Spurious correlation problem. Single-environment extraction might learn:

$$v_1 = v_{\text{formality}} + v_{\text{jargon}}$$
$$v_2 = v_{\text{formality}} + v_{\text{professional}}$$
$$v_3 = v_{\text{formality}} + v_{\text{precise}}$$

Each includes the true factor plus environment-specific confounds.

Invariance solution. Find the component present in all environments:

$$v_{\text{invariant}} = \text{stable component of } \{v_1, v_2, v_3\} \quad (57)$$

The spurious parts ($v_{\text{jargon}}, v_{\text{professional}}, v_{\text{precise}}$) vary across environments and get filtered out.

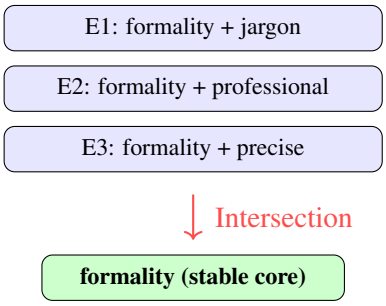


Figure 8: Multi-environment filtering extracts the stable core. Each environment contains the true semantic factor (formality) plus environment-specific confounds. The intersection across environments isolates the invariant component.

D.7 Cross-Layer Intuition: Consistency Check

Setup: Extract steering vectors at layers 8, 12, 16, 20.

Consistency hypothesis. True semantic factors should propagate coherently:

$$v_{12} \approx W_8 v_8, \quad v_{16} \approx W_{12} v_{12}, \quad \text{etc.}$$

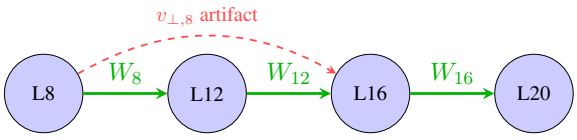
Inconsistency indicates null-space artifacts. If v_8 includes a large null-space component $v_{\perp,8}$:

$$v_8 = v_{\text{true},8} + v_{\perp,8} \quad (58)$$

Then after propagation:

$$W_8 v_8 = W_8 v_{\text{true},8} + W_8 v_{\perp,8} \quad (59)$$

Since $v_{\perp,8} \in \ker(J_8)$ but generically $W_8 v_{\perp,8} \notin \ker(J_{12})$, the null-space component becomes observable at layer 12, causing inconsistency.



Consistent propagation filters null-space artifacts

Figure 9: Cross-layer consistency check. True semantic factors propagate coherently, while null-space artifacts become inconsistent.

Filtering strategy. Project onto the subspace of vectors that are consistent across layers:

$$v_{\text{consistent}} = \arg\min_v \sum_\ell \|W_{\ell-1} v_{\ell-1} - v_\ell\|^2 \quad (60)$$

This filters out layer-specific artifacts (See Figure 9), retaining only the stable semantic direction.

E Mathematical Derivations

E.1 Dimension of Observational Equivalence Class

Setup: For steering vector $v \in \mathbb{R}^d$ and Jacobian $J \in \mathbb{R}^{V \times d}$ with rank r, consider:

$$[v] = \{v' \in \mathbb{R}^d : Jv' = Jv\} \quad (61)$$

Question: How many degrees of freedom exist in the equivalence class of observationally equivalent steering vectors?

Analysis: The equivalence class is an affine subspace:

$$[v] = v + \ker(J) \quad (62)$$

where $\ker(J)$ is a $(d-r)$-dimensional linear subspace.

Parameterization: Let $\{u_1, \ldots, u_{d-r}\}$ be an orthonormal basis for $\ker(J)$. Then:

$$[v] = \left\{ v + \sum_{i=1}^{d-r} \alpha_i u_i : \alpha_i \in \mathbb{R} \right\} \quad (63)$$

The equivalence class has $(d-r)$ degrees of freedom.

Scaling ambiguity: Additionally, v and cv produce equivalent outputs (up to rescaling α). The equivalence class modulo scaling is a $(d-r)$-dimensional projective space.

Illustrative example. Suppose a model has hidden size $d = 4096$ and an effective Jacobian rank $r \approx 3100$. Then the equivalence class dimension is $d - r \approx 996$.

Interpretation: For every identifiable parameter (row-space direction), there are $996/3100 \approx 0.32$

unidentifiable parameters (null-space directions). The under-determination ratio is approximately $1:3$.

E.2 Fisher Information and Cramér-Rao Bound

Setup: Consider the statistical model:

$$o(x; v, \alpha) \sim p(o|x, v, \alpha) \tag{64}$$

where v is the parameter to estimate.

Question: Can we achieve better identifiability by collecting more samples?

Fisher Information Matrix: For the linear Gaussian model $o = Jv + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2 I)$:

$$\begin{aligned}\mathcal{I}(v) = \mathbb{E}_o\Bigg[&\left(\frac{\partial \log p(o \mid x, v, \alpha)}{\partial v}\right) \\ &\left(\frac{\partial \log p(o \mid x, v, \alpha)}{\partial v}\right)^\top \Bigg] \\ = \frac{1}{\sigma^2} J^\top J.\end{aligned} \tag{65}$$

Cramér-Rao Lower Bound: The covariance of any unbiased estimator $\hat{v}$ satisfies:

$$\mathrm{Cov}(\hat{v}) \succeq \mathcal{I}(v)^{-1} = \sigma^2 (J^\top J)^+ \tag{66}$$

where $(J^\top J)^+$ is the Moore-Penrose pseudo-inverse.

Implications: For null-space directions $u_i \in \ker(J)$, the Fisher information is degenerate:

$$u_i^\top (J^\top J) u_i = 0 \tag{67}$$

Therefore, the variance of any unbiased estimator along null-space directions is unbounded:

$$\mathrm{Var}(u_i^\top \hat{v}) \geq u_i^\top \mathcal{I}(v)^{-1} u_i = \infty \tag{68}$$

Conclusion: The Cramér-Rao bound is **infinite** for null-space components, confirming that no finite amount of data can resolve the ambiguity. Non-identifiability is fundamental, not a small-sample problem. More data cannot help because the information geometry has infinite uncertainty in null-space directions.

E.3 Proof that ICA Breaks Gauge Symmetry

Setup: Consider two decompositions:

$$h = A_1 z_1 = A_2 z_2 \tag{69}$$

where both z_1 and z_2 have independent components.

Question: When are these equivalent ($A_1 = A_2$ up to permutation/scaling)?

Analysis: If both decompositions are valid:

$$A_1 z_1 = A_2 z_2 \implies z_1 = A_1^{-1} A_2 z_2 = G z_2 \tag{70}$$

where $G = A_1^{-1} A_2$.

ICA Constraint: If both z_1 and z_2 have independent components:

$$\begin{aligned} I(z_{1,i}; z_{1,j}) = 0 \quad &\text{for } i \neq j, \\ I\big((Gz_2)_i; (Gz_2)_j\big) = 0 \quad &\text{for } i \neq j. \end{aligned} \tag{71}$$

Key Theorem (Comon, 1994): If z_2 has independent components and Gz_2 also has independent components, then G must be a generalized permutation matrix:

$$G = PD \tag{72}$$

where P is a permutation matrix and D is a diagonal scaling matrix.

Implication:

$$\begin{aligned} A_1 z_1 &= A_2 z_2, \\ A_1 G z_2 &= A_2 z_2, \\ A_1 P D &= A_2. \end{aligned} \tag{73}$$

Therefore $A_2 = A_1 PD$, meaning A_2 is A_1 with permuted and scaled columns.

Conclusion: ICA constraints (statistical independence) force G to be a permutation-scaling matrix, breaking arbitrary gauge transformations and establishing identifiability up to unavoidable symmetries. Independence is the structural assumption that eliminates the infinite equivalence class.

F Disclosure of the Use of AI

We used ChatGPT to assist with refining sections of the manuscript for language and clarity and used Claude Code and GitHub Copilot for coding assistance during implementation. All research ideas, technical content, proofs, experiments and final writing decisions are entirely our own and we take full responsibility for the work presented.