generalization to Section F.3.

4.3 PLiCCA-NF

In this section, we introduce an alternative implementation of PLiCCA based on solving a proxy conditional normalizing flow (NF) problem. This is a two-stage approach, as defined in Section 3.4. We therefore assume access to an a priori dimension-reduced representation $W \in \mathbb{R}^d$ of Y such that $Y = \psi(W) + \varepsilon$, together 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 Section G.1) shows that if we specify the model

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{21}$$

then maximizing the likelihood is equivalent to solving

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

where we have added a tuning parameter λ_{NF}, and $J_{\tilde{g}}(w)$ denotes the Jacobian matrix of a smooth function $\tilde{g}$ evaluated at w. We refer to the minimization problem above as the NF problem, because its objective contains the Jacobian log-determinant term arising from the change-of-variables formula, which is characteristic of NF models and which they can handle efficiently (see e.g. Kobyzev et al. (2020) for an introduction).

In analogy with PLiCCA-VAE, we propose estimating $\tilde{g}$ and B from the proxy NF problem. Next, we estimate the remaining model quantities according to Theorem 3.3, using equations (10)–(12). We define the final encoder as $g = \tilde{g} \circ \phi$ and the final decoder as $f = \psi \circ \tilde{g}^{-1}$. We refer to this implementation of PLiCCA as PLiCCA-NF.

In what follows, we assume that W, the latent representation of Y, is isotropic Gaussian. This choice is not critical and only serves to simplify the statements of the results; see Section G.5 for more details.

4.4 Connection between PLiCCA and conditional NFs

To elucidate the connection between PLiCCA and the conditional NF objective, we begin with the natural regression problem suggested by Theorem 3.3 with $\mathcal{C} = \mathcal{C}_{\text{NF}}$:

$$\underset{\substack{g:\mathbb{R}^d\to\mathbb{R}^d, g\in\mathcal{C}_{\text{NF}} \\ B\in\mathbb{R}^{p\times d},\ \Sigma_{g(W)}\succeq aI_d}}{\text{minimize}} \quad \mathbb{E}\left[\left\|g(W) - B^\top X\right\|_2^2\right]. \tag{22}$$

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

$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\text{NF}}, B\in\mathbb{R}^{p\times d} \\ \mathbb{E}[\ln|\det(J_{\tilde{g}}(W))|]\geq b}}{\text{minimize}} \quad \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right], \tag{23}$$

for some constant $b \in \mathbb{R}$. The NF problem we employ in practice corresponds to the Lagrangian form of (23).

The objective in both problems is identical. Although the constraint $\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b$ does not at first appear closely related to the constraint in the regression problem (22), we show that, analogously to the reconstruction term in the conditional VAE, it is closely related to the constraint $\Sigma_{\tilde{g}(W)} \succeq aI_d$. The following results make this connection precise.

Theorem 4.2 *Fix $a > 0$. If $\tilde{g} \in \mathcal{C}_{\text{NF}}$, then*

$$\Sigma_{\tilde{g}(W)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b(a), \tag{24}$$

where $b(a) \equiv \frac{d}{2}\ln(a) - C$ and C is a constant that depends on the bi-Lipschitz constants of $\mathcal{C}_{\text{NF}}$, as well as the Hessian matrices of the coordinates $\tilde{g}_i$ of $\tilde{g}$. See equation (181) for the full expression of C.

Using Theorem 4.2, we can view the NF problem (23) as a relaxation of the regression problem (9). Moreover, since Theorem 3.3 shows that the regression and PLiCCA problems are equivalent, we have the following corollary.

Corollary 4.1 *Fix $a > 0$. The PLiCCA problem,*

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d \\ g\in a^{-1/2}\mathcal{C}_{\text{NF}}}}{\text{maximize}} \quad \sum_{i=1}^{d}\mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2, \tag{25}$$

admits as a relaxation the NF problem

$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\text{NF}}, B\in\mathbb{R}^{p\times d} \\ \mathbb{E}[\ln|\det(J_{\tilde{g}}(W))|]\geq b(a)}}{\text{minimize}} \quad \mathbb{E}\left[\left\|\tilde{g}(W) - B^\top X\right\|_2^2\right]. \tag{26}$$

Remark 4 *In the definition of $\mathcal{C}_{\text{NF}}$, one could explicitly impose an upper bound on a matrix norm of the Hessian of the coordinates of $\tilde{g} \in \mathcal{C}_{\text{NF}}$ as an explicit constraint, but we elect not to, in order to simplify the presentation.*

In the next theorem, Theorem 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 state an auxiliary lemma.

Lemma 4.1 *If* $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$*, then*

$$\mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right] \geq b \implies \det\left(\Sigma_{\tilde{g}(W)}\right) \geq c(b). \tag{27}$$

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

The following theorem can be proved using Lemma 4.1.

Theorem 4.3 *The NF problem,*

$$\underset{\substack{\tilde{g}\in\mathcal{C}_{\mathrm{NF}}, B\in\mathbb{R}^{p\times d} \\ \mathbb{E}[\ln|\det(J_{\tilde{g}}(W))|]\geq b}}{\text{minimize}} \quad \mathbb{E}\left[\left\|\tilde{g}(W) - B^{\top}X\right\|_2^2\right], \tag{28}$$

admits as a relaxation the "geometric PLiCCA" problem

$$\sup_{\substack{\tilde{g}:\mathbb{R}^d\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d} \\ \Sigma_{\tilde{g}(W)}=\Sigma_{T^{\top}X}=I_d \\ \tilde{g}\in c(b)^{-1/2}\mathcal{C}_{\mathrm{NF}}}} \left(\prod_{i=1}^{d} h\left(\rho_i\right)\right)^{1/d}, \tag{29}$$

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

Remark 5 *We can sanity check the results by applying Theorem 4.2 and Lemma 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* $\det\left(\Sigma_{\tilde{g}}(W)\right) \geq a^d$*, the constant* C *is the only source of non-tightness in these inequalities.*

Together, Corollary 4.1 and Theorem 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 provides theoretical justification for using the NF problem as a proxy for PLiCCA.

5 DATA ANALYSIS

We apply PLiCCA to data from 700 subjects from the Human Connectome Project (Van Essen et al., 2013) for training and 303 subjects for validation. For each subject, functional connectivity from resting-state fMRI is represented as a 68×68 covariance matrix capturing the temporal correlations among signals from 68 regions of interest defined by the Desikan–Killiany atlas (Desikan et al., 2006). Standard fMRI preprocessing was applied (Glasser et al., 2013). We vectorize the upper triangular part of each covariance matrix to obtain a 2346-dimensional feature vector, which we treat as the target view.

For each subject, we also observe auxiliary variables, including demographic, psychometric, and behavioral features, represented as a 150-dimensional vector that serves as the auxiliary view. Results for PLiCCA-VAE are shown in Figure 1, where we show the first coordinate of the learned latent representation, namely the one most strongly linearly associated with the auxiliary variables. Results reveal a behavioural positive-negative mode of variation, in line with prior findings in the literature (Smith et al., 2015). Details on training and the choice of architecture are deferred to the appendix. Comparisons of PLiCCA-VAE and PLiCCA-NF with several alternative approaches are also deferred to Section B, which also includes experiments using cortical thickness as the target view.

Additionally, in Section A, we validate the proposed approaches and provide a comparison with related methods on synthetic datasets of images of rings.

6 DISCUSSION

In this work, we introduce PLiCCA, a correlation-based approach to supervised disentanglement. PLiCCA also provides a tool for exploring the dependence structure between two data views. We establish theoretical results for the population formulation and show a nontrivial connection to conditional latent variable models, which enables efficient computation. Finally, we evaluate our approach on synthetic and neuroimaging data and show that it performs favorably relative to alternative models.

PLiCCA models the posterior of the latent variables as $q(z \mid y)$ rather than $q(z \mid y, x)$, a choice related to posterior collapse in the conditional VAE literature (Zhou et al., 2023). This restriction is essential to our notion of supervised disentanglement, since it yields latent representations $U = g(Y)$ that are only a function of Y, but where g is nevertheless informed by X. An interesting future direction is to define an extended notion of supervised disentanglement that allows modeling the posterior as $q(z \mid y, x)$: a step in this direction is the work of Kim et al. (2023), where the posterior is modeled as a mixture of $q(z \mid y, x)$ and $q(z \mid y)$. We hope that PLiCCA will provide a foundation for future theoretical, methodological, and applied developments.

Acknowledgements

This work was partially supported by U.S. NIH grant R03EB033001 and U.S. NSF grant DMS2210064. The

content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

References

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696, 2024. doi: 10.1109/TPAMI.2024.3420937.

Gemma E Moran and Bryon Aragam. Towards interpretable deep generative models via causal representation learning. *Journal of the American Statistical Association: Reviews*, 2026. doi: 10.1080/01621459.2026.2620154.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Cristian Meo, Louis Mahon, Anirudh Goyal, and Justin Dauwels. α-tcvae: On the relationship between disentanglement and diversity. *arXiv preprint arXiv:2411.00588*, 2024.

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. Disentanglement learning via topology. In *International Conference on Machine Learning*, pages 2474–2504. PMLR, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E.J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015. ISSN 15461726. doi: 10.1038/nn.4125.

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pages 50–59. PMLR, 2018.

Toan Pham Van, Tam Minh Nguyen, Ngoc N Tran, Hoai Viet Nguyen, Linh Bao Doan, Huy Quang Dao, and Thanh Ta Minh. Interpreting the latent space of generative adversarial networks using supervised learning. In *2020 International Conference on Advanced Computing and Applications (ACOMP)*, pages 49–54. IEEE, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. In *International Conference on Machine Learning*, pages 4723–4732. PMLR, 2019.

Qingyu Zhao, Ehsan Adeli, Nicolas Honnorat, Tuo Leng, and Kilian M Pohl. Variational autoencoder for regression: Application to brain aging analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 823–831. Springer, 2019.

Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7920–7929, 2020.

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

Thomas Lu, Aboli Marathe, and Ada Martin. Supervising variational autoencoder latent representations with language. In *Proceedings of UniReps:*

the First Workshop on Unifying Representations in Neural Models, pages 267–278. PMLR, 2024.

Kemal Inecik, Aleyna Kara, Antony Rose, Muzlifah Haniffa, and Fabian J Theis. Tardis: Achieving robust and structured disentanglement of multiple covariates. In *International Conference on Research in Computational Molecular Biology*, pages 285–289. Springer, 2025.

Seunghwan An and Jong-June Jeon. Customization of latent space in semi-supervised variational autoencoder. *Pattern Recognition Letters*, 177:54–60, 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Erik Bodin, Alexandru I Stere, Dragos D Margineantu, Carl Henrik Ek, and Henry Moss. Linear combinations of latents in generative models: subspaces and beyond. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19239–19249, 2022.

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

Chenfeng Guo and Dongrui Wu. Canonical correlation analysis (cca) based multi-view learning: An overview. *arXiv preprint arXiv:1907.01693*, 2019.

Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2019.

James Chapman and Hao-Ting Wang. 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, 2021.

Alicja Gosiewska, Anna Kozak, and Przemysław Biecek. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150:113556, 2021.

Yongchun Li, Santanu S Dey, and Weijun Xie. On sparse canonical correlation analysis. *Advances in Neural Information Processing Systems*, 37:10707–10734, 2024.

Anna Bykhovskaya and Vadim Gorin. High-dimensional canonical correlation analysis. *arXiv preprint arXiv:2306.16393*, 2023.

James Buenfil and Eardi Lila. Asymmetric canonical correlation analysis of riemannian and high-dimensional data. *arXiv preprint arXiv:2404.11781*, 2024.

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

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

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

Tomer Michaeli, Weiran Wang, and Karen Livescu. Nonparametric canonical correlation analysis. In *International Conference on Machine Learning*, pages 1967–1976. PMLR, 2016.

Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International journal of neural systems*, 10(05):365–377, 2000.

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

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR, 2013.

Tomer Friedlander and Lior Wolf. Dynamically-scaled deep canonical correlation analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 232–244. Springer, 2023.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092. PMLR, 2015a.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems*, 29, 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6): 863–879, 1992.

Yue Song, Andy Keller, Nicu Sebe, and Max Welling. Latent traversals in generative models as potential flows. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32288–32303, 2023.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Walter Rudin. Principles of mathematical analysis. *3rd ed.*, 1976.

Mateusz Krukowski. Natural proof of the characterization of relatively compact families in l^p–spaces on locally compact groups. *arXiv preprint arXiv:1801.01898*, 2018.

Hyunwoo J Kim, Nagesh Adluru, Barbara B Bendlin, Sterling C Johnson, Baba C Vemuri, and Vikas Singh. Canonical correlation analysis on riemannian manifolds and its applications. In *European conference on computer vision*, pages 251–267. Springer, 2014.

Min Ho Cho, Sebastian Kurtek, and Karthik Bharath. Tangent functional canonical correlation analysis for densities and shapes, with applications to multimodal imaging data. *Journal of multivariate analysis*, 189:104870, 2022.

Wenjia Wang and Yi-Hui Zhou. Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. *Journal of Multivariate Analysis*, 185:104781, 2021.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. In *International Conference on Learning Representations*, 2022.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

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

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Kenton Murray, Jeffery Kinnison, Toan Q. Nguyen, Walter Scheirer, and David Chiang. Auto-sizing the transformer network: Improving speed, efficiency, and performance for low-resource machine translation. In *Proceedings of the Third Workshop on Neural Generation and Translation*, 2019.

Yijia Zheng, Tong He, Yixuan Qiu, and David P Wipf. Learning manifold dimensions with conditional variational autoencoders. *Advances in Neural Information Processing Systems*, 35:34709–34721, 2022.

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

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 3964–3979, 2020.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, July 2006. ISSN 10538119. doi: 10.1016/j.neuroimage.2006.01.021.

Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi,

Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

Guanglin Zhou, Shaoan Xie, Guangyuan Hao, Shiming Chen, Biwei Huang, Xiwei Xu, Chen Wang, Liming Zhu, Lina Yao, and Kun Zhang. Emerging synergies in causality and deep generative models: A survey. *arXiv preprint arXiv:2301.12351*, 2023.

Young-geun Kim, Ying Liu, and Xue-Xin Wei. Covariate-informed representation learning to prevent posterior collapse of ivae. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2660. PMLR, 2023.

Clément Chadebec and Stéphanie Allassonnière. A geometric perspective on variational autoencoders. *Advances in neural information processing systems*, 35:19618–19630, 2022.

Mengjie Chen, Chao Gao, Zhao Ren, and Harrison H Zhou. Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 688–695. IEEE, 2015b.

Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. Scalable and effective deep cca via soft decorrelation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1488–1497, 2018.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. normflows: A pytorch package for normalizing flows. *The Journal of Open Source Software*, 8(86):5361, 2023.

Stephen M. Smith, Christian F. Beckmann, Jesper Andersson, Edward J. Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A. Feinberg, Ludovica Griffanti, Michael P. Harms, Michael Kelly, Timothy Laumann, Karla L. Miller, Steen Moeller, Steve Petersen, Jonathan Power, Gholamreza Salimi-Khorshidi, Abraham Z. Snyder, An T. Vu, Mark W. Woolrich, Junqian Xu, Essa Yacoub, Kamil Uğurbil, David C. Van Essen, and Matthew F. Glasser. Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.039. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811913005338.

Gregory Gundersen, Bianca Dumitrascu, Jordan T Ash, and Barbara E Engelhardt. End-to-end training of deep probabilistic cca on paired biomedical observations. In *Uncertainty in artificial intelligence*, 2019.

Jiwei Zhang, Yi Yu, Suhua Tang, Jianming Wu, and Wei Li. Variational autoencoder with cca for audio–visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3s):1–21, 2023.

Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 1692–1700, 2021.

Ana Lawry Aguila and André Altmann. A tutorial on multi-view autoencoders using the multi-view-ae library. *CoRR*, 2024.

Jia He, Feiyang Pan, Fuzhen Zhuang, and Qing He. Cca-flow: Deep multi-view subspace learning with inverse autoregressive flow. In *Asian Conference on Machine Learning*, pages 177–192. PMLR, 2020.

Mahdi Karami and Dale Schuurmans. Deep probabilistic canonical correlation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8055–8063, 2021.

Lin Qiu, Vernon M Chinchilli, and Lin Lin. Variational interpretable deep canonical correlation analysis. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

Agathe Senellart, Clément Chadebec, and Stéphanie Allassonnière. Improving multimodal joint variational autoencoders through normalizing flows and correlation analysis. *arXiv preprint arXiv:2305.11832*, 2023.

Paris A Karakasis and Nicholas D Sidiropoulos. Revisiting deep generalized canonical correlation analysis. *IEEE Transactions on Signal Processing*, 71:4392–4406, 2023.

Yujia Zheng, Yang Liu, Jiaxiong Yao, Yingyao Hu, and Kun Zhang. Nonparametric factor analysis and beyond. In *International Conference on Artificial Intelligence and Statistics*, pages 424–432. PMLR, 2025.

Lorenzo Basile, Santiago Acevedo, Luca Bortolussi, Fabio Anselmi, and Alex Rodriguez. Intrinsic di-

mension correlation: uncovering nonlinear connections in multimodal representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Roy R Lederman and Ronen Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.

Ronak Mehta and Zaid Harchaoui. A generalization theory for zero-shot prediction. *arXiv preprint arXiv:2507.09128*, 2025.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

De Huang and Joel Tropp. From poincaré inequalities to nonlinear matrix concentration. *Bernoulli*, 23(3), 2021.

S Boucheron, G Lugosi, and P Massart. A non asymptotic theory of independence, 2013.

Paige Bright, Alan Edelman, and Steven G Johnson. Matrix calculus (for machine learning and beyond). *arXiv preprint arXiv:2501.14787*, 2025.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Ingram Olkin and Albert W Marshall. *Inequalities: Theory of Majorization and Its Applications*, volume 143. Academic Press, 2014.

Zoran Kadelburg, Dusan Dukic, Milivoje Lukic, and Ivan Matic. Inequalities of karamata, schur and muirhead, and some applications. *The Teaching of Mathematics*, 8(1):31–45, 2005.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **No.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes.**
   (b) Complete proofs of all theoretical results. **Yes.**
   (c) Clear explanations of any assumptions. **Yes.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes.**
   (b) The license information of the assets, if applicable. **Not Applicable.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**
   (d) Information about consent from data providers/curators. **Not Applicable.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

Supplementary Materials

A Simulations

In order to validate the proposed approaches, we provide a comparison with related methods on synthetic datasets. We take inspiration from the rings and discs example of Chadebec and Allassonnière (2022) and generate datasets of greyscale images of rings. The images are 20×20 pixels and represent the target data view, referred to as Y. Along with Y we generate an auxiliary data view $X \in \mathbb{R}^p$, a higher-dimensional random vector which is correlated with the image Y's latent parameters. The goal of the approaches we compare is to produce a low-dimensional latent representation of Y, $U = g(Y)$, which can be interpreted via its correlation with X. Specifically, we solve the PLiCCA problem defined in (2), which we restate here for convenience:

$$\underset{\substack{g:\mathbb{R}^q \to \mathbb{R}^d,\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d \\ g \in \mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) \theta_i^\top X \right]^2, \tag{30}$$

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

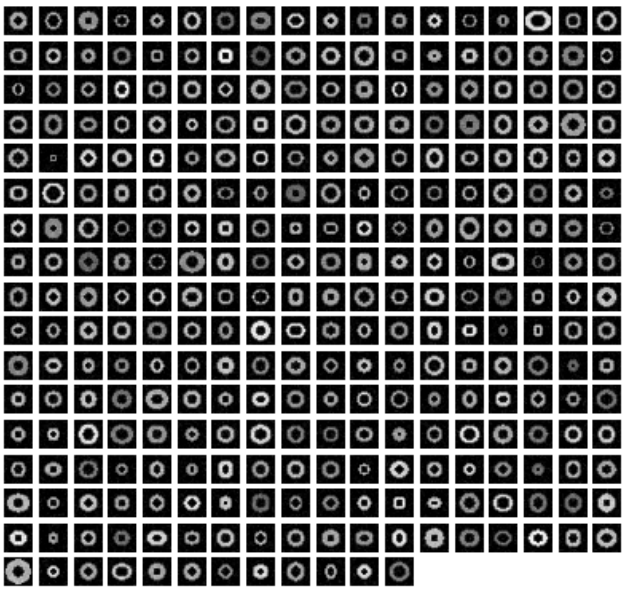

Figure 2: Examples of generated synthetic rings.

In order to construct a dataset where such structure exists, we begin not with images Y but with a simpler representation A which we generate jointly with X, where by construction A and X are correlated. We then map A to Y via a nonlinear map $Y = \phi(A)$, so that in the end we obtain (Y, X) with the desired nonlinear correlation structure.

Concretely, we sample $N = 300$ (A_i, X_i) i.i.d. pairs, where $A_i \in \mathbb{R}^4$ determines the parameters of the ring image Y, and $X \in \mathbb{R}^p$ for $p = 30$ represents the auxiliary data. The rings are parameterized by four parameters, r_1 ="radius of the hole," r_2 ="width of the ring," r_3 ="contrast," which scales the images by a constant, effectively decreasing or increasing the contrast of the image, and r_4 ="ellipticity," compressing the ring along the x-axis.

These parameters are contained in A_i, up to an affine transformation. The joint distribution of (A, X) is Gaussian with $\Sigma_A = I_4$, $\Sigma_X = I_p$, and cross covariance matrix $\Sigma_{AX} = \Sigma_A \left(\gamma_1 \eta_1 \theta_1^\top + \gamma_2 \eta_2 \theta_2^\top\right) \Sigma_X$ chosen so that the canonical vectors and canonical correlations between A and X are determined: $\eta_1 = (0, 0, 1, 0)^\top$, $\eta_2 = (0, 0, 0, 1)$ are the canonical vectors for A, $\theta_1 = (0, 0, 1, 0, \dots 0)^\top, \theta_2 = (0, 0, 0, 1, \dots 0) \in \mathbb{R}^p$ are the canonical vectors for X, and the canonical correlations for each pair, respectively, are set to $\gamma_1 = 0.9$ and $\gamma_2 = 0.7$ (see equation (3) of Chen et al. (2013)). The canonical variable pairs are then $(U_1 = A_3, V_1 = X_3)$ and $(U_2 = A_4, V_2 = X_4)$. This choice of canonical vectors means that X is mostly filled with noise which is uncorrelated with A, except through X_3 and X_4, which are correlated with A_3 and A_4 respectively. Thus, we have selected r_3 = contrast and r_4 = ellipticity as the dimensions along which changes in their values correspond to changes in X_3 and X_4.

We rescale A and call this vector Z, so that the parameters are physically realistic: $Z_1 \in [0.1, 0.5]$, $Z_2 \in [0.1, 0.5]$, $Z_3 \in [0.3, 1.0]$, $Z_4 \in [0.7, 1.3]$. Z is used to construct the ring grey-scale images, which we flatten into $q = 20 \times 20 = 400$ dimensional vectors, $Y_i \in \mathbb{R}^q$. Before applying the scaling via the contrast parameter, a grey-scale value of 1 represents the ring, while a grey-scale value of 0 represents the background. After scaling by the contrast parameter, i.i.d. Gaussian noise with standard deviation 0.003 is added to the pixel values to make the problem more difficult. We view the process of mapping A to Z and subsequently to Y as defining a nonlinear process $\phi : \mathbb{R}^4 \to \mathbb{R}^q$ so that $Y = \phi(A)$, which the methods must invert, via their encoders, in order to recover A and find its correlation with X. Examples of generated rings are shown in Figure 2.

The goal of each method is thus to construct canonical variable pairs (U_1, V_1) and (U_2, V_2), identifying contrast and ellipticity (r_3 and r_4) as the first two latent coordinates of U, and associating them with X_3 and X_4 via V. In a real-world setting, we would not know in advance what the true U_1 and U_2 represent, and would rely on the V_i, linear combinations of X, to interpret the meaning of the disentangled representation of Y, U. Thus, the objective is to learn latent representations for Y, while leveraging the dataset X to guide these representations.

We validate our proposed method against several alternatives, including DCCA, DCCA-NOI, DCCA-SDL, DCCAE, and DVCCA (Andrew et al., 2013; Wang et al., 2015b; Chang et al., 2018; Wang et al., 2015a, 2016). We use the `cca-zoo` package in implementing these alternatives (Chapman and Wang, 2021).

In order to compare the methods of interest, each method is run on the same number of num_trials = 20 i.i.d. datasets, generated as described above. In Table 1 we show the results, where we compute the out-of-sample correlation (sum of the first two canonical correlations) and reconstruction errors on a validation set of size N_val = 2500. Having emphasized the importance of invertibility for interpretability, we include reconstruction error as a proxy for invertibility for the applicable methods (some of the methods we compare are not invertible). For all methods, hyperparameters are chosen via K-fold cross validation, with $K = 4$ or 5. Each method is run for 20000 epochs with a batch size of 50, at which point the in-sample loss has effectively stopped decreasing. Each method uses a $d = 4$ dimensional latent space, the known latent dimension of the image dataset. Our proposed approaches, PLiCCA-VAE and PLiCCA-NF, appear to outperform competing approaches.

Method	**Validation correlation**	**Validation reconstruction error**
DCCA	0.680 (0.033)	–
DCCA-NOI	0.680 (0.037)	–
DCCA-SDL	0.350 (0.031)	–
DCCAE	0.040 (0.009)	0.02297 (0.00047)
DVCCA	0.151 (0.021)	0.10463 (0.00652)
PLiCCA-VAE	0.959 (0.024)	0.00796 (0.00017)
PLiCCA-NF	0.898 (0.033)	0.00794 (0.00012)

Table 1: Comparison of methods by validation correlation and validation reconstruction error. Standard deviations over the trials are included. A hyphen indicates that the method does not reconstruct the target view, and therefore reconstruction error cannot be computed.

All deep approaches use the same encoder (and decoder when applicable) architectures, 3 layer feed-forward networks with ReLu nonlinearities, of sizes $(q, 40, 20, d)$, with symmetric decoders. Notably, for the decoders, their last layer for Y is fed into a sigmoid, reflecting the structure of the underlying image data (grey-scale between 0 and 1). Methods that allow transformations of the auxiliary variables are constrained to be linear.

In training PLiCCA-NF, for every trial we first train an unsupervised VAE with the same architectures to

produce a latent representation of Y of dimension 4, and subsequently train the PLiCCA-NF model for 5000 epochs with a batch size of 50. We use a 4-dimensional normalizing flow with a single affine coupling block, parameterized by a small MLP (input 2, one hidden layer of 4 units, output 2), followed by a swap permutation, with a diagonal Gaussian base distribution 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). We use a lightweight architecture to avoid overfitting, as NFs are very expressive.

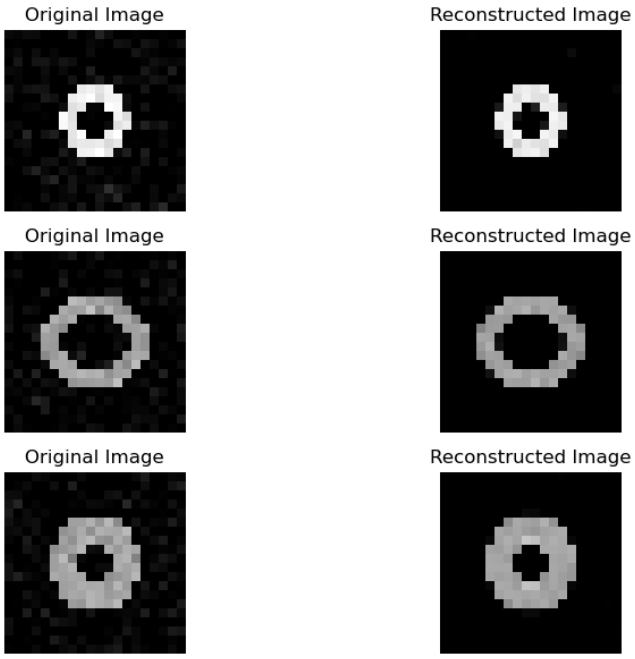

Figure 3: Typical in-sample reconstructions of rings, which were consistent across the methods that performed reconstructions.

Of the methods that performed reconstructions, we found that they were generally successful in reconstructing the noisy images. We show typical reconstructions in Figure 3.

In order to validate PLiCCA-VAE's learned latent embedding $g(Y)$, we examine its latent traversals. For a given ring's embedding coordinates (U_1, U_2), we can perturb one coordinate linearly, keeping the other coordinate fixed, and map these perturbations through the decoder to define a nonlinear trajectory in the space of the target view. We can also examine the associated signatures θ_1 and θ_2. These are shown in Figure 4.

B Data analysis

In the context of our application to the Human Connectome Project (HCP), we validate our proposed method against several alternatives, including DCCA, DCCA-NOI, DCCA-SDL, DCCAE, and DVCCA. To this end, we run two validation studies, each using a different imaging modality: cortical thickness and resting-state functional connectivity. Functional connectivity was computed from resting-state fMRI images preprocessed using the minimal preprocessing HCP pipeline (Glasser et al., 2013), including spatial artifact and distortion removal, as well as mapping onto a common reference template (Smith et al., 2013). We define 68 spatially localized regions of interest (ROIs) using the Desikan–Killiany atlas (Desikan et al., 2006). We then use the fMRI time series to compute a "functional fingerprint" representation: a 68×68 covariance matrix that captures the temporal correlation between the fMRI signals of any two ROIs. We subtract the mean connectivity matrix from each subject-specific matrix and show three examples of the resulting matrices in Figure 1. We then vectorize the upper triangular part of each connectivity matrix, yielding a vector y_k of dimension 2346 for each subject. These are treated as the target view. For each subject, we also observe auxiliary variables, including demographic, psychometric, and behavioral features, represented as a 170-dimensional vector that is used as the auxiliary view. For the results shown in Figure 1, we treat 20 of these as confounders (those identified in Smith

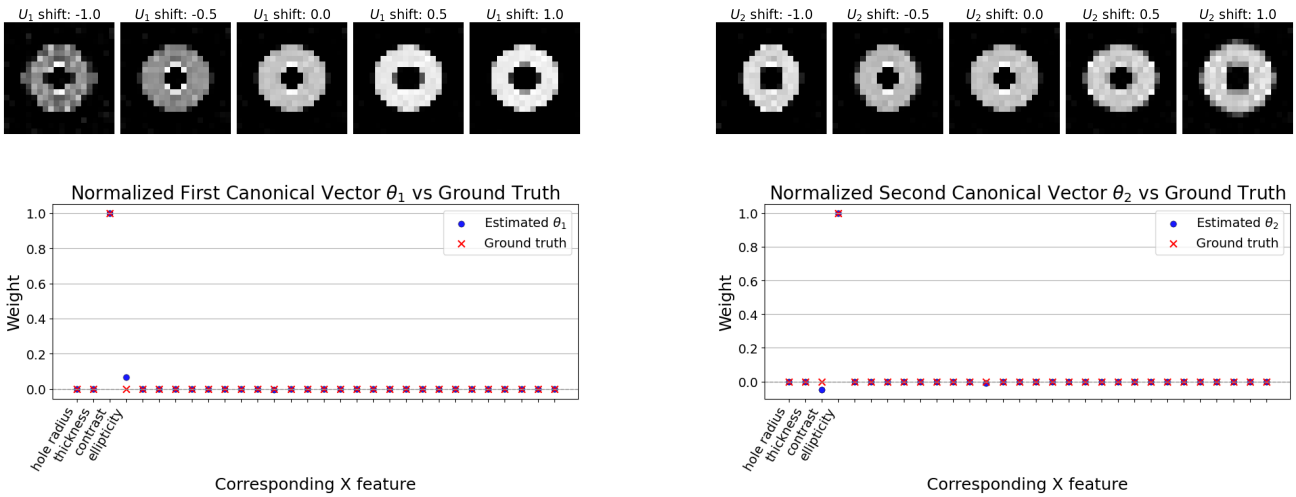

Figure 4: On the left-hand side: the upper plot shows latent traversals of U_1, while the lower plot shows the associated signature θ_1, for a typical run of PLiCCA-VAE. The right-hand side shows the analogous plots of latent traversals of U_2 and signature θ_2. The unlabeled plot axes (indices 5 through p) in the lower plots correspond to uninformative X features. We see that PLiCCA has successfully learned that the contrast of the image r_3 is correlated with X_3, and that the ellipticity of the ring r_4 is correlated with X_4. In practice, without the ground truth, we can use these latent traversals and the associated signatures θ_i to interpret the learned embedding $g(Y)$.

et al. (2015)) and regress them out from the remaining 150 variables, which become our new x_k. However, for the Monte Carlo study shown here, all 170 variables are treated as auxiliary variables, and no step to account for confounding is performed.

A second validation study is conducted by replacing connectivity with cortical thickness. For each subject k, cortical thickness derived from MRI is represented by a 32K-dimensional vector y_k, corresponding to measurements at 32K locations on the cortical surface (see Glasser et al. (2013) for preprocessing details). To account for spatial structure, we expand the cortical thickness data in the basis of the Laplacian of a template cortical surface and keep 1024 coefficients.

We use the same encoder–decoder architecture for all methods that involve a VAE component. For the connectivity data, we employ an MLP with 2346 input features, a single hidden layer of size 256, and a latent output dimension of 128. We use SELU activation functions. A similar architecture is used for the thickness data, with an MLP that takes 1024 input features, has a single hidden layer of size 256, and outputs a 128-dimensional latent representation. The decoder is chosen to be symmetric with respect to the encoder. The log-variance encoder is a two-layer MLP: it maps the input to 256 hidden units, applies a SELU activation, and then projects to a 128-dimensional output. A final Hardtanh is used to bound the log-variance values to the interval $[-6, 2]$.

Methods that include transformations of the auxiliary variables have such a transformation constrained to be linear. DCCA, when applied to the thickness data, failed to provide a solution because of numerical non-invertibility issues. We therefore reduced the number of latent variables until the problem was resolved. As a result, DCCA was trained using only 8 latent features for the cortical thickness data.

Finally, the proposed PLiCCA-NF, which requires a dimension-reduced representation as input, was trained on embeddings obtained from an unsupervised VAE with the same architecture described above and a 128-dimensional latent space. We used a 128-dimensional normalizing flow with six affine coupling blocks, each parameterized by a small MLP (input 64, two hidden layers of 16 units, output 128), alternating with swap permutations. Using a small MLP architecture for the coupling networks was essential to prevent numerical instabilities during training. All methods were trained for 200 epochs using a learning rate of 10^{-3}. The `cca-zoo` implementation was used to run the competing methods. Note that because some subjects are genetically related (e.g., siblings or twins), we first partitioned the dataset into two subsets of maximally unrelated individuals (700 subjects for training and 303 for validation). We then repeatedly subsampled 4/5 of the training set at random

Table 2: Comparison of methods for cortical thickness and functional connectivity in terms of total correlation and reconstruction error. Total correlation denotes the sum of the pairwise correlations among the first 10 identified canonical directions. Standard deviations across trials are reported in parentheses. A hyphen indicates that the method does not reconstruct the target view, and therefore reconstruction error cannot be computed.

Cortical thickness

Method	Total correlation	Reconstruction error
DCCA	0.5500 (0.0951)	–
DCCA-NOI	0.5793 (0.1560)	–
DCCA-SDL	0.6149 (0.0902)	–
DCCAE	0.8319 (0.0296)	0.2038 (0.0104)
DVCCA	0.8823 (0.0538)	0.8028 (0.1475)
PLiCCA-VAE	1.1195 (0.1507)	0.1489 (0.0013)
PLiCCA-NF	0.9536 (0.1329)	0.1753 (0.0508)

Functional connectivity

Method	Total correlation	Reconstruction error
DCCA	1.6879 (0.0971)	–
DCCA-NOI	0.1933 (0.0855)	–
DCCA-SDL	1.8628 (0.0396)	–
DCCAE	0.9466 (0.0879)	0.000251 (0.000009)
DVCCA	0.9912 (0.1294)	0.015607 (0.008683)
PLiCCA-VAE	1.1440 (0.1822)	0.000538 (0.000116)
PLiCCA-NF	1.0753 (0.2969)	0.000458 (0.000008)

and re-evaluated the model. This procedure should mitigate potential spillover effects due to genetic relatedness.

We show the results in Table 2. Among the approaches that guarantee invertibility (and are therefore interpretable), our proposed PLiCCA-VAE performs best for both imaging modalities (Thickness and Connectivity), both in terms of the primary metric—out-of-sample correlation with the identified features of X—and reconstruction error. For functional connectivity, DCCA and DCCA-SDL yield higher correlations. However, these approaches do not aim to provide invertible representations of the data and therefore do not solve the supervised disentanglement problem. Instead, they only aim to learn features that are correlated with linear combinations of X and therefore can encode more information into the first few learned components.

C Additional related works

Our approach is more closely related to Gundersen et al. (2019), who propose learning nonlinear latent representations of each view by maximizing the likelihood of a linear probabilistic CCA model, rather than explicitly maximizing correlation. Zhang et al. (2023) propose a related idea that applies classical CCA to learned latent representations, although their approach 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 (Guo et al., 2019; Lee and Pavlovic, 2021; Aguila and Altmann, 2024). 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 (Wang et al., 2016; He et al., 2020; Karami and Schuurmans, 2021; Qiu et al., 2022; Lyu et al., 2022; Senellart et al., 2023). 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. Deterministic

extensions of nonlinear CCA for solving the multi-view data problem include methods surveyed in Guo and Wu (2019); see Karakasis and Sidiropoulos (2023) for a more recent approach.

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., Khemakhem et al., 2020; Hyvärinen et al., 2023; 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, while the work of Lederman and Talmon (2018) nonlinearly constructs a common latent manifold for two data views via diffusion maps, and are thus tangentially related to our own.

D Background on non-invertible nonlinear CCA 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:

$$L_Y^2(\mathbb{R}^q, \mathbb{R}^d) \equiv \left\{ g : \mathbb{R}^q \to \mathbb{R}^d : \mathbb{E}\left[g(Y)\right] = 0 \text{ and } \forall i = 1, \dots d,\ \mathbb{E}\left[g_i(Y)^2\right] < \infty \right\},$$
$$L_X^2(\mathbb{R}^p, \mathbb{R}^d) \equiv \left\{ f : \mathbb{R}^p \to \mathbb{R}^d : \mathbb{E}\left[f(X)\right] = 0 \text{ and } \forall i = 1, \dots d,\ \mathbb{E}\left[f_i(X)^2\right] < \infty \right\}.$$

Then, the nonlinear, or nonparametric CCA problem, is defined as

$$\underset{\substack{g \in L_Y^2(\mathbb{R}^q, \mathbb{R}^d),\ f \in L_X^2(\mathbb{R}^p, \mathbb{R}^d) \\ \Sigma_{g(Y)} = \Sigma_{f(X)} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) f_i(X)\right]. \tag{31}$$

It has been shown (Breiman and 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 $L_X^2(\mathbb{R}^p, \mathbb{R})$ and $L_Y^2(\mathbb{R}^q, \mathbb{R})$, sometimes referred to as the conditional mean operator (Mehta and 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, (not to be confused with the invertible version of this problem which we propose in this paper, PLiCCA) coined by Michaeli et al. (2016), is defined as follows:

$$\underset{\substack{g \in L_Y^2(\mathbb{R}^q, \mathbb{R}^d),\ T \in \mathbb{R}^{p \times d} \\ \Sigma_{g(Y)} = \Sigma_{T^\top X} = I_d}}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}\left[g_i(Y) \theta_i^\top X\right], \tag{32}$$

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 $\mathbb{E}\left[g(Y)\right] = 0$ and $\mathbb{E}\left[f(X)\right] = 0$; 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_X^2 must be linear, L_X^2 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 $\hat{X} \equiv \mathbb{E}\left[X|Y\right]$. 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 $g(y) = \Sigma_{T^\top \hat{X}}^{-1/2} T^\top \mathbb{E}\left[X|Y = y\right]$. We note that in practice, we do not necessarily use these formulas to compute the solutions, as the conditional expectation $\mathbb{E}\left[X|Y\right]$ may not be easy to estimate.

E Supporting results and proofs for Section 3

E.1 Proof of Theorem 3.1

The following lemma is used in the proof of Lemma E.2 as well as Theorem 3.3. We will also show a more general version of this lemma, Lemma G.5, which is used in the proof of Theorem 4.3.

Lemma E.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*

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \left\|\Sigma_Z^{-1/2} \Sigma_{ZX} \Sigma_X^{-1/2}\right\|_F^2, \tag{33}$$

where $\theta_i \in \mathbb{R}^p$ are the columns of T, and $\eta_i \in \mathbb{R}^d$ 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 E.1
We begin from

$$\sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \sum_{i=1}^{d} \mathbb{E}\left[\theta_i^\top X Z^\top \eta_i\right]^2 \tag{34}$$

$$= \sum_{i=1}^{d} \left(\theta_i^\top \Sigma_{XZ} \eta_i\right)^2. \tag{35}$$

Changing variables from T to $\tilde{T} = \Sigma_X^{1/2} T$, whose columns we denote by $\tilde{\theta}$, we have

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 = \sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \left(\theta_i^\top \Sigma_{XZ} \eta_i\right)^2 \tag{36}$$

$$= \sup_{\substack{\tilde{T} \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \sum_{i=1}^{d} \left(\tilde{\theta}_i^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} \eta_i\right)^2 \tag{37}$$

$$\leq \sup_{\substack{\tilde{T} \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \left\|\tilde{T}^\top \Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2} H\right\|_F^2, \tag{38}$$

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

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} \mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]^2 \leq \sup_{\substack{\tilde{T} \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d}, \\ H^\top H = \tilde{T}^\top \tilde{T} = I_d}} \left\|\tilde{T}^\top U \Lambda V^\top H\right\|_F^2 \tag{39}$$

$$\leq \|\Lambda\|_F^2 \tag{40}$$

$$= \left\|\Sigma_X^{-1/2} \Sigma_{XZ} \Sigma_Z^{-1/2}\right\|_F^2 \tag{41}$$

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. Here, $\mathcal{C}$ can refer to either $\mathcal{C}_{\text{VAE}}$ or $\mathcal{C}_{\text{NF}}$.

Lemma E.2 *The nonlinear solution g maximizes the PLiCCA problem*

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ T\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=\Sigma_{T^\top X}=I_d\\ g\in\mathcal{C}}}{\text{maximize}} \sum_{i=1}^{d}\mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2 \tag{42}$$

if and only if it minimizes the nonlinear regression problem

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=I_d,\ g\in\mathcal{C}}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y)-B^\top X\right\|_2^2\right]. \tag{43}$$

Proof of Lemma E.2
We start from

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=I_d,\ g\in\mathcal{C}}}{\text{minimize}} \mathbb{E}\left[\left\|B^\top X-g(Y)\right\|_2^2\right]. \tag{44}$$

and suppress the constraints in the notation for convenience. We have

$$\begin{aligned}
\inf_{g,B}\mathbb{E}\left[\left\|B^\top X-g(Y)\right\|_2^2\right] &= \inf_{g,B}\mathbb{E}\left[g(Y)^\top g(Y)-2g(Y)^\top B^\top X+X^\top BB^\top X\right] && (45)\\
&= \inf_{g,B}\mathbb{E}\left[\operatorname{tr}\left(g(Y)^\top g(Y)-2g(Y)^\top B^\top X+X^\top BB^\top X\right)\right] && (46)\\
&= \inf_{g,B}\operatorname{tr}\left(\mathbb{E}\left[g(Y)g(Y)^\top-2B^\top Xg(Y)^\top+B^\top XX^\top B\right]\right) && (47)\\
&= \inf_{g,B}\operatorname{tr}\left(\Sigma_{g(Y)}-2B^\top\Sigma_{Xg(Y)}+B^\top\Sigma_X B\right) && (48)\\
&= \inf_{g,B}\operatorname{tr}\left(\Sigma_{g(Y)}\right)-2\operatorname{tr}\left(B^\top\Sigma_{Xg(Y)}\right)+\operatorname{tr}\left(B^\top\Sigma_X B\right). && (49)
\end{aligned}$$

Since this is a least squares problem, we can plug in the optimal B for fixed g, which is simply $B=\Sigma_X^{-1}\Sigma_{Xg(Y)}$:

$$\begin{aligned}
\inf_{g,B}\mathbb{E}\left[\left\|B^\top X-g(Y)\right\|_2^2\right] &= \inf_g\inf_B\mathbb{E}\left[\left\|B^\top X-g(Y)\right\|_2^2\right] && (50)\\
&= \inf_g\inf_B\operatorname{tr}\left(\Sigma_{g(Y)}\right)-2\operatorname{tr}\left(B^\top\Sigma_{Xg(Y)}\right)+\operatorname{tr}\left(B^\top\Sigma_X B\right) && (51)\\
&= \inf_g\left[\operatorname{tr}\left(\Sigma_{g(Y)}\right)-2\operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right)+\operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_X\Sigma_X^{-1}\Sigma_{Xg(Y)}\right)\right] && (52)\\
&= \inf_g\left[\operatorname{tr}\left(\Sigma_{g(Y)}\right)-\operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right)\right], && (53)
\end{aligned}$$

and using the fact that $\Sigma_{g(Y)}=I_d$, we obtain

$$\inf_{g,B}\mathbb{E}\left[\left\|B^\top X-g(Y)\right\|_2^2\right]=\inf_g\left[d-\left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2\right]. \tag{54}$$

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 E.1 to obtain

$$\left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2=\sup_{\substack{T\in\mathbb{R}^{p\times d},H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}}\sum_{i=1}^{d}\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2, \tag{55}$$

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

(54), we obtain

$$\inf_{g,B} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_g \left[d - \left\|\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1/2}\right\|_F^2\right] \tag{56}$$

$$= \inf_g \left[d - \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d \mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right] \tag{57}$$

$$= d - \sup_g \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \left[\sum_{i=1}^d \mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right] \tag{58}$$

$$= d - \sup_g \sup_{\substack{T\in\mathbb{R}^{p\times d},\\ \Sigma_{T^\top X}=I_d}} \left[\sum_{i=1}^d \mathbb{E}\left[g_i(Y)\theta_i^\top X\right]^2\right], \tag{59}$$

where in the last step we have absorbed the optimization over H into g. This completes the proof. □

Proof of Theorem 3.1

Recall that 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 E.2, the maximization problem of interest is equivalent to

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}=I_d,\ g\in\mathcal{C}_{\mathrm{VAE}}}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right]. \tag{60}$$

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 (53)), we obtain an optimization problem over only g,

$$\underset{g\in\mathcal{C}_{\mathrm{VAE}},\Sigma_{g(Y)}=I_d}{\text{minimize}} J(g), \tag{61}$$

where

$$J(g) \equiv d - \operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right). \tag{62}$$

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, an optimal T can be derived from B. This infimum is not $-\infty$ provided that $\mathcal{C}_{\mathrm{VAE}} \cap \{g : \Sigma_{g(Y)} = I_d\} \neq \varnothing$, so that the feasible set is not empty.

We begin with a sequence $\{g_n\} \subset \mathcal{C}_{\mathrm{VAE}}$ for which $J(g_n)$ converges to $\inf_{g\in\mathcal{C}_{\mathrm{VAE}}} J(g)$, and set out to show that $\{g_n\}$ has a subsequence that converges uniformly to a $g^* \in \mathcal{C}_{\mathrm{VAE}}$ 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 $J(g^*) = \inf_{g\in\mathcal{C}_{\mathrm{VAE}},\Sigma_{g(Y)}=I_d} J(g)$ with $g^* \in \mathcal{C}_{\mathrm{VAE}}$ and $\Sigma_{g^*(Y)} = I_d$.

Next, we show continuity of $J(g)$ with respect to the supremum norm. We need to show that if $\{g_n\} \subset \mathcal{C}_{\mathrm{VAE}}$ and $g \in \mathcal{C}_{\mathrm{VAE}}$ 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 Frobenius norm, since the trace function is continuous with respect to the Frobenius norm and because the composition of continuous functions is continuous. We have

$$\left\|\Sigma_{g_n(Y)X} - \Sigma_{g(Y)X}\right\|_F = \left\|\mathbb{E}\left[\left(g_n(Y) - g(Y)\right)X^\top\right]\right\|_F \tag{63}$$

$$\leq \mathbb{E}\left[\left\|\left(g_n(Y) - g(Y)\right)X^\top\right\|_F\right] \tag{64}$$

$$= \mathbb{E}\left[\|g_n(Y) - g(Y)\|_2\,\|X\|_2\right] \tag{65}$$

$$\leq \|g_n - g\|_\infty\,\mathbb{E}\left[\|X\|_2\right]. \tag{66}$$

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 $\mathcal{C}_{\mathrm{VAE}}$ 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 $\mathbb{E}\|Y - f^*(g^*(Y))\|_2^2 \leq \varepsilon$.

It follows directly from the uniform convergence of $\{g_n\}$ that g^* is M-Lipschitz, 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 $\{g_n g_n^\intercal\}$ to show that

$$\lim_{n\to\infty} \mathbb{E}[g_n(Y)] = \mathbb{E}[g^*(Y)] \text{ and} \tag{67}$$

$$\lim_{n\to\infty} \mathbb{E}\left[g_n(Y)g_n(Y)^\intercal\right] = \mathbb{E}\left[g^*(Y)g^*(Y)^\intercal\right]. \tag{68}$$

Therefore, g^* satisfies $\mathbb{E}[g^*(Y)] = 0$ and $\Sigma_{g^*(Y)} = I_d$. Finally, we have

$$\begin{aligned}
\|f_n(g_n(Y)) - f^*(g^*(Y))\|_2 &\leq \|f_n(g_n(Y)) - f_n(g^*(Y))\|_2 + \|f_n(g^*(Y)) - f^*(g^*(Y))\|_2 && (69)\\
&\leq m\,\|g_n(Y) - g^*(Y)\|_2 + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty && (70)\\
&\leq m\,\|g_n - g^*\|_\infty + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty . && (71)
\end{aligned}$$

Here, the notation $f|_S$ refers to the function f whose domain has been restricted to S. Therefore,

$$\mathbb{E}\left[\|(Y - f_n(g_n(Y))) - (Y - f^*(g^*(Y)))\|_2^2\right]^{1/2} \leq m\,\|g_n - g^*\|_\infty + \|f_n|_{K_1} - f^*|_{K_1}\|_\infty , \tag{72}$$

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

$$|\|Y - f_n(g_n(Y))\|_{L^2} - \|Y - f^*(g^*(Y))\|_{L^2}| \leq \|Y - f_n(g_n(Y))) - (Y - f^*(g^*(Y)))\|_{L^2}, \tag{73}$$

so that

$$\lim_{n\to\infty} \mathbb{E}\left[\|Y - f_n(g_n(Y))\|_2^2\right]^{1/2} = \mathbb{E}\left[\|Y - f^*(g^*(Y))\|_2^2\right]^{1/2}, \tag{74}$$

and squaring, we obtain

$$\mathbb{E}\left[\|Y - f^*(g^*(Y))\|_2^2\right] = \lim_{n\to\infty} \mathbb{E}\left[\|Y - f_n(g_n(Y))\|_2^2\right] \leq \varepsilon. \tag{75}$$

Similarly,

$$\mathbb{E}[f^*(g^*(Y))] = \lim_{n\to\infty} \mathbb{E}[f_n(g_n(Y))] = 0. \tag{76}$$

This completes the proof. □

E.2 Proof of Theorem 3.3

The proof is analogous to the proof of Lemma E.2. We start with

$$\underset{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}}{\text{minimize}} \mathbb{E}\left[\left\|B^\intercal X - g(Y)\right\|_2^2\right]. \tag{77}$$

From equation (53), for the optimal B for a fixed g,

$$\mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \operatorname{tr}\left(\Sigma_{g(Y)}\right) - \operatorname{tr}\left(\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\right) \tag{78}$$

$$= \operatorname{tr}\left(\Sigma_{g(Y)}\right) - \operatorname{tr}\left(\Sigma_{g(Y)}\Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}\right) \tag{79}$$

$$= \operatorname{tr}\left(\Sigma_{g(Y)}\left[I_d - \Sigma_{g(Y)}^{-1/2}\Sigma_{g(Y)X}\Sigma_X^{-1}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2}\right]\right). \tag{80}$$

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,

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \operatorname{tr}\left(\Sigma_{g(Y)}D\right) \tag{81}$$

It is straightforward to show that, as a function of $\Sigma_{g(Y)}$ subject to the constraint $\Sigma_{g(Y)} \succeq cI_d$, 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,

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} c\sum_{i=1}^d(1-\gamma_i^2) \tag{82}$$

$$= c\left(d - \sup_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \sum_{i=1}^d \gamma_i^2\right), \tag{83}$$

Now using Lemma E.1, we have

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = c\left(d - \sup_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \Sigma_{g(Y)}\succeq cI_d,\ g\in\mathcal{C}}} \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d \mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2\right). \tag{84}$$

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 E.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:

$$\Sigma_X^{1/2}\Sigma_{Xg(Y)}\Sigma_{g(Y)}^{-1/2} = \tilde{T}\Lambda\tilde{H}^\top, \tag{85}$$

where the solutions are

$$T = \Sigma_X^{-1/2}\tilde{T} \tag{86}$$

$$H = \Sigma_{g(Y)}^{-1/2}\tilde{H}. \tag{87}$$

On the other hand, from its definition, it is straightforward to show that $B = \Sigma_X^{-1}\Sigma_{Xg(Y)}$. Therefore, we have

$$\Sigma_{g(Y)}^{-1/2}B^\top\Sigma_X B\Sigma_{g(Y)}^{-1/2} = \tilde{H}\Lambda\tilde{H}^\top, \tag{88}$$

and

$$T = BH\Lambda^{-1}. \tag{89}$$

which follow from matrix algebra. This completes the proof. □

F Supporting results and proofs for conditional VAEs

F.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

$$p(y|x) = \frac{p(x,y)}{p(x)}\frac{p(x,y,z)}{p(x,y,z)} \tag{90}$$

$$= \frac{p(y,z|x)}{p(z|x,y)}\frac{q(z|y)}{q(z|y)}. \tag{91}$$

The log-likelihood is

$$\ln p(y|x) = \ln\left(\frac{p(y,z|x)}{q(z|y)}\right) + \ln\left(\frac{q(z|y)}{p(z|x,y)}\right) \tag{92}$$

$$= \mathbb{E}_{q(z|y)}\left[\ln\left(\frac{p(y,z|x)}{q(z|y)}\right)\right] + D_{KL}\left(q(z|y), p(z|y,x)\right). \tag{93}$$

Since the KL divergence is nonnegative, we obtain the ELBO lower bound:

$$\text{ELBO} = \mathbb{E}_{q(z|y)}\left[\ln\left(\frac{p(y,z|x)}{q(z|y)}\right)\right]. \tag{94}$$

Choosing to specify the model as

$$Y|Z, X \sim \mathcal{N}\left(f(Z), I_d\right) \tag{95}$$

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{96}$$

it holds that $p(y|z,x) = p(y|z)$, and equivalently $p(y,z|x) = p(y|z)p(z|x)$. From this, the ELBO can be written as

$$\text{ELBO} = \mathbb{E}_{q(z|y)}\left[\ln p(y|z)\right] - D_{KL}\left(q(z|y), p(z|x)\right). \tag{97}$$

Modeling the posterior as $q(z|y) \sim \mathcal{N}\left(g(Y), I_d\right)$, then to maximize the ELBO we can equivalently minimize

$$\frac{1}{2}\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right] + \frac{1}{2}\left\|g(Y) - B^\top X\right\|_2^2, \tag{98}$$

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):

$$\underset{g:\mathbb{R}^q \to \mathbb{R}^d,\ B \in \mathbb{R}^{p\times d}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right] + \mathbb{E}\left[\mathbb{E}_{q(z|Y)}\left[\|Y - f(z)\|_2^2\right]\right], \tag{99}$$

where the outermost expectation is with respect to (X,Y), the observed quantities.

F.2 Proof of Theorem 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

$$h(Y) \equiv -\mathbb{E}\left[\log\left(p(Y)\right)\right]. \tag{100}$$

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

$$I(X;Y) \equiv \mathbb{E}\left[\log\left(\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right)\right]. \tag{101}$$

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

$$h(Y|X) \equiv \mathbb{E}\left[\log\left(f_{X|Y}(X|Y)\right)\right] \tag{102}$$

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 F.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:*

$$h(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Y\right). \tag{103}$$

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*

$$h(Z) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Z\right). \tag{104}$$

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 4.1 and 4.3.

Lemma F.2 (GM-AM Eigenvalue Inequality) *For positive semi-definite $A \in \mathbb{R}^{d\times d}$, we have*

$$\frac{\mathrm{tr}(A)}{d} \geq \det(A)^{1/d}. \tag{105}$$

We have equality when A is a multiple of the identity matrix I_d.

Proof of Lemma F.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 $(\lambda_i)_{i=1}^d$ be the eigenvalues of A. Then Jensen's inequality applied to $f(x) = \log(x)$ gives

$$\log\left(\frac{\mathrm{tr}(A)}{d}\right) = \log\left(\frac{\sum_{i=1}^d \lambda_i}{d}\right) \geq \frac{\sum_{i=1}^d \log\left(\lambda_i\right)}{d} = \log\left(\prod_{i=1}^d \lambda_i^{1/d}\right) = \log\left(\det\left(A\right)^{1/d}\right). \tag{106}$$

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 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

$$R_Y(D) \equiv \inf_{P_{\hat{Y}|Y}:\mathbb{E}\left[\|Y-\hat{Y}\|_2^2 \leq D\right]} I(Y;\hat{Y}) \tag{107}$$

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 $D \geq \mathbb{E}\left[\|Y\|_2^2\right]$ we have $R_Y(D) = 0$, since then $\hat{Y} = 0$ satisfies the given constraint that $\mathbb{E}\left[\left\|Y - \hat{Y}\right\|_2^2 \leq D\right]$ 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 $\varepsilon \sim \mathcal{N}\left(0, I_d\sigma^2_{\text{enc}}\right)$ and ε is independent from Y. Let $\hat{Y} \equiv f(Z)$, i.e. the noisy reconstruction of Y. Then from the definition of $R_Y(\delta)$,

$$R_Y(\delta) \leq I(Y;\hat{Y}). \tag{108}$$

We have that

$$I(Y;\hat{Y}) \leq I(Y;Z) \tag{109}$$

by the data processing inequality (Lemma F.1) since $Y \to Z \to \hat{Y}$ is a Markov chain. Then,

$$\begin{aligned} I(Y;Z) &= I(Z;Y) && (110)\\ &= h(Z) - h(Z|Y) && (111)\\ &= h(Z) - h(Z|g(Y)) && (112)\end{aligned}$$

where we have used the symmetry of mutual information, item 4 of Lemma F.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 F.1 implies that

$$h(Z|g(Y)) = \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log\left(\sigma^2_{\text{enc}}\right). \tag{113}$$

Noting that mutual information is invariant to addition by constants, we can suppose that Z is mean 0 without loss of generality, so that $\Sigma_Z = \Sigma_{g(Y)} + \Sigma_\varepsilon = \Sigma_{g(Y)} + \sigma^2_{\text{enc}} I_d$. Since differential entropy is maximized for Gaussian distributions, Lemma F.1 applied to Z implies that

$$h(Z) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_{g(Y)} + \sigma^2_{\text{enc}} I_d\right). \tag{114}$$

Putting these inequalities together, we have shown that

$$\begin{aligned} R_Y(\delta) &\leq \frac{1}{2}\log\det\left(\Sigma_{g(Y)} + \sigma^2_{\text{enc}} I_d\right) - \frac{d}{2}\log\left(\sigma^2_{\text{enc}}\right) && (115)\\ &= \frac{1}{2}\log\det\left(\frac{1}{\sigma^2_{\text{enc}}}\Sigma_{g(Y)} + I_d\right) + \frac{d}{2}\log\left(\sigma^2_{\text{enc}}\right) - \frac{d}{2}\log\left(\sigma^2_{\text{enc}}\right) && (116)\\ &= \frac{1}{2}\log\det\left(\frac{1}{\sigma^2_{\text{enc}}}\Sigma_{g(Y)} + I_d\right). && (117)\end{aligned}$$

Applying the GM-AM inequality Lemma F.2, this is upper bounded:

$$\begin{aligned} \frac{1}{2}\log\det\left(\frac{1}{\sigma^2_{\text{enc}}}\Sigma_{g(Y)} + I_d\right) &\leq \frac{1}{2}\log\left(\frac{\text{tr}\left(\frac{1}{\sigma^2_{\text{enc}}}\Sigma_{g(Y)} + I_d\right)}{d}\right)^d && (118)\\ &= \frac{d}{2}\log\left(\text{tr}\left(\frac{1}{d\sigma^2_{\text{enc}}}\Sigma_{g(Y)}\right) + 1\right). && (119)\end{aligned}$$

Combining this with equation (117) and rearranging for $\Sigma_{g(Y)}$, we finally have

$$\text{tr}\left(\Sigma_{g(Y)}\right) \geq \sigma^2_{\text{enc}} d\left(e^{\frac{2}{d}R_Y(\delta)} - 1\right), \tag{120}$$

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 $D \geq \mathbb{E}\left[\|Y\|_2^2\right]$ we have $R_Y(D) = 0$, we can check that this translates into $C(D) = 0$ as well. This completes the proof. □

Using equation (117) directly, we also have the tighter bound related to the determinant of $\Sigma_{g(Y)}$:

$$\det\left(\frac{1}{\sigma^2_{\text{enc}}}\Sigma_{g(Y)} + I_d\right) \geq e^{2R_Y(\delta)}. \tag{121}$$

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.

F.3 Generalizing Theorem 4.1 to the case of non-isotropic noise

In Theorem 4.1, modeling the posterior uncertainty as $\sigma_{\text{enc}}^2 I_d$ is a simplification compared to the more general form $\Sigma_{Z|Y}(Y)$, which may depend on the data Y. We adopted this choice because it leads to an intuitive interpretation: the lower bound depends only on the magnitude of the encoder error (captured by the single parameter σ_{enc}^2) and the reconstruction error δ.

However, this result also generalizes to the case of a data-dependent $\Sigma_{Z|Y}(Y)$. Given the general Gaussian encoder

$$q(z \mid y) \sim \mathcal{N}\big(g(y), \Sigma_{Z|Y}(y)\big),$$

the lower bound of Theorem 4.1 becomes

$$\operatorname{tr}(\Sigma_{g(Y)}) \;\geq\; d\Big(\sigma_{\text{geom}}^2 \exp\big(\tfrac{2}{d} R_Y(\delta)\big) - \sigma_{\text{mean}}^2\Big). \tag{122}$$

The proof is essentially equivalent to that of Theorem 4.1. The difference arises from the expressions of $h(Z)$ and $h(Z|g(Y))$ in equation (112), which now involve $\mathbb{E}[\Sigma_{Z|Y}(Y)]$ and $\mathbb{E}\big[\log\det\Sigma_{Z|Y}(Y)\big]$, respectively.

Here, we define the *geometric mean variance* as $\sigma_{\text{geom}}^2 \equiv \exp\big(\frac{1}{d}\mathbb{E}\big[\log\det\Sigma_{Z|Y}(Y)\big]\big)$, and the *mean variance* as $\sigma_{\text{mean}}^2 \equiv \frac{1}{d}\operatorname{tr}\big(\mathbb{E}[\Sigma_{Z|Y}(Y)]\big)$. Although the behavior of the bound is more complex due to the interaction between the mean and geometric variances, the fundamental intuition is the same: higher reconstruction accuracy and higher magnitude noise in the latent space ensure that the encoder g does not collapse to 0.

In the special case when $\Sigma_{Z|Y}(y) = \sigma_{\text{enc}}^2 I_d$, we have $\sigma_{\text{geom}}^2 = \sigma_{\text{mean}}^2 = \sigma_{\text{enc}}^2$, and we recover the original statement of Theorem 4.1.

G Supporting results and proofs for conditional NFs

G.1 Derivation of conditional NF objective

In this case, we already have access to a dimension-reduced Y, which we call $W \in \mathbb{R}^d$.

We introduce the model

$$W = \tilde{f}(Z) \tag{123}$$
$$Z = B^\top X + \varepsilon_2 \tag{124}$$

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

$$p_{W,X}(w,x) = p_{Z,X}\left(\tilde{f}(w), x\right)\left|\det\left(J_{\tilde{f}}(w)\right)\right|, \tag{125}$$

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

$$p_{W|X}(w|x) = p_{Z|X}\left(\tilde{g}(w)|x\right)\left|\det\left(J_{\tilde{g}}(w)\right)\right|, \tag{126}$$

so that the log of the conditional likelihood is

$$\ln p_{W|X}(w|x) = \ln p_{Z|X}\left(\tilde{g}(w)|x\right) + \ln\left|\det\left(J_{\tilde{g}}(w)\right)\right|. \tag{127}$$

Using the model

$$Z|X \sim \mathcal{N}\left(B^\top X, I_d\right), \tag{128}$$

and writing the maximization of the log-likelihood in its population form, we obtain

$$\underset{\tilde{g},B}{\text{maximize}}\; \mathbb{E}\left[-\frac{1}{2}\left\|\tilde{g}(W) - B^\top X\right\|_2^2 + \ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right]. \tag{129}$$

This is equivalent to minimizing

$$\underset{\tilde{g}\in\mathcal{C}_{\text{NF}}, B\in\mathbb{R}^{p\times d}}{\text{minimize}}\; \mathbb{E}\left[\frac{1}{2}\left\|\tilde{g}(W) - B^\top X\right\|_2^2 - \ln\left|\det\left(J_{\tilde{g}}(W)\right)\right|\right], \tag{130}$$

where we have now specified the constraint set for feasible $\tilde{g}$.

G.2 Proof of Theorem 4.2

Lemma G.1 (Gaussian Poincaré Inequality) *For Z Gaussian and isotropic, and $f \in \mathcal{C}_{\text{NF}}$, we have*

$$\Sigma_{f(Z)} \preceq \mathbb{E}\left[J_f(Z)J_f(Z)^\top\right] \tag{131}$$

Proof of Lemma G.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 states that (see Theorem 3.20 of Boucheron et al. (2013)), for differentiable $g : \mathbb{R}^d \to \mathbb{R}$,

$$\text{Var}\left(g(Z)\right) \leq \mathbb{E}\left[\|\nabla g(Z)\|_2^2\right]. \tag{132}$$

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

$$\text{Var}\left(u^\top f(Z)\right) \leq \mathbb{E}\left[\left\|\nabla\left(u^\top f\right)(Z)\right\|_2^2\right]. \tag{133}$$

The left hand side is

$$\begin{aligned}
\text{Var}\left(u^\top f(Z)\right) &= \mathbb{E}\left[\left(u^\top f(Z)\right)^2\right] && (134)\\
&= \mathbb{E}\left[u^\top f(Z)^\top f(Z) u\right] && (135)\\
&= u^\top \Sigma_{f(Z)} u, && (136)
\end{aligned}$$

while the right hand side is

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla\left(u^\top f\right)(Z)\right\|_2^2\right] &= \mathbb{E}\left[\left\|J_f(Z)^\top u\right\|_2^2\right] && (137)\\
&= \mathbb{E}\left[u^\top J_f(Z) J_f(Z)^\top u\right] && (138)\\
&= u^\top \mathbb{E}\left[J_f(Z) J_f(Z)^\top\right] u. && (139)
\end{aligned}$$

Therefore, for all $u \in \mathbb{R}^d$, we have

$$u^\top \Sigma_{f(Z)} u \leq u^\top \mathbb{E}\left[J_f(Z) J_f(Z)^\top\right] u, \tag{140}$$

completing the proof. □

Lemma G.2 *For $\tilde{g} \in \mathcal{C}_{\text{NF}}$ with bi-Lipschitz parameters (m, M), we have*

$$\frac{1}{2}\ln\det\left(\mathbb{E}\left[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top\right]\right) - \frac{1}{4m^4}\mathbb{E}\left[\left\|J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top - \mathbb{E}\left[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top\right]\right\|_F^2\right] \leq \mathbb{E}\left[\ln\left|\det\left(J_{\tilde{g}}(Z)\right)\right|\right] \tag{141}$$

Proof of Lemma G.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 $\mathbb{E}\left[f(V)\right]$ for convex $f : \mathbb{R} \to \mathbb{R}$ and $V \in \mathbb{R}$ a random variable in terms of $\mathbb{E}\left[V\right]$ 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 statement.

For fixed symmetric $H \in \mathbb{R}^{d\times d}$, let

$$\begin{aligned}
f(X) &\equiv \frac{1}{2}\ln\det\left(X\right), && (142)\\
X_t &\equiv X + tH, && (143)\\
h(t) &\equiv f(X_t), && (144)
\end{aligned}$$

for $t \in [0, 1]$. We will use the second order Taylor expansion with integral remainder,

$$h(1) = h(0) + h'(0) + \int_0^1 (1-t)h''(t)dt, \tag{145}$$

and with a lower bound on $h''(t) \geq l$ on $[0,1]$ and integrating $(1-t)$, we have

$$h(1) \geq h(0) + h'(0) + \frac{1}{2}l. \tag{146}$$

Let $A \equiv J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top$ for notational convenience. Choosing $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$,

$$\begin{aligned} h(1) &= f(X_1) && (147)\\ &= \frac{1}{2}\ln\det\left(\mathbb{E}[A] + (A - \mathbb{E}[A])\right) && (148)\\ &= \frac{1}{2}\ln\det\left(J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top\right) && (149)\\ &= \ln\det|J_{\tilde{g}}(Z)|. && (150)\end{aligned}$$

Similarly,

$$\begin{aligned} h(0) &= \frac{1}{2}\ln\det\left(\mathbb{E}[A]\right) && (151)\\ &= \frac{1}{2}\ln\det\left(\mathbb{E}\left[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top\right]\right), && (152)\end{aligned}$$

so that, taking the expectation, we have

$$\mathbb{E}[\ln\det|J_{\tilde{g}}(Z)|] \geq \frac{1}{2}\ln\det\left(\mathbb{E}\left[J_{\tilde{g}}(Z)J_{\tilde{g}}(Z)^\top\right]\right) + \mathbb{E}[h'(0)] + \frac{1}{2}\mathbb{E}[l]. \tag{153}$$

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), $f'(X)[H] = \frac{1}{2}\operatorname{tr}\left(X^{-1}H\right)$, so that

$$h'(t) = \frac{1}{2}\operatorname{tr}\left(X_t^{-1}H\right), \tag{154}$$

and, for $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$, $h'(0) = \frac{1}{2}\operatorname{tr}\left(\mathbb{E}[A]^{-1}(A - \mathbb{E}[A])\right)$, so that

$$\mathbb{E}[h'(0)] = 0. \tag{155}$$

Now we evaluate $h''(t)$. Call $h'(t) \equiv g(X_t)$, so that $g(X) = \frac{1}{2}\operatorname{tr}\left(X^{-1}H\right)$. We again identify $h''(t)$ as a directional derivative, in this case of g in the direction H:

$$h''(t) = g'(X_t)[H]. \tag{156}$$

A short computation shows that, using equation (124) of Petersen et al. (2008),

$$g'(X)[H] = -\frac{1}{2}\operatorname{tr}\left(X^{-1}HX^{-1}H\right), \tag{157}$$

so for $H = A - \mathbb{E}[A]$ and $X = \mathbb{E}[A]$, we have

$$h''(t) = -\frac{1}{2}\operatorname{tr}\left(X_t^{-1}(A - \mathbb{E}[A])X_t^{-1}(A - \mathbb{E}[A])\right). \tag{158}$$

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,

$$\begin{aligned} -h''(t) &= \frac{1}{2}\left\|X_t^{-1/2}(A - \mathbb{E}[A])X_t^{-1/2}\right\|_F^2 && (159)\\ &\leq \frac{1}{2}\left\|X_t^{-1/2}\right\|_2^4 \|A - \mathbb{E}[A]\|_F^2 && (160)\\ &\leq \frac{1}{2m^4}\|A - \mathbb{E}[A]\|_F^2. && (161)\end{aligned}$$

where we have used the property of the Frobenius 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 4.2

Now, we will relax the constraint that $\Sigma_{g(Y)} \succeq aI_d$ for $g \in \mathcal{C}_{\text{NF}}$. We start with Lemma G.1, which says that

$$\Sigma_{g(Y)} \preceq \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]. \tag{162}$$

This along with $\Sigma_{g(Y)} \succeq aI_d$ implies

$$aI_d \preceq \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]. \tag{163}$$

From Lemma G.2, we have

$$\frac{1}{2}\ln\det\left(\mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right) - \frac{1}{4m^4}\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right] \leq \mathbb{E}\left[\ln\left|\det\left(J_g(Y)\right)\right|\right]. \tag{164}$$

Using $aI_d \preceq \mathbb{E}[J_g(Y)J_g(Y)^\top]$, we can lower bound the first term in equation (164):

$$\frac{d}{2}\ln(a) \leq \frac{1}{2}\ln\det\left(\mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right). \tag{165}$$

Thus far we have shown that

$$\Sigma_{g(Y)} \succeq aI_d \implies \mathbb{E}\left[\ln\left|\det\left(J_g(Y)\right)\right|\right] \geq \frac{d}{2}\ln(a) - \frac{1}{4m^4}\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right], \tag{166}$$

where we can think of the second term in equation (164) as the 'variance' of $J_g(Y)J_g(Y)^\top$. We now derive an upper bound for $\mathbb{E}\left[\|J_g(Y)J_g(Y)^\top - \mathbb{E}[J_g(Y)J_g(Y)^\top]\|_F^2\right]$. We denote the matrix function $J_g(y)J_g(y)^\top - \mathbb{E}[J_g(Y)J_g(Y)^\top] \equiv F(y) \in \mathbb{R}^{d\times d}$. We apply the one-dimensional Gaussian Poincare inequality (equation (132)) to each entry of $F(y)$:

$$\begin{aligned}
\text{Var}\left(F_{ij}(Y)\right) &\leq \mathbb{E}\left[\|\nabla F_{ij}(Y)\|_2^2\right] && (167)\\
&= \mathbb{E}\left[\sum_{k=1}^{d}\left(\frac{\partial F_{ij}(Y)}{\partial y_k}\right)^2\right]. && (168)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left\|J_g(Y)J_g(Y)^\top - \mathbb{E}\left[J_g(Y)J_g(Y)^\top\right]\right\|_F^2\right] &= \mathbb{E}\left[\|F(Y)\|_F^2\right] && (169)\\
&= \sum_{i,j=1}^{d}\text{Var}\left(F_{ij}(Y)\right) && (170)\\
&\leq \sum_{i,j,k}^{d}\mathbb{E}\left[\left(\frac{\partial F_{ij}(Y)}{\partial y_k}\right)^2\right] && (171)\\
&= \sum_{k=1}^{d}\mathbb{E}\left[\left\|\frac{\partial F(Y)}{\partial y_k}\right\|_F^2\right] && (172)
\end{aligned}$$

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 et al. (2008), we have

$$\begin{aligned}
\frac{\partial F(y)}{\partial y_k} &= \frac{\partial J_g(y)J_g(y)^\top}{\partial y_k} && (173)\\
&= \frac{\partial J_g(y)}{\partial y_k}J_g(y)^\top + J_g(y)\frac{\partial J_g(y)}{\partial y_k}^\top, && (174)
\end{aligned}$$

so that by the triangle inequality, the fact that $\|A\|_F = \|A^\top\|_F$, and the fact that $\|AB\|_F \le \|A\|_2 \|B\|_F$,

$$\left\| \frac{\partial F(y)}{\partial y_k} \right\|_F \le 2 \left\| J_g(y) \frac{\partial J_g(y)}{\partial y_k}^\top \right\|_F \tag{175}$$

$$\le 2 \|J_g(y)\|_2 \left\| \frac{\partial J_g(y)}{\partial y_k} \right\|_F . \tag{176}$$

Putting these inequalities together, we have

$$\mathbb{E}\left[\left\| J_g(Y) J_g(Y)^\top - \mathbb{E}\left[J_g(Y) J_g(Y)^\top \right] \right\|_F^2 \right] \le \mathbb{E}\left[4 \sum_{k=1}^d \|J_g(Y)\|_2^2 \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 \right] \tag{177}$$

$$= \mathbb{E}\left[4 \|J_g(Y)\|_2^2 \sum_{k=1}^d \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 \right]. \tag{178}$$

Since $g \in \mathcal{C}_{\mathrm{NF}}$, the bound $\|J_g(Y)\|_2^2 \le M^2$ holds. We can rewrite

$$\sum_{k=1}^d \left\| \frac{\partial J_g(Y)}{\partial y_k} \right\|_F^2 = \sum_{i=1}^d \|H_{g_i}(Y)\|_F^2 , \tag{179}$$

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

$$\mathbb{E}\left[\left\| J_g(Y) J_g(Y)^\top - \mathbb{E}\left[J_g(Y) J_g(Y)^\top \right] \right\|_F^2 \right] \le 4M^2 \sum_{i=1}^d \mathbb{E}\left[\|H_{g_i}(Y)\|_F^2 \right]. \tag{180}$$

Denoting our lower bound on the right hand side of equation (166) as $b(a)$, we finally have

$$b(a) \equiv \frac{d}{2} \ln(a) - \frac{M^2}{m^4} \sum_{i=1}^d \mathbb{E}\left[\|H_{g_i}(Y)\|_F^2 \right], \tag{181}$$

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. □

G.3 Proof of Lemma 4.1

Lemma 4.1 follows immediately from the following Lemma G.3, since then

$$b \le \mathbb{E}\left[\ln \left| \det \left(J_{\tilde{g}}(W) \right) \right| \right] \tag{182}$$

$$\le \frac{1}{2} \log \det \left(\Sigma_{\tilde{g}(W)} \right), \tag{183}$$

and therefore,

$$b \le \mathbb{E}\left[\ln \left| \det \left(J_{\tilde{g}}(W) \right) \right| \right] \implies e^{2b} \equiv C \le \det \left(\Sigma_{\tilde{g}(W)} \right). \tag{184}$$

Lemma G.3 *For $\tilde{g} \in \mathcal{C}_{\mathrm{NF}}$ and isotropic Gaussian $W \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}\left[\log \left| \det \left(J_{\tilde{g}}(W) \right) \right| \right] \le \frac{1}{2} \log \det \left(\Sigma_{\tilde{g}(W)} \right). \tag{185}$$

Proof of Lemma G.3
We denote $\tilde{g}$ by g and W by Z for notational ease. We begin with by using[1], for $g \in \mathcal{C}_{\mathrm{NF}}$,

$$h(g(Y)) = h(Y) + \mathbb{E}\left[\log \left| \det \left(J_g(Y) \right) \right| \right]. \tag{186}$$

[1] See https://statproofbook.github.io/P/dent-noninv

Having assumed that Y is an isotropic multivariate Gaussian, we can use Lemma F.1 for the expression of its differential entropy $h(Y) \equiv -\mathbb{E}\left[\log\left(p(Y)\right)\right]$:

$$h(Y) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Y\right) = \frac{d}{2}\log(2\pi e). \tag{187}$$

Lemma F.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

$$h(Z) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_Z\right). \tag{188}$$

Combining these facts together, picking $Z = g(Y)$, we have

$$\begin{aligned} h(g(Y)) &= h(Y) + \mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] && (189)\\ &= \frac{d}{2}\log(2\pi e) + \mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] && (190)\\ &\leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma_{g(Y)}\right). && (191) \end{aligned}$$

This implies that

$$\mathbb{E}\left[\log\left|\det\left(J_g(Y)\right)\right|\right] \leq \frac{1}{2}\log\det\left(\Sigma_{g(Y)}\right), \tag{192}$$

completing the proof. □

G.4 Proof of Theorem 4.3

We provide some machinery needed for the proof of Lemma G.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 possible 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 G.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*

$$\sum_{i=1}^{d} f(y_i) \leq \sum_{i=1}^{d} f(x_i). \tag{193}$$

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 E.1.

Lemma G.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*

$$\sup_{\substack{T \in \mathbb{R}^{p\times d}, H \in \mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z} = \Sigma_{T^\top X} = I_d}} \sum_{i=1}^{d} f\left(\mathbb{E}\left[\eta_i^\top Z \theta_i^\top X\right]\right) = \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{194}$$

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 6 *Taking $f(x) = x^2$, we recover Lemma E.1.*

Proof of Lemma G.5
We begin by using a similar argument to the proof of Lemma E.1. We have, using a change of variables $T = \Sigma_X^{-1/2}\tilde{T}$, $H = \Sigma_Z^{-1/2}\tilde{H}$,

$$\sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d f\left(\mathbb{E}\left[\eta_i^\top Z\theta_i^\top X\right]\right) = \sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d \left(\theta_i^\top \Sigma_{XZ}\eta_i\right)^2 \tag{195}$$

$$= \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d f\left(\tilde{\theta}_i^\top \Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_i\right) \tag{196}$$

$$\le \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\tilde{\theta}_i^\top \Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_j\right) \tag{197}$$

where in the last inequality we have used that $f(x) \ge 0$. Then,

$$\sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\tilde{\theta}_i^\top \Sigma_X^{-1/2}\Sigma_{XZ}\Sigma_Z^{-1/2}\tilde{\eta}_j\right) = \sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\left(\tilde{T}^\top U\Lambda V^\top \tilde{H}\right)_{ij}\right), \tag{198}$$

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 $R = V^\top\tilde{H}$:

$$\sup_{\substack{\tilde{T}\in\mathbb{R}^{p\times d}, \tilde{H}\in\mathbb{R}^{d\times d},\\ \tilde{H}^\top\tilde{H}=\tilde{T}^\top\tilde{T}=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\left(\tilde{T}^\top U\Lambda V^\top \tilde{H}\right)_{ij}\right) = \sup_{\substack{L\in\mathbb{R}^{p\times d}, R\in\mathbb{R}^{d\times d},\\ L^\top L=R^\top R=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\left(L^\top\Lambda R\right)_{ij}\right) \tag{199}$$

$$= \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}} \sup_{\substack{R\in\mathbb{R}^{d\times d},\\ R^\top R=I_d}} \sum_{i=1}^d \sum_{j=1}^d f\left(\left(K^\top R\right)_{ij}\right), \tag{200}$$

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 $t_i \equiv \left(\left(k_i^\top r_j\right)^2\right)_{j=1}^d \in \mathbb{R}^d$. Then, $(K^\top R)_{ij} = k_i^\top r_j = \sqrt{\left(k_i^\top r_j\right)^2} = \sqrt{t_{ij}}$, 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 G.4, to t and w, applied to $f(\sqrt{x})$ which was assumed to be convex and increasing. We obtain, for $i = 1, \dots d$,

$$\sum_{j=1}^d f\left(\left(K^\top R\right)_{ij}\right) = \sum_{j=1}^d f\left(\sqrt{t_{ij}}\right) \tag{201}$$

$$\le \sum_{j=1}^d f\left(\sqrt{w_{ij}}\right) \tag{202}$$

$$= f\left(\sqrt{\|k_i\|_2^2}\right) \tag{203}$$

where we have used that $f(0) = 0$ in the final equality. Therefore,

$$\sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}}\sup_{\substack{R\in\mathbb{R}^{d\times d},\\ R^\top R=I_d}}\sum_{i=1}^{d}\sum_{j=1}^{d} f\left(\left(K^\top R\right)_{ij}\right) \leq \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}}\sum_{i=1}^{d} f\left(\sqrt{\|k_i\|_2^2}\right) \tag{204}$$

$$= \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}}\sum_{i=1}^{d} f\left(\sqrt{l_i^\top \Lambda\Lambda^\top l_i}\right) \tag{205}$$

since $k_i = \Lambda^\top l_i$ where l_i is the ith column of L. Letting $b(L) \equiv \left(l_i^\top \Lambda\Lambda^\top l_i\right)_{i=1}^d$, 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 $\lambda_{1:d}\left(\Lambda\Lambda^\top\right) \equiv \left(\gamma_i^2\right)_{i=1}^d$ 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 $\lambda_{1:d}\left(\Lambda\Lambda^\top\right)$ 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

$$\sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}}\sum_{i=1}^{d} f\left(\sqrt{l_i^\top \Lambda\Lambda^\top l_i}\right) \leq \sup_{\substack{L\in\mathbb{R}^{p\times d}\\ L^\top L=I_d}}\sum_{i=1}^{d} f\left(\sqrt{\lambda_i(B(L))}\right) \tag{206}$$

$$\leq \sum_{i=1}^{d} f\left(\sqrt{\lambda_i\left(\Lambda\Lambda^\top\right)}\right) \tag{207}$$

$$= \sum_{i=1}^{d} f\left(\sqrt{\gamma_i^2}\right) \tag{208}$$

$$= \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{209}$$

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

$$\sup_{\substack{T\in\mathbb{R}^{p\times d}, H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top Z}=\Sigma_{T^\top X}=I_d}}\sum_{i=1}^{d} f\left(\mathbb{E}\left[\eta_i^\top Z\theta_i^\top X\right]\right) \leq \sum_{i=1}^{d} f\left(\gamma_i\right), \tag{210}$$

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 (197) 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 4.3

We denote $\tilde{g}$ by g and W by Y for ease of notation. We begin with

$$\underset{\substack{g\in\mathcal{C}_{\mathrm{NF}}, B\in\mathbb{R}^{p\times d}\\ \mathbb{E}[\ln|\det(J_g(Y))|]\geq b}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right], \tag{211}$$

the equivalent formulation of the normalizing flow problem. From Theorem 4.1, we know that this problem is a constrained problem relative to

$$\underset{\substack{g\in\mathcal{C}_{\mathrm{NF}}, B\in\mathbb{R}^{p\times d}\\ \det\left(\Sigma_{g(Y)}\right)\geq c}}{\text{minimize}} \mathbb{E}\left[\left\|g(Y) - B^\top X\right\|_2^2\right], \tag{212}$$

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 up until equation (81), we have

$$\inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ \det\left(\Sigma_{g(Y)}\right)\geq c,\ g\in\mathcal{C}_{\mathrm{NF}}}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] = \inf_{\substack{g:\mathbb{R}^q\to\mathbb{R}^d\\ \det\left(\Sigma_{g(Y)}\right)\geq c,\ g\in\mathcal{C}_{\mathrm{NF}}}} \operatorname{tr}\left(\Sigma_{g(Y)}D\right), \tag{213}$$

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 F.2, we have

$$\begin{aligned}
\operatorname{tr}\left(\Sigma_{g(Y)}D\right) &= \operatorname{tr}\left(\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}\right) && (214)\\
&\geq d\det\left(\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}\right)^{1/d} && (215)\\
&= d\det\left(\Sigma_{g(Y)}\right)^{1/d}\det\left(D\right)^{1/d}, && (216)
\end{aligned}$$

with equality when $\Sigma_{g(Y)}^{1/2}D\Sigma_{g(Y)}^{1/2}$ is a multiple of I_d. Therefore,

$$\begin{aligned}
\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] &\geq \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} d\det\left(\Sigma_{g(Y)}\right)^{1/d}\det\left(D\right)^{1/d} && (217)\\
&\geq \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} dc^{1/d}\det\left(D\right)^{1/d}. && (218)
\end{aligned}$$

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

$$\begin{aligned}
\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d,\ B\in\mathbb{R}^{p\times d}\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \mathbb{E}\left[\left\|B^\top X - g(Y)\right\|_2^2\right] &= \inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} dc^{1/d}\det\left(D\right)^{1/d} && (219)\\
&= dc^{1/d}\inf_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \left(\prod_{i=1}^d\left(1-\gamma_i^2\right)\right)^{1/d}. && (220)
\end{aligned}$$

Therefore, minimizing the relaxed NF problem (212) is equivalent to maximizing

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sum_{i=1}^d \log\left(\frac{1}{1-\gamma_i^2}\right). \tag{221}$$

Applying Lemma G.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

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sum_{i=1}^d \log\left(\frac{1}{1-\gamma_i^2}\right) = \sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sup_{\substack{T\in\mathbb{R}^{p\times d},H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \sum_{i=1}^d \log\left(\frac{1}{1-\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2}\right), \tag{222}$$

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

$$\sup_{\substack{g:\mathbb{R}^d\to\mathbb{R}^d\\ g\in\mathcal{C}_{\mathrm{NF}},\ \det\left(\Sigma_{g(Y)}\right)\geq c}} \sup_{\substack{T\in\mathbb{R}^{p\times d},H\in\mathbb{R}^{d\times d},\\ \Sigma_{H^\top g(Y)}=\Sigma_{T^\top X}=I_d}} \left(\prod_{i=1}^d \frac{1}{1-\mathbb{E}\left[\eta_i^\top g(Y)\theta_i^\top X\right]^2}\right)^{1/d}, \tag{223}$$

and writing this problem over $h(y)\equiv H^\top g(y)$ completes the proof. □

G.5 Note on the necessity of Gaussian assumptions

We note that the Gaussian assumption on W, the dimension reduced representation of Y, serves to simplify the statements of the results in Section 4.3, but in fact is not necessary. In showing Theorem 4.2 (and Corollary 4.1), we only use the Gaussian assumption to show the Gaussian Poincaré inequality. There are many distributions for which Poincaré inequalities hold, and only change the result by a constant (see for example Section 2.7 of Huang and Tropp (2021)).

The only other usage of the Gaussian assumption is in showing Lemma 4.1 (subsequently Theorem 4.3). However, its proof does not depend on the Gaussian assumption: it suffices to assume that the entropy of W is lower bounded. The same result holds, except that the expression for $c(b)$ has an additional constant.