# Stability Analysis of Canonical Correlation Analysis

Alisha Pham and Jianyang Xiao

March 2026

**Abstract**

Canonical Correlation Analysis (CCA) is a statistical method for identifying linear relationships between two datasets. For these relationships to be meaningful, the underlying patterns must be consistent across datasets measuring the same features. Understanding the conditions under which CCA produces stable and reliable patterns is therefore essential. In this paper, we extend the simulation framework and results of Helmer et al. (2024), building on their generative modeling approach to simulate synthetic datasets and systematically assess the stability of single and multiple canonical pairs under diverse and controlled conditions.

## 1 Introduction

### 1.1 Motivation and Related Literature

Canonical Correlation Analysis (CCA) is a multivariate method used to study the relationship between two sets of variables by finding linear combinations of each set that are maximally correlated. It is widely used in settings where both sides of the analysis are high-dimensional, such as relating brain imaging measures to behavioral, cognitive, or clinical variables. In such applications, CCA is attractive because it summarizes complex cross-set relationships into a small number of interpretable canonical pairs.

Several recent studies have highlighted both the usefulness of CCA and the practical challenges involved in its application. Zhuang et al. (2020) provide a technical review of CCA and its variants in neuroscience, emphasizing its broad applicability and important methodological considerations. Yang et al. (2021) study the stability of CCA in brain behavior analyzes and show that the subject-to-variable ratio and correlation strength have substantial effects on stability. Most relevant to the present work, Helmer et al. (2024) develop a generative simulation framework and show that when the sample size is small relative to the number of

features, CCA associations and weight patterns can be highly unstable and inaccurate.

Despite these contributions, there remains limited controlled simulation evidence on how classical CCA recovers individual canonical pairs across both the single-pair and multiple-pair settings under systematically varied sample size, feature dimension, and correlation structure. Existing work has mainly emphasized overall stability, subject-to-variable ratios, or broad methodological guidance. The present study addresses this gap by extending the simulation framework of Helmer et al. (2024) and evaluating component-wise recovery using multiple metrics, including canonical correlation error, weight error, score error, loading error, stability, and observed correlation behavior.

## 1.2 Canonical Correlation Analysis

Consider two sets of variables represented by the random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, with corresponding covariance matrices $\Sigma_X \in \mathbb{R}^{p\times p}$ and $\Sigma_Y \in \mathbb{R}^{q\times q}$, and cross-covariance matrix $\Sigma_{XY} \in \mathbb{R}^{p\times q}$. Canonical Correlation Analysis (CCA) seeks $k$ pairs of linear combinations of $X$ and $Y$ that are maximally correlated, where $k \leq \min(p, q)$.

Specifically, let $W_X \in \mathbb{R}^{p\times k}$ and $W_Y \in \mathbb{R}^{q\times k}$ denote the canonical weight matrices, whose $i$th columns, $W_{X,i}$ and $W_{Y,i}$, are the $i$th weight vector pair. The corresponding canonical variates are defined by $U = XW_X \in \mathbb{R}^{n\times k}$ and $V = YW_Y \in \mathbb{R}^{n\times k}$, so that the $i$th canonical pair is given by $U_i = XW_{X,i}$ and $V_i = YW_{Y,i}$. For each $i = 1, \ldots, k$, CCA finds $W_{X,i}$ and $W_{Y,i}$ to maximize the correlation

$$r_i = \operatorname{corr}(U_i, V_i) = \operatorname{corr}(XW_{X,i},\, YW_{Y,i}) = \frac{W_{X,i}^\top \Sigma_{XY} W_{Y,i}}{\sqrt{\left(W_{X,i}^\top \Sigma_X W_{X,i}\right)\left(W_{Y,i}^\top \Sigma_Y W_{Y,i}\right)}},$$

with resulting canonical correlations satisfying $r_1 \geq r_2 \geq \cdots \geq r_k$. To make the solution identifiable and nonredundant, two sets of constraints are imposed. First, scale invariance is removed by requiring each canonical variate to have unit variance:

$$\operatorname{var}(U_i) = W_{X,i}^\top \Sigma_X W_{X,i} = 1, \qquad \operatorname{var}(V_i) = W_{Y,i}^\top \Sigma_Y W_{Y,i} = 1.$$

Second, each successive pair must capture a new relationship between the two sets of variables, so the canonical variates are required to be uncorrelated with those from previously extracted pairs:

$$\operatorname{corr}(U_i, U_j) = W_{X,i}^\top \Sigma_X W_{X,j} = 0, \qquad \operatorname{corr}(V_i, V_j) = W_{Y,i}^\top \Sigma_Y W_{Y,j} = 0, \qquad j < i.$$

In practice, the population covariance matrices are unknown and must be estimated from column-centered data matrices $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ obtained from $n$ joint realizations of the features. The empirical covariance estimators are $\hat{\Sigma}_X = \frac{1}{n-1}X^\top X, \hat{\Sigma}_Y = \frac{1}{n-1}Y^\top Y$ and $\hat{\Sigma}_{XY} = \frac{1}{n-1}X^\top Y$.
The empirical CCA problem for the $i$th canonical pair is therefore

$$\max_{W_{X,i}\in\mathbb{R}^p,\, W_{Y,i}\in\mathbb{R}^q} \quad W_{X,i}^\top \hat{\Sigma}_{XY} W_{Y,i}$$

subject to $W_{X,i}^\top \hat{\Sigma}_X W_{X,i} = 1$, $W_{Y,i}^\top \hat{\Sigma}_Y W_{Y,i} = 1$, and $W_{X,i}^\top \hat{\Sigma}_X W_{X,j} = 0$, $W_{Y,i}^\top \hat{\Sigma}_Y W_{Y,j} = 0$, for $j < i$.

## 2 Data Generation Method

We define the feature dimensions $p, q$, sample size $n$, and number of canonical correlations $k$. The true correlations between the $k$ canonical variate pairs are denoted by $r_{\text{true},1}, \ldots, r_{\text{true},k}$, and we define $S = \text{diag}(r_{\text{true},1}, \ldots, r_{\text{true},k})$. To control the within-set covariance structure, we define $\Sigma_X$ and $\Sigma_Y$ as diagonal matrices whose entries follow a power-law decay:

$$\Sigma_X = c_x \cdot \text{diag}\left(1, \frac{1}{2^{\alpha_x}}, \ldots, \frac{1}{p^{\alpha_x}}\right), \qquad \Sigma_Y = c_y \cdot \text{diag}\left(1, \frac{1}{2^{\alpha_y}}, \ldots, \frac{1}{q^{\alpha_y}}\right).$$

In our simulations, we set $c_x = c_y = \alpha_x = \alpha_y = 1$.

To generate population weight matrices satisfying the canonical constraints, we first draw random matrices $M_X \in \mathbb{R}^{p\times k}$ and $M_Y \in \mathbb{R}^{q\times k}$, with entries generated independently from the standard normal distribution. We then apply QR decomposition to obtain matrices $U \in \mathbb{R}^{p\times k}$ and $V \in \mathbb{R}^{q\times k}$, whose columns are orthonormal. The population weight matrices are defined by $W_X = \Sigma_X^{-1/2}U \in \mathbb{R}^{p\times k}$ and $W_Y = \Sigma_Y^{-1/2}V \in \mathbb{R}^{q\times k}$. By construction,

$$W_X^\top \Sigma_X W_X = U^\top \Sigma_X^{-1/2} \Sigma_X \Sigma_X^{-1/2} U = U^\top U = I_k,$$

and similarly, $W_Y^\top \Sigma_Y W_Y = I_k$.

In addition to the QR-based construction above, an alternative method was used. Random matrices $\widetilde{W}_X \in \mathbb{R}^{p\times k}$ and $\widetilde{W}_Y \in \mathbb{R}^{q\times k}$ were first generated with i.i.d. $N(0,1)$ entries. They were then normalized with respect to the covariance-induced inner products:

$$W_X = \widetilde{W}_X \left(\widetilde{W}_X^\top \Sigma_X \widetilde{W}_X\right)^{-1/2}, \qquad W_Y = \widetilde{W}_Y \left(\widetilde{W}_Y^\top \Sigma_Y \widetilde{W}_Y\right)^{-1/2}.$$

This guarantees $W_X^\top \Sigma_X W_X = I_k$ and $W_Y^\top \Sigma_Y W_Y = I_k$. The corresponding matrices were then defined by $U = \Sigma_X^{1/2} W_X$ and $V = \Sigma_Y^{1/2} W_Y$.

To avoid configurations where the signal is overly concentrated in few low variance dimensions, we impose a feasibility constraint on the population weights, we require the orthonormalized weights, $\tilde{W}_X = W_X(W_X^T W_X)^{-1/2}$ and $\tilde{W}_Y = W_Y(W_Y^T W_Y)^{-1/2}$, to satisfy

$$\mathrm{tr}(\tilde{W}_X^T \Sigma_X \tilde{W}_X) \geq \frac{k\ \mathrm{tr}(\Sigma_X)}{2p} \quad \text{and} \quad \mathrm{tr}(\tilde{W}_Y^T \Sigma_Y \tilde{W}_Y) \geq \frac{k\ \mathrm{tr}(\Sigma_Y)}{2q}.$$

If a generated pair $(U, V)$ or $(W_X, W_Y)$ fails to satisfy the criteria, the matrices are discarded and regenerated.

We define the between set covariance as $\Sigma_{XY} = \Sigma_X^{1/2} U S V^T \Sigma_Y^{1/2}$ and the joint covariance as $\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$. We draw $n$ i.i.d samples from the joint multivariate normal distribution, $\binom{x_i}{y_i} \sim \mathcal{N}_{p+q}(\mathbf{0}, \Sigma)$ and store them row-wise in the datasets $X = \begin{pmatrix} x_1^T \\ \cdots \\ x_n^T. \end{pmatrix}$ and $Y = \begin{pmatrix} y_1^T \\ \cdots \\ y_n^T \end{pmatrix}$.

# 3 Stability Analysis Method

To evaluate the stability of CCA, synthetic datasets are generated from known population models, and the estimated canonical correlations, weight vectors, scores, and loadings are compared with their true population values. In addition to parameter recovery, the stability of the estimated weight and loading vectors is quantified, and the statistical power of the estimated canonical correlations is evaluated. This recovery is examined by systematically varying three experimental factors: the correlation level, the feature dimension, and the sample size.

The analysis is carried out component-wise. In the single canonical pair case, recovery is assessed for one canonical direction. In the multiple canonical pair case, the same metrics are applied separately to each canonical pair in order to study how recovery changes across components.

This component-wise approach was adopted because the main goal is to evaluate the estimation accuracy of each canonical direction individually. An alternative would be to compare the estimated and true canonical subspaces jointly using principal angles. While such a subspace-based analysis may be useful for measuring overall agreement between sets of directions, it does not isolate the recovery of each canonical pair. For this reason, the present study focuses on component-wise recovery.

### 3.1 Metrics for a Single Canonical Pair

In the case of a single canonical pair, we consider a single population canonical correlation $\rho$ and the pair of the associated weight vectors $w_X \in \mathbb{R}^p$ and $w_Y \in \mathbb{R}^q$. The score vectors, $t_X = Xw_X \in \mathbb{R}^n$ and $t_Y = Yw_Y \in \mathbb{R}^n$, as well as the loading vectors $\ell_X = \text{corr}(X, t_X) \in \mathbb{R}^p$ and $\ell_Y = \text{corr}(Y, t_Y) \in \mathbb{R}^q$ with correlations calculated between each dataset feature and the score vector. The loading vector $\ell$ measures the marginal similarity between each feature and the latent score and is defined as:

$$\ell_X = \frac{\text{diag}(\Sigma_{XX})^{-1/2}\Sigma_{XX}w_X}{\sqrt{w_X^T\Sigma_{XX}w_X}},$$

illustrating that loadings are the weights transformed by the covariance structure, normalized and scaled.

We use hat notation to denote the estimated values of these quantities and investigate the following stability metrics.

1. The relative error of the estimated canonical correlation $\hat{r}$ with respect to the true correlation $r_{true}$ measures how accurately the association strength between $X$ and $Y$ is estimated:

$$\Delta r = \frac{\hat{r} - r_{\text{true}}}{r_{\text{true}}}.$$

2. The weight error measures the discrepancy between the estimated weights $\hat{w}_X$ and $\hat{w}_Y$, and the true weights $w_X$ and $w_Y$. It is computed as the maximum deviation across both sets using the cosine similarity $\text{cossim}(\hat{w}, w) = \frac{\hat{w}^T w}{||\hat{w}||||w||}$:

$$\Delta w = \max(1 - |\text{cossim}(\hat{w}_X, w_X)|, \;\; 1 - |\text{cossim}(\hat{w}_Y, w_Y)|).$$

   Cosine similarity is employed to compare the alignment between estimated and true weight vectors in a scale invariant manner. By taking the absolute value, we account for the fact that weight pairs are equivalent up to a sign. This similarity measure is then converted into an error metric by taking its complement.

3. Denote the test datasets by $X^{(test)}$ and $Y^{(test)}$. The score error quantifies the discrepancy between the true test scores $t_X^{(test)} = X^{(test)}w_X$ and $t_Y^{(test)} = Y^{(test)}w_Y$ and the estimated test scores $\hat{t}_X^{(test)}$ and $\hat{t}_X^{(test)}$:

$$\Delta t = \max\left(1 - \left|\text{corr}\left(\hat{t}_X^{(test)}, t_X^{(test)}\right)\right|, \;\; 1 - \left|\text{corr}\left(\hat{t}_Y^{(test)}, t_Y^{(test)}\right)\right|\right)$$

   To make the scores comparable across repeated simulations, the metric is calculated using a common

test set.

4. The loading error measures the discrepancy between the true loadings $\ell_X = \mathrm{corr}(X^{(test)}, t_X^{(test)})$ and $\ell_Y = \mathrm{corr}(Y^{(test)}, t_Y^{(test)})$ and the estimated loadings $\hat{\ell}_X^{(test)}$ and $\hat{\ell}_Y^{(test)}$:

$$\Delta\ell = \max\left(1 - \left|\mathrm{corr}\left(\hat{\ell}_X^{(test)}, \ell_X^{(test)}\right)\right|,\ \ 1 - \left|\mathrm{corr}\left(\hat{\ell}_Y^{(test)}, \ell_Y^{(test)}\right)\right|\right).$$

5. Weight stability is quantified as the absolute cosine similarity between estimates obtained from two independently drawn datasets:

$$s_X = \left|\mathrm{cossim}\left(w_X^{(1)}, w_X^{(2)}\right)\right| \quad \text{and} \quad s_Y = \left|\mathrm{cossim}\left(w_Y^{(1)}, w_Y^{(2)}\right)\right|.$$

   Here, $w_X^{(1)}$ and $w_X^{(2)}$ denote the weights estimated from the first and second datasets, respectively (and analogously for $w_Y$). When multiple pairs of datasets are available, stability is computed for each pair and then averaged across all pairs. The stability of the loadings is defined analogously.

6. Power measures how often the observed canonical correlation is sufficiently large to be detected as statistically significant relative to a null distribution. It is estimated through the following procedure.

   For each generated dataset pair $(X, Y)$ drawn from a specified multivariate normal distribution, we compute the estimated canonical correlation $\hat{r}$. To construct the null distribution, we fix $X$ and independently permute the rows of $Y$ $M$ times. For each permutation, we compute the canonical correlations, resulting in $M$ null correlations $\hat{r}_1^{\mathrm{null}}, \ldots, \hat{r}_M^{\mathrm{null}}$. We then estimate the empirical $p$-value as

$$P = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}\big(\hat{r}_i^{\mathrm{null}} \geq \hat{r}\big).$$

   For a chosen significance level $\alpha$, the observed correlation is declared significant if $P < \alpha$.

   The experiment is repeated $N$ times, yielding p-values $P^{(1)}, \ldots, P^{(N)}$. The empirical power is defined as the proportion of dataset pairs for which the observed canonical correlation is significant:

$$\mathrm{Power} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(P^{(j)} < \alpha).$$

### 3.2 Metrics for Multiple Canonical Pairs

When analyzing $k$ canonical pairs, the strengths of the associations between the two sets of variables are summarized by the population canonical correlation vector $r_{true} = (r_{\mathrm{true},1}, \ldots, r_{\mathrm{true},k})$. The corresponding

population weight matrices are $W_X \in \mathbb{R}^{p\times k}$ and $W_Y \in \mathbb{R}^{q\times k}$, whose $i$th columns correspond to the $i$th canonical pair. Similarly, the score matrices are denoted by $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{n\times k}$, and the loading matrices by $L_X \in \mathbb{R}^{p\times k}$ and $L_Y \in \mathbb{R}^{q\times k}$.

To evaluate stability in the multiple-pair setting, the metrics introduced in the previous subsection are applied separately to each canonical pair. That is, the relative association error, weight error, score error, loading error, and stability are computed component-wise in order to assess how recovery changes across the successive canonical directions.

To compute power in this setting, a permutation-based procedure is again applied to each canonical correlation estimate. To account for multiple testing across the $k$ components, a Bonferroni correction is used, so that the $i$th observed canonical correlation is declared significant if its $p$-value is below $\alpha/k$. The empirical power for component $i$ is then defined as the proportion of the $N$ simulated datasets in which that component is detected as significant.

An alternative approach would be to compare the estimated and true canonical subspaces jointly using principal angles. However, the objective of the present study is to evaluate the recovery of each canonical pair individually, so the analysis is carried out component-wise.

# 4 Simulation Results

## 4.1 Samples per Feature is a Key Determinant in Stability

Building on the framework established by Helmer et al. (2024), we employ their benchmark analysis to evaluate whether samples per feature remains a key determinant of stability in single and multiple canonical pair analysis under our proposed generation approach.

**Single Canonical Pair** For the single canonical pair setting, we consider a fixed number of features, $p = q = 8$, and investigate two population correlation sizes, $r_{\text{true}} \in \{0.1, 0.5\}$. To ensure that our findings generalize across a range of covariance structures, we generated 10 distinct population models for each correlation. For every model, we sampled 100 independent datasets across the sample per feature ratios $n/p \in \{3, 8, 16, 32, 64, 128, 256, 300\}$. We then applied CCA to each dataset to estimate the canonical correlations and weights which were used to compute our stability metrics: weight, score, and loading error, averaging the results across all datasets. To assess statistical power, we generated 100 datasets per model and 100 permutations per dataset. This procedure for calculating power is maintained for all subsequent figures. For

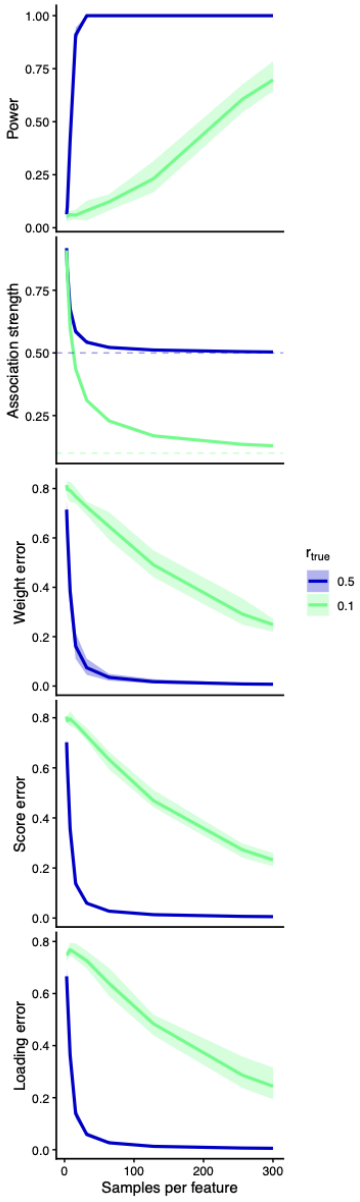



Figure 1: Stability metrics for the single canonical pair setting as a function of samples per feature. The figure shows statistical power, estimated association strength, weight error, score error, and loading error for two population canonical correlations, $r_{\text{true}} \in \{0.1, 0.5\}$, with $p = q = 8$. As samples per feature increase, power rises and the error metrics decrease, with substantially faster convergence when the true correlation is larger.

computational efficiency, statistical power was capped at 1.0 for all cases where $n/p \geq 512$. This threshold was validated across a subset of models for each figure, confirming that power consistently converges at this point. Full computation for these values remains an objective for future work.

We present the results in Figure 1. In this single canonical pair setting, the results demonstrate a stark contrast between medium ($r_{\text{true}} = 0.5$) and low ($r_{\text{true}} = 0.1$) correlation values. Under the stronger correlation, the model reaches near perfect power and minimal error rates by 32 to 64 samples per feature, while under weak correlation, errors remain high and power fails to reach 1.0 at max ratio of 300. This highlights that while the samples per feature ratio is a useful heuristic, the correlation magnitude serves a decisive role in determining when a model achieves stability in the univariate setting.

Building on this, we observe that the figure provided by Helmer et al. (2024) reflects a more complex simulation environment with more diverse covariance structures, evidenced by substantially wider confidence intervals, specifically for $r_{\text{true}} = 0.1$. Additionally, the power curve for $r_{\text{true}} = 0.1$ reaches convergence earlier around 250 samples per feature which suggests that their method allows for easier detection of the presence of an effect.

**Multiple Canonical Pair** For the multivariate case ($k = 3$), we investigated two population correlation triplets, $r_{\text{true}} \in \{(0.25, 0.20, 0.15), (0.55, 0.50, 0.45)\}$, and followed a procedure similar to the univariate setting, using $p = q = 8$, 10 population models, 100 independent datasets per model and samples per feature ratios of $\{3, 8, 16, 32, 64, 128, 256, 512, 1000\}$. The results are presented in Figure 2.

The multicanonical analysis confirms that samples per feature remains a dominant predictor of stability as each metric converges towards its asymptotic value as samples per feature increases. Across both correlation triplets, a clear hierarchical pattern emerges, with stability achieved more rapidly for lower components. This delayed convergence is consistent with our conclusions from Figure 1 as correlation magnitude decreases by 0.05 for each successive component. However, this hierarchical effect becomes even more pronounced: for the same correlation, $r_{\text{true}} = 0.5$, the multicanonical setting exhibits slower stabilization than in the unicanonical setting. This suggests that estimating later canonical pairs requires a higher samples per feature ratio to achieve the same level of error.

A plausible hypothesis for why the second canonical component takes longer to stabilize in the multicanonical setting is that it must be estimated conditionally on the first component. As each successive pair must satisfy an orthogonality constraint relative to the previous pairs, the range of values it can take is limited. As a result, errors in estimating the first component can propagate to the second, making its estimation less stable and requiring a higher samples per feature ratio for reliable convergence.

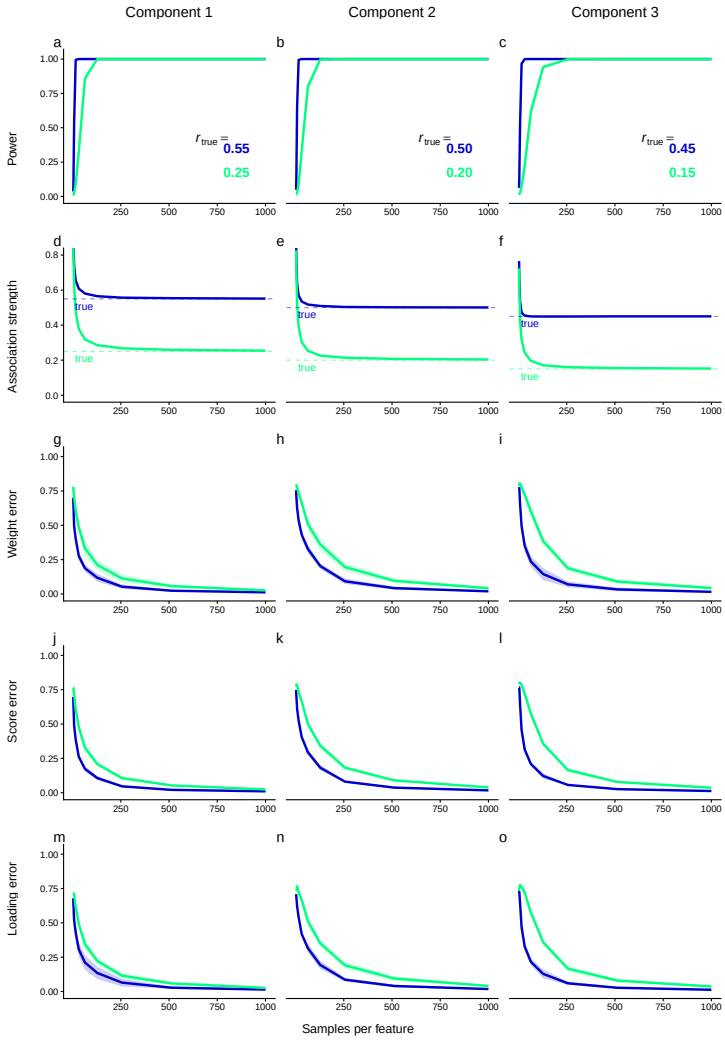



Figure 2: Stability metrics for the multiple canonical pair setting as a function of samples per feature. The columns correspond to the first, second, and third canonical components, and the rows show statistical power, estimated association strength, weight error, score error, and loading error for two population correlation vectors, $(0.55, 0.50, 0.45)$ and $(0.25, 0.20, 0.15)$, with $p = q = 8$. As samples per feature increase, power rises and the error metrics decrease for all three components, with faster convergence for larger canonical correlations and earlier components.

## 4.2 Effect of Varying Total Number of Features

In this section, we evaluate the stability of CCA across varying correlation strengths, feature dimensions, and sample sizes. These plots illustrate the effect of these parameters on the reliability of the canonical estimates and establish the thresholds required for accurate recovery in both the unicanonical and multicanonical case.

**Single Canonical Pair** For the univariate setting, we set the feature dimensions of both datasets to be equal and vary them incrementally: $p = q \in \{4, 8, 16, 32\}$. We also vary the correlation strengths: $r_{\text{true}} \in \{0.1, 0.3, 0.7\}$. For each condition, we generated 10 population models and simulated 100 datasets per model. We then calculated the stability metrics: power, relative association error, weight error, score error and loading error. We show the results in Figure 3.

In this univariate setting, the results indicate that across various feature dimensionality and correlation settings, samples per feature remain a primary factor of CCA stability. Consistent with our conclusions from Figure 1, we observe that stability is highly dependent on the true population correlation. We further observe that feature dimensions have a minor effect on stability compared to correlation in the univariate setting, as the line types within each colour group remain closely clustered. In terms of statistical power, larger feature dimensions generally yield higher power than smaller dimensions, except in settings with low correlation and limited samples per feature. In contrast, for relative association error, higher feature dimensions consistently produce larger errors than lower dimensional ones. For weight, score, and loading errors, however, no single combination of feature dimensionality and sample ratio consistently dominates, as the trajectories for different $p = q$ values frequently intersect. Notably, weight, score, and loading errors follow a remarkably similar decay pattern.

The following provides a proposal as to why these patterns in relative association and power occur. Larger feature dimensions result in higher relative association error because, in high dimensional spaces, there are many more ways for the model to overfit the data. This overfitting introduces an upward bias in the estimated correlation $\hat{r}$, producing greater relative error compared to lower dimensional models at the same sample density. By a similar mechanism, larger feature dimensions increase power, as more features provide additional opportunities to detect correlation.

Comparing the univariate setting to the figure provided by Helmer et al. (2024), we observe that while both studies yield similar overall results, our approach demonstrates a more uniform scaling behavior across weight, score, and loading error. In contrast, the weight error in the Helmer et al. analysis exhibits a distinct and more variable decay compared to score and loading error, suggesting that our generative framework produces

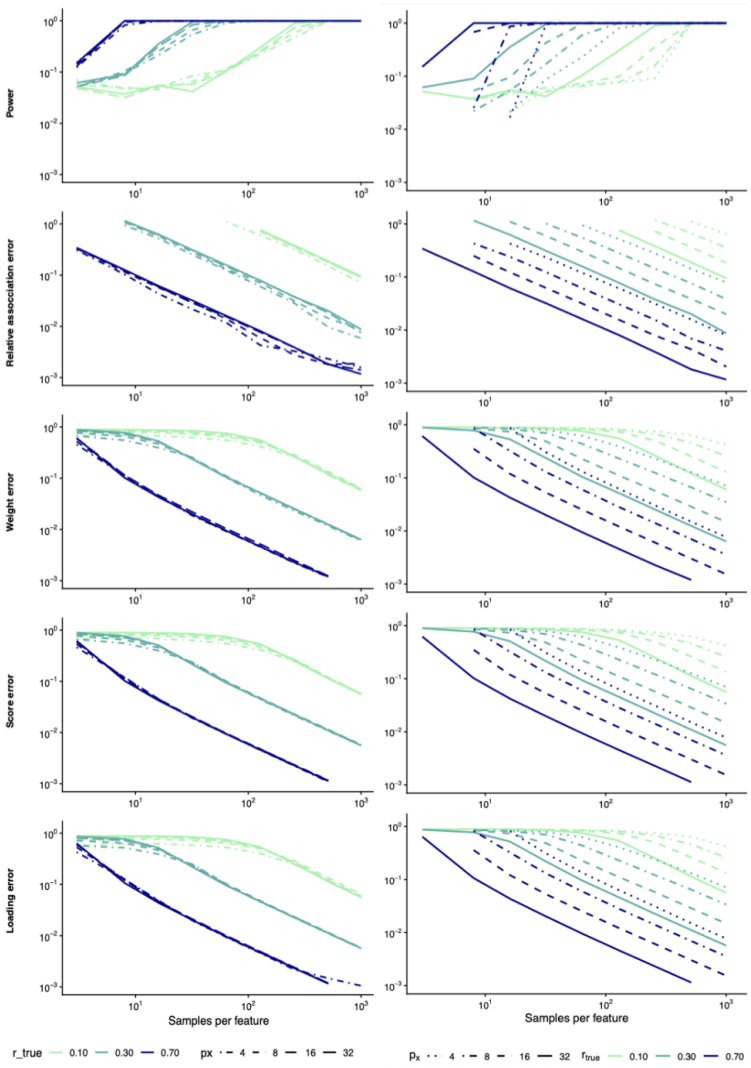



Figure 3: Stability performance in the single canonical pair setting under varying feature dimensions. The left panel shows results for balanced dimensions with $p = q \in \{4, 8, 16, 32\}$, while the right panel shows results for asymmetric dimensions with $p + q = 64$ and $p \in \{4, 8, 16, 32\}$, across $r_{\text{true}} \in \{0.1, 0.3, 0.7\}$. Across both panels, larger samples per feature improve power and reduce relative association, weight, score, and loading errors, with stronger correlations converging faster and greater asymmetry generally leading to poorer stability.

covariance structures that are less prone to inducing variability in weight estimates than those generated by Helmer et al.

Several patterns observed in this figure warrant further investigation. These include the similarity in the behavior of weight, score, and loading errors, the apparent minor influence of feature dimensions on univariate canonical stability, and whether the higher power observed for lower feature dimensions at low sample ratios and weak correlations reflects a genuine effect or a coincidental trend.

**Multiple Canonical Pair** For the multivariate case ($k = 3$), we investigated three population correlation triplets, $r_{\text{true}} \in \{(0.25, 0.20, 0.15), (0.55, 0.50, 0.45), (0.80, 0.75, 0.70)\}$. For each triplet, we generated 15 population models and sampled 100 independent datasets per model across varying feature dimensions ($p = q \in \{4, 8, 16, 32\}$). For each dataset, we computed the primary stability metrics and plotted their resulting averages in Figure 4.

In the multivariate canonical setting, stability follows a clear hierarchy: as component order increases or correlation strength decreases, the curves shift rightward, suggesting that reliably estimating later canonical pairs requires a substantially higher samples per feature ratio. Consistent with the univariate setting, weight, score, and loading errors maintain a consistent pattern to one another. However, in the multivariate case, varying $p$ and $q$ produces a much greater spread across these metrics, suggesting that feature dimensionality plays a significantly larger role in determining stability in multiple canonical pairs than in single canonical pair analysis. Furthermore, for high correlation ($r_{\text{true}} = (0.80, 0.75, 0.70)$), the diversity in power across varied feature dimensions is most pronounced in the first component and progressively decreases with each subsequent component. Finally, the relative association error reveals a complex transition across components, while larger feature dimensions surprisingly yield a lower relative error for the first canonical pair, this trend reverses in the second and third components. This reversal is most extreme in the third component, where lower feature dimensions achieve substantially smaller relative association errors compared to larger dimensions.

In the following, we propose why the patterns in this figure occur. The increased impact of feature dimensionality in multicanonical analysis is likely due to the requirement that each successive canonical pair be estimated conditionally on the previous pairs and satisfy orthogonality constraints. As feature dimensionality increases, the space of possible linear combinations grows, introducing more degrees of freedom in the weight estimates. While this can allow better alignment with the true canonical directions, it also increases the variability of the estimates, particularly for later components where errors from earlier components can propagate. Consequently, feature dimensionality has a more pronounced effect on stability when multiple

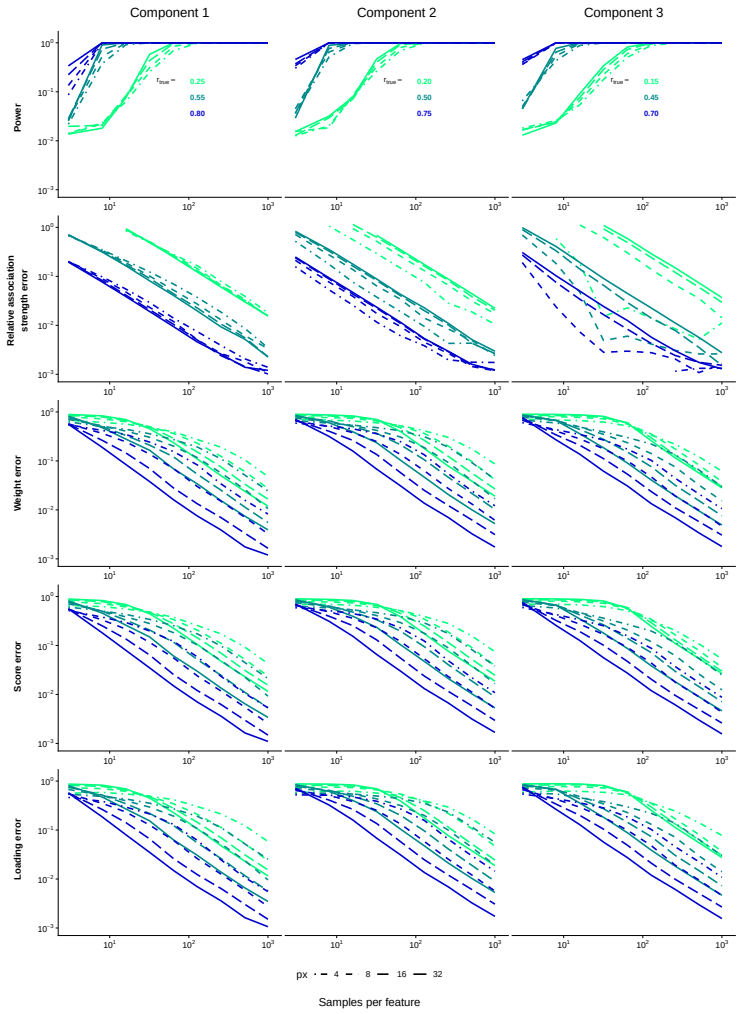



Figure 4: Stability performance in the multiple canonical pair setting under varying balanced feature dimensions. The columns correspond to the first, second, and third canonical components, and the rows show power, relative association error, weight error, score error, and loading error for population correlation vectors $(0.25, 0.20, 0.15)$, $(0.55, 0.50, 0.45)$, and $(0.80, 0.75, 0.70)$ with $p = q \in \{4, 8, 16, 32\}$. Across all components, larger samples per feature improve power and reduce error, while stronger correlations and earlier components converge more quickly; increasing feature dimension generally requires more samples for stable recovery.

canonical pairs are analyzed compared to the univariate case.

Regarding statistical power, we observe a feature advantage that is unique to high signal regimes. We hypothesize that when correlation is strong, larger feature dimensions improve the efficiency of signal detection; this effect diminishes as correlation decreases or component order increases. Finally, the observation that relative association error is lower for higher feature dimensions in the first component remains an area for further investigation, as it contradicts the typical overfitting patterns observed in other settings.

### 4.3 Effect of Feature Asymmetry

In this section, we further evaluate the stability of CCA across varying correlation strengths and feature dimensions. Specifically, the following plots illustrate how stability is influenced by asymmetry in the feature dimensions.

**Single Canonical Pair** In the univariate setting, we evaluate the correlation strengths, $r_{\text{true}} \in \{0.1, 0.3, 0.7\}$. We fix the total number of feature dimensions, $p + q = 64$, and vary one of the feature dimensions, $p \in \{4, 8, 16, 32\}$. Under each condition, 10 population models were generated and 100 datasets were simulated per model. Stability metrics were then computed for each dataset. Results are reported only for models satisfying the criteria, $n \geq \max(p, q)$ and presented in Figure 3.

For the univariate setting, asymmetry emerges as a key determinant of stability. Within a single colour group, varying levels of asymmetry results in clear differences in metric performance, with higher asymmetry consistently resulting in poorer outcomes and requiring a higher sample to feature ratio to achieve comparable errors. This figure also reinforces previously observed patterns, including the correlation effect and the consistent trends in weight, score, and loading errors.

In comparing our results to the figure from Helmer et al.(2024), we observe that asymmetry between feature dimensions has a significantly more pronounced impact on stability in our work compared to that of Helmer's analysis. While their curves remain relatively tightly grouped within each correlation level, our results show a much larger fanning out effect, where increased asymmetry causes substantial performance degradation. This disparity highlights that dimensionality imbalance is a more critical factor in stability than previously suggested by their model.

**Multiple Canonical Pair** For the multivariate case, we evaluated the population correlation triplets, $r_{\text{true}} \in \{(0.25, 0.20, 0.15), (0.55, 0.50, 0.45), (0.80, 0.75, 0.70)\}$, and varied the feature dimensions of one dataset, $p \in \{4, 8, 16, 32\}$, subject to the constraint that $p+q = 64$. Under each parameter combination, we generated

15 population models and sampled 100 independent datasets per model. For each dataset, we computed the primary stability metrics and plotted their resulting averages. Results are shown only for models satisfying the criteria, $n \geq \max(p, q)$ and presented in Figure 5.

In this multivariate setting, we observe the same asymmetry pattern identified in the univariate case, where asymmetrical feature dimensions consistently lead to higher error rates and diminished stability across all metrics. This confirms that the penalty for dimensionality imbalance remains a robust factor regardless of the number of components being estimated. Additionally, a successive rightward shift across components is evident in all plots, illustrating the hierarchical pattern observed in earlier figures.

## 4.4 Weight Error, Stability, and PC1 Similarity

**Single Canonical Pair** For the univariate CCA setting, a one-dimensional population CCA model was first generated with $p = q = 100$, $r_{\text{true}} = 0.3$, and $\alpha_X = \alpha_Y = 1$. For Figure 6(a), sample covariance matrices were then generated at each samples-per-feature level from the corresponding Wishart distribution, partitioned into $S_{XX}$, $S_{YY}$, and $S_{XY}$, and used to compute the one-dimensional sample CCA solution. The plotted curves show the difference between the estimated sample $X$-weight and the population $X$-weight across principal components for three representative sample-size regimes. For Figure 6(b), the same univariate setup was used to evaluate weight stability across repeated samples: for each sample size, sample covariance matrices were repeatedly drawn, the sample CCA weights were recomputed, and stability was summarized by the average pairwise cosine similarity across estimated weights. This follows the design of the paper, except that Wishart draws were used in place of direct sampling from the multivariate normal distribution. For Figure 6(c), the analysis was repeated for $r_{\text{true}} = 0.1$ and $r_{\text{true}} = 0.5$. For each setting, sample covariance matrices were drawn from the corresponding Wishart distribution at each samples-per-feature level, partitioned into $S_{XX}$, $S_{YY}$, and $S_{XY}$, and used to compute the one-dimensional sample CCA solution. The estimated sample $X$-weight was then compared with the first principal component axis of $X$, and the resulting PC1 similarity was summarized across repeated runs. In the paper, it reports CCA PC1 similarity across synthetic datasets with varying feature dimensions and $r_{\text{true}}$.

The reproduced results match the main pattern reported in the paper. In Figure 6(a), the estimated weight is highly variable when the sample size is small, but the discrepancy from the population weight becomes much smaller as samples per feature increase. In Figure 6(b), weight stability is close to zero at small sample sizes and increases steadily toward one as more observations are available. This is consistent with the paper's conclusion that, in univariate CCA, weight estimates are unstable at low sample-to-feature ratios but become increasingly accurate and stable with larger sample sizes. In Figure 6(c), PC1 similarity for CCA remains

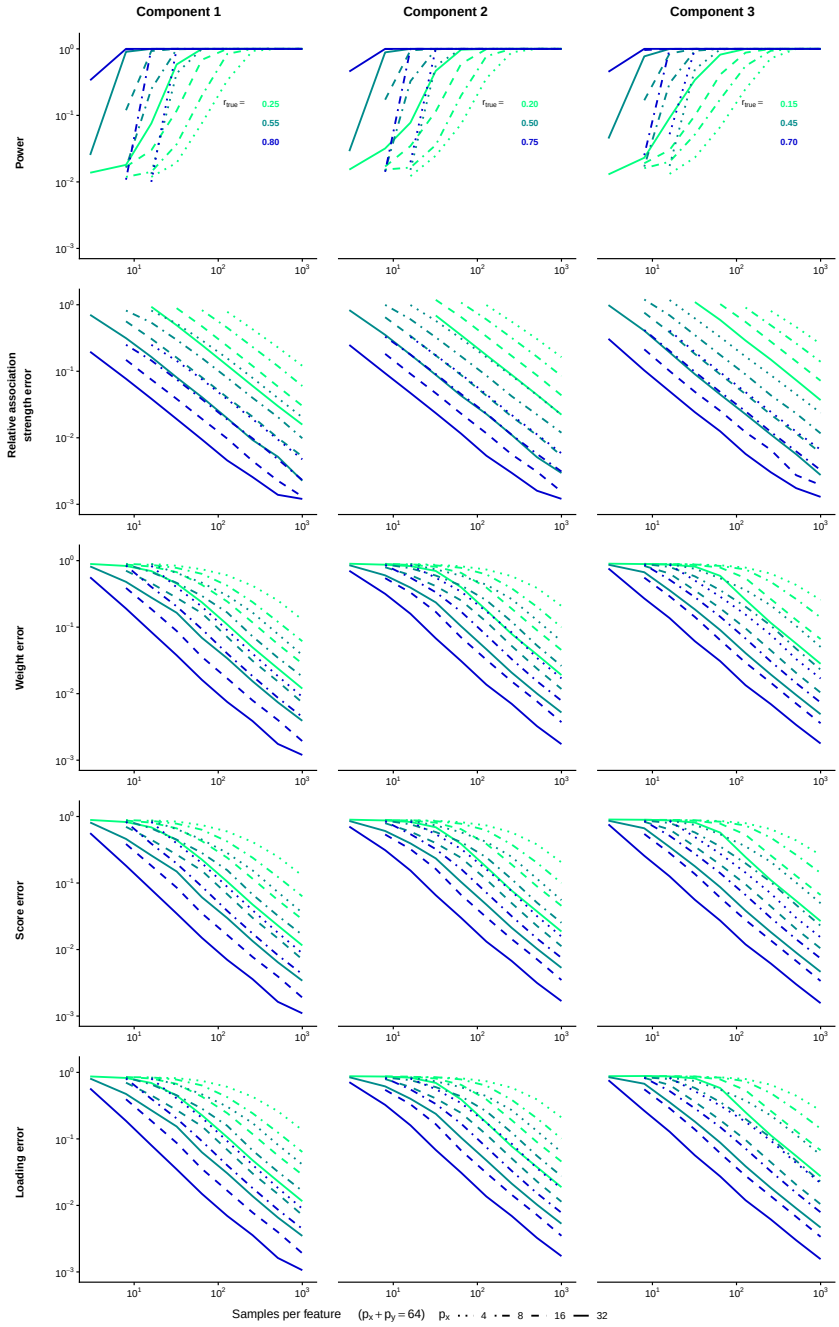
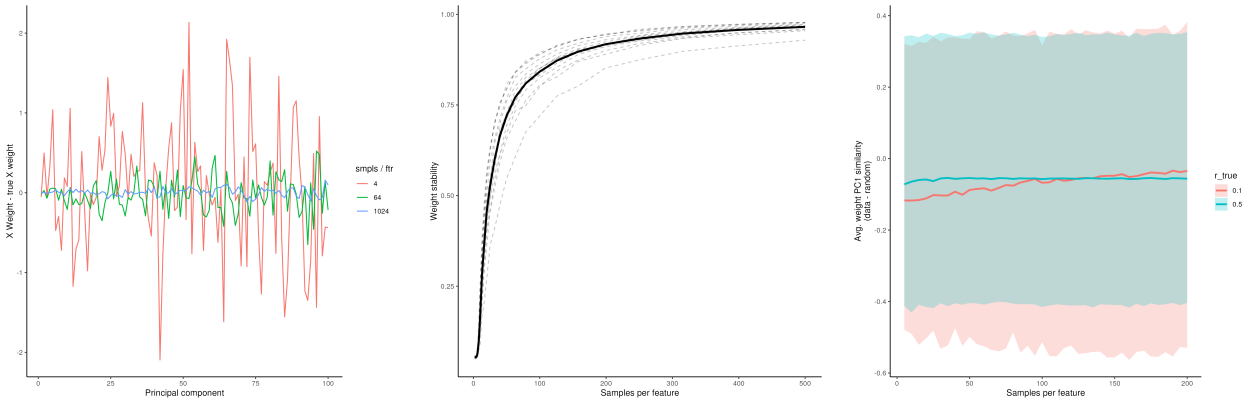



Figure 5: Stability performance in the multiple canonical pair setting under asymmetric feature dimensions. The columns correspond to the first, second, and third canonical components, and the rows show power, relative association error, weight error, score error, and loading error for population correlation vectors $(0.25, 0.20, 0.15)$, $(0.55, 0.50, 0.45)$, and $(0.80, 0.75, 0.70)$ with $p + q = 64$ and $p \in \{4, 8, 16, 32\}$. Across all components, larger samples per feature improve power and reduce error, while stronger correlations and earlier components converge more quickly; increasing asymmetry in the feature dimensions generally leads to poorer stability and requires more samples for accurate recovery.




((a)) Weight error in the single canonical pair setting with $p = q = 100$, $r_{\text{true}} = 0.3$, and power-law decay parameters $\alpha_X = \alpha_Y = 1$. The curves show the componentwise difference between the estimated sample $X$-weight and the true population $X$-weight under 4, 64, and 1024 samples per feature. As samples per feature increase, the estimated weight becomes much closer to the population weight.

((b)) Weight stability in the single canonical pair setting with $p = q = 100$, $r_{\text{true}} = 0.3$, and $\alpha_X = \alpha_Y = 1$. Stability is measured by the average cosine similarity across repeated estimated $X$-weights as samples per feature increase. The stability rises from near zero to near one, indicating that the estimated weights become more reproducible with larger sample sizes.

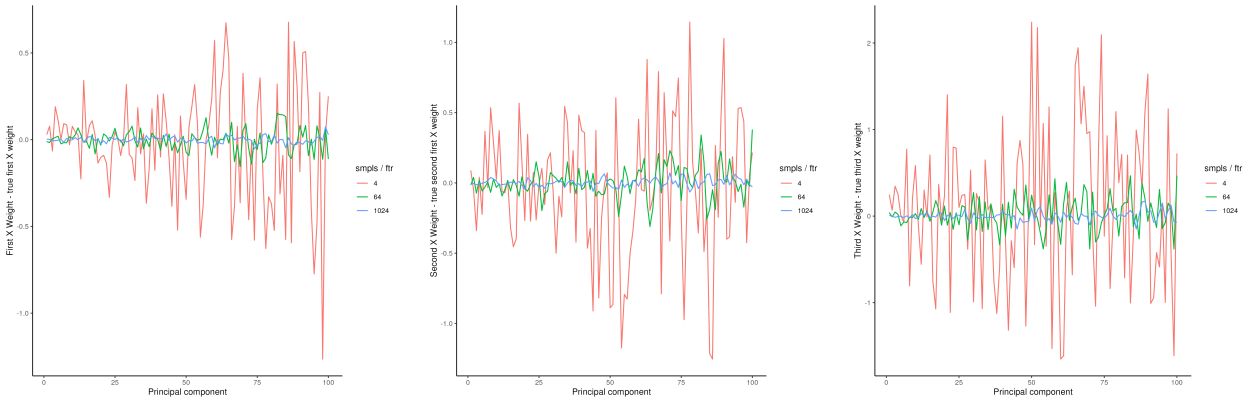
((c)) Weight PC1 similarity for univariate CCA with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and $r_{\text{true}} \in \{0.1, 0.5\}$. The plotted quantity compares the estimated sample $X$-weight with the first principal component axis across samples per feature. In both cases, the similarity remains low, indicating weak alignment of CCA weights with PC1.

Figure 6: Reproduction of univariate CCA results for weight error, weight stability, and weight PC1 similarity. The three panels summarize how the estimated $X$-weight behaves as sample size increases in the single canonical pair setting. Together, they show that larger samples per feature improve accuracy and stability, while PC1 similarity remains weak.

low across the sample-size range, indicating that the estimated CCA weights do not show a strong systematic bias toward the first principal component axis. This matches the paper's conclusion that CCA shows weak PC1 alignment.

**Multiple Canonical Pair Weight Error** For the multivariate CCA setting, a three-dimensional population CCA model was first generated with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and population correlations $(0.7, 0.5, 0.3)$ for the first, second, and third canonical directions. At each samples-per-feature level, a sample covariance matrix was then drawn from the corresponding Wishart distribution and partitioned into $S_{XX}$, $S_{YY}$, and $S_{XY}$. A three-dimensional sample CCA solution was computed from this sample covariance matrix, and the estimated $X$-weights were compared with the corresponding population $X$-weights mode by mode. Thus, three separate panels are shown, corresponding to the first, second, and third canonical weight vectors, where each curve plots the componentwise difference between the estimated sample weight and the true population weight under 4, 64, and 1024 samples per feature. This follows the same logic as the weight error figure in the paper, except that Wishart draws were used instead of direct sampling from

the multivariate normal model.




((a)) Weight error for the first canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. The curves show the componentwise difference between the estimated sample $X$-weight and the true first population $X$-weight under 4, 64, and 1024 samples per feature. As samples per feature increase, the estimated weight becomes much closer to the population weight.

((b)) Weight error for the second canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. The curves show the componentwise difference between the estimated sample $X$-weight and the true second population $X$-weight under 4, 64, and 1024 samples per feature. The discrepancy decreases as samples per feature increase, although convergence is slower than for the first direction.

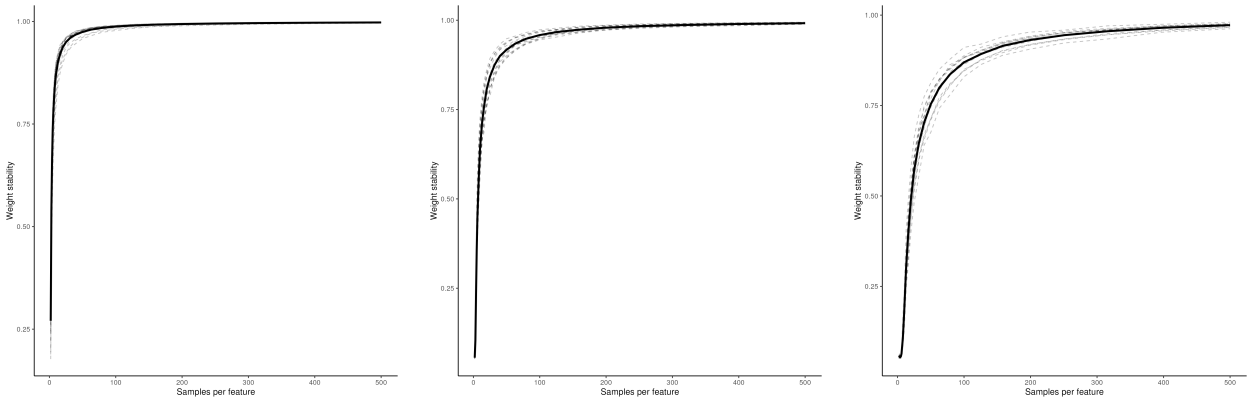
((c)) Weight error for the third canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. The curves show the componentwise difference between the estimated sample $X$-weight and the true third population $X$-weight under 4, 64, and 1024 samples per feature. The error decreases with larger samples per feature, but this weakest component remains the noisiest at intermediate sample sizes.

Figure 7: Weight error for the first three canonical directions in the multiple canonical pair setting. The panels compare estimated and population $X$-weights for components 1, 2, and 3 when the population correlations are $(0.7, 0.5, 0.3)$. Together, they show that larger samples per feature improve recovery for all three components, with faster convergence for stronger canonical correlations.

The same qualitative pattern as in the univariate case is observed in all three panels. When the sample size is small, the estimated weights are highly variable and can deviate substantially from the population weights. As the samples-per-feature ratio increases, these discrepancies shrink and the estimated weights become much closer to their population targets. The improvement is visible for all three canonical directions, although the lower-correlation modes remain somewhat noisier at intermediate sample sizes than the strongest mode. Overall, these plots agree with the paper's main finding that CCA weight estimates are unstable at low sample-to-feature ratios but become much more accurate once the sample size is sufficiently large, with stronger population associations converging faster than weaker ones.

**Multiple Canonical Pair Weight Stability** For the multivariate CCA setting, a three-dimensional population CCA model was first generated with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and population correlations $(0.7, 0.5, 0.3)$ for the first, second, and third canonical directions. For each samples-per-feature level, sample covariance matrices were drawn from the corresponding Wishart distribution and partitioned into $S_{XX}$,

$S_{YY}$, and $S_{XY}$. A three-dimensional sample CCA solution was then computed from each sampled covariance matrix. Weight stability was evaluated separately for each canonical direction: for a fixed mode and sample size, the estimated $X$-weights from repeated runs were compared pairwise using cosine similarity, and these pairwise similarities were averaged to obtain the stability value. Thus, three separate panels are shown, corresponding to the first, second, and third canonical weight vectors. This follows the same idea as the weight stability figure in the paper, except that Wishart draws were used instead of direct sampling from the multivariate normal model.




((a)) Weight stability for the first canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. Stability is measured by the average cosine similarity across repeated estimates of the first sample $X$-weight. The stability is already high at small sample sizes and approaches one rapidly as samples per feature increase.

((b)) Weight stability for the second canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. Stability is measured by the average cosine similarity across repeated estimates of the second sample $X$-weight. The stability increases steadily with samples per feature and converges slightly more slowly than for the first direction.

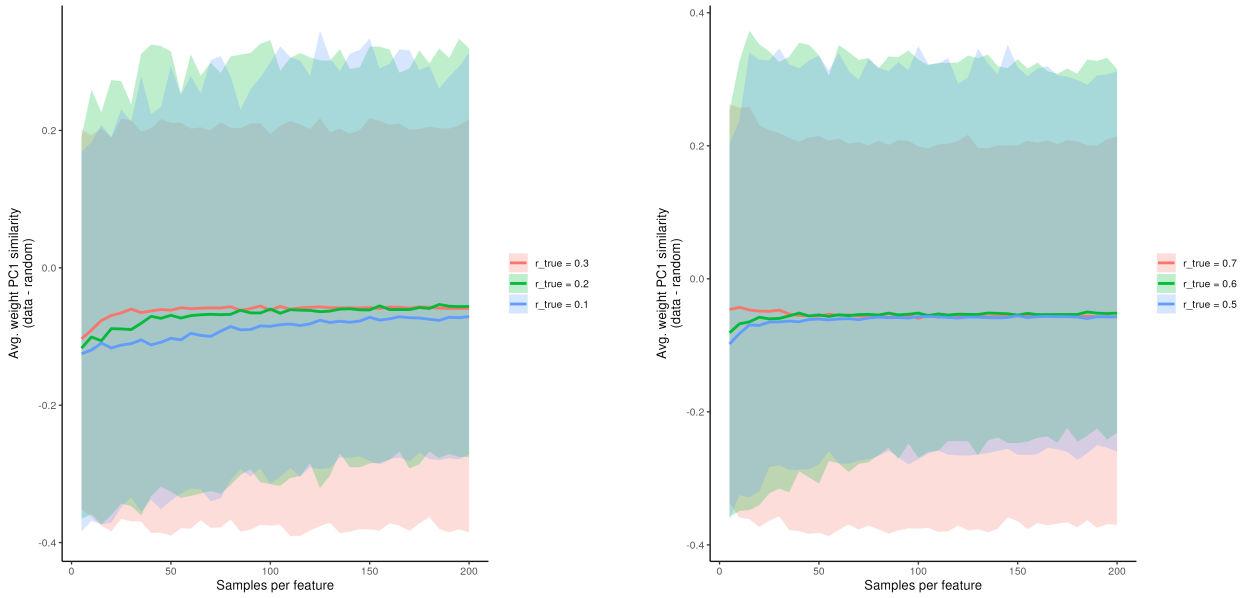
((c)) Weight stability for the third canonical direction in the multiple canonical pair setting with $p = q = 100$, $r_{\text{true}} = (0.7, 0.5, 0.3)$, and $\alpha_X = \alpha_Y = 1$. Stability is measured by the average cosine similarity across repeated estimates of the third sample $X$-weight. This weakest component has the slowest convergence, but its stability still approaches one as samples per feature become large.

Figure 8: Weight stability for the first three canonical directions in the multiple canonical pair setting. The panels show how reproducibility of the estimated $X$-weights changes with samples per feature when the population correlations are $(0.7, 0.5, 0.3)$. Together, they show that stability improves for all three components with larger sample size, with faster convergence for stronger canonical correlations.

The three panels show the same overall pattern as in the univariate case, but with different rates of convergence across modes. In all three cases, weight stability increases with samples per feature and approaches one as the sample size becomes large. The first canonical direction, which has the strongest population correlation, stabilizes the fastest and remains highly stable across the whole range. The second direction also converges strongly, but slightly more gradually. The third direction, associated with the weakest population correlation, shows the slowest rise and remains the least stable at intermediate sample sizes, although it still approaches near-perfect stability when the sample size is sufficiently large. These results agree with the paper's main conclusion for CCA that weight stability is low when the sample size is small and increases

toward one with more observations, with stronger between-set associations becoming stable more quickly than weaker ones.

**Multiple Canonical Pair Weight PC1 Similarity** For the multivariate CCA setting, two three-dimensional population CCA models were generated with $p = q = 100$ and $\alpha_X = \alpha_Y = 1$. The first used a lower-correlation configuration, $r_{\text{true}} = (0.3, 0.2, 0.1)$, and the second used a higher-correlation configuration, $r_{\text{true}} = (0.7, 0.6, 0.5)$. For each samples-per-feature level, a sample covariance matrix was drawn from the corresponding Wishart distribution, partitioned into $S_{XX}$, $S_{YY}$, and $S_{XY}$, and used to compute the three-dimensional sample CCA solution. For each of the three canonical directions, the estimated sample $X$-weight was compared with the first principal component axis of $X$, and the corresponding random-vector baseline was subtracted. These relative PC1 similarities were then summarized across repeated runs, producing three curves in each plot, one for each canonical direction.




((a)) Relative weight PC1 similarity in the lower-correlation multivariate CCA setting with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and $r_{\text{true}} = (0.3, 0.2, 0.1)$. The three curves correspond to the first, second, and third canonical directions, and the plotted quantity compares the estimated sample $X$-weights with the first principal component axis after subtracting a random-vector baseline. The similarity remains small across samples per feature, indicating weak alignment with PC1 in the lower-correlation setting.

((b)) Relative weight PC1 similarity in the higher-correlation multivariate CCA setting with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and $r_{\text{true}} = (0.7, 0.6, 0.5)$. The three curves correspond to the first, second, and third canonical directions, and the plotted quantity compares the estimated sample $X$-weights with the first principal component axis after subtracting a random-vector baseline. The similarity again remains small across samples per feature, indicating weak alignment with PC1 even when the population correlations are larger.

Figure 9: Weight PC1 similarity for multivariate CCA under lower- and higher-correlation settings. The two panels compare relative PC1 similarity when $r_{\text{true}} = (0.3, 0.2, 0.1)$ and when $r_{\text{true}} = (0.7, 0.6, 0.5)$. Together, they show that the estimated multivariate CCA weights remain only weakly aligned with the first principal component axis across the full sample-size range.

Both panels below show that the relative PC1 similarity remains small throughout the sample-size range, indicating that the estimated CCA weights do not display a strong systematic tendency to align with the first principal component axis. In the lower-correlation setting, the three curves are slightly more negative and show somewhat more variation at smaller sample sizes. In the higher-correlation setting, the curves are slightly closer to zero and more stable, but the overall pattern remains the same. Thus, even when the population correlations are increased, the multivariate CCA weights still do not show the strong PC1 bias highlighted in the paper. This agrees with the paper's weight PC1 similarity figure, where PC1 similarity is reported to be weak for CCA.

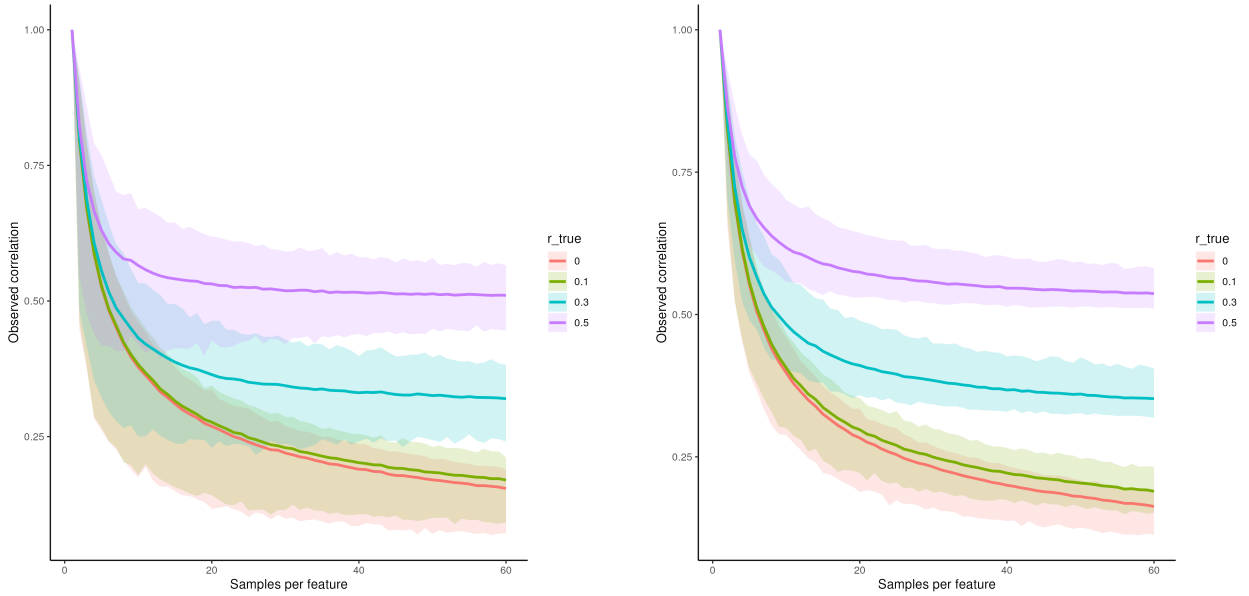
### 4.5 Observed Correlation

For the univariate version, one-dimensional population CCA models were generated with $p = q = 100$ and $\alpha_X = \alpha_Y = 1$. The analysis was repeated over a grid of true between-set correlations, and for each value and each samples-per-feature level, sample covariance matrices were drawn from the corresponding Wishart distribution and partitioned into $S_{XX}$, $S_{YY}$, and $S_{XY}$. A one-dimensional sample CCA solution was then computed, and the observed canonical correlation was recorded. Repeating this many times for each setting gave the mean observed correlation and its uncertainty band as a function of samples per feature. This reproduces the synthetic prediction part underlying Figure 6(a), but not the literature dots or Figure 6(b), since those require the external database of published CCA studies used in the paper's meta-analysis.

For the multivariate version, the same procedure was applied to a higher-dimensional CCA model. A multivariate population CCA was first generated, sample covariance matrices were then drawn from the corresponding Wishart distribution at each samples-per-feature level, and a sample CCA solution was recomputed from each draw. The resulting observed canonical correlations were summarized in the same way as in the univariate case. In practice, the multivariate plot shows essentially the same pattern as the univariate one, so the two panels are very similar.

Both panels show the same qualitative pattern as the smooth curves in the paper's observed correlation figure. When the samples-per-feature ratio is small, the observed correlation is strongly inflated relative to the true value; as the sample size increases, the observed correlation decreases toward its population target. Stronger true correlations remain higher across the whole range, while weaker or null correlations fall more substantially as the number of samples per feature grows. This agrees with the paper's conclusion that many reported CCAs are compatible with a wide range of true correlations and that the number of samples per feature is a strong determinant of the observed correlation. The fact that the univariate and multivariate reproductions look nearly identical here suggests that, for this figure, the dominant behavior is

driven mainly by the sample-size-to-dimension ratio rather than by whether the model is one-dimensional or three-dimensional. Since the literature dots and Figure 6(b) depend on the paper's external meta-analysis dataset and matching procedure, they are not included in the present reproduction.




((a)) Observed correlation in the single canonical pair setting with $p = q = 100$, $\alpha_X = \alpha_Y = 1$, and $r_{\text{true}} \in \{0, 0.1, 0.3, 0.5\}$. The curves show the mean observed canonical correlation as a function of samples per feature, with shaded bands indicating variability across repeated simulations. When samples per feature are small, the observed correlations are strongly inflated relative to their population values, and they decrease toward the true values as sample size increases.

((b)) Observed correlation in the multiple canonical pair setting under the same range of $r_{\text{true}}$ values. The curves again show mean observed canonical correlation versus samples per feature, with shaded bands indicating variability across repeated simulations. The same inflation-at-small-sample pattern appears here, and the overall trend is very similar to the single canonical pair case.

Figure 10: Observed correlation as a function of samples per feature in the Figure 6 reproduction analysis. The two panels compare the single canonical pair and multiple canonical pair settings for the same range of $r_{\text{true}}$ values. Together, they show that observed correlations are overestimated when samples per feature are small and move toward their population targets as sample size increases.

# 5 Code

The code supporting the analyses presented in this study is openly available in the following GitHub repository: `https://github.com/phamali/Stability-Analysis-of-CCA.git`. The repository contains all scripts used for data generation and stability assessment procedures. It also includes documentation to facilitate replication of the results and further extension of the methods presented in this work.

# 6 References

Helmer, M., Warrington, S., Mohammadi-Nejad, A.-R., Ji, J. L., Howell, A., Rosand, B., Anticevic, A., Sotiropoulos, S. N., and Murray, J. D. (2024). On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications Biology*, 7(1):217.

Yang, Q., Zhang, X., Song, Y., Liu, F., Qin, W., Yu, C., and Liang, M. (2021). Stability test of canonical correlation analysis for studying brain-behavior relationships: The effects of subject-to-variable ratios and correlation strengths. *Human Brain Mapping*, 42(8):2374–2392.

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