Provable Affine Identifiability of Nonlinear CCA Under Latent Distributional Priors

Zhiwei Han **Stefan Matthes** **Hao Shen**
fortiss GmbH, Munich, Germany
Technical university of Munich
{han,matthes,shen}@fortiss.org

Abstract

In this work, we establish the sufficient conditions under which nonlinear Canonical Correlation Analysis (CCA) recovers ground-truth latent factors up to an affine transformation. By transporting the analysis from the observation space to the source space, we extend classical statistical results on orthogonal polynomial expansions of bivariate distributions to representation learning, proving affine identifiability under specific distributional priors. We formally demonstrate that whitening is strictly necessary to ensure the boundedness and well-conditioning of the learned mappings. Furthermore, we bridge the gap between theory and practice by proving that ridge-regularized empirical CCA converges to its population counterpart in the finite-sample regime. Finally, our findings provide a rigorous theoretical foundation explaining the empirical success of recent correlation-based non-contrastive learning methods. Experiments on synthetic and rendered image datasets, alongside systematic ablations, validate the predicted recovery behavior and illustrate the failure modes that arise when the assumptions are violated.

1 INTRODUCTION

Identifying explanatory factors from raw sensory data is a fundamental challenge in machine learning. Originated from the problem of blind source separation, independent component analysis (ICA) and its nonlinear variations (Comon, 1994; Hyvärinen and Oja, 2000; Hyvärinen et al., 2019) aim to identify independent latent factors that improve downstream interpretability, controllability, fairness, and sample efficiency (Bengio et al., 2013; Higgins et al., 2017; Locatello et al., 2019). Unfortunately, directly learning these factors from nonlinear mixtures is unidentifiable without additional structural constraints or weak supervision (Hyvärinen and Pajunen, 1999; Locatello et al., 2019, 2020).

Canonical Correlation Analysis (CCA) (Hotelling, 1936) is another classic method, which aims to learn representations by maximizing the cross-correlation between paired views. While nonlinear CCA and its extensions (Bach and Jordan, 2002; Fukumizu et al., 2007; Andrew et al., 2013; Hardoon and Shawe-Taylor, 2011; Ermolov et al., 2021; Zbontar et al., 2021; Bardes et al., 2022) demonstrate empirical performance comparable to strong contrastive learning (Gutmann and Hyvärinen, 2010; van den Oord et al., 2018) baselines, the theoretical understanding of nonlinear CCA identifiability lags behind. Contrastive learning now offers robust block- and factor-level identifiability guarantees by leveraging specific data or model assumptions (von Kügelgen et al., 2021; Yao et al., 2024; Zimmermann et al., 2021; Daunhawer et al., 2023; Matthes et al., 2023; Hyvärinen and Morioka, 2016; Brehmer et al., 2022; Lachapelle et al., 2022). In contrast, existing analyses for nonlinear CCA either assume restricted post-nonlinear mixing (Lyu and Fu, 2020) or only guarantee recovery up to broad, arbitrary invertible transformations (Lyu et al., 2022; Karakasis and Sidiropoulos, 2023). This discrepancy raises a fundamental open question: *Under what conditions can nonlinear CCA provide strictly stronger identifiability guarantees, up to simpler transformations that are comparable to those of contrastive learning?*

This work answers the above question by revisiting the classical theory of bivariate distributions (Lancaster, 1958; Eagleson, 1964). We establish the sufficient conditions under which nonlinear CCA provably recovers

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

ground-truth latent variables up to an affine transformation. Crucially, we formalize the role of representation whitening in ensuring the boundedness and well-conditioning strictly necessary for this identifiability. We empirically validate these theoretical guarantees and systematically ablate our core assumptions to characterize failure modes when these conditions are violated. We summarize our contributions as follows:

1. We characterize the invariance of the population CCA objective and its maximizers under reparameterization via composition with the generators, enabling direct analysis in the latent source space (Proposition 1).

2. We extend the classical theory of bivariate distributions to representation learning and derive affine identifiability results for nonlinear CCA in the population setting across a broad class of latent distributional priors (Theorem 1).

3. We analyze ridge-regularized empirical CCA and show its convergence to the population counterpart, linking the identifiability results to the finite-sample regime (Theorem 2).

4. We assess these theoretical predictions across candidate latent distributions using both a fully controlled synthetic dataset and a rendered 3D image dataset.

5. We study the failure modes of nonlinear CCA under partial violations of the core assumptions through systematic ablations.

2 RELATED WORK

Disentangled Representation Learning. Disentangled representation learning aims to isolate independent explanatory factors of variation, originating from blind source separation (Cardoso, 1998; Comon, 1994). Because disentanglement is provably impossible for i.i.d. nonlinear mixtures without structural assumptions (Locatello et al., 2019; Hyvärinen and Pajunen, 1999), research has focused on exploiting suitable inductive biases to invert nonlinear generative processes. One prevalent approach leverages shared or co-observed variables, such as temporal dynamics (Hyvärinen and Morioka, 2016), nonstationarity (Hyvärinen and Morioka, 2017), auxiliary labels (Hyvärinen et al., 2019; Khemakhem et al., 2020a), weak supervision (Locatello et al., 2020; Shu et al., 2020), and multi-view observations (Daunhawer et al., 2023; Lyu and Fu, 2020; von Kügelgen et al., 2021). A complementary direction restricts the model class or source distribution, enforcing structural constraints like mechanism sparsity (Lachapelle et al., 2022), conformal mappings (Buchholz et al., 2022), or specific latent priors (e.g., exponential families (Hyvärinen et al., 2019), causal structures (Shen et al., 2022), or energy-based models (Khemakhem et al., 2020b)). Our work aligns with the latter by imposing latent distributional priors within a two-view setting.

Nonlinear CCA. Nonlinear CCA extends classical CCA to arbitrary nonlinear mappings via kernels (Fukumizu et al., 2007) or deep neural networks (Andrew et al., 2013; Benton et al., 2019). While these methods excel at maximizing cross-view correlations, achieving strictly identifiable representations was not their original objective. Recent efforts have explored the identifiability of nonlinear CCA-based models (Lyu and Fu, 2020; Lyu et al., 2022; Karakasis and Sidiropoulos, 2023; Sidiropoulos and Sørensen, 2022). However, existing theoretical guarantees of factor-level identifiability rely heavily on restrictive post-nonlinear mixing assumptions (Lyu and Fu, 2020) or only guarantee recovery up to broad equivalence classes, such as arbitrary invertible transformations (Lyu et al., 2022; Karakasis and Sidiropoulos, 2023). In contrast, modern contrastive learning frameworks successfully invert the data-generating process to recover latent variables up to strict affine or permutation ambiguities (Matthes et al., 2023; Zimmermann et al., 2021; Daunhawer et al., 2023). A detailed theoretical comparison between the identifiability frameworks of nonlinear CCA and contrastive learning is provided in Appendix B. An equivalently rigorous, affine identifiability theory for nonlinear CCA remains absent.

We bridge this gap by revisiting classical bivariate distribution theory, which establishes that linear mappings uniquely maximize canonical correlations under Gaussian priors (Lancaster, 1958), a result later generalized to broader distribution families (Eagleson, 1964). Building on these foundations, we extend this classical theory to modern representation learning. By proving the affine identifiability of nonlinear CCA under a family of latent distributional priors, we cement CCA as a theoretically principled alternative to contrastive learning for exact latent recovery, moving beyond mere view alignment.

3 CANONICAL CORRELATION ANALYSIS FOR DISENTANGLED REPRESENTATION

Notations. Throughout this work, scalars are denoted by plain letters, (random) vectors or vector-valued functions by bold lowercase, matrices by bold uppercase symbols, and sets/spaces by calligraphic let-

ters.

3.1 Data Generating Process

We consider a latent pair $(\mathbf{s}, \mathbf{s}') \in \mathcal{S} \times \mathcal{S}$, where $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ and $d_\mathcal{S} \geq 2$. For the main population identifiability theorem, we assume that $(\mathbf{s}, \mathbf{s}')$ is a non-degenerate jointly Gaussian pair and parameterize its dependence through its canonical correlations, which are the natural quantities for CCA. Our primary analysis focuses on the Gaussian case with extensions to other latent distribution families admitting orthogonal polynomial expansions. Formal proofs for these Lancaster-type distributions (Lancaster, 1958), namely the negative binomial, Gamma, Poisson, and hypergeometric cases, are deferred to Appendix F.

Assumption 1 (Non-degenerate Joint Gaussian Latent Pair)**.** *The latent pair $(\mathbf{s}, \mathbf{s}') \in \mathcal{S} \times \mathcal{S}$ is jointly Gaussian with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\mu}_{\mathbf{s}'})$ and block covariance*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{ss}} & \boldsymbol{\Sigma}_{\mathbf{ss}'} \\ \boldsymbol{\Sigma}_{\mathbf{s}'\mathbf{s}} & \boldsymbol{\Sigma}_{\mathbf{s}'\mathbf{s}'} \end{bmatrix}, \quad \textit{where} \quad \boldsymbol{\Sigma}_{\mathbf{ss}} \succ 0,\ \boldsymbol{\Sigma}_{\mathbf{s}'\mathbf{s}'} \succ 0.$$

Let $\mathbf{K} := \boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2} \boldsymbol{\Sigma}_{\mathbf{ss}'} \boldsymbol{\Sigma}_{\mathbf{s}'\mathbf{s}'}^{-1/2}$ be the normalized cross-covariance matrix. We assume the singular values of $\mathbf{K}$, i.e., the canonical correlations between $\mathbf{s}$ and $\mathbf{s}'$, satisfy:

$$1 > \rho_1 \geq \rho_2 \geq \cdots \geq \rho_{d_\mathcal{S}} > 0.$$

The spectral bounds on ρ_i guarantee a full-rank correlated subspace while excluding deterministic equivalence between the views. We then give the practical implication of this assumption and present the corresponding latent model for better structural clarity.

Remark 1 (Canonical Additive Latent Model). Let $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top$ be the singular value decomposition, where $\boldsymbol{\Lambda} = \mathrm{diag}(\rho_1, \ldots, \rho_{d_\mathcal{S}})$. Defining the canonical coordinates

$$\bar{\mathbf{s}} := \mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2} (\mathbf{s} - \boldsymbol{\mu}_\mathbf{s}), \quad \bar{\mathbf{s}}' := \mathbf{V}^\top \boldsymbol{\Sigma}_{\mathbf{s}'\mathbf{s}'}^{-1/2} (\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}),$$

yields $\mathrm{Cov}(\bar{\mathbf{s}}) = \mathrm{Cov}(\bar{\mathbf{s}}') = \mathbf{I}_{d_\mathcal{S}}$ and $\mathrm{Cov}(\bar{\mathbf{s}}, \bar{\mathbf{s}}') = \boldsymbol{\Lambda}$. Consequently, there exist mutually independent, centered Gaussian vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ such that $(\bar{\mathbf{s}}, \bar{\mathbf{s}}') := (\mathbf{a} + \mathbf{c}, \mathbf{b} + \mathbf{c})$ with $\mathrm{Cov}(\mathbf{c}) = \boldsymbol{\Lambda}$ and $\mathrm{Cov}(\mathbf{a}) = \mathrm{Cov}(\mathbf{b}) = \mathbf{I}_{d_\mathcal{S}} - \boldsymbol{\Lambda}$. Thus, an additive shared-private interpretation is inherently guaranteed (Cramér, 1936) in the whitened canonical space, even if the cross-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{ss}'}$ is asymmetric.

To construct the joint distributions in our experiments, we employ the explicit additive generative model as follows:

$$\mathbf{s} = \mathbf{a} + \mathbf{c}, \qquad \mathbf{s}' = \mathbf{b} + \mathbf{c}, \tag{1}$$

where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are mutually independent latent vectors. We emphasize that Equation 1 serves as an experimental device rather than a prerequisite for our population identifiability theorem.

Let $\mathbf{g}\colon \mathcal{S} \to \mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ and $\mathbf{g}'\colon \mathcal{S} \to \mathcal{X}' \subset \mathbb{R}^{d_{\mathcal{X}'}}$ be injective and Borel-measurable mappings that generate the high-dimensional observed pairs $(\mathbf{x}, \mathbf{x}') = (\mathbf{g}(\mathbf{s}), \mathbf{g}'(\mathbf{s}'))$, where $\mathcal{X}, \mathcal{X}'$ denote the observation spaces of two modalities with $d_\mathcal{S} \ll d_\mathcal{X}, d_{\mathcal{X}'}$. Under Assumption 1, our objective is to learn encoders $\mathbf{f}\colon \mathcal{X} \to \mathcal{Z}$ and $\mathbf{f}'\colon \mathcal{X}' \to \mathcal{Z}$ that invert the generative process to recover the true latent sources $(\mathbf{s}, \mathbf{s}')$. To establish baseline affine identifiability, we assume a dimension-matched latent space ($d_\mathcal{Z} = d_\mathcal{S}$), deferring the analysis of mismatched regimes ($d_\mathcal{Z} \neq d_\mathcal{S}$) to the ablation study.

3.2 Canonical Correlation Analysis

Nonlinear CCA aims to learn a pair of nonlinear encoders $\mathbf{f}\colon \mathcal{X} \to \mathcal{Z}$ and $\mathbf{f}'\colon \mathcal{X}' \to \mathcal{Z}$ that maximize the sum of the singular values of their normalized cross-covariance. To ensure well-conditioned representations, we restrict the search space to zero-mean, identity-covariance encoders:

Assumption 2 (Whitened Encoder Classes)**.** *Let $\mathcal{H}_\mathcal{X} \subset L^2(P_\mathbf{x}; \mathbb{R}^{d_\mathcal{Z}})$ be a Hilbert space of square-integrable vector-valued functions. For any base encoder $\mathbf{f} \in \mathcal{H}_\mathcal{X}$ with a positive definite covariance matrix $\boldsymbol{\Sigma}_\mathbf{f} \succ 0$, let $\mathbf{W}_\mathbf{f} \in \mathrm{GL}(d_\mathcal{Z})$ be a whitening matrix satisfying $\mathbf{W}_\mathbf{f} \boldsymbol{\Sigma}_\mathbf{f} \mathbf{W}_\mathbf{f}^\top = \mathbf{I}_{d_\mathcal{Z}}$. We define the whitened encoder class on $\mathcal{X}$ as*

$$\tilde{\mathcal{F}}_\mathcal{X} := \Big\{ \mathbf{W}_\mathbf{f}\big(\mathbf{f}(\mathbf{x}) - \mathbb{E}[\mathbf{f}(\mathbf{x})]\big) : \mathbf{f} \in \mathcal{H}_\mathcal{X},\ \boldsymbol{\Sigma}_\mathbf{f} \succ 0 \Big\},$$

and assume it is closed under orthogonal transformations, i.e., $\mathbf{Q}\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_\mathcal{X}$ for all $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_\mathcal{X}$ and $\mathbf{Q} \in O(d_\mathcal{Z})$. The corresponding class $\tilde{\mathcal{F}}'_{\mathcal{X}'}$ is defined analogously for $\mathcal{H}_{\mathcal{X}'} \subset L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_\mathcal{Z}})$.

CCA Population Objective. Given these feasible encoder classes, nonlinear CCA maximizes the following population objective:

$$\max_{\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_\mathcal{X},\, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}} J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = \sum_{i=1}^{d_\mathcal{Z}} \sigma_i\big(\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x}), \tilde{\mathbf{f}}'(\mathbf{x}'))\big), \tag{2}$$

where $\sigma_i(\cdot)$ denotes the i-th largest singular value and $\mathrm{Cov}(\cdot)$ the population covariance. Due to the orthogonal invariance of the objective and the post-orthogonal closure of the encoder classes, every population maximizer generates an equivalent $O(d_\mathcal{Z}) \times O(d_\mathcal{Z})$-orbit of maximizers. In general, this does not imply that all maximizers lie in a single orbit. Stronger single-orbit

statement is obtained later under the additional assumptions of Theorem 1 as shown in Appendix D.2 and Corollary 4.

3.3 Population Affine Identifiability

In this section, we derive precise conditions under which solving nonlinear CCA leads to provable affine identification of the ground-truth latent factors at the population level. To leverage the distributional priors of the source, we begin by transporting the CCA problem from the observation domain to the underlying source domain. By composing the observation space encoders with the generators, $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$ and $\mathbf{h}' = \mathbf{f}' \circ \mathbf{g}'$, we re-express the learned latent via representation maps $\mathbf{h}, \mathbf{h}'$ and ground-truth latent pair. This change of variables preserves the first- and second-order structure of the CCA problem, so the whitening constraints and the CCA objective are unaffected. We formally justify the invariance of transporting CCA problems from observation spaces to source space.

Proposition 1 (Reparameterization Invariance and Representational Universality)**.** *Let's assume $\mathcal{S}, \mathcal{X}, \mathcal{X}'$ be standard Borel spaces, and let $\mathbf{g}, \mathbf{g}'$ be injective, Borel-measurable mappings. Further assume the base encoders underlying the whitened classes $\tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathcal{F}}'_{\mathcal{X}'}$ in Assumption 2 are dense in $L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ and $L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$, respectively. Define whitened representable latent classes:*

$$\tilde{\mathcal{F}}_{\mathcal{S}} := \{\tilde{\mathbf{f}} \circ \mathbf{g} : \ \tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}\}, \qquad \tilde{\mathcal{F}}'_{\mathcal{S}} := \{\tilde{\mathbf{f}}' \circ \mathbf{g}' : \ \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}\},$$

and the whitened feasible latent classes:

$$\hat{\mathcal{F}}_{\mathcal{S}} := \{\, \hat{\mathbf{h}} \in L^2(P_{\mathbf{s}}; \mathbb{R}^{d_{\mathcal{Z}}}) : \mathbb{E}[\hat{\mathbf{h}}] = 0, \mathrm{Cov}(\hat{\mathbf{h}}) = \mathbf{I}_{d_{\mathcal{Z}}} \,\},$$

$$\hat{\mathcal{F}}'_{\mathcal{S}} := \{\, \hat{\mathbf{h}}' \in L^2(P_{\mathbf{s}'}; \mathbb{R}^{d_{\mathcal{Z}}}) : \mathbb{E}[\hat{\mathbf{h}}'] = 0, \mathrm{Cov}(\hat{\mathbf{h}}') = \mathbf{I}_{d_{\mathcal{Z}}} \,\}.$$

Consider the source-space CCA objective defined over the feasible whitened latent classes on $\mathcal{S} \times \mathcal{S}$: for all $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$ and $\hat{\mathbf{h}}' \in \hat{\mathcal{F}}'_{\mathcal{S}}$,

$$J_{\mathcal{S}}(\hat{\mathbf{h}}, \hat{\mathbf{h}}') = \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\big(\mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{s}),\, \hat{\mathbf{h}}'(\mathbf{s}'))\big), \tag{3}$$

where $\sigma_i(\cdot)$ denotes the i-th largest singular value. Then the following properties hold.

1. ***Objective Preservation.*** *$\forall (\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$, $J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = J_{\mathcal{S}}\big(\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}'\big)$.*
2. ***Maximizer Correspondence.*** *A pair $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ maximizes J over $\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$ if and only if $(\tilde{\mathbf{f}}^* \circ \mathbf{g}, \tilde{\mathbf{f}}'^* \circ \mathbf{g}')$ maximizes $J_{\mathcal{S}}$ over $\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}$.*
3. ***Representation Universality.*** *For any $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$ and $\hat{\mathbf{h}}' \in \hat{\mathcal{F}}'_{\mathcal{S}}$, and any $\epsilon, \epsilon' > 0$, there exist representable maps $\tilde{\mathbf{h}} \in \tilde{\mathcal{F}}_{\mathcal{S}}$ and $\tilde{\mathbf{h}}' \in \tilde{\mathcal{F}}'_{\mathcal{S}}$ such that $\mathbb{E}[\|\hat{\mathbf{h}}(\mathbf{s}) - \tilde{\mathbf{h}}(\mathbf{s})\|^2] < \epsilon^2$ and $\mathbb{E}[\|\hat{\mathbf{h}}'(\mathbf{s}') - \tilde{\mathbf{h}}'(\mathbf{s}')\|^2] < \epsilon'^2$. Consequently, since $J_{\mathcal{S}}$ is continuous under the product L^2 topology,*

$$\sup_{\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}} J = \sup_{\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}} J_{\mathcal{S}} = \sup_{\hat{\mathcal{F}}_{\mathcal{S}} \times \hat{\mathcal{F}}'_{\mathcal{S}}} J_{\mathcal{S}}.$$

Proof sketch. The pushforward identity preserves all moments, so the CCA objective and whitening constraints are invariant under composition with $\mathbf{g}, \mathbf{g}'$. Injectivity on standard Borel spaces yields measurable inverses, allowing any target in $\hat{\mathcal{F}}_{\mathcal{S}}$ or $\hat{\mathcal{F}'}_{\mathcal{S}}$ to be pulled back and approximated by the dense base encoders. See Appendix D.1 for details. □

By Proposition 1, the observation-space CCA problem is isometrically equivalent to the representable source-space problem, and this representable problem attains the same optimal value as the full feasible source-space problem. To establish affine identifiability, it suffices to characterize the maximizers of $J_{\mathcal{S}}$ over $\hat{\mathcal{F}}_{\mathcal{S}} \times \hat{\mathcal{F}}'_{\mathcal{S}}$, any observation-space maximizer induces one such source-space maximizer.

Assumption 3 (First-Order Canonical Dominance)**.** *Under Assumption 1, we further assume that the smallest canonical correlation $\rho_{d_{\mathcal{S}}}$ is strictly larger than the product of any two canonical correlations, i.e., $\rho_{d_{\mathcal{S}}} > \rho_i \rho_j, \forall 1 \le i \le j \le d_{\mathcal{S}}$. Equivalently, $\rho_{d_{\mathcal{S}}} > \rho_1^2$, where ρ_1 is the largest canonical correlation.*

We then state our central population identifiability result as follows:

Theorem 1 (Population Affine Identifiability)**.** *Assume the conditions of Proposition 1, Assumption 1, and Assumption 3, and let $d_{\mathcal{Z}} = d_{\mathcal{S}}$. For the whitened encoder classes in Assumption 2, any population maximizer pair $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ of Equation 2 identifies the marginally whitened latent factors up to orthogonal transformations. Specifically, there exist orthogonal matrices $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$ such that*

$$\tilde{\mathbf{h}}^*(\mathbf{s}) := (\tilde{\mathbf{f}}^* \circ \mathbf{g})(\mathbf{s}) = \mathbf{Q}\, \boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2} (\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}),$$

$$\tilde{\mathbf{h}}'^*(\mathbf{s}') := (\tilde{\mathbf{f}}'^* \circ \mathbf{g}')(\mathbf{s}') = \mathbf{Q}'\, \boldsymbol{\Sigma}_{\mathbf{s's'}}^{-1/2} (\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}).$$

Equivalently, the corresponding unwhitened representation maps recover $\mathbf{s}$ and $\mathbf{s}'$ up to invertible affine transformations.

Proof sketch. By Proposition 1, the CCA optimization can be transported to the source space, allowing us to characterize the optimal representation mappings directly in terms of the ground-truth latents. For joint Gaussian priors, Mehler-Hermite expansion (Mehler,

1866) effectively diagonalizes the CCA objective, revealing that the canonical correlations of the learned features are bounded by combinations of the true latent canonical correlations. under Assumption 3, the CCA objective is uniquely maximized by selecting only the first-order Hermite polynomials and all higher-order nonlinearities vanish. A complete proof is deferred to Appendix D.3. □

At the population optimum under Gaussian priors with $d_{\mathcal{Z}} = d_{\mathcal{S}}$, the learned encoders successfully invert the data-generating process up to an affine ambiguity, i.e., centering, scaling, and an orthogonal rotation. In the proof of Theorem 1, we formally prove that explicit whitening or unit-variance regularization, as employed by non-contrastive methods like Barlow Twins, is strictly necessary to ensure the boundedness and well-conditioning of the learned mappings. This provides a partial but rigorous explanation for the effectiveness of whitening-based non-contrastive methods. While our theoretical guarantees require matched dimensions ($d_{\mathcal{S}} = d_{\mathcal{Z}}$), we extend our analysis to mismatched regimes, i.e., $d_{\mathcal{S}} \neq d_{\mathcal{Z}}$ and finite-sample convergence in the subsequent empirical sections. A detailed discussion of the role of canonical-correlation separation and the consequences of relaxing this condition is provided in Appendix C.

3.4 Empirical Estimation and Statistical Consistency

While population identifiability characterizes the infinite-sample optimum of Equation 2, practical representation learning requires statistical consistency with finite data. To bridge this gap, we introduce a regularized empirical CCA estimator and analyze its statistical consistency in the finite-sample regime. We prove that as the regularization parameter decays asymptotically, the empirical maximizers strictly converge to the population solution, formally linking theoretical identifiability to finite-sample learnability.

To construct this empirical objective, we parameterize the nonlinear encoders $\mathbf{f}_{\boldsymbol{\theta}} \colon \mathbb{R}^{d_{\mathcal{X}}} \to \mathbb{R}^{d_{\mathcal{Z}}}$ and $\mathbf{f}'_{\boldsymbol{\theta}'} \colon \mathbb{R}^{d_{\mathcal{X}'}} \to \mathbb{R}^{d_{\mathcal{Z}}}$ using neural networks with sufficient capacity to act as universal function approximators. Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ denote their respective weights. The regularized finite-sample objective is formulated as follows:

Empirical Objective. Let $\mathbf{X} \in \mathbb{R}^{n \times d_{\mathcal{X}}}$ and $\mathbf{X}' \in \mathbb{R}^{n \times d_{\mathcal{X}'}}$ denote observation matrices comprising n mutually independent samples drawn from $P_{\mathbf{x}\mathbf{x}'}$. We define the learned representations as $\mathbf{Z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{X})$ and $\mathbf{Z}' = \mathbf{f}'_{\boldsymbol{\theta}'}(\mathbf{X}') \in \mathbb{R}^{n \times d_{\mathcal{Z}}}$. Let $\mathbf{Z}_c$ and $\mathbf{Z}'_c$ denote their respective mean-centered counterparts. The empirical auto- and cross-covariance matrices are given by $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}} := \frac{1}{n-1}\mathbf{Z}_c^\top \mathbf{Z}_c$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'} := \frac{1}{n-1}\mathbf{Z}_c'^\top \mathbf{Z}'_c$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}'} := \frac{1}{n-1}\mathbf{Z}_c^\top \mathbf{Z}'_c$. To ensure well-conditioning in the finite-sample regime, we apply a ridge penalty $\epsilon > 0$ to construct the regularized whitening matrices:

$$\hat{\mathbf{W}}_{\mathbf{z}} := (\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}} + \epsilon \mathbf{I})^{-1/2}, \quad \hat{\mathbf{W}}_{\mathbf{z}'} := (\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'} + \epsilon \mathbf{I})^{-1/2}.$$

The empirical CCA objective maximizes the nuclear norm, i.e., the sum of singular values, of the normalized cross-covariance $\hat{\mathbf{K}} := \hat{\mathbf{W}}_{\mathbf{z}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}'} \hat{\mathbf{W}}_{\mathbf{z}'}$, formulated via the variational trace characterization:

$$\begin{aligned} \max_{\boldsymbol{\theta}, \boldsymbol{\theta}'} \hat{J}(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \max_{\boldsymbol{\theta}, \boldsymbol{\theta}'} \sum_{k=1}^{d_{\mathcal{Z}}} \sigma_k(\hat{\mathbf{K}}) \\ &= \max_{\substack{\boldsymbol{\theta}, \boldsymbol{\theta}' \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I} \\ \mathbf{V}^\top \mathbf{V} = \mathbf{I}}} \operatorname{tr}\left(\mathbf{U}^\top \hat{\mathbf{K}} \mathbf{V}\right). \end{aligned} \tag{4}$$

The following theorem establishes our central statistical consistency result. By demonstrating the uniform convergence of the empirical covariance operators and the stability of both the regularized whitening and spectral mappings, we prove the asymptotic consistency of the empirical maximizers.

Theorem 2 (Consistency of Empirical Maximizers)**.** *Assume the conditions of Theorem 1 hold with $d_{\mathcal{Z}} = d_{\mathcal{S}}$. Let $(\tilde{\mathbf{f}}_{\boldsymbol{\theta}}, \tilde{\mathbf{f}}'_{\boldsymbol{\theta}'})$ denote a pair of empirical whitened encoders and consider the regularized empirical objective $\hat{J}$ in Equation 4 with ridge penalty $\epsilon \to 0^+$. Suppose the following standard consistency assumptions hold as n goes to infinity:*

A1. ***Realizability and Orbit Uniqueness.*** *The population maximizer of J in Equation 2 exists in the whitened encoder classes in Assumption 2 and is unique up to postorthogonal transformations.*

A2. ***Capacity and Nondegeneracy.*** *The second moments of $(\tilde{\mathbf{f}}_{\boldsymbol{\theta}}, \tilde{\mathbf{f}}'_{\boldsymbol{\theta}'})$ are uniformly bounded, and the empirical covariances $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'}$ are uniformly symmetric positive definite.*

A3. ***Uniform Convergence and Ridge Schedule.*** *The empirical auto- and cross-covariances converge uniformly at a rate faster than the ridge penalty, i.e., $\|\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}} - \boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}}\| = o_p(\epsilon), \|\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'} - \boldsymbol{\Sigma}_{\mathbf{z}'\mathbf{z}'}\| = o_p(\epsilon), \|\hat{\boldsymbol{\Sigma}}_{\mathbf{z}\mathbf{z}'} - \boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}'}\| = o_p(\epsilon)$.*

A4. ***Approximate Maximization.*** *The empirical whitened encoders $(\tilde{\mathbf{f}}_{\boldsymbol{\theta}}, \tilde{\mathbf{f}}'_{\boldsymbol{\theta}'})$ are δ_n-maximizers of $\hat{J}$ for some sequence $\delta_n \to 0^+$, i.e.,*

$$\hat{J}(\tilde{\mathbf{f}}_{\boldsymbol{\theta}}, \tilde{\mathbf{f}}'_{\boldsymbol{\theta}'}) \geq \sup_{\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}}} \hat{J}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') - \delta_n, \quad \delta_n \to 0.$$

Then, as $n \to \infty$, the following convergence properties hold in probability:

1. ***Objective Consistency.***

$$\sup_{\tilde{\mathbf{f}}\in\tilde{\mathcal{F}}_{\mathcal{X}},\tilde{\mathbf{f}}'\in\tilde{\mathcal{F}}'_{\mathcal{X}'}} \left|\hat{J}(\tilde{\mathbf{f}},\tilde{\mathbf{f}}') - J(\tilde{\mathbf{f}},\tilde{\mathbf{f}}')\right| \xrightarrow{\mathbb{P}} 0.$$

2. ***Estimator Consistency.*** *For any population maximizer $(\tilde{\mathbf{f}}^\star,\tilde{\mathbf{f}}'^\star)$, there exist orthogonal matrices $\mathbf{Q},\mathbf{Q}'\in O(d_{\mathcal{Z}})$ such that the optimal empirical encoders $(\tilde{\mathbf{f}}^*_{\hat{\boldsymbol{\theta}}^*},\tilde{\mathbf{f}}'^*_{\hat{\boldsymbol{\theta}}'^*})$ satisfy:*

$$\left\|\mathbf{Q}\tilde{\mathbf{f}}^*_{\hat{\boldsymbol{\theta}}^*} - \tilde{\mathbf{f}}^\star\right\|_{L^2(P_{\mathbf{x}})} + \left\|\mathbf{Q}'\tilde{\mathbf{f}}'^*_{\hat{\boldsymbol{\theta}}'^*} - \tilde{\mathbf{f}}'^\star\right\|_{L^2(P_{\mathbf{x}'})} \xrightarrow{\mathbb{P}} 0.$$

3. ***Latent Recovery.*** *By Proposition 1 and Theorem 1, the composed encoders recover the ground-truth latents up to orthogonal transformations:*

$$\inf_{\mathbf{Q},\mathbf{Q}'\in O(d_{\mathcal{Z}})} \Big(\left\|\mathbf{Q}\left(\tilde{\mathbf{f}}^*_{\hat{\boldsymbol{\theta}}^*}\circ\mathbf{g}\right) - \tilde{\mathbf{h}}^*\right\|_{L^2(P_{\mathbf{s}})} + \left\|\mathbf{Q}'\left(\tilde{\mathbf{f}}'^*_{\hat{\boldsymbol{\theta}}'^*}\circ\mathbf{g}'\right) - \tilde{\mathbf{h}}'^*\right\|_{L^2(P_{\mathbf{s}'})}\Big) \xrightarrow{\mathbb{P}} 0,$$

Proof sketch. The complete proof, deferred to Appendix D.5, proceeds in four steps. First, uniform convergence of the empirical covariances establishes the asymptotic stability of the ridge-regularized whitening operators. Second, this stability ensures the empirical normalized cross-covariance $\hat{\mathbf{K}}$ converges uniformly to its population counterpart $\mathbf{K}$ in operator norm; the continuity of the CCA objective then guarantees uniform convergence of the empirical objective (Theorem 2.1). Third, given the uniqueness of the population maximizer up to orthogonal transformations, standard M-estimation arguments on the quotient space yield estimator consistency (Theorem 2.2). Finally, coupling this consistency with reparameterization invariance (Proposition 1) and affine identifiability (Theorem 1) proves that the learned encoders recover the true latents up to an orthogonal ambiguity (Theorem 2.3). □

An interpretation of Assumptions A3–A4 and their role in the consistency argument is provided in Appendix C.

4 EXPERIMENTS

In this section, we empirically validate our theoretical guarantees for nonlinear CCA. We evaluate affine identifiability across diverse latent distributions, verify reparameterization invariance and finite-sample consistency, and systematically ablate our core assumptions. We adopt the experimental setup from previous works (Zimmermann et al., 2021; Matthes et al., 2023) on disentangled representations and extend it to a multi-modal setting. The code is available at `https://github.com/ZhiweiHan9277/AISTATS2026_Provable_Affine_Identifiability`.

4.1 Experimental Setup

Datasets. We evaluate our theoretical claims on a synthetic dataset and the 3DIdent image dataset (Zimmermann et al., 2021). For the synthetic setup, we adapt the online sampling mechanism from prior work (Zimmermann et al., 2021; Matthes et al., 2023), generalizing their single-generator framework to two independent generators to simulate multi-modal observations. 3DIdent comprises images of a 3D object rendered from 11 latent factors, e.g., position, rotation and illumination, yielding (image, latent) tuples.

To map our continuous generative process to the finite 3DIdent dataset, we employ a nearest-neighbor matching strategy. We first independently sample $\mathbf{a},\mathbf{b},\mathbf{c}$ to construct continuous latent pairs $(\mathbf{s},\mathbf{s}')$ as in Equation 1. We then retrieve the corresponding rendered images $(\mathbf{x},\mathbf{x}')$ by finding the Euclidean nearest neighbors of $\mathbf{s}$ and $\mathbf{s}'$ in the 3DIdent ground-truth latent space. To emulate distinct multi-view modalities, we retain the original color image for $\mathbf{x}$ and use a grayscale conversion for $\mathbf{x}'$, preserving shared generative factors while marginalizing out color.

Baselines. We benchmark the CCA-like objectives and strong non-contrastive SSL baselines to test affine identifiability under a unified encoder architecture and a family of distributions aligned with our theory.

- **SwAV** (Caron et al., 2020) learns view-invariant features via online clustering with prototype assignments.
- **BarlowTwins** (Zbontar et al., 2021) enforces invariance and minimizes redundancy through cross-correlation diagonalization.
- **VICReg** (Bardes et al., 2022) encourages invariance while regularizing variance and covariance to decorrelate neural representations.
- **W-MSE** (Ermolov et al., 2021) aligns whitened multi-view neural representations by minimizing mean-squared error.
- **DGCCA** (Benton et al., 2019) is a nonlinear extension of generalized CCA, optimizing a shared latent that reconstructs all view projections.
- **rDGCCA** (Karakasis and Sidiropoulos, 2023) improves DGCCA by extracting shared representations under conditionally independent private latents, offering theoretical recovery guarantees.
- **DeepCCA** (Andrew et al., 2013) maximizes the CCA objective with whitened multi-view representations.

Methods	Gaussian		Negative Binomial		Gamma		Poisson		Hypergeometric	
	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$
SwAV	32.96 ±4.05	37.88 ±0.45	33.17 ±3.67	28.85 ±2.62	29.37 ±4.85	28.38 ±2.74	36.43 ±0.88	35.23 ±1.69	12.88 ±8.25	12.84 ±8.66
BarlowTwins	79.44±0.26	79.51 ±0.3	67.22 ±5.67	68.33 ±4.11	61.99 ±2.4	61.98 ±2.43	55.77 ±5.82	55.29 ±4.59	25.47 ±0.49	23.9 ±1.53
VICReg	14.54 ±1.14	13.05 ±1.1	13.83 ±1.82	12.82 ±1.08	12.27 ±0.56	13.61 ±0.82	11.88 ±0.48	11.37 ±0.82	18.94 ±2.37	16.7 ±0.64
W-MSE	99.17 ±0.21	99.4 ±0.04	**99.07 ±0.04**	99.23 ±0.04	99.11 ±0.15	99.34 ±0.05	**99.18 ±0.08**	**99.4 ±0.04**	**98.97 ±0.04**	**99.02 ±0.32**
DGCCA	21.9 ±0.79	18.17 ±1.4	21.45 ±1.52	23.9 ±0.86	22.0 ±0.73	21.17 ±1.65	22.16 ±1.43	18.94 ±1.2	23.81 ±1.09	23.95 ±1.25
rDGCCA	90.57 ±2.89	82.48 ±3.7	90.62 ±3.33	78.51 ±2.73	90.35 ±3.23	83.35 ±4.16	92.97 ±2.26	83.33 ±1.99	91.74 ±3.04	86.93 ±2.89
DeepCCA	**99.4 ±0.05**	**99.41 ±0.03**	99.05 ±0.04	**99.23 ±0.03**	**99.12 ±0.14**	**99.35 ±0.03**	99.18 ±0.09	99.36 ±0.03	98.91 ±0.15	98.37 ±0.1

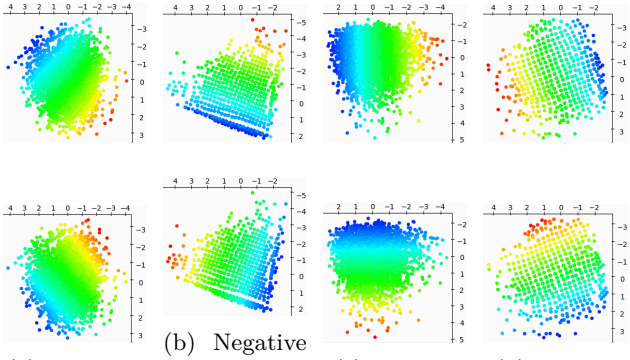
Table 1: Comparison of the coefficient of determination $R^2\uparrow(\%)$ of both encoders $\mathbf{f}, \mathbf{f}'$ on synthetic data ($d_{\mathcal{S}} = d_{\mathcal{Z}} = 10$).

We exclude Kernel CCA due to poor high-dimensional scalability. Among the evaluated baselines, only DeepCCA faithfully enforces exact whitening and maximizes canonical correlations. We therefore use its converged solution as a proxy for population CCA with the data generated from online sampling of the underlying distributions. All baselines are implemented using their official repositories when available, otherwise we reimplement by following their paper specifications.

Metrics. Following prior work (Hyvärinen and Morioka, 2016; Zimmermann et al., 2021; Matthes et al., 2023), we assess affine recoverability by regressing ground-truth latents onto learned latents and reporting the coefficient of determination (R^2) averaged across dimensions. For settings with partial distributional violations or dimensional mismatch, we evaluate subspace error using the mean (PA_{mean}) and maximum ($PA_{\max}$) principal angles between the learned span and the ground-truth canonical subspace, in line with subspace identification studies (Ma and Li, 2020; Gao et al., 2017; Cai and Zhang, 2018). To quantify the alignment between two encoders up to orthogonal transformations, we measure the orbit distance defined as $\min_{\mathbf{Q},\mathbf{Q}'\in O(d_{\mathcal{Z}})} \frac{1}{n}\left(\|\hat{\mathbf{Z}} - \mathbf{Z}\mathbf{Q}\|_F^2 + \|\hat{\mathbf{Z}}' - \mathbf{Z}'\mathbf{Q}'\|_F^2\right)$, where n is the batch size, $\mathbf{Z}, \mathbf{Z}'$ and $\hat{\mathbf{Z}}, \hat{\mathbf{Z}}'$ are the learned latents of two sets of encoders.

Methods	$R^2\uparrow(\%)$		$PA_{\text{mean}}\downarrow(^\circ)$		$PA_{\max}\downarrow(^\circ)$	
	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$
SwAV	22.45±5.59	26.21±5.99	67.22±3.64	65.94±4.26	89.28±0.58	89.28±0.41
BarlowTwins	85.20±0.38	86.07±0.48	17.39±0.61	16.76±0.67	73.69±3.86	70.17±4.29
VICReg	22.34±1.37	24.45±1.39	66.13±1.46	65.40±1.31	89.07±0.75	89.74±0.95
W-MSE	97.40±0.80	97.00±0.87	8.82±0.31	9.65±1.06	13.98±1.97	13.00±2.01
DGCCA	0.37±0.77	0.47±1.36	87.70±1.21	86.72±1.73	89.97±0.21	89.42±0.28
rDGCCA	89.13±0.89	79.98±0.70	11.39±0.13	19.51±0.33	57.78±5.61	82.55±7.12
DeepCCA	**97.80±0.23**	**97.79±0.35**	**8.33±0.46**	**8.33±0.54**	**11.11±0.91**	**10.76±1.04**

Table 2: Comparison of R^2, the mean principal angle PA_{mean}, and the maximal principal angle $PA_{\max}$ of both encoders on 3DIdent ($d_{\mathcal{S}} = d_{\mathcal{Z}} = 7$).



(a) Gaussian (b) Negative Binomial (c) Gamma (d) Poisson

Figure 1: Whitened latent learned representations by DeepCCA on synthetic data ($d_{\mathcal{S}} = d_{\mathcal{Z}} = 2$). Color gradients in different rows illustrate the variations along a single coordinate in $\mathcal{S}$.

4.2 Validation of Theoretical Findings

Affine Identifiability. For the synthetic experiments ($d_{\mathcal{S}} = 10$), latent pairs $(\mathbf{s}, \mathbf{s}')$ are generated via Equation 1 by sampling $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ independently from five candidate distributions, and observations $(\mathbf{x}, \mathbf{x}')$ are produced by independent decoders $\mathbf{g}$ and $\mathbf{g}'$. The R^2 scores in Table 1 support the predicted affine identifiability under this additive model: DeepCCA and W-MSE consistently achieve near-perfect recovery across all distributions. rDGCCA forms a weaker second tier with noticeable asymmetry across the two views, while Barlow Twins deteriorates substantially on the hypergeometric setting. SwAV, VICReg, and DGCCA remain far from exact recovery. Detailed distribution settings and subspace recovery errors are deferred to Appendix H.

On 3DIdent (Table 2), DeepCCA again performs best, with W-MSE close behind in both R^2 and principal-angle errors. rDGCCA and Barlow Twins provide partial recovery but remain clearly below these two methods, whereas DGCCA fails outright and SwAV and VICReg also perform poorly. Because 3DIdent is constructed by nearest-neighbor matching in latent space rather than from exact generative pairs, exact recovery is inherently harder. We therefore also report subspace recovery errors as a more robust measure of representation alignment under finite-sample and matching noise.

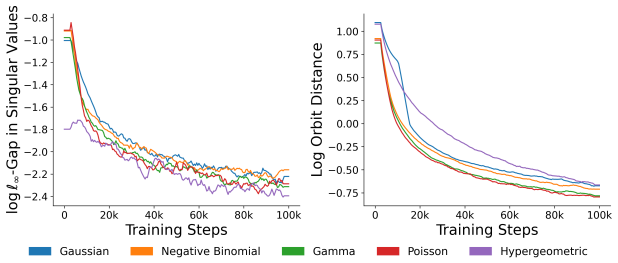


Figure 2: The ℓ_∞-gap in singular values and log orbit distance over training steps in Gaussian case.

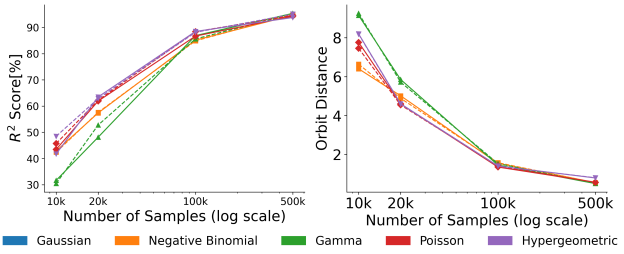


Figure 3: $R^2 \uparrow$ and Orbit distance over changing sample size in fixed dataset setup.

Figure 1 visualizes the 2D whitened representations learned by DeepCCA across four distributions, omitting Hypergeometric due to representation collapse. The source coordinates manifest as smoothly varying, mutually orthogonal color gradients. As theoretically expected, these gradients are rotated relative to the canonical axes, empirically validating that whitening identifies the true subspace only up to an orthogonal transformation.

Reparameterization Invariance. This experiment is not a downstream-performance benchmark and therefore does not admit a meaningful chance-level baseline. The relevant reference point is exact agreement, i.e., zero singular-value gap and zero orbit distance before the logarithmic transform. Our goal is only to test Proposition 1, namely whether CCA trained in observation space and in source space converges to the same $O(d_{\mathcal{Z}})$-orbit and canonical spectrum. To this end, we independently optimize two CCA encoders under identical synthetic configurations: one operating on the observation space and another directly on the latent source space. Figure 2 tracks the l_∞-norm singular-value gap and the orbit distance between the two learned encoders for a given view. Smaller values signify closer alignment, exact invariance implies a zero singular value gap and zero orbit distance prior to logarithmic transformation. Across five random seeds, both metrics rapidly converge and remain consistently small throughout training. The median orbit distance remains below 0.2, while the l_∞ singular-value gap stays below 0.01. Crucially, this demonstrates that the observation-space maximizer of the CCA objective tightly aligns with its source-space counterpart, providing strong empirical corroboration for Proposition 1.

Finite-sample Consistency. We evaluated the finite-sample consistency of CCA by varying the sample size n while keeping the remaining synthetic setup fixed, utilizing $\epsilon = 0.01 \cdot n^{-1/4}$ balancing bias and stability. We fix the source dimension to $d_{\mathcal{S}} = 10$ and report both the R^2 score and the orbit distance between the empirical and population CCA maximizers for a single view in Figure 3. As the sample size increases, the R^2 improves steadily across all distributions, exceeding 90% at roughly $4 \cdot 10^5$ samples, while orbit distance rapidly decays to near zero, indicating convergence of the learned representation to the population solution up to an orthogonal transformation. These results demonstrate that empirical CCA recovers the affine structure of the latent sources as the sample size grows as stated in Theorem 2.

4.3 Ablation Study

We provide an ablation study to isolate the effects of source dimensionality, first-order canonical dominance and latent–dimension mismatch on affine identifiability. Following the synthetic experimental setup, we vary only one parameter at a time while keeping all others fixed.

Source Dimension. CCA consistently achieves affine identifiability in low-dimensional settings ($d_{\mathcal{S}} = 2, 10, 20$), as evidenced by small principal angles between the recovered and ground-truth subspaces and high R^2 scores, as shown in the left panel of Figure 4. As the source dimension increases ($d_{\mathcal{S}} = 30, 40$), CCA continues to recover a substantial portion of the correlated subspace, but the recovery quality gradually deteriorates. Both the mean and variance of the principal angles increase, indicating reduced estimation stability in higher dimensions, as illustrated in the right panel of Figure 4.

First-order Canonical Dominance. We introduce the *first-order canonical dominance ratio*, $\rho_{d_{\mathcal{S}}}/\rho_1^2$, to characterize the weakest canonical correlation's strength relative to the squared strongest correlation. Assuming source canonical correlations are uniformly sampled from $[\rho_{d_{\mathcal{S}}}, \rho_1]$, Figure 5 illustrates the impact of this ratio on affine identifiability. Increasing the dominance ratio consistently improves R^2 scores across latent dimensions, indicating enhanced source recoverability. Notably, identifiability exhibits a threshold

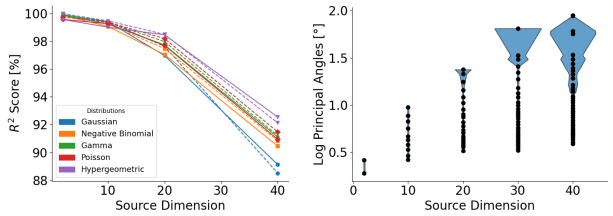


Figure 4: Ablation over the source dimension $d_{\mathcal{S}}$ ($d_{\mathcal{S}} = d_{\mathcal{Z}}$). Left: $R^2 \uparrow$. Solid lines denote encoder $\mathbf{f}$ and dashed lines encoder $\mathbf{f}'$. Right: log principal angles of $\mathbf{f}$ in the Gaussian case. Black dots denote principal angles and shaded region indicates the log-standard deviation.

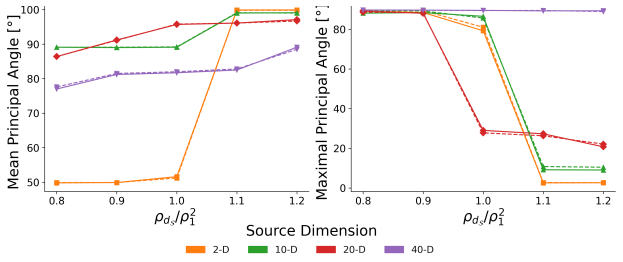


Figure 5: Ablation over the first order canonical ratio $\rho_{d_{\mathcal{S}}}/\rho_1^2$ ($d_{\mathcal{S}} = d_{\mathcal{Z}}$). Left: $R^2 \uparrow$. Right: $PA_{max} \downarrow$. Colors denote different source dimensions.

effect at 1. For $\rho_{d_{\mathcal{S}}}/\rho_1^2 < 1$, elevated maximum principal angles ($PA_{\max}$) reveal that at least one canonical direction remains poorly identifiable. Once the ratio exceeds 1, $PA_{\max}$ rapidly converges to zero, confirming full recovery of all canonical directions and strict alignment with the ground-truth source subspace.

Encoder	Gaussian	Negative Binomial	Gamma	Poisson	Hypergeometric
$\mathbf{f}$	98.93±0.04	98.9±0.03	99.09±0.03	99.12±0.03	98.56±0.05
$\mathbf{f}'$	98.94±0.05	99.17±0.03	99.33±0.02	99.31±0.017	98.56±0.03

Table 3: $R^2 \uparrow$ [%] for both encoders $\mathbf{f}$ and $\mathbf{f}'$ in the overcomplete setup ($d_{\mathcal{S}} = 10,\ d_{\mathcal{Z}} = 13$).

Dimension Mismatch. We investigate both under- and overcomplete learning regimes. In the overcomplete case ($d_{\mathcal{S}} < d_{\mathcal{Z}}$), consistently high R^2 scores (Table 3) demonstrate that the true source subspace is recoverable despite redundant latent coordinates, though the stability of these redundancies remains unclear. Conversely, in the undercomplete regime ($d_{\mathcal{S}} > d_{\mathcal{Z}}$), insufficient capacity precludes full $d_{\mathcal{S}}$-dimensional recovery. Although four canonical directions are recovered with minimal principal angle deviation (Figure 6), affine identifiability cannot be broadly guaranteed.

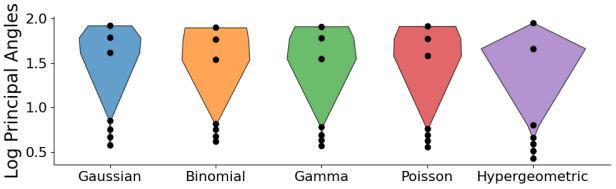


Figure 6: Log principal angles of encoder $\mathbf{f}$ in the undercomplete setup ($d_{\mathcal{S}} = 10,\ d_{\mathcal{Z}} = 7$). Black dots denote principal angles and the shaded region indicates the log-standard deviation.

Discussion on Whitening. Whitening, enforcing unit variance and decorrelation across latent dimensions for self-supervised learning w (Hua et al., 2021; Weng et al., 2022, 2024), is critical for affine identifiability as it ensures well-conditioned representations and stabilizes canonical direction estimation. Empirically, methods explicitly enforcing whitening, e.g., DeepCCA and W-MSE, achieve superior affine identifiability. Furthermore, even relaxed decorrelation constraints, as in Barlow Twins, yield substantial improvements over unregularized baselines. These empirical results corroborate the theoretical analysis detailed in Appendix E.

5 CONCLUSION

We characterized the sufficient conditions under which nonlinear CCA achieves affine identifiability, advancing the theoretical understanding of correlation-based non-contrastive learning. Crucially, our framework provides a principled explanation for the recent remarkable empirical performance of non-contrastive methods: it proves that explicit whitening and unit-variance regularizers, e.g., in Barlow Twins, natively drive the recovery of ground-truth latent factors under (near-)Gaussian priors. Although the affine guarantee is naturally weaker than the permutation-level identifiability of contrastive methods, CCA provides practical advantages by circumventing negative sampling and exhibiting robustness to small batch sizes. Finally, our proof of finite-sample consistency guarantees that scaling dataset size directly improves empirical source recovery and downstream disentanglement.

References

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Bardes, A., Ponce, J., and LeCun, Y. (2022). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., and Arora, R. (2019). Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 1–6. Association for Computational Linguistics.

Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. S. (2022). Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331.

Buchholz, S., Besserve, M., and Schölkopf, B. (2022). Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 16946–16961.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89.

Cardoso, J.-F. (1998). Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Cramér, H. (1936). Über eine eigenschaft der normalen verteilungsfunktion. *Mathematische Zeitschrift*, 41:405–414.

Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., and Vogt, J. E. (2023). Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*.

Eagleson, G. K. (1964). Polynomial expansions of bivariate distributions. *The Annals of Mathematical Statistics*, 35(3):1208–1215.

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. (2021). Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(14):361–383.

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

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, pages 297–304. PMLR.

Hardoon, D. R. and Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

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

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., and Zhao, H. (2021). On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608.

Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, pages 3772–3780.

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, pages 460–469. PMLR.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Hyvärinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR.

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

Khemakhem, I., Kingma, D., Monti, R., and Hyvärinen, A. (2020a). Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Khemakhem, I., Monti, R., Kingma, D., and Hyvärinen, A. (2020b). ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 33, pages 12768–12778.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. (2022). Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR.

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

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR.

Lyu, Q. and Fu, X. (2020). Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Transactions on Signal Processing*, 68:2697–2712.

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

Ma, Z. and Li, X. (2020). Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. *Bernoulli*, 26(1):432–470.

Matthes, S., Han, Z., and Shen, H. (2023). Towards a unified framework of contrastive learning for disentangled representations. In *Advances in Neural Information Processing Systems*, volume 36, pages 67459–67470.

Mehler, F. G. (1866). Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung. *Journal für die reine und angewandte Mathematik*, 66:161–176.

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

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. (2020). Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*.

Sidiropoulos, N. D. and Sørensen, M. (2022). Canonical identification of autoregressive nonlinear systems. In *56th Asilomar Conference on Signals, Systems, and Computers*, pages 1021–1025. IEEE.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467.

Weng, X., Huang, L., Zhao, L., Anwer, R. M., Khan, S. H., and Khan, F. S. (2022). An investigation into whitening loss for self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 29748–29760.

Weng, X., Ni, Y., Song, T., Luo, J., Anwer, R. M., Khan, S. H., Khan, F. S., and Huang, L. (2024). Modulate your spectrum in self-supervised learning. In *International Conference on Learning Representations*.

Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., and Locatello, F. (2024). Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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] The mathematical setting (data generating process, CCA objective, whitened encoder classes) and all assumptions (Assumptions 1–3) are formally stated in Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Population identifiability properties are analyzed in Theorem 1, and finite-sample convergence is analyzed in Theorem 2. Training times and computational costs are reported in Appendix G.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Source code is available on GitHub.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] All assumptions (Assumptions 1–3 and conditions A1–A4 in Theorem 2) are explicitly stated before each result.
   (b) Complete proofs of all theoretical results. [Yes] Proof sketches are provided in the main text; complete proofs are given in the supplementary material (Appendix D).
   (c) Clear explanations of any assumptions. [Yes] Each assumption is followed by a remark explaining its role, e.g., the spectral gap condition in Assumption 3 is explained via the Hermite expansion.

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] Source code is available on GitHub, and all datasets used are publicly available.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Training configurations (optimizer, learning rate, batch size, iterations, encoder architectures, distribution setups) are detailed in Appendix G.
   (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] All metrics (R^2, PA_{mean}, PA_{max}, orbit distance) are defined in Section 4.1. Error bars report standard deviations over five independent random seeds.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Experiments were run on NVIDIA RTX A5000 GPUs, as described in Appendix G.

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] The 3DIdent dataset (Zimmermann et al., 2021) and all baseline methods are properly cited.
   (b) The license information of the assets, if applicable. [Yes] The 3DIdent dataset is released under the Creative Commons Attribution 4.0 International (CC-BY-4.0) license.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Source code is available on GitHub.
   (d) Information about consent from data providers/curators. [Not Applicable] We use only publicly available datasets and open-source code.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] The datasets used (synthetic data and rendered 3D object images) do not contain personally identifiable information or offensive content.

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] No crowdsourcing or human subjects research was conducted.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable] No crowdsourcing or human subjects research was conducted.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable] No crowdsourcing or human subjects research was conducted.

Supplementary Material

A OVERVIEW OF THE SUPPLEMENTARY MATERIAL

This supplementary material is organized as follows. Appendix B compares the identifiability frameworks of nonlinear CCA and contrastive learning. Appendix D contains complete proofs of the main theoretical results and Appendix E isolates the role of whitening in the identifiability argument. Appendix F provides proof sketches for additional latent distribution families. Appendix G describes implementation details and Appendix H reports additional experimental results. Accompanying source code is provided to support reproducibility and independent verification.

B EXTENDED RELATED WORK: COMPARISON OF IDENTIFIABILITY FRAMEWORKS: NONLINEAR CCA VERSUS CONTRASTIVE LEARNING

While both nonlinear Canonical Correlation Analysis (CCA) and Contrastive Learning (CL), e.g., InfoNCE, serve as foundational paradigms for multi-view representation learning, their underlying theoretical mechanisms for achieving identifiability differ fundamentally. Recent works have established strong identifiability guarantees for contrastive methods (Hyvärinen et al., 2019; Zimmermann et al., 2021; Matthes et al., 2023). However, unifying CL and CCA under a single theoretical framework remains challenging due to divergent structural assumptions, regularization mechanisms, and proof techniques. We systematically compare these two theoretical frameworks below.

1. Distributional Assumptions

- **Contrastive Learning:** CL identifiability typically assumes the ground-truth latents are sampled from an exponential family distribution. Crucially, it relies on non-Gaussianity (often requiring that at most one latent dimension is Gaussian, following classical ICA theory) or imposes no constraints on the marginal distributions (Hyvärinen et al., 2019). CL models remain identifiable even when the latent spaces are strictly bounded or distributed on a hypersphere (e.g., von Mises-Fisher distributions).
- **Nonlinear CCA:** In contrast, our nonlinear CCA framework requires strong joint distributional assumptions across all latent dimensions, specifically relying on families that admit an orthogonal polynomial expansion (such as the joint Gaussian priors driving the Mehler expansion). Consequently, nonlinear CCA becomes theoretically non-identifiable under bounded or hyperspherical latent spaces—precisely the regimes where CL theories thrive.

2. Regularization Mechanisms

- **Contrastive Learning:** CL objectives inherently regularize the geometry of the representation space. This is typically achieved via unit-norm constraints on the embeddings (e.g., temperature-scaled InfoNCE) or by assuming bounded latent spaces, which prevents the representations from diverging.
- **Nonlinear CCA:** CCA relies strictly on representation whitening to achieve well-conditioning. As demonstrated in our theoretical analysis, unit variance ensures the compactness of the representation functions, while zero cross-correlation guarantees orthogonality between the components in the function space. Without whitening, the orthogonal polynomial expansion collapses, and the CCA objective cannot reliably separate first-order components from higher-order artifacts.

3. Proof Techniques

- **Contrastive Learning:** The standard proof mechanism for CL demonstrates that, under exponential family generative models, the contrastive loss forces the learned representation to act as an isometry. By preserving the underlying metric structure of the data-generating process, the learned features asymptotically align with the true latents (Zimmermann et al., 2021).
- **Nonlinear CCA:** Our analysis of CCA operates via function expansion in an infinitesimal function space. The proof mechanism relies on decomposing the representation mapping into an orthogonal polynomial basis (e.g., normalized Hermite polynomials). Identifiability is established by proving that the CCA objective uniquely isolates the first-order basis components that maximize canonical correlations, provided that structural conditions—such as the First-Order Canonical Dominance (Assumption 3)—hold.

4. Identifiability Guarantees

- **Contrastive Learning:** Because CL forces metric preservation, it yields exceptionally strong global identifiability conclusions. When the order of the sufficient statistics is known, CL can achieve permutation or block-permutation disentanglement (up to sign flips) (Matthes et al., 2023).
- **Nonlinear CCA:** CCA achieves strictly *affine* identifiability; the ground-truth representation is recovered up to a linear shift and an orthogonal rotation (after whitening). While this represents a broader equivalence class than permutation disentanglement, our theory directly explains the empirical success of purely non-contrastive, correlation-based methods (e.g., Barlow Twins (Zbontar et al., 2021), W-MSE (Ermolov et al., 2021)). It proves that exact latent recovery is possible under minimal architectural complexity, entirely circumventing the need for negative sampling.

Remark 2. Modern non-contrastive pipelines often incorporate architectural mechanisms such as momentum encoders and stop-gradient pathways (e.g., BYOL, SimSiam). These dynamics are not yet rigorously captured by current identifiability frameworks for either ICA/CL or CCA-type methods, representing an important frontier for future theoretical unification.

C Further Remarks on Theoretical Results

Remark 3 (Role of Canonical-Correlation Separation). The condition in Assumption 3 is precisely the threshold between strictly affine recovery and contamination by higher-order nonlinearities. It requires that the weakest first-order canonical correlation strictly exceed the largest correlation induced by any second-order term in the normalized Hermite expansion. Since nonlinear CCA ranks candidate directions solely by their inter-view correlations, irrespective of polynomial degree, this strict separation prevents any higher-order Hermite component from entering the top-$d_{\mathcal{Z}}$ representation. Consequently, the learned subspace is supported only on first-order terms, which is exactly what yields strictly affine identifiability.

If this separation condition is relaxed, the ordering argument no longer applies. Higher-order terms associated with strongly correlated latent factors can then outrank first-order terms associated with weaker factors, so the learned representation generally becomes a degree-mixed polynomial subspace determined jointly by the full canonical spectrum $\{\rho_i\}_{i=1}^{d_{\mathcal{S}}}$ and the representation dimension $d_{\mathcal{Z}}$. Nevertheless, because the canonical correlations of Hermite components decay exponentially with the polynomial degree, enlarging the representation budget eventually brings all first-order terms into the retained subspace in the limit $d_{\mathcal{Z}} \to \infty$, albeit entangled with higher-order artifacts. See Appendix D.4.

Remark 4 (Interpretation of A3–A4). Assumptions A3 and A4 do not impose further structural restrictions on the latent data generation process. Instead, they constitute standard conditions for establishing the argmax consistency of M-estimators applied to the empirical CCA objective. Specifically, Assumption A3 gives a uniform law of large numbers alongside a ridge schedule that ensures stable empirical whitening operators. Assumption A4 requires the optimization procedure to yield approximate maximizers of the empirical objective. We formulate these conditions uniformly to facilitate a transparent application of standard argmax theorems, weaker pointwise variants, complemented by compactness and continuity requirements, would also suffice at the expense of increased technical complexity.

D Proofs of Main Theoretical Results

D.1 Proof of Proposition 1

Lemma 1 (Pushforward identities and whitening preservation)**.** *Let $(\mathbf{s}, \mathbf{s}')$ be square-integrable random vectors on $\mathcal{S} \times \mathcal{S}$ with joint law $P_{\mathbf{ss}'}$. Let $\mathbf{g}\colon \mathcal{S} \to \mathcal{X}$ and $\mathbf{g}'\colon \mathcal{S} \to \mathcal{X}'$ be Borel-measurable, and define*

$$\mathbf{x} = \mathbf{g}(\mathbf{s}), \qquad \mathbf{x}' = \mathbf{g}'(\mathbf{s}').$$

Denote

$$P_{\mathbf{x}} = \mathbf{g}_{\#}P_{\mathbf{s}}, \qquad P_{\mathbf{x}'} = \mathbf{g}'_{\#}P_{\mathbf{s}'}, \qquad P_{\mathbf{xx}'} = (\mathbf{g}, \mathbf{g}')_{\#}P_{\mathbf{ss}'}.$$

Then the following properties hold.

1. ***Square integrability and isometry.*** *For any $\boldsymbol{\phi} \in L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ and $\boldsymbol{\phi}' \in L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$,*

$$\boldsymbol{\phi} \circ \mathbf{g} \in L^2(P_{\mathbf{s}}; \mathbb{R}^{d_{\mathcal{Z}}}), \qquad \boldsymbol{\phi}' \circ \mathbf{g}' \in L^2(P_{\mathbf{s}'}; \mathbb{R}^{d_{\mathcal{Z}}}),$$

   and

$$\|\boldsymbol{\phi} \circ \mathbf{g}\|_{L^2(P_{\mathbf{s}})} = \|\boldsymbol{\phi}\|_{L^2(P_{\mathbf{x}})}, \qquad \|\boldsymbol{\phi}' \circ \mathbf{g}'\|_{L^2(P_{\mathbf{s}'})} = \|\boldsymbol{\phi}'\|_{L^2(P_{\mathbf{x}'})}.$$

2. ***Expectation and covariance preservation.*** *For all $\boldsymbol{\phi} \in L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ and $\boldsymbol{\phi}' \in L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$,*

$$\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})] = \mathbb{E}[\boldsymbol{\phi}(\mathbf{g}(\mathbf{s}))], \qquad \mathbb{E}[\boldsymbol{\phi}'(\mathbf{x}')] = \mathbb{E}[\boldsymbol{\phi}'(\mathbf{g}'(\mathbf{s}'))],$$

$$\mathrm{Cov}(\boldsymbol{\phi}(\mathbf{x})) = \mathrm{Cov}(\boldsymbol{\phi}(\mathbf{g}(\mathbf{s}))), \qquad \mathrm{Cov}(\boldsymbol{\phi}'(\mathbf{x}')) = \mathrm{Cov}(\boldsymbol{\phi}'(\mathbf{g}'(\mathbf{s}'))),$$

   and

$$\mathrm{Cov}(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}'(\mathbf{x}')) = \mathrm{Cov}(\boldsymbol{\phi}(\mathbf{g}(\mathbf{s})), \boldsymbol{\phi}'(\mathbf{g}'(\mathbf{s}'))).$$

3. ***Whitening preservation.*** *If $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$ and $\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$, then*

$$\tilde{\mathbf{f}} \circ \mathbf{g} \in \hat{\mathcal{F}}_{\mathcal{S}}, \qquad \tilde{\mathbf{f}}' \circ \mathbf{g}' \in \hat{\mathcal{F}}'_{\mathcal{S}}.$$

   Equivalently,

$$\tilde{\mathcal{F}}_{\mathcal{S}} \subseteq \hat{\mathcal{F}}_{\mathcal{S}}, \qquad \tilde{\mathcal{F}}'_{\mathcal{S}} \subseteq \hat{\mathcal{F}}'_{\mathcal{S}}.$$

Proof. Since $P_{\mathbf{x}} = \mathbf{g}_{\#}P_{\mathbf{s}}$ and $P_{\mathbf{x}'} = \mathbf{g}'_{\#}P_{\mathbf{s}'}$, the defining property of pushforward measures yields, for every nonnegative or integrable Borel function φ,

$$\int_{\mathcal{X}} \varphi(x)\, dP_{\mathbf{x}}(x) = \int_{\mathcal{S}} \varphi(\mathbf{g}(s))\, dP_{\mathbf{s}}(s), \qquad \int_{\mathcal{X}'} \varphi(x')\, dP_{\mathbf{x}'}(x') = \int_{\mathcal{S}} \varphi(\mathbf{g}'(s'))\, dP_{\mathbf{s}'}(s').$$

For Part 1, apply this identity with $\varphi(x) = \|\boldsymbol{\phi}(x)\|^2$ and $\varphi(x') = \|\boldsymbol{\phi}'(x')\|^2$. Then

$$\mathbb{E}\big[\|\boldsymbol{\phi}(\mathbf{g}(\mathbf{s}))\|^2\big] = \mathbb{E}\big[\|\boldsymbol{\phi}(\mathbf{x})\|^2\big] < \infty,$$

and similarly for $\boldsymbol{\phi}'$. The stated L^2-norm equalities follow immediately.

For Part 2, the same pushforward identity with vector-valued integrands gives the expectation formulas. The covariance identities then follow from

$$\mathrm{Cov}(\mathbf{u}, \mathbf{v}) = \mathbb{E}[\mathbf{u}\mathbf{v}^\top] - \mathbb{E}[\mathbf{u}]\mathbb{E}[\mathbf{v}]^\top.$$

Indeed, since $P_{\mathbf{xx}'} = (\mathbf{g}, \mathbf{g}')_{\#}P_{\mathbf{ss}'}$, we have

$$\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}'(\mathbf{x}')^\top] = \mathbb{E}[\boldsymbol{\phi}(\mathbf{g}(\mathbf{s}))\boldsymbol{\phi}'(\mathbf{g}'(\mathbf{s}'))^\top].$$

Part 3 is immediate from Part 2: if $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{f}}'$ are centered and whitened under $P_{\mathbf{x}}$ and $P_{\mathbf{x}'}$, then

$$\mathbb{E}[(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})] = \mathbf{0}, \qquad \mathrm{Cov}((\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})) = \mathbf{I}_{d_{\mathcal{Z}}},$$

and likewise for $\tilde{\mathbf{f}}' \circ \mathbf{g}'$. Hence the compositions belong to the feasible source-space classes. □

Proof of Proposition 1. Throughout this proof, functions in L^2 are identified up to almost-sure equality.

1. Objective preservation. Let $(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$. By Lemma 1.2,

$$\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x}), \tilde{\mathbf{f}}'(\mathbf{x}')) = \mathrm{Cov}((\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s}), (\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')).$$

Therefore the two cross-covariance matrices have the same singular values, and hence

$$J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = J_{\mathcal{S}}(\tilde{\mathbf{f}} \circ \mathbf{g}, \tilde{\mathbf{f}}' \circ \mathbf{g}').$$

2. Maximizer correspondence on representable classes. Define

$$\Psi : \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'} \to \tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}, \qquad \Psi(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = (\tilde{\mathbf{f}} \circ \mathbf{g}, \tilde{\mathbf{f}}' \circ \mathbf{g}').$$

By definition of $\tilde{\mathcal{F}}_{\mathcal{S}}$ and $\tilde{\mathcal{F}}'_{\mathcal{S}}$, the map Ψ is surjective. Moreover, for any $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2 \in \tilde{\mathcal{F}}_{\mathcal{X}}$,

$$\|\tilde{\mathbf{f}}_1 - \tilde{\mathbf{f}}_2\|^2_{L^2(P_{\mathbf{x}})} = \|(\tilde{\mathbf{f}}_1 - \tilde{\mathbf{f}}_2) \circ \mathbf{g}\|^2_{L^2(P_{\mathbf{s}})} = \|\tilde{\mathbf{f}}_1 \circ \mathbf{g} - \tilde{\mathbf{f}}_2 \circ \mathbf{g}\|^2_{L^2(P_{\mathbf{s}})}$$

by Lemma 1.1, and the same holds on the primed side. Hence Ψ is an isometric bijection modulo null sets. Combining this with Part 1 gives

$$(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*) \in \arg\max_{\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}} J \quad \Longleftrightarrow \quad (\tilde{\mathbf{f}}^* \circ \mathbf{g}, \tilde{\mathbf{f}}'^* \circ \mathbf{g}') \in \arg\max_{\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}} J_{\mathcal{S}}.$$

3. Representation universality. We prove the claim for $\hat{\mathcal{F}}_{\mathcal{S}}$; the primed case is identical. Since $\mathcal{S}$ and $\mathcal{X}$ are standard Borel spaces and $\mathbf{g}$ is injective and Borel measurable, the Lusin–Souslin theorem implies that $\mathbf{g}(\mathcal{S}) \subseteq \mathcal{X}$ is Borel and that

$$\mathbf{g}^{-1} : \mathbf{g}(\mathcal{S}) \to \mathcal{S}$$

is Borel measurable.

Fix $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$. Define the pullback $\bar{\mathbf{h}} : \mathcal{X} \to \mathbb{R}^{d_{\mathcal{Z}}}$ by

$$\bar{\mathbf{h}}(x) = \begin{cases} \hat{\mathbf{h}}(\mathbf{g}^{-1}(x)), & x \in \mathbf{g}(\mathcal{S}), \\ \mathbf{0}, & x \notin \mathbf{g}(\mathcal{S}). \end{cases}$$

Then $\bar{\mathbf{h}}$ is Borel measurable. Since $P_{\mathbf{x}}$ is supported on $\mathbf{g}(\mathcal{S})$,

$$\bar{\mathbf{h}}(\mathbf{g}(\mathbf{s})) = \hat{\mathbf{h}}(\mathbf{s}) \qquad \text{for } P_{\mathbf{s}}\text{-a.e. } \mathbf{s}.$$

By Lemma 1.1,

$$\|\bar{\mathbf{h}}\|_{L^2(P_{\mathbf{x}})} = \|\hat{\mathbf{h}}\|_{L^2(P_{\mathbf{s}})} < \infty,$$

so $\bar{\mathbf{h}} \in L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$. Moreover, by Lemma 1.2 and the fact that $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$,

$$\mathbb{E}[\bar{\mathbf{h}}(\mathbf{x})] = \mathbf{0}, \qquad \mathrm{Cov}(\bar{\mathbf{h}}(\mathbf{x})) = \mathbf{I}_{d_{\mathcal{Z}}}.$$

By density of the base encoder class $\mathcal{H}_{\mathcal{X}}$ in $L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$, there exists a sequence $\mathbf{f}_n \in \mathcal{H}_{\mathcal{X}}$ such that

$$\|\mathbf{f}_n - \bar{\mathbf{h}}\|_{L^2(P_{\mathbf{x}})} \to 0.$$

Set

$$\boldsymbol{\mu}_n := \mathbb{E}[\mathbf{f}_n(\mathbf{x})], \qquad \boldsymbol{\Sigma}_n := \mathrm{Cov}(\mathbf{f}_n(\mathbf{x})).$$

Because convergence in L^2 implies convergence of first and second moments,

$$\boldsymbol{\mu}_n \to \mathbf{0}, \qquad \boldsymbol{\Sigma}_n \to \mathbf{I}_{d_{\mathcal{Z}}}.$$

In particular, $\boldsymbol{\Sigma}_n \succ 0$ for all sufficiently large n. For such n, choose the symmetric whitener

$$\mathbf{W}_n := \boldsymbol{\Sigma}_n^{-1/2}, \qquad \tilde{\mathbf{f}}_n := \mathbf{W}_n(\mathbf{f}_n - \boldsymbol{\mu}_n) \in \tilde{\mathcal{F}}_{\mathcal{X}}.$$

Since the map $A \mapsto A^{-1/2}$ is continuous on the cone of symmetric positive definite matrices and $\mathbf{\Sigma}_n \to \mathbf{I}_{d_\mathcal{Z}}$, we have $\mathbf{W}_n \to \mathbf{I}_{d_\mathcal{Z}}$. Hence

$$\begin{aligned}\|\tilde{\mathbf{f}}_n - \bar{\mathbf{h}}\|_{L^2(P_\mathbf{x})} &\le \|(\mathbf{W}_n - \mathbf{I}_{d_\mathcal{Z}})(\mathbf{f}_n - \boldsymbol{\mu}_n)\|_{L^2(P_\mathbf{x})} + \|\mathbf{f}_n - \bar{\mathbf{h}}\|_{L^2(P_\mathbf{x})} + \|\boldsymbol{\mu}_n\| \\ &\le \|\mathbf{W}_n - \mathbf{I}_{d_\mathcal{Z}}\|\,\|\mathbf{f}_n - \boldsymbol{\mu}_n\|_{L^2(P_\mathbf{x})} + \|\mathbf{f}_n - \bar{\mathbf{h}}\|_{L^2(P_\mathbf{x})} + \|\boldsymbol{\mu}_n\| \to 0.\end{aligned}$$

Now define

$$\tilde{\mathbf{h}}_n := \tilde{\mathbf{f}}_n \circ \mathbf{g} \in \tilde{\mathcal{F}}_\mathcal{S}.$$

By Lemma 1.1,

$$\|\tilde{\mathbf{h}}_n - \hat{\mathbf{h}}\|_{L^2(P_\mathbf{s})} = \|\tilde{\mathbf{f}}_n - \bar{\mathbf{h}}\|_{L^2(P_\mathbf{x})} \to 0.$$

Therefore, for every $\epsilon > 0$, there exists n such that

$$\mathbb{E}\Big[\|\tilde{\mathbf{h}}_n(\mathbf{s}) - \hat{\mathbf{h}}(\mathbf{s})\|^2\Big] < \epsilon^2.$$

Repeating the same construction for $\hat{\mathbf{h}}' \in \hat{\mathcal{F}}'_\mathcal{S}$ proves the density claim on the primed side.

4. Continuity of $J_\mathcal{S}$ and equality of suprema. Let $(\mathbf{h}, \mathbf{h}'), (\mathbf{u}, \mathbf{u}') \in \hat{\mathcal{F}}_\mathcal{S} \times \hat{\mathcal{F}}'_\mathcal{S}$, and define

$$\Delta := \mathrm{Cov}(\mathbf{h}(\mathbf{s}), \mathbf{h}'(\mathbf{s}')) - \mathrm{Cov}(\mathbf{u}(\mathbf{s}), \mathbf{u}'(\mathbf{s}')).$$

Since all four maps are centered,

$$\Delta = \mathbb{E}[(\mathbf{h} - \mathbf{u})(\mathbf{s})\,\mathbf{h}'(\mathbf{s}')^\top] + \mathbb{E}[\mathbf{u}(\mathbf{s})\,(\mathbf{h}' - \mathbf{u}')(\mathbf{s}')^\top].$$

Hence, by Cauchy–Schwarz,

$$\|\Delta\|_F \le \|\mathbf{h} - \mathbf{u}\|_{L^2(P_\mathbf{s})}\,\|\mathbf{h}'\|_{L^2(P_{\mathbf{s}'})} + \|\mathbf{u}\|_{L^2(P_\mathbf{s})}\,\|\mathbf{h}' - \mathbf{u}'\|_{L^2(P_{\mathbf{s}'})}.$$

Because every element of $\hat{\mathcal{F}}_\mathcal{S}$ and $\hat{\mathcal{F}}'_\mathcal{S}$ is whitened,

$$\|\mathbf{h}\|_{L^2(P_\mathbf{s})} = \|\mathbf{u}\|_{L^2(P_\mathbf{s})} = \|\mathbf{h}'\|_{L^2(P_{\mathbf{s}'})} = \sqrt{\mathrm{tr}(\mathbf{I}_{d_\mathcal{Z}})} = \sqrt{d_\mathcal{Z}}.$$

Therefore,

$$\|\Delta\|_F \le \sqrt{d_\mathcal{Z}}\Big(\|\mathbf{h} - \mathbf{u}\|_{L^2(P_\mathbf{s})} + \|\mathbf{h}' - \mathbf{u}'\|_{L^2(P_{\mathbf{s}'})}\Big).$$

Since $J_\mathcal{S}$ is the nuclear norm of the cross-covariance matrix,

$$|J_\mathcal{S}(\mathbf{h}, \mathbf{h}') - J_\mathcal{S}(\mathbf{u}, \mathbf{u}')| \le \|\Delta\|_* \le \sqrt{d_\mathcal{Z}}\|\Delta\|_F \le d_\mathcal{Z}\Big(\|\mathbf{h} - \mathbf{u}\|_{L^2(P_\mathbf{s})} + \|\mathbf{h}' - \mathbf{u}'\|_{L^2(P_{\mathbf{s}'})}\Big).$$

Thus $J_\mathcal{S}$ is continuous under the product L^2 topology on $\hat{\mathcal{F}}_\mathcal{S} \times \hat{\mathcal{F}}'_\mathcal{S}$.

Because $\tilde{\mathcal{F}}_\mathcal{S}$ is dense in $\hat{\mathcal{F}}_\mathcal{S}$ and $\tilde{\mathcal{F}}'_\mathcal{S}$ is dense in $\hat{\mathcal{F}}'_\mathcal{S}$, this continuity implies

$$\sup_{\tilde{\mathcal{F}}_\mathcal{S} \times \tilde{\mathcal{F}}'_\mathcal{S}} J_\mathcal{S} = \sup_{\hat{\mathcal{F}}_\mathcal{S} \times \hat{\mathcal{F}}'_\mathcal{S}} J_\mathcal{S}.$$

Together with Part 1,

$$\sup_{\tilde{\mathcal{F}}_\mathcal{X} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}} J = \sup_{\tilde{\mathcal{F}}_\mathcal{S} \times \tilde{\mathcal{F}}'_\mathcal{S}} J_\mathcal{S} = \sup_{\hat{\mathcal{F}}_\mathcal{S} \times \hat{\mathcal{F}}'_\mathcal{S}} J_\mathcal{S}.$$

Finally, if $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ maximizes J over $\tilde{\mathcal{F}}_\mathcal{X} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$, then by Part 2 its composition maximizes $J_\mathcal{S}$ over $\tilde{\mathcal{F}}_\mathcal{S} \times \tilde{\mathcal{F}}'_\mathcal{S}$; since the supremum over representable classes equals the supremum over the full feasible classes, the same composed pair is also a maximizer of $J_\mathcal{S}$ over $\hat{\mathcal{F}}_\mathcal{S} \times \hat{\mathcal{F}}'_\mathcal{S}$. This proves the proposition. □

D.2 Proof of the rotation-invariance of CCA

In this appendix, we formalize the rotational symmetry underlying the population CCA objective in Equation 2. Throughout, for any feasible pair $(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$, we write

$$\mathbf{C}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}'} := \mathrm{Cov}\big(\tilde{\mathbf{f}}(\mathbf{x}), \tilde{\mathbf{f}}'(\mathbf{x}')\big) \in \mathbb{R}^{d_{\mathcal{Z}} \times d_{\mathcal{Z}}}.$$

Proposition 2 (Orthogonal invariance of the whitened CCA objective)**.** *Under Assumption 2, let* $(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$ *be any feasible pair, and let* $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$*. Then*

$$\mathbf{Q}\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \qquad \mathbf{Q}'\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'},$$

and the population CCA objective is invariant under these post-transformations:

$$J(\mathbf{Q}\tilde{\mathbf{f}},\, \mathbf{Q}'\tilde{\mathbf{f}}') = J(\tilde{\mathbf{f}},\, \tilde{\mathbf{f}}').$$

Equivalently, J is constant on every $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit.

Proof. By construction of the whitened encoder classes, every feasible encoder is centered and whitened. Hence

$$\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})] = \mathbf{0}, \qquad \mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x})) = \mathbf{I}_{d_{\mathcal{Z}}},$$

and analogously,

$$\mathbb{E}[\tilde{\mathbf{f}}'(\mathbf{x}')] = \mathbf{0}, \qquad \mathrm{Cov}(\tilde{\mathbf{f}}'(\mathbf{x}')) = \mathbf{I}_{d_{\mathcal{Z}}}.$$

We first verify feasibility after orthogonal post-transformation. Since $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$, we have

$$\mathbb{E}[\mathbf{Q}\tilde{\mathbf{f}}(\mathbf{x})] = \mathbf{Q}\,\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})] = \mathbf{0},$$

and

$$\mathrm{Cov}(\mathbf{Q}\tilde{\mathbf{f}}(\mathbf{x})) = \mathbf{Q}\,\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x}))\,\mathbf{Q}^\top = \mathbf{Q}\mathbf{I}_{d_{\mathcal{Z}}}\mathbf{Q}^\top = \mathbf{I}_{d_{\mathcal{Z}}}.$$

Thus $\mathbf{Q}\tilde{\mathbf{f}}$ is again zero-mean and whitened. By the post-orthogonal closure in Assumption 2, $\mathbf{Q}\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$. The same argument yields $\mathbf{Q}'\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$.

Next, using the zero-mean property of the whitened encoders, the transformed cross-covariance satisfies

$$\mathbf{C}_{\mathbf{Q}\tilde{\mathbf{f}},\,\mathbf{Q}'\tilde{\mathbf{f}}'} = \mathrm{Cov}\big(\mathbf{Q}\tilde{\mathbf{f}}(\mathbf{x}),\, \mathbf{Q}'\tilde{\mathbf{f}}'(\mathbf{x}')\big) = \mathbf{Q}\,\mathbf{C}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}'}\,\mathbf{Q}'^\top.$$

Let

$$\mathbf{C}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}'} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

be a singular value decomposition, where $\mathbf{U}, \mathbf{V} \in O(d_{\mathcal{Z}})$ and $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{d_{\mathcal{Z}}})$ with $\sigma_1 \geq \cdots \geq \sigma_{d_{\mathcal{Z}}} \geq 0$. Then

$$\mathbf{C}_{\mathbf{Q}\tilde{\mathbf{f}},\,\mathbf{Q}'\tilde{\mathbf{f}}'} = (\mathbf{Q}\mathbf{U})\,\mathbf{D}\,(\mathbf{Q}'\mathbf{V})^\top.$$

Since $\mathbf{QU}$ and $\mathbf{Q}'\mathbf{V}$ are again orthogonal, this is a singular value decomposition of $\mathbf{C}_{\mathbf{Q}\tilde{\mathbf{f}},\,\mathbf{Q}'\tilde{\mathbf{f}}'}$. Therefore,

$$\sigma_i\Big(\mathbf{C}_{\mathbf{Q}\tilde{\mathbf{f}},\,\mathbf{Q}'\tilde{\mathbf{f}}'}\Big) = \sigma_i\Big(\mathbf{C}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}'}\Big), \qquad i = 1, \ldots, d_{\mathcal{Z}}.$$

Summing over i yields

$$J(\mathbf{Q}\tilde{\mathbf{f}},\, \mathbf{Q}'\tilde{\mathbf{f}}') = \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\Big(\mathbf{C}_{\mathbf{Q}\tilde{\mathbf{f}},\,\mathbf{Q}'\tilde{\mathbf{f}}'}\Big) = \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\Big(\mathbf{C}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}'}\Big) = J(\tilde{\mathbf{f}},\, \tilde{\mathbf{f}}'),$$

which proves the claim. □

Corollary 1 (Orthogonal orbit of equivalent maximizers)**.** *If* $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ *is a maximizer of Equation 2, then for every* $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$*, the pair*

$$(\mathbf{Q}\tilde{\mathbf{f}}^*,\, \mathbf{Q}'\tilde{\mathbf{f}}'^*)$$

is also a maximizer. In particular, each maximizer generates an equivalent $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit of solutions.

Proof. This follows immediately from Proposition 2, since the objective value is unchanged by orthogonal post-transformations and the feasible set is closed under them. □

Remark 5 (Why whitening leaves an orthogonal ambiguity). The preceding result shows that, after whitening, the residual linear ambiguity is orthogonal. Indeed, let $\tilde{\mathbf{f}}$ be any whitened encoder and let $\mathbf{A} \in \mathrm{GL}(d_{\mathcal{Z}})$. Then

$$\mathrm{Cov}(\mathbf{A}\tilde{\mathbf{f}}(\mathbf{x})) = \mathbf{A}\,\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x}))\,\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top.$$

Hence $\mathbf{A}\tilde{\mathbf{f}}$ remains whitened if and only if

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_{d_{\mathcal{Z}}},$$

that is, if and only if $\mathbf{A} \in O(d_{\mathcal{Z}})$. Therefore whitening eliminates arbitrary anisotropic rescalings and shears, leaving precisely the orthogonal symmetry established above.

Proposition 3 (Orthogonal invariance does not imply uniqueness modulo orbit)**.** *Under Assumption 2 alone, the set of maximizers of Equation 2 need not be a single $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit.*

Proof. Take $d_{\mathcal{Z}} = 1$, $\mathcal{X} = \mathcal{X}' = \mathbb{R}$, and let

$$X = X' \sim \mathcal{N}(0, 1).$$

Let the base classes be

$$\mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{X}'} = L^2(P_X; \mathbb{R}),$$

so the corresponding whitened classes are exactly the centered, unit-variance scalar functions.

For any feasible pair $(\tilde{f}, \tilde{f}')$,

$$J(\tilde{f}, \tilde{f}') = \left|\mathrm{Cov}(\tilde{f}(X), \tilde{f}'(X))\right| \le \sqrt{\mathrm{Var}(\tilde{f}(X))\,\mathrm{Var}(\tilde{f}'(X))} = 1$$

by Cauchy–Schwarz. Hence every pair of the form (h, h), where

$$\mathbb{E}[h(X)] = 0, \qquad \mathrm{Var}(h(X)) = 1,$$

is a population maximizer.

In particular, the two feasible functions

$$h_1(X) = X, \qquad h_2(X) = \frac{X^2 - 1}{\sqrt{2}}$$

give two maximizers (h_1, h_1) and (h_2, h_2). Since $O(1) = \{\pm 1\}$ and $h_2 \neq \pm h_1$ in $L^2(P_X)$, these two maximizers are not in the same orbit. □

Remark 6 (Scope of Proposition 2). Proposition 2 and Corollary 1 show only that each maximizer generates an equivalent $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit of maximizers. They do not imply that all maximizers belong to a single orbit.

The stronger single-orbit statement used later in the paper requires the additional distributional and spectral assumptions of Theorem 1 (see Corollary 4).

D.3 Proof of Theorem 1

Throughout this subsection, write

$$d := d_{\mathcal{S}} = d_{\mathcal{Z}}.$$

Lemma 2 (Hermite–Mehler expansion of a bivariate Gaussian pair)**.** *Let (S, S') be jointly Gaussian with means $\mu_S, \mu_{S'}$, variances $\sigma_S^2, \sigma_{S'}^2$, and correlation coefficient $t \in (-1, 1)$. Define the standardized coordinates*

$$X := \frac{S - \mu_S}{\sigma_S}, \qquad Y := \frac{S' - \mu_{S'}}{\sigma_{S'}}, \qquad U := \frac{X}{\sqrt{2}}, \qquad V := \frac{Y}{\sqrt{2}}.$$

Let H_n denote the physicists' Hermite polynomials and define

$$\psi_n(z) := \frac{1}{\sqrt{2^n n!}}\, H_n(z), \qquad n \in \mathbb{N}_0.$$

Then $\{\psi_n\}_{n\geq 0}$ is an orthonormal basis of $L^2(\nu)$, where

$$\nu(dz) = \pi^{-1/2} e^{-z^2}\, dz,$$

and the joint density of (S, S') admits the Hermite–Mehler expansion

$$p_{S,S'}(s,s') = \frac{1}{2\pi\,\sigma_S\sigma_{S'}} e^{-(u^2+v^2)} \sum_{n=0}^{\infty} t^n\, \psi_n(u)\psi_n(v), \tag{5}$$

where

$$u = \frac{s-\mu_S}{\sqrt{2}\,\sigma_S}, \qquad v = \frac{s'-\mu_{S'}}{\sqrt{2}\,\sigma_{S'}}.$$

Equivalently, the joint density of (U, V) is

$$p_{U,V}(u,v) = \frac{1}{\pi} e^{-(u^2+v^2)} \sum_{n=0}^{\infty} t^n\, \psi_n(u)\psi_n(v).$$

In particular,

$$\mathbb{E}[\psi_m(U)\psi_n(V)] = t^n\, \delta_{mn}, \qquad m, n \in \mathbb{N}_0.$$

Proof. The classical Mehler formula for the physicists' Hermite polynomials states that, for $|t| < 1$,

$$\sum_{n=0}^{\infty} \frac{t^n}{2^n n!} H_n(u)H_n(v) = \frac{1}{\sqrt{1-t^2}} \exp\!\left(\frac{2tuv - t^2(u^2+v^2)}{1-t^2}\right).$$

Multiplying both sides by $\pi^{-1}e^{-(u^2+v^2)}$ gives

$$\frac{1}{\pi} e^{-(u^2+v^2)} \sum_{n=0}^{\infty} \frac{t^n}{2^n n!} H_n(u)H_n(v) = \frac{1}{\pi\sqrt{1-t^2}} \exp\!\left(-\frac{u^2 - 2tuv + v^2}{1-t^2}\right),$$

which is precisely the density of (U, V). Since

$$\psi_n(u)\psi_n(v) = \frac{1}{2^n n!} H_n(u)H_n(v),$$

this proves the expansion for $p_{U,V}$, and the formula for $p_{S,S'}$ follows by the change of variables $(s, s') \mapsto (u, v)$, whose Jacobian determinant is $(2\sigma_S\sigma_{S'})^{-1}$.

Orthogonality is immediate from

$$\int_{\mathbb{R}} \psi_m(z)\psi_n(z)\, \nu(dz) = \delta_{mn}.$$

Finally,

$$\begin{aligned}
\mathbb{E}[\psi_m(U)\psi_n(V)] &= \int_{\mathbb{R}^2} \psi_m(u)\psi_n(v)\, p_{U,V}(u,v)\, du\, dv \\
&= \int_{\mathbb{R}^2} \psi_m(u)\psi_n(v) \left(\sum_{k=0}^{\infty} t^k \psi_k(u)\psi_k(v)\right) \nu(du)\nu(dv) \\
&= \sum_{k=0}^{\infty} t^k \left(\int_{\mathbb{R}} \psi_m(u)\psi_k(u)\, \nu(du)\right) \left(\int_{\mathbb{R}} \psi_n(v)\psi_k(v)\, \nu(dv)\right) \\
&= t^n\, \delta_{mn}.
\end{aligned}$$

□

Proposition 4 (Multivariate Hermite–Mehler expansion in canonical coordinates)**.** *Let* $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^d \times \mathbb{R}^d$ *be jointly Gaussian such that* $\sqrt{2}\,\mathbf{u}$ *and* $\sqrt{2}\,\mathbf{v}$ *are standard Gaussian and*

$$\mathrm{Cov}(\sqrt{2}\,\mathbf{u}, \sqrt{2}\,\mathbf{v}) = \mathrm{diag}(\rho_1, \ldots, \rho_d), \qquad 1 > \rho_1 \geq \cdots \geq \rho_d > 0.$$

Define, for every multi-index $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$,

$$\Psi_{\mathbf{n}}(\mathbf{z}) := \prod_{i=1}^{d} \psi_{n_i}(z_i), \qquad \rho^{\mathbf{n}} := \prod_{i=1}^{d} \rho_i^{n_i}.$$

Then the coordinate pairs (u_i, v_i) *are mutually independent across* $i = 1, \ldots, d$*, and the joint density of* $(\mathbf{u}, \mathbf{v})$ *is*

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u}, \mathbf{v}) = \frac{1}{\pi^d} e^{-(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)} \sum_{\mathbf{n} \in \mathbb{N}_0^d} \rho^{\mathbf{n}}\, \Psi_{\mathbf{n}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v}). \tag{6}$$

Moreover, for all multi-indices $\mathbf{m}, \mathbf{n} \in \mathbb{N}_0^d$,

$$\mathbb{E}[\Psi_{\mathbf{m}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v})] = \rho^{\mathbf{n}}\, \delta_{\mathbf{mn}}, \qquad \delta_{\mathbf{mn}} := \prod_{i=1}^{d} \delta_{m_i n_i}.$$

Proof. Because $(\sqrt{2}\,\mathbf{u}, \sqrt{2}\,\mathbf{v})$ is jointly Gaussian with block covariance

$$\begin{pmatrix} \mathbf{I}_d & \mathrm{diag}(\rho_1, \ldots, \rho_d) \\ \mathrm{diag}(\rho_1, \ldots, \rho_d) & \mathbf{I}_d \end{pmatrix},$$

the pairs (u_i, v_i) are mutually independent across i. Hence

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{d} p_{u_i, v_i}(u_i, v_i).$$

Applying Lemma 2 coordinatewise,

$$p_{u_i, v_i}(u_i, v_i) = \frac{1}{\pi} e^{-(u_i^2 + v_i^2)} \sum_{n_i = 0}^{\infty} \rho_i^{n_i} \psi_{n_i}(u_i) \psi_{n_i}(v_i).$$

Taking the product over $i = 1, \ldots, d$ yields

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u}, \mathbf{v}) = \frac{1}{\pi^d} e^{-(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)} \prod_{i=1}^{d} \left(\sum_{n_i = 0}^{\infty} \rho_i^{n_i} \psi_{n_i}(u_i) \psi_{n_i}(v_i) \right),$$

which expands into Equation 6. The covariance identity follows by Fubini and the one-dimensional orthogonality:

$$\begin{aligned} \mathbb{E}[\Psi_{\mathbf{m}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v})] &= \prod_{i=1}^{d} \mathbb{E}[\psi_{m_i}(u_i) \psi_{n_i}(v_i)] \\ &= \prod_{i=1}^{d} \rho_i^{n_i} \delta_{m_i n_i} = \rho^{\mathbf{n}}\, \delta_{\mathbf{mn}}. \end{aligned}$$

□

Corollary 2 (Orthogonality of multivariate normalized Hermite polynomials)**.** *The family*

$$\left\{ \Psi_{\mathbf{n}} : \mathbf{n} \in \mathbb{N}_0^d \right\}$$

is an orthonormal basis of $L^2(\nu_d)$*, where*

$$\nu_d(d\mathbf{z}) = \pi^{-d/2} e^{-\|\mathbf{z}\|^2}\, d\mathbf{z}.$$

In particular,

$$\Psi_{\mathbf{0}} \equiv 1, \qquad \Psi_{\mathbf{e}_i}(\mathbf{z}) = \psi_1(z_i) = \sqrt{2}\, z_i, \quad i = 1, \ldots, d.$$

Proof. For $\mathbf{m}, \mathbf{n} \in \mathbb{N}_0^d$,

$$\begin{aligned}\int_{\mathbb{R}^d} \Psi_{\mathbf{m}}(\mathbf{z})\Psi_{\mathbf{n}}(\mathbf{z})\,\nu_d(d\mathbf{z}) &= \prod_{i=1}^{d} \int_{\mathbb{R}} \psi_{m_i}(z_i)\psi_{n_i}(z_i)\,\nu(dz_i) \\ &= \prod_{i=1}^{d} \delta_{m_i n_i} = \delta_{\mathbf{mn}}.\end{aligned}$$

Completeness follows from the standard tensor-product construction of orthonormal bases. The identity for $\Psi_{\mathbf{e}_i}$ uses $H_1(z) = 2z$. □

Lemma 3 (Diagonal selection under orthonormal row constraints)**.** *Let $N \in \mathbb{N}$, let $\{d_j\}_{j\geq 1}$ be a nonnegative nonincreasing sequence, and let*

$$\mathbf{A} = (a_{ij})_{1\leq i\leq N,\, j\geq 1}, \qquad \mathbf{B} = (b_{ij})_{1\leq i\leq N,\, j\geq 1}$$

be real matrices with countably many columns such that

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_N, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I}_N.$$

Let

$$\mathbf{D} = \operatorname{diag}(d_1, d_2, \dots).$$

Then

$$\left\|\mathbf{A}\mathbf{D}\mathbf{B}^\top\right\|_* \;\leq\; \sum_{i=1}^{N} d_i.$$

Moreover, equality is attained, for example, by

$$\mathbf{A} = \mathbf{B} = \begin{bmatrix}\mathbf{I}_N & 0 & 0 & \cdots\end{bmatrix}.$$

Proof. Let $\mathbf{a}_j, \mathbf{b}_j \in \mathbb{R}^N$ denote the j-th columns of $\mathbf{A}^\top$ and $\mathbf{B}^\top$, respectively. Then

$$\sum_{j\geq 1} \mathbf{a}_j\mathbf{a}_j^\top = \mathbf{I}_N, \qquad \sum_{j\geq 1} \mathbf{b}_j\mathbf{b}_j^\top = \mathbf{I}_N,$$

hence

$$\sum_{j\geq 1} \|\mathbf{a}_j\|^2 = N, \qquad \sum_{j\geq 1} \|\mathbf{b}_j\|^2 = N.$$

Also, since $\mathbf{A}^\top\mathbf{A} \preceq \mathbf{I}$ and $\mathbf{B}^\top\mathbf{B} \preceq \mathbf{I}$, we have

$$0 \leq \|\mathbf{a}_j\|^2 \leq 1, \qquad 0 \leq \|\mathbf{b}_j\|^2 \leq 1 \qquad \text{for all } j \geq 1.$$

By the variational characterization of the nuclear norm,

$$\left\|\mathbf{A}\mathbf{D}\mathbf{B}^\top\right\|_* = \max_{\mathbf{R},\mathbf{S}\in O(N)} \operatorname{tr}\big(\mathbf{R}\mathbf{A}\mathbf{D}\mathbf{B}^\top\mathbf{S}^\top\big) = \max_{\mathbf{R},\mathbf{S}\in O(N)} \sum_{j\geq 1} d_j \,\langle \mathbf{R}\mathbf{a}_j, \mathbf{S}\mathbf{b}_j\rangle.$$

Using Cauchy–Schwarz and $2ab \leq a^2 + b^2$,

$$\langle \mathbf{R}\mathbf{a}_j, \mathbf{S}\mathbf{b}_j\rangle \leq \|\mathbf{a}_j\|\,\|\mathbf{b}_j\| \leq \frac{1}{2}\Big(\|\mathbf{a}_j\|^2 + \|\mathbf{b}_j\|^2\Big).$$

Therefore

$$\left\|\mathbf{A}\mathbf{D}\mathbf{B}^\top\right\|_* \leq \frac{1}{2}\sum_{j\geq 1} d_j\|\mathbf{a}_j\|^2 + \frac{1}{2}\sum_{j\geq 1} d_j\|\mathbf{b}_j\|^2.$$

Now, if $\{u_j\}_{j\geq 1}$ satisfies $0 \leq u_j \leq 1$ and $\sum_{j\geq 1} u_j = N$, then the monotonicity of $\{d_j\}_{j\geq 1}$ implies

$$\sum_{j\geq 1} d_j u_j \leq \sum_{i=1}^{N} d_i;$$

indeed, the maximum is achieved by taking $u_1 = \cdots = u_N = 1$ and $u_j = 0$ for $j > N$. Applying this with $u_j = \|\mathbf{a}_j\|^2$ and $u_j = \|\mathbf{b}_j\|^2$ yields

$$\sum_{j\geq 1} d_j \|\mathbf{a}_j\|^2 \leq \sum_{i=1}^{N} d_i, \qquad \sum_{j\geq 1} d_j \|\mathbf{b}_j\|^2 \leq \sum_{i=1}^{N} d_i.$$

Hence

$$\left\|\mathbf{A}\mathbf{D}\mathbf{B}^\top\right\|_* \leq \sum_{i=1}^{N} d_i.$$

Finally, for $\mathbf{A} = \mathbf{B} = \begin{bmatrix}\mathbf{I}_N & 0 & 0 & \cdots\end{bmatrix}$ we have

$$\mathbf{A}\mathbf{D}\mathbf{B}^\top = \operatorname{diag}(d_1, \ldots, d_N),$$

whose nuclear norm is exactly $\sum_{i=1}^{N} d_i$. □

Corollary 3 (Strict-gap characterization of maximizers)**.** *Under the assumptions of Lemma 3, assume in addition that*

$$d_N > d_{N+1}.$$

If

$$\left\|\mathbf{A}\mathbf{D}\mathbf{B}^\top\right\|_* = \sum_{i=1}^{N} d_i,$$

then

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_1 & 0 & 0 & \cdots\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}\mathbf{B}_1 & 0 & 0 & \cdots\end{bmatrix}$$

for some $\mathbf{A}_1, \mathbf{B}_1 \in O(N)$.

Proof. If equality holds in Lemma 3, then equality must hold in both weighted-sum bounds

$$\sum_{j\geq 1} d_j \|\mathbf{a}_j\|^2 \leq \sum_{i=1}^{N} d_i, \qquad \sum_{j\geq 1} d_j \|\mathbf{b}_j\|^2 \leq \sum_{i=1}^{N} d_i.$$

Because $d_N > d_{N+1}$ and

$$0 \leq \|\mathbf{a}_j\|^2 \leq 1, \qquad \sum_{j\geq 1} \|\mathbf{a}_j\|^2 = N,$$

the only way to achieve

$$\sum_{j\geq 1} d_j \|\mathbf{a}_j\|^2 = \sum_{i=1}^{N} d_i$$

is to place all mass on the first N coordinates, i.e.

$$\|\mathbf{a}_j\|^2 = 1 \text{ for } 1 \leq j \leq N, \qquad \|\mathbf{a}_j\|^2 = 0 \text{ for } j > N.$$

The same argument gives

$$\|\mathbf{b}_j\|^2 = 1 \text{ for } 1 \leq j \leq N, \qquad \|\mathbf{b}_j\|^2 = 0 \text{ for } j > N.$$

Hence all columns indexed by $j > N$ vanish, so

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_1 & 0 & 0 & \cdots\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}\mathbf{B}_1 & 0 & 0 & \cdots\end{bmatrix}$$

with $\mathbf{A}_1, \mathbf{B}_1 \in \mathbb{R}^{N\times N}$. Since $\mathbf{A}\mathbf{A}^\top = \mathbf{B}\mathbf{B}^\top = \mathbf{I}_N$, we obtain

$$\mathbf{A}_1\mathbf{A}_1^\top = \mathbf{I}_N, \qquad \mathbf{B}_1\mathbf{B}_1^\top = \mathbf{I}_N.$$

Because $\mathbf{A}_1$ and $\mathbf{B}_1$ are square, this means $\mathbf{A}_1, \mathbf{B}_1 \in O(N)$. □

Proof of Theorem 1. By Proposition 1, it suffices to characterize the maximizers of the source-space objective over

$$\hat{\mathcal{F}}_{\mathcal{S}} \times \hat{\mathcal{F}}'_{\mathcal{S}}.$$

Let

$$\mathbf{K} := \mathbf{\Sigma}_{\mathbf{ss}}^{-1/2} \mathbf{\Sigma}_{\mathbf{ss}'} \mathbf{\Sigma}_{\mathbf{s}'\mathbf{s}'}^{-1/2} = \mathbf{U}_0 \, \mathrm{diag}(\rho_1, \dots, \rho_d) \, \mathbf{V}_0^\top$$

be a singular value decomposition, where $\mathbf{U}_0, \mathbf{V}_0 \in O(d)$ and

$$1 > \rho_1 \geq \cdots \geq \rho_d > 0.$$

Define the canonical Gaussian coordinates

$$\mathbf{u} := \frac{1}{\sqrt{2}} \, \mathbf{U}_0^\top \mathbf{\Sigma}_{\mathbf{ss}}^{-1/2} (\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}), \qquad \mathbf{v} := \frac{1}{\sqrt{2}} \, \mathbf{V}_0^\top \mathbf{\Sigma}_{\mathbf{s}'\mathbf{s}'}^{-1/2} (\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}).$$

Then $\sqrt{2}\,\mathbf{u}$ and $\sqrt{2}\,\mathbf{v}$ are standard Gaussian, and

$$\mathrm{Cov}(\sqrt{2}\,\mathbf{u}, \sqrt{2}\,\mathbf{v}) = \mathrm{diag}(\rho_1, \dots, \rho_d).$$

Hence Proposition 4 applies.

The maps

$$\mathbf{s} = \boldsymbol{\mu}_{\mathbf{s}} + \sqrt{2}\, \mathbf{\Sigma}_{\mathbf{ss}}^{1/2} \mathbf{U}_0 \, \mathbf{u}, \qquad \mathbf{s}' = \boldsymbol{\mu}_{\mathbf{s}'} + \sqrt{2}\, \mathbf{\Sigma}_{\mathbf{s}'\mathbf{s}'}^{1/2} \mathbf{V}_0 \, \mathbf{v}$$

are affine bijections. Therefore, maximizing over centered, whitened maps of $\mathbf{s}$ and $\mathbf{s}'$ is equivalent to maximizing over centered, whitened maps of $\mathbf{u}$ and $\mathbf{v}$.

Let $(\bar{\mathbf{h}}, \bar{\mathbf{h}}')$ be any feasible pair in these canonical coordinates. Since $\{\Psi_{\mathbf{n}}\}_{\mathbf{n} \in \mathbb{N}_0^d}$ is an orthonormal basis of $L^2(\nu_d)$ by Corollary 2, each component admits an L^2 expansion

$$\bar{h}_r(\mathbf{u}) = \sum_{\mathbf{n} \in \mathcal{I}} a_{r,\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{u}), \qquad \bar{h}'_q(\mathbf{v}) = \sum_{\mathbf{n} \in \mathcal{I}} b_{q,\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{v}), \qquad \mathcal{I} := \mathbb{N}_0^d \setminus \{\mathbf{0}\},$$

for $r, q = 1, \dots, d$. There is no $\mathbf{n} = \mathbf{0}$ term because the maps are centered. Let

$$\mathbf{A} = (a_{r,\mathbf{n}})_{1 \leq r \leq d,\, \mathbf{n} \in \mathcal{I}}, \qquad \mathbf{B} = (b_{q,\mathbf{n}})_{1 \leq q \leq d,\, \mathbf{n} \in \mathcal{I}}$$

denote the corresponding coefficient matrices. Orthogonality of the basis and the whitening constraints imply

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_d, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I}_d.$$

Next, by Proposition 4,

$$\mathbb{E}[\Psi_{\mathbf{m}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v})] = \rho^{\mathbf{n}} \delta_{\mathbf{mn}}, \qquad \rho^{\mathbf{n}} := \prod_{i=1}^{d} \rho_i^{n_i}.$$

Hence the cross-covariance matrix of $(\bar{\mathbf{h}}, \bar{\mathbf{h}}')$ is

$$\mathbf{C} := \mathrm{Cov}\big(\bar{\mathbf{h}}(\mathbf{u}), \bar{\mathbf{h}}'(\mathbf{v})\big) = \mathbf{A} \, \mathrm{diag}\big(\rho^{\mathbf{n}}\big)_{\mathbf{n} \in \mathcal{I}} \, \mathbf{B}^\top. \tag{7}$$

Now enumerate the multi-indices in $\mathcal{I}$ as $\mathcal{I} = \{\mathbf{n}^{(1)}, \mathbf{n}^{(2)}, \dots\}$ so that

$$d_j := \rho^{\mathbf{n}^{(j)}}$$

is nonincreasing. Every first-order multi-index $\mathbf{e}_i$ gives

$$\rho^{\mathbf{e}_i} = \rho_i.$$

On the other hand, if $|\mathbf{n}| := \sum_{i=1}^d n_i \geq 2$, then

$$\rho^{\mathbf{n}} = \prod_{i=1}^{d} \rho_i^{n_i} \leq \rho_1^{|\mathbf{n}|} \leq \rho_1^2 < \rho_d$$

by Assumption 3. Therefore the top d diagonal entries of (7) are exactly

$$\rho_1, \dots, \rho_d,$$

corresponding to the first-order indices $\mathbf{e}_1, \dots, \mathbf{e}_d$, and moreover there is a strict spectral gap

$$d_d = \rho_d > d_{d+1}.$$

Applying Lemma 3 with $N = d$ to (7) yields

$$J_{\mathcal{S}}(\bar{\mathbf{h}}, \bar{\mathbf{h}}') = \|\mathbf{C}\|_* \le \sum_{i=1}^{d} \rho_i.$$

This upper bound is attained by the purely first-order maps

$$\bar{\mathbf{h}}_{\text{lin}}(\mathbf{u}) = \sqrt{2}\,\mathbf{u}, \qquad \bar{\mathbf{h}}'_{\text{lin}}(\mathbf{v}) = \sqrt{2}\,\mathbf{v},$$

because these maps are centered and whitened, and their cross-covariance is

$$\text{Cov}(\sqrt{2}\,\mathbf{u}, \sqrt{2}\,\mathbf{v}) = \text{diag}(\rho_1, \dots, \rho_d),$$

whose singular values are exactly $\rho_1, \dots, \rho_d$. Hence every population maximizer must satisfy

$$\|\mathbf{C}\|_* = \sum_{i=1}^{d} \rho_i.$$

By the strict-gap characterization in Corollary 3, this forces all coefficients outside the first-order block to vanish. Thus there exist orthogonal matrices $\mathbf{A}_1, \mathbf{B}_1 \in O(d)$ such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & 0 & 0 & \cdots \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & 0 & 0 & \cdots \end{bmatrix}.$$

Using Corollary 2,

$$\Psi_{\mathbf{e}_i}(\mathbf{u}) = \sqrt{2}\,u_i, \qquad \Psi_{\mathbf{e}_i}(\mathbf{v}) = \sqrt{2}\,v_i.$$

Therefore every maximizer has the form

$$\bar{\mathbf{h}}^*(\mathbf{u}) = \mathbf{A}_1(\sqrt{2}\,\mathbf{u}), \qquad \bar{\mathbf{h}}'^*(\mathbf{v}) = \mathbf{B}_1(\sqrt{2}\,\mathbf{v}).$$

Substituting the definitions of $\mathbf{u}$ and $\mathbf{v}$ gives

$$\bar{\mathbf{h}}^*(\mathbf{u}) = \mathbf{A}_1 \mathbf{U}_0^\top \boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2}(\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}),$$

$$\bar{\mathbf{h}}'^*(\mathbf{v}) = \mathbf{B}_1 \mathbf{V}_0^\top \boldsymbol{\Sigma}_{\mathbf{s's'}}^{-1/2}(\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}).$$

Since $\mathbf{A}_1\mathbf{U}_0^\top$ and $\mathbf{B}_1\mathbf{V}_0^\top$ are orthogonal, there exist $\mathbf{Q}, \mathbf{Q}' \in O(d)$ such that

$$\tilde{\mathbf{h}}^*(\mathbf{s}) = \mathbf{Q}\,\boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2}(\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}), \qquad \tilde{\mathbf{h}}'^*(\mathbf{s}') = \mathbf{Q}'\,\boldsymbol{\Sigma}_{\mathbf{s's'}}^{-1/2}(\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}).$$

Recalling that

$$\tilde{\mathbf{h}}^* = \tilde{\mathbf{f}}^* \circ \mathbf{g}, \qquad \tilde{\mathbf{h}}'^* = \tilde{\mathbf{f}}'^* \circ \mathbf{g}',$$

this proves the whitened statement in Theorem 1. Finally, if

$$\tilde{\mathbf{h}}^* = \mathbf{W}_{\mathbf{h}}\big(\mathbf{h}^* - \mathbb{E}[\mathbf{h}^*]\big), \qquad \tilde{\mathbf{h}}'^* = \mathbf{W}_{\mathbf{h}'}\big(\mathbf{h}'^* - \mathbb{E}[\mathbf{h}'^*]\big)$$

are the corresponding population-whitened maps, then

$$\mathbf{h}^*(\mathbf{s}) = \mathbf{W}_{\mathbf{h}}^{-1}\mathbf{Q}\,\boldsymbol{\Sigma}_{\mathbf{ss}}^{-1/2}(\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}) + \mathbb{E}[\mathbf{h}^*],$$

$$\mathbf{h}'^*(\mathbf{s}') = \mathbf{W}_{\mathbf{h}'}^{-1}\mathbf{Q}'\,\boldsymbol{\Sigma}_{\mathbf{s's'}}^{-1/2}(\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}) + \mathbb{E}[\mathbf{h}'^*].$$

Hence the unwhitened representation maps are affine, which proves the theorem. □

Why the strict gap is needed. The argument above uses the strict inequality

$$\rho_d > \rho_1^2$$

exactly once: it ensures that the first-order block $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ is separated from every higher-order multi-index. If this fails, then a second-order term such as $2\mathbf{e}_1$ can have coefficient $\rho_1^2 \geq \rho_d$, and a population maximizer need not be supported entirely on the first-order block. In that regime, nonlinear CCA can mix linear and higher-order terms, and strict affine identifiability is no longer guaranteed.

Corollary 4 (Single-orbit structure of population maximizers under affine identifiability)**.** *Assume the hypotheses of Theorem 1, and let $(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star)$ be any population maximizer of Equation 2. Then the full set of population maximizers is exactly*

$$\left\{(\mathbf{Q}\tilde{\mathbf{f}}^\star, \mathbf{Q}'\tilde{\mathbf{f}}'^\star) : \mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})\right\}$$

as a subset of $L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}}) \times L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$. In particular, under the assumptions of Theorem 1, all population maximizers lie in a single $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit.

Proof. Let $(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}')$ be any other population maximizer of Equation 2. By Theorem 1, there exist orthogonal matrices $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_1', \mathbf{R}_2' \in O(d_{\mathcal{Z}})$ such that

$$(\tilde{\mathbf{f}}^\star \circ \mathbf{g})(\mathbf{s}) = \mathbf{R}_1\, \mathbf{\Sigma}_{\mathbf{ss}}^{-1/2}(\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}), \qquad (\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s}) = \mathbf{R}_2\, \mathbf{\Sigma}_{\mathbf{ss}}^{-1/2}(\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}}),$$

and similarly

$$(\tilde{\mathbf{f}}'^\star \circ \mathbf{g}')(\mathbf{s}') = \mathbf{R}_1'\, \mathbf{\Sigma}_{\mathbf{s's'}}^{-1/2}(\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}), \qquad (\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}') = \mathbf{R}_2'\, \mathbf{\Sigma}_{\mathbf{s's'}}^{-1/2}(\mathbf{s}' - \boldsymbol{\mu}_{\mathbf{s}'}).$$

Define

$$\mathbf{Q} := \mathbf{R}_2\mathbf{R}_1^\top \in O(d_{\mathcal{Z}}), \qquad \mathbf{Q}' := \mathbf{R}_2'\mathbf{R}_1'^\top \in O(d_{\mathcal{Z}}).$$

Then

$$(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s}) = \mathbf{Q}\,(\tilde{\mathbf{f}}^\star \circ \mathbf{g})(\mathbf{s}), \qquad (\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}') = \mathbf{Q}'\,(\tilde{\mathbf{f}}'^\star \circ \mathbf{g}')(\mathbf{s}').$$

By the pullback isometry from Lemma 1.1,

$$\|\tilde{\mathbf{f}} - \mathbf{Q}\tilde{\mathbf{f}}^\star\|_{L^2(P_{\mathbf{x}})} = \|(\tilde{\mathbf{f}} - \mathbf{Q}\tilde{\mathbf{f}}^\star) \circ \mathbf{g}\|_{L^2(P_{\mathbf{s}})} = 0,$$

and likewise

$$\|\tilde{\mathbf{f}}' - \mathbf{Q}'\tilde{\mathbf{f}}'^\star\|_{L^2(P_{\mathbf{x}'})} = \|(\tilde{\mathbf{f}}' - \mathbf{Q}'\tilde{\mathbf{f}}'^\star) \circ \mathbf{g}'\|_{L^2(P_{\mathbf{s}'})} = 0.$$

Thus every population maximizer lies in the $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$-orbit of $(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star)$.

The reverse inclusion is exactly Corollary 1. □

D.4 Preservation of Higher-Order Nonlinearities Beyond First-Order Dominance

In the main text, Assumption 3 requires first-order canonical dominance, specifically $\rho_{d_{\mathcal{S}}} > \rho_1^2$, to guarantee strictly affine identifiability. This section provides a self-contained illustration of the representation learning dynamics when this assumption is relaxed. Specifically, we demonstrate how the canonical spectrum interacts with the representation capacity $d_{\mathcal{Z}}$ to determine the survival of higher-order nonlinearities.

Theoretical Mechanism. As established in the proof of Theorem 1, the normalized orthogonal polynomial expansion (e.g., the Mehler expansion for Gaussian priors) diagonalizes the CCA objective. Under this expansion, a multivariate polynomial term of multi-index degree $\mathbf{n} = (n_1, \dots, n_{d_{\mathcal{S}}})$ corresponds to a canonical correlation given by $\prod_{i=1}^{d_{\mathcal{S}}} \rho_i^{n_i}$. Because nonlinear CCA strictly prioritizes orthogonal components that maximize the sum of inter-view correlations, it greedily selects the $d_{\mathcal{Z}}$ terms with the largest corresponding values of $\prod_{i=1}^{d_{\mathcal{S}}} \rho_i^{n_i}$, regardless of their polynomial degree. Consequently, higher-order nonlinear terms of strongly correlated source components can overshadow the first-order (linear) terms of more weakly correlated components.

Infinite Capacity Guarantee. Because the true canonical correlations satisfy $0 < \rho_i < 1$ for all $i \in \{1, \dots, d_{\mathcal{S}}\}$, the correlation of any higher-order term diminishes exponentially as its total polynomial degree $|\mathbf{n}| \to \infty$. Therefore, while a limited capacity $d_{\mathcal{Z}}$ can lead to the truncation of weakly correlated linear terms, the asymptotic limit of infinite representation capacity ($d_{\mathcal{Z}} \to \infty$) guarantees the eventual recovery of all first-order linear terms. However, in this unconstrained regime, the desired linear features are inherently entangled with a multitude of higher-order artifacts.

A Concrete Example. To rigorously illustrate this phenomenon, consider a two-dimensional latent space ($d_{\mathcal{S}} = 2$) with ground-truth canonical correlations $\rho_1 = 0.9$ and $\rho_2 = 0.5$. Notice that this violates the first-order dominance condition since $\rho_2 < \rho_1^2$ ($0.5 < 0.81$).

Evaluating the canonical correlations for the lowest-degree polynomial terms yields:

- First-order terms: s_1 yields $\rho_1 = 0.9$, and s_2 yields $\rho_2 = 0.5$.
- Higher-order terms of s_1: s_1^2 yields $\rho_1^2 = 0.81$, s_1^3 yields $\rho_1^3 = 0.729$, s_1^4 yields $\rho_1^4 = 0.6561$, s_1^5 yields $\rho_1^5 = 0.59049$ and s_1^6 yields $\rho_1^6 = 0.531441$.

Sorting these available orthogonal terms in descending order of their canonical correlations generates the sequence: $\{0.9, 0.81, 0.729, 0.6561, 0.59049, 0.531441, 0.5, \dots\}$.

If we constrain the representation capacity to match the source dimension ($d_{\mathcal{Z}} = 2$), nonlinear CCA greedily selects the top two components. In this case, it captures $\{s_1, s_1^2\}$, introducing a second-order nonlinearity of s_1 and entirely failing to recover the linear term s_2.

If the representation capacity is expanded to $d_{\mathcal{Z}} = 7$, the algorithm selects the top seven components: $\{s_1, s_1^2, s_1^3, s_1^4, s_1^5, s_1^6, s_2\}$. Here, the linear term s_2 is successfully recovered, consistent with the infinite capacity guarantee, but the learned representation is now dominated by higher-order nonlinearities up to the sixth degree. This example clearly demonstrates why the strict spectral gap $\rho_{d_{\mathcal{S}}} > \rho_1^2$ is mathematically necessary to isolate purely affine representations under finite capacity.

D.5 Proof of Theorem 2

We write $\|\cdot\|$ for the spectral norm and $\|\cdot\|_*$ for the nuclear norm.

Lemma 4 (Stability of ridge whitening near the identity)**.** *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be symmetric positive semidefinite and assume $\|\mathbf{A} - \mathbf{I}_d\| \le \frac{1}{2}$. Then, for every $\epsilon \in [0, 1]$,*

$$\big\|(\mathbf{A} + \epsilon\mathbf{I}_d)^{-1/2} - \mathbf{I}_d\big\| \;\le\; C\,\|\mathbf{A} - \mathbf{I}_d\| + \big|(1+\epsilon)^{-1/2} - 1\big|,$$

where $C > 0$ is an absolute constant. In particular,

$$\big\|(\mathbf{A} + \epsilon\mathbf{I}_d)^{-1/2} - \mathbf{I}_d\big\| = O\big(\|\mathbf{A} - \mathbf{I}_d\| + \epsilon\big).$$

Proof. Since $\mathbf{A}$ is symmetric, it admits an eigendecomposition $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \dots, \lambda_d)\mathbf{U}^\top$. The condition $\|\mathbf{A} - \mathbf{I}_d\| \le \frac{1}{2}$ implies $\lambda_i \in [\frac{1}{2}, \frac{3}{2}]$ for all i. Therefore

$$(\mathbf{A} + \epsilon\mathbf{I}_d)^{-1/2} = \mathbf{U}\,\mathrm{diag}\big((\lambda_i + \epsilon)^{-1/2}\big)_{i=1}^{d}\mathbf{U}^\top,$$

and hence

$$\big\|(\mathbf{A} + \epsilon\mathbf{I}_d)^{-1/2} - (1+\epsilon)^{-1/2}\mathbf{I}_d\big\| = \max_{1\le i\le d}\big|(\lambda_i + \epsilon)^{-1/2} - (1+\epsilon)^{-1/2}\big|.$$

Now apply the mean value theorem to the scalar map $t \mapsto (t+\epsilon)^{-1/2}$ on $[\frac{1}{2}, \frac{3}{2}]$:

$$\big|(\lambda_i + \epsilon)^{-1/2} - (1+\epsilon)^{-1/2}\big| \le \sup_{t\in[\frac{1}{2},\frac{3}{2}],\,\epsilon\in[0,1]} \frac{1}{2(t+\epsilon)^{3/2}}\,|\lambda_i - 1| \le C\,|\lambda_i - 1|.$$

Taking the maximum over i gives

$$\big\|(\mathbf{A} + \epsilon\mathbf{I}_d)^{-1/2} - (1+\epsilon)^{-1/2}\mathbf{I}_d\big\| \le C\,\|\mathbf{A} - \mathbf{I}_d\|.$$

Finally,

$$\left\|(1+\epsilon)^{-1/2}\mathbf{I}_d - \mathbf{I}_d\right\| = \left|(1+\epsilon)^{-1/2} - 1\right|,$$

and the claim follows by the triangle inequality. □

Proof. For each n, let

$$(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) := (\tilde{\mathbf{f}}_{\boldsymbol{\theta}_n}, \tilde{\mathbf{f}}'_{\boldsymbol{\theta}'_n})$$

be any empirical δ_n-maximizer from Assumption A4. Fix a feasible pair

$$(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}.$$

Write

$$\mathbf{z} := \tilde{\mathbf{f}}(\mathbf{x}), \qquad \mathbf{z}' := \tilde{\mathbf{f}}'(\mathbf{x}').$$

Since the feasible classes in Assumption 2 are whitened,

$$\boldsymbol{\Sigma}_{\mathbf{zz}} = \mathbf{I}_{d_{\mathcal{Z}}}, \qquad \boldsymbol{\Sigma}_{\mathbf{z}'\mathbf{z}'} = \mathbf{I}_{d_{\mathcal{Z}}}.$$

Hence the population normalized cross-covariance simplifies to

$$\mathbf{K} = \boldsymbol{\Sigma}_{\mathbf{zz}'}, \qquad J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = \|\mathbf{K}\|_*.$$

For the empirical quantities, define

$$\hat{\mathbf{W}}_{\mathbf{z}} := (\hat{\boldsymbol{\Sigma}}_{\mathbf{zz}} + \epsilon\mathbf{I})^{-1/2}, \qquad \hat{\mathbf{W}}_{\mathbf{z}'} := (\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'} + \epsilon\mathbf{I})^{-1/2},$$

and

$$\hat{\mathbf{K}} := \hat{\mathbf{W}}_{\mathbf{z}}\, \hat{\boldsymbol{\Sigma}}_{\mathbf{zz}'}\, \hat{\mathbf{W}}_{\mathbf{z}'}.$$

Then

$$\hat{J}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = \|\hat{\mathbf{K}}\|_*.$$

Step 1: stability of empirical whitening. Let

$$\Delta_n := \sup_{\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}},\, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}} \Big(\|\hat{\boldsymbol{\Sigma}}_{\mathbf{zz}} - \mathbf{I}\| \vee \|\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'} - \mathbf{I}\| \vee \|\hat{\boldsymbol{\Sigma}}_{\mathbf{zz}'} - \boldsymbol{\Sigma}_{\mathbf{zz}'}\| \Big).$$

By Assumption A3,

$$\Delta_n = o_p(\epsilon),$$

hence also $\Delta_n = o_p(1)$ because $\epsilon \to 0$.

On the event $\{\Delta_n \le \frac{1}{2}\}$, every eigenvalue of $\hat{\boldsymbol{\Sigma}}_{\mathbf{zz}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}'\mathbf{z}'}$ lies in $[\frac{1}{2}, \frac{3}{2}]$, uniformly over the feasible classes. Therefore Lemma 4 yields

$$\sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \left\|\hat{\mathbf{W}}_{\mathbf{z}} - \mathbf{I}\right\| \le C \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{zz}} - \mathbf{I}\right\| + \left|(1+\epsilon)^{-1/2} - 1\right| = o_p(\epsilon) + O(\epsilon) = o_p(1),$$

and similarly

$$\sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \left\|\hat{\mathbf{W}}_{\mathbf{z}'} - \mathbf{I}\right\| = o_p(1).$$

In particular,

$$\sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \left\|\hat{\mathbf{W}}_{\mathbf{z}}\right\| + \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \left\|\hat{\mathbf{W}}_{\mathbf{z}'}\right\| = O_p(1).$$

Step 2: uniform convergence of the normalized cross-covariance. For every feasible pair, $\mathbf{z}$ and $\mathbf{z}'$ are whitened. Hence

$$\|\mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'}\| \le 1.$$

Indeed, for arbitrary unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d_{\mathcal{Z}}}$,

$$\left|\mathbf{u}^\top \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'} \mathbf{v}\right| = \left|\mathrm{Cov}(\mathbf{u}^\top \mathbf{z}, \mathbf{v}^\top \mathbf{z}')\right| \le \sqrt{\mathrm{Var}(\mathbf{u}^\top \mathbf{z})\,\mathrm{Var}(\mathbf{v}^\top \mathbf{z}')} = 1,$$

so the operator norm is at most 1. Therefore, on the event $\{\Delta_n \le \frac{1}{2}\}$,

$$\sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'}\| \le \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'}\| + \Delta_n \le 1 + \Delta_n \le \frac{3}{2}.$$

Now expand

$$\begin{aligned}
\hat{\mathbf{K}} - \mathbf{K} &= \hat{\mathbf{W}}_{\mathbf{z}}\, \hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'}\, \hat{\mathbf{W}}_{\mathbf{z}'} - \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'} \\
&= (\hat{\mathbf{W}}_{\mathbf{z}} - \mathbf{I})\, \hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'}\, \hat{\mathbf{W}}_{\mathbf{z}'} + (\hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'} - \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'})\, \hat{\mathbf{W}}_{\mathbf{z}'} + \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'}\, (\hat{\mathbf{W}}_{\mathbf{z}'} - \mathbf{I}).
\end{aligned}$$

Taking suprema over the feasible classes and using Step 1 gives

$$\begin{aligned}
\sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{K}} - \mathbf{K}\| &\le \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{W}}_{\mathbf{z}} - \mathbf{I}\| \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'}\| \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{W}}_{\mathbf{z}'}\| \\
&\quad + \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{\Sigma}}_{\mathbf{z}\mathbf{z}'} - \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'}\| \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{W}}_{\mathbf{z}'}\| \\
&\quad + \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\mathbf{\Sigma}_{\mathbf{z}\mathbf{z}'}\| \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{W}}_{\mathbf{z}'} - \mathbf{I}\| \\
&= o_p(1).
\end{aligned}$$

Step 3: objective consistency. Since the population and empirical objectives are the nuclear norms of $\mathbf{K}$ and $\hat{\mathbf{K}}$, respectively,

$$\left|\hat{J}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') - J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}')\right| = \left|\|\hat{\mathbf{K}}\|_* - \|\mathbf{K}\|_*\right| \le \|\hat{\mathbf{K}} - \mathbf{K}\|_*.$$

Using $\|\mathbf{A}\|_* \le d_{\mathcal{Z}} \|\mathbf{A}\|$ for $\mathbf{A} \in \mathbb{R}^{d_{\mathcal{Z}} \times d_{\mathcal{Z}}}$, we obtain

$$\sup_{\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}},\, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}} \left|\hat{J}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') - J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}')\right| \le d_{\mathcal{Z}} \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \|\hat{\mathbf{K}} - \mathbf{K}\| \xrightarrow{\mathbb{P}} 0.$$

This proves Claim 1.

Step 4: estimator consistency up to orthogonal transformations. Let

$$(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star) \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$$

be any population maximizer, and write

$$J^\star := J(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star).$$

Define its orbit

$$\mathcal{O}^\star := \left\{(\mathbf{Q}\tilde{\mathbf{f}}^\star, \mathbf{Q}'\tilde{\mathbf{f}}'^\star) :\ \mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})\right\}.$$

We equip the quotient by this orthogonal action with the distance

$$d_{\mathcal{Q}}\big((\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'), \mathcal{O}^\star\big) := \inf_{\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})} \Big(\|\mathbf{Q}\tilde{\mathbf{f}} - \tilde{\mathbf{f}}^\star\|_{L^2(P_{\mathbf{x}})} + \|\mathbf{Q}'\tilde{\mathbf{f}}' - \tilde{\mathbf{f}}'^\star\|_{L^2(P_{\mathbf{x}'})}\Big).$$

By Assumption A4,

$$\hat{J}(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) \ge \hat{J}(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star) - \delta_n.$$

Therefore

$$\begin{aligned} 0 \le J^\star - J(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) &\le \big|J(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star) - \hat{J}(\tilde{\mathbf{f}}^\star, \tilde{\mathbf{f}}'^\star)\big| \\ &\quad + \big|\hat{J}(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) - J(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n)\big| + \delta_n \\ &\le 2 \sup_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'} \big|\hat{J}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') - J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}')\big| + \delta_n \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

where the last step uses Claim 1 and $\delta_n \to 0$.

We now invoke the orbit-separation part of Assumption A1. If

$$d_{\mathcal{Q}}\big((\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n), \mathcal{O}^\star\big) \not\xrightarrow{\mathbb{P}} 0,$$

then there exist $\eta, \alpha > 0$ and a subsequence (not relabeled) such that

$$\mathbb{P}\Big(d_{\mathcal{Q}}\big((\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n), \mathcal{O}^\star\big) \ge \eta\Big) \ge \alpha \qquad \text{for all } n.$$

On this event, Assumption A1 gives

$$J^\star - J(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) \ge \kappa(\eta) > 0,$$

which contradicts

$$J^\star - J(\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n) \xrightarrow{\mathbb{P}} 0.$$

Hence

$$d_{\mathcal{Q}}\big((\hat{\mathbf{f}}_n, \hat{\mathbf{f}}'_n), \mathcal{O}^\star\big) \xrightarrow{\mathbb{P}} 0.$$

Since $O(d_{\mathcal{Z}}) \times O(d_{\mathcal{Z}})$ is compact and the distance functional is continuous, the infimum in $d_{\mathcal{Q}}$ is attained. Equivalently, there exist random orthogonal matrices $\mathbf{Q}_n, \mathbf{Q}'_n \in O(d_{\mathcal{Z}})$ such that

$$\|\mathbf{Q}_n \hat{\mathbf{f}}_n - \tilde{\mathbf{f}}^\star\|_{L^2(P_{\mathbf{x}})} + \|\mathbf{Q}'_n \hat{\mathbf{f}}'_n - \tilde{\mathbf{f}}'^\star\|_{L^2(P_{\mathbf{x}'})} \xrightarrow{\mathbb{P}} 0.$$

This proves Claim 2.

Step 5: latent recovery. Let

$$\tilde{\mathbf{h}}^\star := \tilde{\mathbf{f}}^\star \circ \mathbf{g}, \qquad \tilde{\mathbf{h}}'^\star := \tilde{\mathbf{f}}'^\star \circ \mathbf{g}'.$$

By the pullback isometry established in the proof of Proposition 1, for every $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$,

$$\|\mathbf{Q}(\hat{\mathbf{f}}_n \circ \mathbf{g}) - \tilde{\mathbf{h}}^\star\|_{L^2(P_{\mathbf{s}})} = \|\mathbf{Q}\hat{\mathbf{f}}_n - \tilde{\mathbf{f}}^\star\|_{L^2(P_{\mathbf{x}})},$$

and likewise

$$\|\mathbf{Q}'(\hat{\mathbf{f}}'_n \circ \mathbf{g}') - \tilde{\mathbf{h}}'^\star\|_{L^2(P_{\mathbf{s}'})} = \|\mathbf{Q}'\hat{\mathbf{f}}'_n - \tilde{\mathbf{f}}'^\star\|_{L^2(P_{\mathbf{x}'})}.$$

Therefore Claim 2 immediately implies

$$\begin{aligned} \inf_{\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})} \Big(&\|\mathbf{Q}(\hat{\mathbf{f}}_n \circ \mathbf{g}) - \tilde{\mathbf{h}}^\star\|_{L^2(P_{\mathbf{s}})} \\ &+ \|\mathbf{Q}'(\hat{\mathbf{f}}'_n \circ \mathbf{g}') - \tilde{\mathbf{h}}'^\star\|_{L^2(P_{\mathbf{s}'})}\Big) \xrightarrow{\mathbb{P}} 0. \end{aligned}$$

Finally, by Theorem 1, $\tilde{\mathbf{h}}^\star$ and $\tilde{\mathbf{h}}'^\star$ are the marginally whitened latent factors up to orthogonal transformations. This proves Claim 3. □

E Necessity of whitening

For precision, the necessity claim has to be stated carefully. Centering and unit covariance play different roles, and only the latter is the genuinely restrictive whitening condition. What is mathematically indispensable is a scale-fixing normalization equivalent to whitening. In our formulation this normalization is imposed by restricting the encoder classes to zero-mean, identity-covariance maps. The proof of Theorem 1 uses this normalization in two logically distinct ways: centering removes the constant polynomial mode, while unit covariance yields the orthonormal coefficient geometry required by the diagonal selection argument.

We first record that normalized CCA can always be represented on whitened encoders.

Proposition 5 (Whitening is without loss of generality for normalized CCA)**.** *Let* $\mathbf{h} \in L^2(P_{\mathbf{s}}; \mathbb{R}^{d_{\mathcal{Z}}})$ *and* $\mathbf{h}' \in L^2(P_{\mathbf{s}'}; \mathbb{R}^{d_{\mathcal{Z}}})$ *be centered maps with*

$$\mathbf{\Sigma}_{\mathbf{h}} := \mathrm{Cov}\big(\mathbf{h}(\mathbf{s})\big) \succ 0, \qquad \mathbf{\Sigma}_{\mathbf{h}'} := \mathrm{Cov}\big(\mathbf{h}'(\mathbf{s}')\big) \succ 0.$$

Define their whitened versions by

$$\tilde{\mathbf{h}} := \mathbf{\Sigma}_{\mathbf{h}}^{-1/2}\mathbf{h}, \qquad \tilde{\mathbf{h}}' := \mathbf{\Sigma}_{\mathbf{h}'}^{-1/2}\mathbf{h}'.$$

If

$$J_{\mathcal{S}}^{\mathrm{cca}}(\mathbf{h}, \mathbf{h}') := \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\Big(\mathbf{\Sigma}_{\mathbf{h}}^{-1/2}\mathrm{Cov}\big(\mathbf{h}(\mathbf{s}), \mathbf{h}'(\mathbf{s}')\big)\mathbf{\Sigma}_{\mathbf{h}'}^{-1/2}\Big),$$

then

$$J_{\mathcal{S}}^{\mathrm{cca}}(\mathbf{h}, \mathbf{h}') = J_{\mathcal{S}}(\tilde{\mathbf{h}}, \tilde{\mathbf{h}}'),$$

where $J_{\mathcal{S}}$ *is the whitened source-space objective in Equation 3. In particular, every candidate pair for normalized CCA admits an equivalent representation by whitened encoders.*

Proof. Since $\mathbf{h}$ and $\mathbf{h}'$ are centered,

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s})\big) = \mathbf{\Sigma}_{\mathbf{h}}^{-1/2}\mathbf{\Sigma}_{\mathbf{h}}\mathbf{\Sigma}_{\mathbf{h}}^{-1/2} = \mathbf{I}_{d_{\mathcal{Z}}}, \qquad \mathrm{Cov}\big(\tilde{\mathbf{h}}'(\mathbf{s}')\big) = \mathbf{\Sigma}_{\mathbf{h}'}^{-1/2}\mathbf{\Sigma}_{\mathbf{h}'}\mathbf{\Sigma}_{\mathbf{h}'}^{-1/2} = \mathbf{I}_{d_{\mathcal{Z}}}.$$

Hence $\tilde{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$ and $\tilde{\mathbf{h}}' \in \hat{\mathcal{F}}'_{\mathcal{S}}$. Moreover,

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s}), \tilde{\mathbf{h}}'(\mathbf{s}')\big) = \mathbf{\Sigma}_{\mathbf{h}}^{-1/2}\mathrm{Cov}\big(\mathbf{h}(\mathbf{s}), \mathbf{h}'(\mathbf{s}')\big)\mathbf{\Sigma}_{\mathbf{h}'}^{-1/2}.$$

Taking singular values and summing them proves the identity. □

The next proposition shows why some such normalization is necessary.

Proposition 6 (Without scale normalization the covariance objective is ill-posed)**.** *Define the unnormalized objective*

$$J_{\mathcal{S}}^{\mathrm{raw}}(\mathbf{h}, \mathbf{h}') := \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\Big(\mathrm{Cov}\big(\mathbf{h}(\mathbf{s}), \mathbf{h}'(\mathbf{s}')\big)\Big)$$

over centered square-integrable encoder pairs. If there exists one pair $(\mathbf{h}, \mathbf{h}')$ *such that* $J_{\mathcal{S}}^{\mathrm{raw}}(\mathbf{h}, \mathbf{h}') > 0$*, then*

$$\sup_{\mathbf{h}, \mathbf{h}'} J_{\mathcal{S}}^{\mathrm{raw}}(\mathbf{h}, \mathbf{h}') = +\infty.$$

Proof. For any $\lambda > 0$,

$$\mathrm{Cov}\big(\lambda\mathbf{h}(\mathbf{s}), \lambda\mathbf{h}'(\mathbf{s}')\big) = \lambda^2\,\mathrm{Cov}\big(\mathbf{h}(\mathbf{s}), \mathbf{h}'(\mathbf{s}')\big),$$

so

$$J_{\mathcal{S}}^{\mathrm{raw}}(\lambda\mathbf{h}, \lambda\mathbf{h}') = \lambda^2 J_{\mathcal{S}}^{\mathrm{raw}}(\mathbf{h}, \mathbf{h}').$$

If $J_{\mathcal{S}}^{\mathrm{raw}}(\mathbf{h}, \mathbf{h}') > 0$, letting $\lambda \to \infty$ proves the claim. Under Assumption 1, this hypothesis holds by taking the first pair of canonical variates and, if $d_{\mathcal{Z}} > 1$, padding the remaining coordinates with zeros. □

The preceding propositions isolate the formal necessity of whitening. In the proof of Theorem 1, whitening enters at three specific places.

1. Basis alignment and exclusion of the constant mode. The Hermite-Mehler expansions in Lemma 2 and Proposition 4 are written in the standardized source coordinates introduced there. This basis alignment is determined by the latent Gaussian law after source standardization and is independent of the encoder parameterization. For the encoder expansions, the only degree-zero condition needed is centering. Indeed, if

$$\tilde{\mathbf{h}}_r(\mathbf{s}) = \sum_{\mathbf{n} \in \mathbb{N}^{d_{\mathcal{S}}}} \alpha_{r,\mathbf{n}}\Psi_{\mathbf{n}}(\mathbf{u}), \qquad \tilde{\mathbf{h}}'_q(\mathbf{s}') = \sum_{\mathbf{n} \in \mathbb{N}^{d_{\mathcal{S}}}} \beta_{q,\mathbf{n}}\Psi_{\mathbf{n}}(\mathbf{v}),$$

then $\Psi_{\mathbf{0}}$ is the unique constant basis element and $\mathbb{E}[\Psi_{\mathbf{n}}] = 0$ for every $\mathbf{n} \neq \mathbf{0}$. Hence

$$\mathbb{E}[\tilde{\mathbf{h}}_r(\mathbf{s})] = 0 \iff \alpha_{r,\mathbf{0}} = 0, \qquad \mathbb{E}[\tilde{\mathbf{h}}'_q(\mathbf{s}')] = 0 \iff \beta_{q,\mathbf{0}} = 0.$$

Thus the constant mode disappears because the encoders are centered. The unit-covariance part of whitening is used later.

2. Feasible-set geometry for the CCA objective. After applying Proposition 5, we may work entirely with whitened encoders. Writing

$$\tilde{\mathbf{h}}_r(\mathbf{s}) = \sum_{\mathbf{n}\neq\mathbf{0}} \alpha_{r,\mathbf{n}} \Psi_{\mathbf{n}}(\mathbf{u}), \qquad \tilde{\mathbf{h}}'_q(\mathbf{s}') = \sum_{\mathbf{n}\neq\mathbf{0}} \beta_{q,\mathbf{n}} \Psi_{\mathbf{n}}(\mathbf{v}),$$

and collecting coefficients into matrices $\mathbf{A} = (\alpha_{r,\mathbf{n}})$ and $\mathbf{B} = (\beta_{q,\mathbf{n}})$, Corollary 2 gives

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s})\big) = \mathbf{A}\mathbf{A}^\top, \qquad \mathrm{Cov}\big(\tilde{\mathbf{h}}'(\mathbf{s}')\big) = \mathbf{B}\mathbf{B}^\top.$$

Hence whitening is exactly the constraint

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_{d_{\mathcal{Z}}}, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I}_{d_{\mathcal{Z}}}.$$

Under these constraints, the cross-covariance diagonalizes as

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s}), \tilde{\mathbf{h}}'(\mathbf{s}')\big) = c\,\mathbf{A}\,\mathrm{diag}(t_{\mathbf{n}})\,\mathbf{B}^\top,$$

and Lemma 3 applies directly. Without the unit-covariance constraints, the coefficient matrices are no longer row-orthonormal, and the proof no longer reduces to selecting the largest diagonal entries $\{|t_{\mathbf{n}}|\}$. This is the precise point where whitening enters the identifiability argument.

3. Scale fixing and collapse avoidance. Proposition 6 already shows that without any scale-fixing normalization the raw covariance objective is unbounded. After passing to the whitened representation, the remaining first-order block is also well-conditioned. Once Assumption 3 excludes all higher-order indices, the optimizer is confined to the first-order block $\{\mathbf{e}_1, \ldots, \mathbf{e}_{d_{\mathcal{S}}}\}$. Let $\mathbf{A}_{\mathrm{lin}}$ and $\mathbf{B}_{\mathrm{lin}}$ denote the corresponding $d_{\mathcal{Z}} \times d_{\mathcal{S}}$ coefficient matrices. Because the full coefficient matrices are whitened, we still have

$$\mathbf{A}_{\mathrm{lin}}\mathbf{A}_{\mathrm{lin}}^\top = \mathbf{I}_{d_{\mathcal{Z}}}, \qquad \mathbf{B}_{\mathrm{lin}}\mathbf{B}_{\mathrm{lin}}^\top = \mathbf{I}_{d_{\mathcal{Z}}}.$$

When $d_{\mathcal{Z}} = d_{\mathcal{S}}$, this forces $\mathbf{A}_{\mathrm{lin}}, \mathbf{B}_{\mathrm{lin}} \in O(d_{\mathcal{Z}})$. Therefore the recovered linear maps are full-rank and perfectly conditioned in the whitened coordinates: no output coordinate can vanish, duplicate another one, or absorb an arbitrary scale. This is exactly what yields the orthogonal ambiguities in Theorem 1. Without whitening, even after higher-order terms are removed, the remaining linear coefficient matrices need only be invertible (if at all) and may be arbitrarily ill-conditioned.

Remark 7. The discussion above yields the precise form of the necessity claim used in the main text. The indispensable ingredient is a scale-fixing normalization equivalent to whitening. In our formulation it is imposed as a hard whitening constraint on the encoder classes; alternatively, it can be implemented inside the objective as in standard normalized CCA, in which case Proposition 5 shows that the analysis may still be carried out on whitened encoders. What fails without such normalization is twofold: the raw covariance objective is unbounded by Proposition 6, and the row-orthonormal coefficient geometry required by Lemma 3 disappears. Hence the affine identifiability proof does not go through without whitening or an equivalent variance normalization.

F Proof Sketches for the Remaining Candidate Distributions

The Gaussian proof of Theorem 1 relies on three Gaussian-specific ingredients: the scalar Hermite–Mehler expansion in Lemma 2, its tensor-product extension in Proposition 4, and the resulting orthogonality statement in Corollary 2. For the remaining candidate priors, the same optimization argument carries over once these ingredients are replaced by the corresponding scalar Lancaster expansion and the associated orthonormal polynomial basis (Lancaster, 1958; Eagleson, 1964). The only family-dependent objects are therefore the one-dimensional orthogonal polynomials and the associated diagonal coefficient sequence.

F.1 Unified Proof Template

We state the common reduction at the level needed by the CCA proof. Throughout this subsection, let $I_i = \mathbb{N}_0$ for Poisson, negative binomial, and gamma marginals, and let $I_i = \{0, \dots, m_i\}$ for hypergeometric marginals, where m_i is the maximal polynomial degree permitted by the finite support of the i-th coordinate. In the hypergeometric case, all sums below are therefore finite.

Proposition 7 (Unified Lancaster reduction for the remaining candidate priors)**.** *Assume that, for each coordinate $i \in \{1, \dots, d_{\mathcal{S}}\}$, the pair (s_i, s_i') has a bivariate Lancaster law with common marginal ν_i, orthonormal polynomial basis $\{\psi_{i,n}\}_{n \in I_i} \subset L^2(\nu_i)$ satisfying $\psi_{i,0} \equiv 1$, and coefficient sequence $\{\lambda_{i,n}\}_{n \in I_i}$ with $\lambda_{i,0} = 1$ such that*

$$dP_{s_i, s_i'}(u, v) = \left(\sum_{n \in I_i} \lambda_{i,n} \, \psi_{i,n}(u) \psi_{i,n}(v) \right) d\nu_i(u) \, d\nu_i(v). \tag{8}$$

Assume moreover that the coordinate pairs $\{(s_i, s_i')\}_{i=1}^{d_{\mathcal{S}}}$ are independent across i. Define the tensor-product basis and multivariate Lancaster coefficients by

$$\Psi_{\mathbf{n}}(\mathbf{s}) := \prod_{i=1}^{d_{\mathcal{S}}} \psi_{i,n_i}(s_i), \qquad \lambda_{\mathbf{n}} := \prod_{i=1}^{d_{\mathcal{S}}} \lambda_{i,n_i}, \qquad \mathbf{n} = (n_1, \dots, n_{d_{\mathcal{S}}}) \in \mathcal{I} := \prod_{i=1}^{d_{\mathcal{S}}} I_i.$$

Let $\tilde{\mathbf{h}}, \tilde{\mathbf{h}}'$ be whitened latent maps in source space and expand their coordinates as

$$\tilde{h}_r(\mathbf{s}) = \sum_{\mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}} \alpha_{r,\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{s}), \qquad \tilde{h}'_q(\mathbf{s}') = \sum_{\mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}} \beta_{q,\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{s}'),$$

for $r, q \in \{1, \dots, d_{\mathcal{Z}}\}$, and define the coefficient matrices

$$\mathbf{A} = (\alpha_{r,\mathbf{n}})_{1 \le r \le d_{\mathcal{Z}},\, \mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}}, \qquad \mathbf{B} = (\beta_{q,\mathbf{n}})_{1 \le q \le d_{\mathcal{Z}},\, \mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}}.$$

Then:

1. ***Whitening constraints.*** *Because the tensor-product basis is orthonormal and the constant mode is excluded, the whitening conditions imply*

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_{d_{\mathcal{Z}}}, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I}_{d_{\mathcal{Z}}}.$$

2. ***Diagonal cross-covariance.*** *The source-space cross-covariance diagonalizes as*

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s}), \tilde{\mathbf{h}}'(\mathbf{s}')\big) = \mathbf{A} \, \mathrm{diag}\big((\lambda_{\mathbf{n}})_{\mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}}\big) \, \mathbf{B}^\top.$$

3. ***Reduction to diagonal selection.*** *Consequently, by Lemma 3, the source-space CCA objective equals the sum of the $d_{\mathcal{Z}}$ largest absolute values among $\{\lambda_{\mathbf{n}} : \mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\}\}$.*

If $d_{\mathcal{Z}} = d_{\mathcal{S}}$ and

$$\min_{1 \le i \le d_{\mathcal{S}}} |\lambda_{i,1}| > \sup_{\substack{\mathbf{n} \in \mathcal{I} \setminus \{\mathbf{0}\} \\ \mathbf{n} \notin \{\mathbf{e}_1, \dots, \mathbf{e}_{d_{\mathcal{S}}}\}}} |\lambda_{\mathbf{n}}|, \tag{9}$$

then every population maximizer uses only the degree-one tensor basis functions. Equivalently, there exist orthogonal matrices $\mathbf{Q}, \mathbf{Q}' \in O(d_{\mathcal{Z}})$ such that

$$\tilde{\mathbf{h}}^*(\mathbf{s}) = \mathbf{Q} \left(\psi_{1,1}(s_1), \dots, \psi_{d_{\mathcal{S}},1}(s_{d_{\mathcal{S}}}) \right)^\top,$$

$$\tilde{\mathbf{h}}'^*(\mathbf{s}') = \mathbf{Q}' \left(\psi_{1,1}(s_1'), \dots, \psi_{d_{\mathcal{S}},1}(s_{d_{\mathcal{S}}}') \right)^\top.$$

Hence, whenever each first-degree basis function $\psi_{i,1}$ is affine in its argument, the optimal source-space maps are affine up to orthogonal transformations.

Proof sketch. **1. Reduce to the latent space.** By Proposition 1, it suffices to maximize $J_{\mathcal{S}}$ over whitened source-space maps $(\tilde{\mathbf{h}}, \tilde{\mathbf{h}}')$.

2. Expand the encoders in the orthonormal polynomial basis. Since $\{\Psi_{\mathbf{n}}\}_{\mathbf{n}\in\mathcal{I}}$ is an orthonormal tensor-product basis of $L^2(P_{\mathbf{s}})$ and whitening enforces zero mean, the constant term $\mathbf{n} = \mathbf{0}$ is absent from both expansions. The identity-covariance constraints then yield $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ and $\mathbf{B}\mathbf{B}^\top = \mathbf{I}$.

3. Diagonalize the cross-covariance. Independence across coordinates and the scalar Lancaster expansion (8) imply

$$\mathbb{E}\big[\Psi_{\mathbf{m}}(\mathbf{s})\Psi_{\mathbf{n}}(\mathbf{s}')\big] = \lambda_{\mathbf{n}}\,\delta_{\mathbf{mn}}, \qquad \mathbf{m}, \mathbf{n} \in \mathcal{I}.$$

Substituting the encoder expansions gives

$$\mathrm{Cov}\big(\tilde{\mathbf{h}}(\mathbf{s}), \tilde{\mathbf{h}}'(\mathbf{s}')\big) = \mathbf{A}\,\mathrm{diag}\big((\lambda_{\mathbf{n}})_{\mathbf{n}\in\mathcal{I}\setminus\{\mathbf{0}\}}\big)\,\mathbf{B}^\top.$$

4. Reduce CCA to selecting diagonal entries. By Lemma 3, the CCA objective is maximized by selecting the $d_{\mathcal{Z}}$ largest absolute values among $\{\lambda_{\mathbf{n}} : \mathbf{n} \neq \mathbf{0}\}$. Under (9), those top coefficients are exactly the degree-one indices $\mathbf{e}_1, \ldots, \mathbf{e}_{d_{\mathcal{S}}}$.

5. Conclude affinity. Thus every maximizer lies in the span of the first-degree basis functions. Since the first-degree orthonormal polynomial of a one-dimensional orthogonal polynomial system is always affine in the underlying scalar variable, the optimal source-space maps are affine. Pushing them back to observation space via Proposition 1 yields the same conclusion for the learned encoders. □

Remark 8. The non-Gaussian argument does *not* require the Gaussian-specific identity $\lambda_{\mathbf{n}} = \prod_i \rho_i^{n_i}$. The proof only needs diagonalization in an orthonormal polynomial basis and an ordering condition such as (9). Whenever a particular family admits the geometric specialization $\lambda_{i,n} = \rho_i^n$, the dominance condition reduces to the Gaussian-style separation $\rho_{d_{\mathcal{S}}} > \rho_1^2$.

F.2 Instantiations for the Four Families

For each family below, the first-degree orthonormal polynomial is precisely the standardized coordinate,

$$\psi_1(x) = \frac{x - \mu}{\sigma},$$

with μ and σ^2 the marginal mean and variance. In particular,

$$\lambda_{i,1} = \mathbb{E}[\psi_{i,1}(s_i)\psi_{i,1}(s_i')] = \mathrm{Corr}(s_i, s_i').$$

Therefore, once higher-degree terms are excluded by (9), Proposition 7 immediately yields affine recovery.

Poisson (λ)**.** The marginal law is

$$\nu(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x \in \mathbb{N}_0.$$

The associated orthogonal family is the Charlier family $C_m(x;\lambda)$. After normalization, the first-degree term is

$$\psi_1(x) = \frac{x - \lambda}{\sqrt{\lambda}}.$$

Negative Binomial (r, p)**.** We use the convention

$$\nu(x) = \binom{r + x - 1}{x} p^r (1 - p)^x, \qquad x \in \mathbb{N}_0, \quad 0 < p < 1.$$

The associated orthogonal family is the Meixner family $M_m(x;\beta,c)$ with $\beta = r$ and $c = 1 - p$. Its mean and variance are

$$\mu = \frac{r(1-p)}{p}, \qquad \sigma^2 = \frac{r(1-p)}{p^2},$$

hence

$$\psi_1(x) = \frac{x - \mu}{\sigma}.$$

Hypergeometric (N, K, n)**.** The marginal law is

$$\nu(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \qquad x = \max\{0, n-(N-K)\}, \ldots, \min\{n, K\}.$$

The associated orthogonal family is the Hahn family associated with the hypergeometric weight; its degree is finite because the support is finite. Its mean and variance are

$$\mu = n\frac{K}{N}, \qquad \sigma^2 = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1},$$

so the first-degree normalized polynomial is

$$\psi_1(x) = \frac{x-\mu}{\sigma}.$$

Gamma (k, θ)**.** The marginal law is

$$d\nu(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \mathbf{1}_{(0,\infty)}(x)\, dx.$$

The associated orthogonal family is the generalized Laguerre family $L_m^{(k-1)}(x/\theta)$. Its mean and variance are

$$\mu = k\theta, \qquad \sigma^2 = k\theta^2,$$

hence

$$\psi_1(x) = \frac{x - k\theta}{\sqrt{k}\,\theta}.$$

F.3 Connection to the Gaussian Proof

The relationship to the Gaussian case is now transparent.

- **Only the basis changes.** Replace the Hermite–Mehler system by the appropriate Lancaster system: Charlier for Poisson, Meixner for negative binomial, Hahn for hypergeometric, and generalized Laguerre for gamma.
- **The diagonalization step is identical.** Once the joint law is expanded in an orthonormal polynomial basis, the cross-covariance of the encoder coefficients is diagonal, and the CCA objective reduces to a diagonal selection problem exactly as in the Gaussian proof.
- **The optimization step is unchanged.** Lemma 3 still implies that the maximizer selects the $d_{\mathcal{Z}}$ largest absolute Lancaster coefficients.
- **Affine recovery again follows from first-degree dominance.** If the degree-one coefficients dominate all higher-degree coefficients, only first-degree basis elements survive at the optimum. Since those first-degree basis elements are affine in the latent coordinates, the conclusion of Theorem 1 carries over verbatim.

G Implementation Details

Training Configurations. We adapt the experimental protocols of (Zimmermann et al., 2021; Matthes et al., 2023) to our dual-encoder framework. All models are optimized via Adam (Kingma and Ba, 2015) with a constant learning rate of 10^{-4}. For the synthetic dataset, we train for 10^5 iterations with a batch size of 1024, requiring approximately 1.5 hours for $d_{\mathcal{S}} = 40$ on a single NVIDIA RTX A5000 GPU. For 3DIdent, we train for 10^4 iterations with a batch size of 512, requiring roughly 4 hours. Empirically, we observe that CCA-based objectives converge 2–3× faster than standard contrastive losses. With the exception of qualitative visualizations, all reported metrics are averaged across five independent random seeds. Comprehensive implementation details and extended results are provided in Appendix H.

For encoder architectures, we adopt the same design as (Zimmermann et al., 2021) for both synthetic and 3DIdent datasets, except that for synthetic data we employ a lighter residual network. This network comprises two hidden layers of sizes $10 \cdot d_{\mathcal{S}}$ and $20 \cdot d_{\mathcal{S}}$, followed by three residual blocks and an output layer. Each residual block contains two layers of width $20 \cdot d_{\mathcal{S}}$. We apply leaky-ReLU activations and batch normalization to all hidden layers.

The decoders are implemented as approximately invertible multi-layer perceptrons. The encoders $\mathbf{f}$ and $\mathbf{f}'$ are parameterized independently using residual connections and batch normalization for optimization stability.

Training the synthetic model with $d_{\mathcal{S}} = 10$ on a single RTX A5000 GPU takes approximately one hour, whereas experiments on 3DIdent require roughly four hours using eight RTX A5000 GPUs.

Candidate Distribution Setups. For each latent dimension, we generate a paired source $(\mathbf{s}, \mathbf{s}') \in \mathbb{R}^{d_{\mathcal{S}}} \times \mathbb{R}^{d_{\mathcal{S}}}$ according to the additive latent model $\mathbf{s} = \mathbf{a} + \mathbf{c}$, $\mathbf{s}' = \mathbf{b} + \mathbf{c}$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are independent random vectors. Unless otherwise stated, the coordinates are independent across dimensions except for Gaussian case as stated in Assumption 1.

Joint Gaussian. We sample $(\mathbf{s}, \mathbf{s}') \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with block covariance

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_n & \mathbf{A}\,\mathrm{diag}(\boldsymbol{\rho})\,\mathbf{B}^\top \\ \mathbf{B}\,\mathrm{diag}(\boldsymbol{\rho})\,\mathbf{A}^\top & \mathbf{I}_n \end{pmatrix},$$

where $\mathbf{A}, \mathbf{B}$ are independent random orthogonal matrices (constructed via QR decomposition of Gaussian matrices with $\det = +1$). The canonical correlations $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ are linearly spaced in $[0.3, 0.5]$. This yields standard-normal marginals and the desired cross-covariance structure $\mathbf{A}\,\mathrm{diag}(\boldsymbol{\rho})\,\mathbf{B}^\top$.

Joint Gamma. For each coordinate j, we draw three independent random variables $c_j \sim \mathrm{Gamma}(k = 1, \mathrm{rate} = 1)$ and $a_j, b_j \sim \mathrm{Gamma}(k = 2, \mathrm{rate} = 1)$, and set

$$s_j = a_j + c_j, \qquad s'_j = b_j + c_j.$$

Then $\mathbb{E}[s_j] = \mathbb{E}[s'_j] = 3$, $\mathrm{Var}(s_j) = \mathrm{Var}(s'_j) = 3$, and $\mathrm{Cov}(s_j, s'_j) = 1$.

Joint Poisson. For each coordinate j, $c_j \sim \mathrm{Poisson}(1)$ and $a_j, b_j \sim \mathrm{Poisson}(2)$ independently, and

$$s_j = a_j + c_j, \qquad s'_j = b_j + c_j.$$

This yields $\mathbb{E}[s_j] = \mathbb{E}[s'_j] = 3$, $\mathrm{Var}(s_j) = \mathrm{Var}(s'_j) = 3$, and $\mathrm{Cov}(s_j, s'_j) = 1$.

Joint Negative Binomial. Each coordinate is generated as $c_j \sim \mathrm{NegBin}(r = 1, p = 0.5)$ and $a_j, b_j \sim \mathrm{NegBin}(r = 2, p = 0.5)$, and we set

$$s_j = a_j + c_j, \qquad s'_j = b_j + c_j.$$

Under the PyTorch parameterization, $\mathbb{E}[\mathrm{NegBin}(r, p)] = r(1-p)/p$ and $\mathrm{Var} = r(1-p)/p^2$. Hence $\mathbb{E}[s_j] = 3$, $\mathrm{Var}(s_j) = 6$, and $\mathrm{Cov}(s_j, s'_j) = 2$. Note that this construction uses the *negative binomial* distribution (rather than the bounded binomial).

Joint Hypergeometric. For each dimension j, we first sample population and success counts $N_j \sim \mathrm{Uniform}\{10, \ldots, 20\}$ and $M_j \sim \mathrm{Uniform}\{1, \ldots, N_j - 1\}$, and fix $n_{1j} = n_{2j} = 1$. We draw without replacement $n_{1j} + n_{2j} = 2$ items from a population of N_j containing M_j successes, and define

$$s_j = \#\text{successes in first draw}, \qquad s'_j = \#\text{successes in second draw}.$$

The marginals are $\mathrm{Hypergeometric}(N_j, M_j, n = 1)$, equivalently $\mathrm{Bernoulli}(p_j = M_j/N_j)$, and the two draw indicators are negatively correlated with

$$\mathrm{Cov}(s_j, s'_j) = -\frac{M_j(N_j - M_j)}{N_j^2(N_j - 1)}.$$

Implementation Notes. Each sampling routine draws $\mathbf{s}$ and caches the paired $\mathbf{s}'$ for retrieval during conditional evaluation; it thus implements a paired sampler rather than a conditional generator. All coordinates are sampled independently across j.

H Further Experimental Details and Additional Experimental Results

Table 4 and Table 5 gives the subspace errors measured by maximal and mean principal angels on synthetic data across five candidate distributions.

Methods	Gaussian		Negative Binomial		Gamma		Poisson		Hypergeometric	
	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$
SwAV	89.26 ±0.55	88.88 ±0.45	89.64 ±0.2	89.65 ± 0.26	89.49 ±0.25	89.32 ±0.46	89.67 ±0.27	89.5 ± 0.33	86.48 ±0.27	89.5 ±0.33
BarlowTwins	89.44±0.86	89.51 ±0.63	85.53 ± 0.86	85.36 ± 0.59	86.16 ± 1.15	88.6 ± 0.58	87.73 ±0.86	87.43 ±0.91	88.93 ±0.51	89.59 ±0.58
VICReg	88.99 ±0.54	89.07 ±0.61	89.36 ±0.48	89.16 ±0.73	89.6 ±0.24	89.03 ±0.59	89.66 ±0.41	88.38 ±0.4	88.95 ±0.43	89.29 ±0.22
W-MSE	**7.83 ±0.31**	**8.11 ±0.24**	**8.5 ±0.18**	**8.17 ±0.19**	8.4 ±0.37	**8.01 ±0.32**	**7.63 ± 0.42**	7.61 ± 0.2	8.72 ±0.22	8.5 ± 0.18
DGCCA	89.33 ± 1.19	88.17 ±0.76	89.12 ±0.51	89.38 ±0.68	88.60 ±0.61	89.06 ±0.62	89.21 ±0.61	89.3 ± 0.62	89.21 ±0.7	88.39 ±0.59
DeepCCA	8.88 ±0.28	8.22 ±0.23	8.79 ±0.66	9.65 ±0.3	**7.93 ±0.4**	8.81 ±0.45	11.43 ±0.59	**7.37 ±0.45**	**8.61 ±0.24**	**7.39 ±0.45**

Table 4: Comparison of the maximal principal angles $PA_{max}\downarrow(^\circ)$ of both encoders $\mathbf{f}, \mathbf{f}'$ on synthetic data ($d_{\mathcal{S}} = d_{\mathcal{Z}} = 10$).

Methods	Gaussian		Negative Binomial		Gamma		Poisson		Hypergeometric	
	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$	$\mathbf{f}$	$\mathbf{f}'$
SwAV	54.95 ±1.75	55.18 ±0.95	59.13 ±10.12	62.48 ± 9.98	56.68 ±0.59	55.17 ± 0.6	57.57 ±0.46	56.39 ±0.33	63.21 ±9.22	68.92 ±8.69
BarlowTwins	33.91±0.16	32.78 ±0.23	30.27 ±0.18	28.66 ±0.15	33.6 ± 0.22	32.29 ± 0.17	42.18 ± 0.15	40.58 ± 0.22	60.77 ± 0.29	60.72 ± 0.23
VICReg	68.47 ±0.44	67.89 ±0.63	67.56 ±0.45	68.91 ±0.32	74.27 ±0.18	70.39 ±0.25	75.46 ±0.27	72.95 ±0.31	65.8 ±0.43	64.75 ±0.29
W-MSE	5.12 ±0.09	5.15 ±0.03	5.21 ± 0.09	**4.74 ±0.07**	4.75 ±0.06	**4.57 ±0.1**	**4.69 ± 0.1**	4.37 ±0.04	5.38 ±0.05	**5.13 ±0.02**
DGCCA	67.22 ± 0.38	65.91 ±0.71	65.65 ±0.23	64.89 ±0.17	67.27 ±0.11	66.12 ±0.32	68.24 ± 0.25	68.82 ±0.33	63.12 ± 0.23	63.58 ±0.17
DeepCCA	**5.06 ±0.08**	**4.92 ±0.09**	**5.00 ±0.11**	5.16 ±0.05	**4.65 ±0.1**	4.86 ±0.08	5.18 ±0.07	**4.06 ±0.08**	**5.37 ±0.07**	6.17 ±0.07

Table 5: Comparison of the mean principal angles $PA_{\text{mean}}\downarrow(^\circ)$ of both encoders $\mathbf{f}, \mathbf{f}'$ on synthetic data ($d_{\mathcal{S}} = d_{\mathcal{Z}} = 10$).